Fall 10-5-2017

Investigating Statistics Teachers' Knowledge of Probability in the Context of Hypothesis Testing

Jason Mark Asis Dolor
Portland State University

Investigating Statistics Teachers' Knowledge of Probability in the Context of Hypothesis Testing

by

Jason Mark Asis Dolor

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Mathematics Education

Dissertation Committee:
Jennifer Noll, Chair
Michael Shaughnessy
Steven Boyce
Robert Fountain
Swapna Mukhopadhyay

Portland State University
2017

ABSTRACT

In the last three decades, there has been a significant growth in the number of undergraduate students taking introductory statistics. As a result, there is a need by universities and community colleges to find well-qualified instructors and graduate teaching assistants to teach the growing number of statistics courses. Unfortunately, research has shown that even teachers of introductory statistics struggle with concepts they are employed to teach. The data presented in this research sheds light on the statistical knowledge of graduate teaching assistants (GTAs) and community college instructors (CCIs) in the realm of probability by analyzing their work on surveys and task-based interviews on the *p*-value. This research could be useful for informing professional development programs to better support present and future teachers of statistics.

ACKNOWLEDGEMENTS

The completion of this dissertation would not have been possible without the support of a number of individuals. Foremost, I would like to thank my advisor, Jennifer Noll, for all her support, encouragement, and guidance throughout my years as a doctoral student. In all the years working together, you have always encouraged me to never give up. Your support has helped me grow as an educator, researcher, and individual. I would also like to thank the members of my committee who have taken time from their busy lives to make this process possible: Michael Shaughnessy, Steven Boyce, Robert Fountain, Swapna Mukhopadhyay, and the late Ronald Narode. You all have been inspirational as teachers and colleagues throughout the years.

I would like to also thank my math education family at Portland State University. To Dana Kirin, thank you for all your amazing work throughout this dissertation process. To Briana Mills and Sonya Redmond, thank you for providing professional and emotional support through my dissertation studies. To Lewis Lum, thank you for leading me down the road of a mathematician and educator.

Finally, I would also like to thank my family. To my brothers, thank you for encouraging me to pursue higher goals. To my parents, your constant support and encouragement has helped me continue to grow into who I am today. I also want to thank God for giving me the opportunity share this success with so many amazing people.

TABLE OF CONTENTS

ABSTRACT i
ACKNOWLEDGEMENTS ii
List of Tables vii
List of Figures viii
1. INTRODUCTION 1
2. RATIONALE 7
2.1. Why Study Undergraduate Statistics Education? 7
2.2. Why Study Hypothesis Testing? 9
2.3. Why Study GTAs and CCIs? 12
2.3.1. Graduate teaching assistants. 14
2.3.2. Community college instructors. 16
2.3.3. Summary 18
3. LITERATURE REVIEW 20
3.1. Frameworks for Mathematical and Statistical Knowledge for Teaching 20
3.2. Research on Hypothesis Testing in the Classroom 29
3.2.1. Traditional hypothesis testing. 30
3.2.2. Research on students' and teachers' understanding of NHST. 38
3.2.2.1. The role of conditional probability in hypothesis testing. 40
3.2.2.2. The role of sampling distributions in hypothesis testing. 45
3.2.3. New pedagogical approaches to hypothesis testing. 50
4. THEORETICAL FRAMEWORK 57
4.1. A Theoretical Framework for Assessing SKT for p-value. 57
4.2. Categorizing SKT for P-value. 60
4.2.1. Category 1 – Definition of a p-value. 61
4.2.2. Category 2 – Symbolic representation of a p-value. 62
4.2.3. Category 3 – Relationship of a p-value with sampling distributions. 64
4.2.4. Category 4 – Relating a p-value to the level of significance. 68
5. METHODOLOGY 71
5.1. Data Collection 72
5.1.1. Assessment survey. 72
5.1.1.1 Content Question 1 (The Drug Test Task). 74

5.1.1.2. Content Question 2 (The Graduate Student Task). 76
5.1.1.3. Content Question 3 (The Car Task). 78
5.1.1.4. Content Question 4 (Helper-Hinderer Task) 81
5.1.2. Interview. 86
5.1.2.1. Follow-up survey questions. 87
5.1.3. Survey and interview participants. 87
5.2. Analysis 93
5.2.1. Survey analysis. 93
5.2.2. Interview analysis. 95
5.2.2.1. Post-survey and post-interview analysis. 96
5.2.2.2. Transcription analysis. 96
5.2.2.3. Coding scheme. 96
5.2.2.4. Chronicling emerging themes. 97
5.3. Validity 98
6. Results 99
6.1. Teachers' Understanding of the p-value's Magnitude. 100
6.1.1. Survey results of the Graduate Student Task 102
6.1.1.1. Quantitative results of the Graduate Student Task. 102
6.1.1.2. Qualitative results for the Graduate Student Task. 104
6.1.2. Interview Results of the Graduate Student Task. 115
6.1.2.1. Evidence of procedural understanding of a p-value's magnitude 118
6.1.2.2. Evidence of conceptual understanding of a p-value's magnitude. 120
6.1.2.3. Evidence of hybrid understanding of a p-value's magnitude. 121
6.1.3. Summary of the Graduate Student Task 126
6.2. Computing p-values with Samplings Distributions. 128
6.2.1. Survey Results: Computing p-values with sampling distributions. 130
6.2.2. Interview Data: Computing p-values with sampling distributions. 143
6.2.2.1. Struggling with empirical sampling distributions. 146
6.2.2.2. Preferring theoretical over empirical probability. 148
6.2.2.3. Favoring simulations. 152
6.2.3. Summary: Computation of the p-value using sampling distributions. 155
6.3. Teachers' Understanding of Conditional Probably and the p-value. 157

6.3.1. Survey results: verbal interpretations of the p-value. 157
6.3.2. Survey results: symbolic representations of the p-value. 166
6.3.3. Interview results of the p-value conditional nature 172
6.3.3.1. Struggling with verbal and symbolic representations of a p-value. 176
6.3.3.2. Relating verbal and symbolic representations of a p-value. 182
6.3.3.3. Conflicting views of the symbolic representations of a p-value. 192
6.3.3.4. Developing an understanding of conditional reasoning. 196
6.3.4. Summary: Conditional probability and p-values. 199
7. Discussions 201
7.1. SKT of Probability on the concept of the p-value. 203
7.1.1. Teachers' understanding of p-value's magnitude. 203
7.1.1.1. Connecting CCK and SCK of a p-value's magnitude. 204
7.1.1.2. A misconception of the p-value's magnitude. 207
7.1.2. Teachers' understanding of p-value's computation. 208
7.1.2.1. Correct computations with empirical sampling distributions. 209
7.1.2.2. Misconceptions when using empirical sampling distributions. 210
7.1.2.3. Theoretical over empirical methods. 212
7.1.3. Teachers' understanding of conditional probability. 213
7.1.3.1. Verbal interpretations of the p-value. 213
7.1.3.2. Symbolic representations of the p-value. 215
7.1.3.3. Reasons for relating verbal and symbolic representations. 217
7.2. Implications and Future Research 219
7.2.1. Implications for undergraduate education statistics curriculum 219
7.2.1.1. Simulations and statistical inference in the classroom. 219
7.2.1.2. Informal and formal statistical inference in the classroom. 223
7.2.1.3. Conditional probability and statistical inference. 224
7.2.2. Implications for the preK-12 classrooms. 226
7.2.3. Implications for professional development for graduate students and instructors. 228
7.2.4. Implications for future research. 231
7.3. Limitations of the Research 234
8. Conclusions 236

REFERENCES 239
APPENDIX A: TASK DESCRIPTIONS 251
APPENDIX B: SURVEY QUESTIONS 253
APPENDIX C: INTERVIEW QUESTIONS 258

LIST OF TABLES

Table 1. Significance Testing Procedure for a College Proportion. 31
Table 2. Table for Type-I/II Errors in Hypothesis Testing..... 34
Table 3. Critical Region Hypothesis Test for a Population Proportion 35
Table 4. Null Hypothesis Significance Test for a Population Proportion..... 37
Table 5. Theoretical Framework for Analyzing Categories of SKT for the P-value..... 58
Table 6. Categories of Understanding for the P-value..... 61
Table 7. Demographics for the Graduate Student Participants..... 89
Table 8. Demographics for the Instructors. 91
Table 9. Demographics of the Interview Participants..... 92
Table 10. Results for the Multiple-Choice Portion of Graduate Student Task..... 103
Table 11. Examples of SSHP..... 105
Table 12. Examples of SSLP. 106
Table 13. Examples of SSL. 107
Table 14. Examples of SSD..... 109
Table 15. Examples of SSPPS. 110
Table 16. Examples of Hybrid Thinking of the Magnitude of the P-value. 112
Table 17. Results of Categories for the Magnitude of the P-value..... 113
Table 18. Responses for the Interview Participants on the Graduate Student Task. 116
Table 19. Results of P-value Computation for Helper-Hinderer Task. 130
Table 20. Examples of CRF..... 132
Table 21. Examples of CRFANH. 133
Table 22. Alternative Computations of the P-value for the Helper-Hinderer Task..... 134
Table 23. Examples of CIOS. 135
Table 24. Examples of CUR. 137
Table 25. Examples of CLS. 139
Table 26. Examples of CTP. 140
Table 27. Examples of CHTM..... 141
Table 28. Summary of Categories for the Computation of the P-value..... 142
Table 29. Categories of Interviewee Responses for the Computation of the P-value. ... 144
Table 30. Accurate and Inaccurate Feedback for Student Interpretation A..... 161
Table 31. Accurate and Inaccurate Feedback for Student Interpretation D..... 162
Table 32. Results for Student Verbal Interpretations A and D. 162
Table 33. Conditional and Non-Conditional Feedback for Student Interpretation B. 163
Table 34. Results for Student Verbal Interpretations B, C, E and F..... 165
Table 35. Results for the Car Task on the Symbolic Representations of the P-value. ... 171
Table 36. Results of the Drug Task for the Interviewees. 172
Table 37. Results of the Car Task for the Interviewees..... 174
Table 38. Theoretical Framework for Analyzing Categories of SKT for the P-value.... 202

LIST OF FIGURES

Figure 1. "Domain map for mathematical knowledge for teaching (Hill et al, 2008, p. 377)." 22
Figure 2. "Hypothesized structure of statistical knowledge for teaching (Groth, 2007, p. 429)." 26
Figure 3. "Set of proposed indicators to assess SKT (Gonzalez, 2012, p. 5)." 27
Figure 4. "Model for Statistical Knowledge for Teaching (Noll, 2007, p. 70)." 29
Figure 5. Theoretical sampling distribution with critical value and critical region. 34
Figure 6. "Questionnaire on the interpretations of a p-value (Haller & Krauss, 2002, p. 5)." 45
Figure 7. "Model of Hypothesis Testing with Empirical Sampling Distributions (Dolor & Noll, 2015, p. 64)." 53
Figure 8. Image of p-value in a theoretical sampling distribution. 66
Figure 9. Empirical sampling distribution with population parameter p = 0.50. 66
Figure 10. Relating a p-value with a level of significance in a sampling distribution. 69
Figure 11. Drug Test Task 76
Figure 12. Graduate Student Task. 78
Figure 13. Car Task. 81
Figure 14. Helper-Hinderer Task. 83
Figure 15. Empirical sampling distribution with a correct p-value region marked. 84
Figure 16. Empirical sampling distribution with only the observed value marked. 85
Figure 17. Graduate Student Task 101
Figure 18. Helper-Hinder Task 130
Figure 19. Empirical Sampling Distribution with the correct p-value region. 133
Figure 20. Empirical sampling distribution illustrating CIOS. 137
Figure 21. Drug Test Task. 159
Figure 22. Car Task. 168

1. INTRODUCTION

Statistical analysis has always been an important component in all areas of research that use data. This is especially relevant in our technological world where the flow of data has increased because of computing technologies and social media. This makes a strong understanding of statistics important and necessary to every citizen if they want to make informed decisions in such a world. The last few decades of academia have shown a dramatic rise in the number of statistic courses in all educational levels. In elementary education, Common Core State Standards (CCSS) by the National Governors Association for Best Practices Center (2010) require that mathematics classrooms incorporate statistical concepts, such as measures of center (e.g. mean, median and mode), variation (e.g. standard deviation, IQR and range), and notions of distribution into the mathematics curriculum. Secondary education requires that, in addition to learning about analyzing graphs and measures of center and variation, a standard mathematics curriculum incorporate lessons on statistical inference and simulations. Traditional four-year universities and community college also require introductory statistics as mandatory coursework for several undergraduate degrees (e.g. Business, Psychology, Economics, and Sociology.). As the number of students now transitioning from high-school to undergraduate education continues to grow, there has been a surge in the number of introductory statistics courses now being offered in four-year universities and community colleges (Blair, Kirkman, & Maxwell 2013). It is therefore vital for the statistics education community to have appropriate methods to assess the quality of instruction in our introductory statistics courses.

A topic of interest in the statistics education community in recent few years is students' and teachers' understanding of statistical inference. *Inferential statistics* is defined as "a branch of statistics that consists of generalizing from samples to populations, performing hypothesis testing, determining relationships among variables, and making predictions (Bluman 2012, p. 809)." Documents such as Common Core School Standards (CCSS) states that secondary school students should (a) "Understand statistics as a process for making inferences about population parameters based on a random sample from that population; and (b) Decide if a specified model is consistent with results from a given data-generating process. (p. 50)" Similar statements are reverberated by the American Statistical Association (2007) in their release of Guidelines for Assessment and Instruction in Statistics Education (GAISE) for PreK-12 stating that,

> "Instructional programs from pre-kindergarten through grade 12 should enable all students to: formulate questions that can be addressed with data; collect, organize, and display relevant data to answer them; select and use appropriate statistical methods to analyze data; develop and evaluate inferences and predictions that are based on data; and understand and apply basic concepts of probability (Franklin, Kader, Mewborn, Moreno, Peck, Perry, & Schaefer 2007, p. 5)."

The GAISE report for undergraduate education also highlights that, "students should understand the basic ideas of statistical inference which include the concept of statistical significance, significance levels, and *p*-values (Aliaga et al., 2010, p.12)." Similar ideas are also mentioned in the Principles and Standards for School Mathematics released by the National Council of Teachers of Mathematics (NCTM) that states,

> "Students are expected to develop and evaluate inferences and predictions about data. These include: (a) use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions; (b) understand how sample statistics reflect the value of population parameters and use sampling distributions as the basis for informal inference; (c) evaluate published reports that are based on data by examining the design of the study, the appropriateness of the data analysis, and the validity of the conclusions; (d) understand how basic statistical techniques are used to monitor process characteristics in the workplace (NCTM, 2000)."

The CCSS, GAISE, and NCTM documents illustrate the overwhelming influence of statistical inference in the current statistics curriculum. A concept that is particularly important in the study of statistical inference is hypothesis testing. *Hypothesis testing* is defined as "a decision-making process for evaluating claims about a population (Bluman, 2012 p. 809)." The hypothesis test is formally taught as a component in secondary school as part of an advance placement statistics courses and collegiate statistics curricula. Hypothesis testing is also a statistical tool used by researchers (e.g., education, psychology, science, etc.) to analyze and justify claims about data (Metz, 2010). It is therefore fundamental that students and teachers of statistics understand the concepts related to hypothesis testing. This is particularly relevant for teachers whose professional responsibility is the instruction of this concept.

Unfortunately, research from the statistics education community has shown that both students and teachers struggle understanding hypothesis testing concepts (see

Batanero, 2000; Batanero & Diaz, 2006; Castro Sotos et al, 2007; Falk, 1986; Garfield & Ben-Zvi, 2008; Haller & Krauss, 2002; Thompson, Liu & Saldahna, 2007; Vallecillos, 2002; Vallecillos & Batanero, 1997). Common misconceptions in the research literature can be traced to concepts of probability in hypothesis testing. Research has shown that even students who can correctly apply formal inference procedures and compute probability (e.g. *p*-value) in the context of hypothesis testing are unable to appropriately interpret the meaning behind the procedures (Batanero, 2000). This has led statistics education researchers to question both traditional methods of statistical instruction (e.g. Cobb, 2007; Haller & Krauss, 2002) and the statistical knowledge of teachers (e.g. Thompson, Liu & Saldahna, 2007).

In the classroom, the development of student knowledge can be highly influenced by teachers' content and pedagogical knowledge (Shulman, 1986). Research suggests it is important that statistics education research focus on the knowledge of statistics teachers since it has a direct influence on a teacher's ability to teach (Ball, Hill & Bass, 2005; Hill, Sleep, Lewis, Ball & Lester, 2007; Liu & Thompson, 2009; Shulman, 1986). Shaughnessy (2007) highlights that while there has been significant growth of statistics education literature in the last few decades specifically related to student learning, more research should be done analyzing the statistical knowledge of teachers. Ten years later this is still an area in need of research.

Another area lacking significant research in mathematics and statistics education is the study of graduate teaching assistants (GTA) and community college instructors (CCI). In recent years, many four-year universities and community colleges have shifted

the responsibility of teaching introductory statistics to GTAs or adjunct instructors rather than full-time faculty members (Blair, Kirkman, & Maxwell 2013). There has also been a rise in new pedagogical approaches by the statistics education research community that have affected the teaching of statistical inference in the classroom (Erickson, 2006). Because of these changes to statistics education curriculum, it is necessary for the research community to re-evaluate future and current statistics teachers to determine whether they possess the relevant statistical knowledge for teaching.

This research hopes to help fill in the gap in the statistics education research on teachers by focusing on the population of GTAs and CCIs who may be responsible for the education of a vast majority of statistics students at the undergraduate level, add to the current research on their content knowledge of hypothesis testing and probability. In particular, the research questions proposed in this study are: What knowledge do CCIs and GTAs have about probability in the context of hypothesis testing? In particular,

1) How do CCIs and GTAs understand the concept of a *p*-value?
2) How do CCIs and GTAs understand the role of conditional probability in the context of hypothesis testing?
3) How do CCIs and GTAs understand formal/informal concepts of probability, simulations and sampling distributions when reasoning about tasks related to hypothesis testing?

To help answer these questions this dissertation begins with a rationale for studying GTAs and CCIs understanding of hypothesis testing in undergraduate education. Second, research from the mathematics and statistics education community on theoretical

frameworks concerning teacher knowledge is discussed. Third, relevant literature on traditional hypothesis testing presented in introductory statistics courses and research found in the statistics education community on teacher and student understanding on concepts related to probability in hypothesis testing is examined. Included in this discussion are current trends in the statistics education community regarding new pedagogical approaches to the learning of hypothesis testing. Fourth, data collection methods, the process of data analysis and issues of validity are addressed. Fifth, results of the data analysis are presented. Finally, this dissertation concludes with directions for future research.

2. RATIONALE

The goal of this research is to analyze the knowledge of CCIs and GTAs understanding of probability in hypothesis testing. This section highlights the rationale for my study by first focusing on the importance of studying undergraduate statistics education. I then discuss the importance of studying hypothesis testing. Finally, I discuss my rationale for studying the population of CCIs and GTAs.

2.1. Why Study Undergraduate Statistics Education?

In order to handle the flood of data currently being produced today, research in all areas of study require trained individuals to have a deep understanding of statistical techniques. As early as elementary school, students are introduced to concepts of statistics to develop basic analytical skills. For example, teaching students basic concepts of measures of center and variation help students see patterns and make sense of data. Students' statistical development is continued in secondary school where they are expected to cover additional topics in statistics with a focus on data analysis and statistical inference as a component of a mathematics curriculum. Lessons on analyzing graphs and modeling data extend student thinking towards making inferences about a population based on sample information.

Undergraduate students are first enrolled in a formal introductory statistics course where they not only review material from elementary and secondary education, but are introduced to more formal techniques of statistical inference such as hypothesis testing.[1] In undergraduate education, the majority of the student population is expected to have at

[1] This is to the exception of students who have taken AP statistic courses in high-school.

least one course in introductory statistics. In majors where analyzing data is a necessity (e.g. psychology, economics, education, etc.) students are expected to develop a deeper understanding of statistical methods as a means to draw valid conclusions from empirical data. Having a deeper understanding of statistical methods will also prepare students to assess the quality of statistical research as they progress in their future careers.

With the overwhelming need to develop a society of statistically illiterate individuals, universities and community colleges have increased the number of entry-level statistics courses now being offered. This is illustrated in a survey by the Conference Board of Mathematical Sciences (CBMS) which has shown an increased enrollment for elementary-level statistics in the fall 2010 by 56% from the levels of fall 2005 in four-year colleges and 17% in two-year colleges (Blair, Kirkman, & Maxwell, 2013). Because of the large growth of student enrollment, it becomes more important that the research community reflect on the current quality of introductory statistics courses.

Unfortunately, a large part of the research literature indicates that many adults struggle using statistics to make informed decisions about quantitative information (Noll, 2007). Furthermore, understanding of statistics is a necessary component for equal participation in a democratic society where public issues, managing personal finance, and making decisions all require some level of statistical competence (Noll, 2007). It should therefore be the goal of all statistics teachers to ensure students are knowledgeable of statistical methods in order to be informed members of society.

The need for a statistically literate society has caused a shift in the way introductory statistics has generally been taught in colleges (Aliaga et al., 2010). A

recommendation by the GAISE report is that undergraduate statistics should emphasize the importance of *statistical literacy* and *statistical thinking.* Statistical literacy is viewed as "understanding the basic language of statistics (e.g. knowing what statistical terms and symbols mean and being able to read statistical graphs) and fundamental ideas of statistics (Aliaga et al., 2012, p. 14)." Statistical thinking is defined as "the type of thinking that statisticians use when approaching or solving statistical problems (Aliaga et al., 2012, p. 14)." This includes understanding the need for data, the importance of data production, the omnipresence of variability in data, and the quantification and explanation of variability (Cobb & Moore, 1997). These two perspectives on statistics education and cumulative research in statistics education have led many researchers to rethink traditional curriculum. Cobb (2007) suggests transitioning from traditional methods of statistics education that relied heavily on theory towards more practical and innovative approaches that use new educational resources such as computer simulations. Furthermore, the use of simulations is more common to practice of statisticians today. With the research community shifting statistics education towards these new pedagogical practices, the study of undergraduate statistics education becomes a ripe field of research.

2.2. Why Study Hypothesis Testing?

The concept of hypothesis testing is one of many important concepts in the study of statistics and is one of great interest to the statistics education research community in recent years. A basic understanding of hypothesis testing is required in order to deeply understand advanced methods of statistical analysis (e.g. regression analysis and ANOVA). The standard approach to learning introductory hypothesis testing is simply a

basis for various types of hypothesis tests used by statisticians depending on the type of data analysis being performed (e.g. sample means, proportions, standard deviations, etc.). For example, a hypothesis test on data for a single proportion will look procedurally different than a hypothesis test with multi-categorical data or data for linear regression. There are, however, underlying concepts of hypothesis testing whose conceptual meaning remains the same regardless of the data being analyzed or hypothesis test being performed (e.g. *p*-value or level of significance). For example, calculus students must understand the basic idea of integration before learning other integration techniques. Similarly, statistics students need to understand the concept of hypothesis testing before learning advanced statistical concepts like categorical or linear regression analysis. This is especially important for students majoring in subjects focused on scientific research where data analysis is a norm in their future career. As a result, students taking introductory statistics are required to learn hypothesis testing as part of the major focus.

The approach to teaching hypothesis testing has not changed much since its introduction into statistics curriculum with most current introductory statistics textbooks still taking a traditional and procedural approach (e.g. Bluman, 2012; Brase & Brase, 2012). Historically, statisticians developed the traditional method of teaching hypothesis testing because the ability to perform large-scale simulations was impossible in the early 20th century due to technological limitations. Thus, many hypothesis test procedures are based on theorems from advanced probability (Cobb, 2007). In the last decade, the statistics education community began to re-evaluate the teaching of hypothesis testing pushing for newer curriculum. Researchers such as Cobb (2007) recommend a re-

evaluation of our current approach to the teaching of hypothesis testing in undergraduate education, claiming that the traditional approach of teaching introductory statistics needs to progress to appropriately match the technology of the current world.

Today many individuals have access to computers and calculators to run simulations both at home and in the classroom. Furthermore, the practices of modern statistics have shifted to using computer simulations as a way to analyze data, which may not be easily predicted using traditional statistical methods.[2] This had led statistics education researchers to develop new pedagogical approaches to the teaching and learning of hypothesis testing that emphasizes the importance of technology and simulations (e.g., Chance, del Mas & Garfield, 2004; Chance, B. et al., 2007; delMas, R., Garfield, J. & Chance, B., 1999; Erickson, 2006; Garfield & Everson, 2009; Garfield & Zvi, 2008; Heid et al., 2005; Weinberg, Wesner & Pfaff, 2011; Zieffler, Garfield, delMas & Reading, 2008; Zieffler, Garfield, et al., 2008). These new approaches have started to impact curriculum (e.g. GAISE) by emphasizing simulations as a vital component in undergraduate statistics education. With these new alternatives to teaching statistics slowly being incorporated into the current standards, it is vital as a research community to reassess the knowledge of statistics teachers (current and future) to determine whether they possess the knowledge, experience, and training to transition with the new curriculum.

[2] Examples include Monte Carlo or bootstrapping methods.

2.3. Why Study GTAs and CCIs?

Researchers argue there are many factors that can influence a student's educational development, but many education researchers agree that one of the most crucial factors is the teacher (see Ball, Hill & Bass 2005; Hill, Sleep, Lewis, Ball & Lester, 2007; Liu & Thompson, 2009; Shulman, 1986). As the number of statistics courses in universities and community colleges continues to grow, universities and community colleges must find instructors to teach these courses. While many advanced mathematics and statistics courses are taught by full-time faculty, many universities and community colleges have steadily relied on part-time faculty and/or graduate students to cover a large number of introductory statistics courses needed by a vast number of the student population. The Conference Board of Mathematical Science (2010) indicates that the number of part-time faculty in mathematics and statistics at community colleges has increased by 29% from 2005 to 2010. As a result, teachers in introductory statistics (and other entry-level mathematics courses) could potentially have various levels of statistical knowledge and background.

In the last few decades, there has been a large body of research targeting the classroom practices of K-12 teachers, but little has been done to study the teaching practice of collegiate instructors (Speer, Smith, & Horvath, 2010). Speer et al. (2010) highlight the lack of empirical research on the practices of instructors of mathematics at the collegiate level saying that,

> "Some mathematicians have written about their teaching others have analyzed aspects of their teaching and their students' learning in innovative collegiate

> courses, and a diverse body of other scholarship mentions collegiate mathematics teaching, very little research has focused directly on *teaching practice* – what teachers do and think daily, in class and out, as they perform their teaching work (p. 99)."

Unlike pre-college teachers, college teachers should possess robust levels of content knowledge but also likely spend significantly less time with students. In most cases, a collegiate teacher spends approximately 3 hours of class time per week for 10-15 weeks with students in a single academic term. This limitation on classroom hours makes lesson planning challenging for many instructors, which could lead to difficulties in managing content, and experimenting with new learning activities. In addition, many college teachers also do not have background in education or learning and have no set curriculum. Furthermore, collegiate teachers are typically less constrained in their classroom planning and decision-making. According to Speer et al, (2010) "collegiate teachers make judgments and decisions, before, during, and after teaching, based on their sense of the content, what their students do and do not understand, and what is possible in the time remaining in their courses (p. 101)."

The fact that the current undergraduate faculties possess potential gaps in education practices, subject matter knowledge, lack of a set curriculum, and a need for more accountability introductory classes shows a greater need for professional development for these individuals in terms of effective teaching practices. A lack of empirical research on collegiate instructors and the potential impact teachers have on student learning makes the study of introductory statistics teachers a ripe population for

research. Unfortunately, little is known about the population of collegiate teachers and their understanding of statistics. This research hopes to fill the gap in the statistics education literature on the knowledge of statistics teachers in hypothesis testing by focusing on two populations who could potentially teach (or are currently teaching) introductory statistics: graduate teaching assistants and community college instructors.

2.3.1. Graduate teaching assistants. The jobs of a GTA can vary greatly in the university community. Presently, GTAs are expected to fill the role of student, grader, tutor, research assistant, and instructor in college institutions. In a large-scale study conducted by Diamond & Gray (1987), they analyzed the demographics and efficacies of GTAs from multiple colleges across the nation. The study suggests that GTAs played a significant role in student education but only a select few universities offer minimal support to GTAs in regard to training and professional development. The result is GTAs who might be asked to work as instructors with very few possessing appropriate teaching experience, knowledge, and/or training. The study has also indicated that some GTAs are potentially placed to teach courses that are outside of their realm of expertise. As a result, GTAs who might possess a strong mathematics background may be asked to teach a statistics course even though they might have minimal statistical knowledge or background in statistics. Luft et. al. (2004) highlight that even though GTAs play an important role in the instruction of undergraduate students, some are often poorly prepared for teaching.

When responsibilities of GTAs take on instructional roles in the collegiate community, the expectation by many students is that GTAs have a robust knowledge of

mathematics and statistics similar to full-time faculty. This is most relevant when GTAs take on the roles that have direct contact with students (e.g. tutors and instructors). A study by Golde & Dore (2001) has shown that approximately half of all GTAs are placed in classrooms with little or no instruction in teaching. This had led members of the mathematics and statistics education community to realize the important role GTAs play in the collegiate community and the serious need to better understand the mathematical and statistical knowledge of this population (Speer, Gutmann, & Murphy, 2010). An example of an empirical study that has addressed GTAs knowledge of statistical inference was conducted by Noll (2007) who focused on GTAs understanding of confidence intervals and variability. In her research, Noll discovered that some GTAs struggle understanding the role of sampling distribution in the creation of confidence intervals. She also found that GTAs had difficulty attending to multiple attributes of a distribution, instead GTAs overly focused on either measures of center or measures of variability. Noll's research is one of the few that analyzes the statistical knowledge of GTAs, but a large body of research focuses on the training (or lack thereof) of GTAs. An example is DeChenne (2010) who investigated the teaching effectiveness of GTAs and the types of professional development that supported their success as teachers finding that factors such as teaching experience, teaching self-efficacy, GTA training and a departmental teaching climate affect the behavioral outcome of teaching effectiveness.

Finally, it is important to consider the future careers of most GTAs. Many GTAs who graduate have the potential to become future community colleges instructors and tenure-track faculty members. In some cases, a GTA position may be the first teaching

experience for many future faculty members. Researchers (e.g. Speer, Gutman, & Murphy, 2010; Luft et al., 2004) have discussed the significance of early experiences in solidifying beliefs, developing practices and setting patterns of social learning for new teachers. Speer, Gutman, & Murphy (2010) highlight that "the time spent as a TA is the time which young mathematicians will develop teaching practices they likely will carry with them into their careers as faculty members. (p. 76)." Thus, a GTA's professional development could provide rich opportunities to shape their instructional practice. This also becomes a time when GTAs become interested in education and the importance of teaching. Once hired as faculty, they also end up having several roles in their institutions making it harder to focus just on their teaching. It is therefore important that the research community pays special attention to these future educators and analyzes how their knowledge of teaching their subjects develops in the early stages of their instructional career.

2.3.2. Community college instructors. Studies have shown a dramatic rise in community college enrollment in the last few years (Blair, Kirkman, & Maxwell, 2013). In the Fall of 2005, almost 44% of all undergraduates in the United States were in community colleges (Lutzer, Rodi, Kirkman, & Maxwell, 2007) with the trend seemingly continuing to increase. This is not surprising considering the cost to enroll in community college classrooms and much diverse population (including non-traditional students) makes it appealing to students. In many cases, students typically take community college courses as a stepping-stone to future courses in four-year universities. In fact, between the Fall of 1995 to 2010, there has been roughly a 41% growth in the mathematics and

statistics enrollment in community colleges. It is therefore common that many students take credits in introductory statistics at community colleges, which they then transfer to four-year universities.

The educational background of CCIs varies widely from bachelors to doctorate instructors with several of the community college faculty consisting of part-time instructors. Similar to GTAs, CCIs (part-time/full-time) are sometimes asked to teach an introductory statistics course even though the CCIs might possess minimal background in statistics. While many community colleges require and support professional development for its full-time faculty, little is known for part-time faculty. This is troubling considering that full-time and part-time faculty members teach similar courses in community colleges. Furthermore, the expectation of faculty responsibilities is quite different in community colleges compared to four-year universities. In four-year universities, faculty members have a shared responsibility between their role as instructors and researchers. The general environment of community colleges leans more towards instruction of the students with a focus towards developmental mathematics courses (Mesa, Wladis, & Watkins, 2014). As a result, most community college faculty's instructional coursework is the teaching of courses ranging from intermediate algebra to calculus. Faculty members in four-year universities typically teach a broader range of courses ranging from intermediate algebra to high-level mathematics and statistics courses. The differences in the expected instructional coursework of the faculty suggest potential differences in the mathematical and statistical knowledge of faculties in the different academic settings (Mesa, Wladis, & Watkins, 2014).

Unfortunately, there is little research regarding community college teachers' knowledge of statistical inference or community college teachers' mathematical knowledge in general (Mesa, Sitomer, Strom, & Yonatta, 2012). At present, how community college mathematics classes should be taught are made by instructors, administrators, and policy makers in the absence of research-based evidence (Mesa, Wladis, & Watkins, 2014). According to Mesa, Wladis, & Watkins (2014), "mathematics education researchers need to concentrate more extensively on issues of mathematical learning in the community college context and collaborate with the practitioners who have expertise in the teaching and learning of mathematics in this setting (p. 185)." Conducting research on the statistical knowledge of community college teachers can help administrators make decisions that better support their faculty's professional development needs.

2.3.3. Summary. Because of the similar instructional roles CCIs share with GTAs in regards to instructional coursework, it is vital that the statistics education community focus on the types of knowledge of these educators. Since there is minimal research but a greater need by the educational community to further study the work of GTAs and CCIs, this study is important because it takes a first step at filling in the gap in the statistics education literature regarding these two populations. This research may aid policymakers and administrators to make research based decisions to better support present and upcoming statistics instructors. Furthermore, individuals who coordinate and develop professional development programs for these two populations will gain better

understanding of the kinds of concepts to be covered in courses to prepare current and future statistics instructors.

3. LITERATURE REVIEW

The overarching goal of this research is to study the statistical knowledge of GTAs and CCIs by investigating their understanding of probability in the context of hypothesis testing. In order to do so, it is necessary to discuss prior research regarding the knowledge of statistics teachers to set a foundation for analyzing the statistical knowledge of GTAs and CCIs. In particular, I discuss current theoretical frameworks on the knowledge of statistics teachers. I then discuss research literature regarding the traditional approach of hypothesis testing in statistics classrooms, misconceptions held by students and teachers due to the traditional approach, and current shifts in pedagogical practices related to hypothesis testing suggested by the statistics education community.

3.1. Frameworks for Mathematical and Statistical Knowledge for Teaching

There has been a long history in mathematics education research focused on investigating the knowledge of mathematics teachers. A foundational piece on mathematical knowledge for teaching was by Shulman (1986) who theorized that teachers of mathematics required various forms of knowledge. For example, simply knowing how to add numbers and teaching someone how to add numbers require different forms of teacher knowledge. Teachers must account for the knowledge of the student and what factors or methods would be appropriate to support a student's learning of a mathematical concept. This led Shulman to introduce three important categories of teacher knowledge. The first is *subject matter knowledge for teaching*. This includes knowing both the underlying theory of a mathematical concept and the procedures and meaning behind the concept. For example, knowing how to compute an arithmetic mean

and understanding the mean as a mathematical 'balance point' are two different forms of subject matter knowledge. The second category of knowledge defined by Shulman is *pedagogical knowledge*. This knowledge encompasses knowledge of student understanding regarding a mathematical concept and instructional methods that support students learning. For example, teachers should be aware of the prior knowledge and potential misconceptions students might have in regard to a *p*-value. Knowing various ways students misinterpret the definition of the *p*-value or how to introduce a task on *p*-values that is accessible to an introductory statistics classroom are types of knowledge align with effective teaching practices. Teachers should be aware of the limitations students might have learning a mathematical/statistical concept and know ways of presenting material that best fits a student's prior knowledge. The third knowledge is *curricula knowledge* of teachers, which encompass knowledge of materials and tools for teaching. For example, knowing how simulations might be a useful tool in developing students' understanding of sampling distributions, probability and inference.

The work of Shulman (1986) was highly influential in the mathematics and statistics education community. Prior to his work, there was little research regarding the concepts of pedagogical knowledge and subject matter knowledge. Shulman's (1986) research on mathematical knowledge for teaching was further expanded in the work of Hill, Ball, and Schilling (2008) whose framework for teacher knowledge has been widely accepted in the mathematics education community (Figure 1). The framework of Hill et al. (2008) categorizes teachers' knowledge into two main types: subject matter knowledge and pedagogical content knowledge.

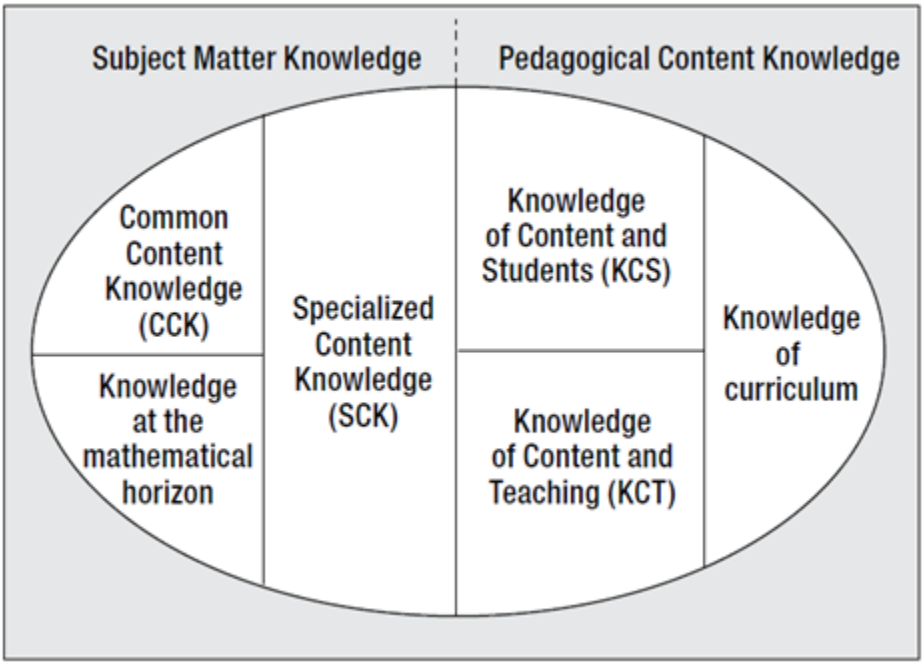


Figure 1. "Domain map for mathematical knowledge for teaching (Hill et al, 2008, p. 377)."

Subject matter knowledge consists of three underlying categories. *Common content knowledge* is mathematical knowledge teachers are responsible for developing in students. An example of this could be a teacher instructing students in the procedures related to hypothesis testing. This might include teaching students how to compute a *p*-value, correctly identifying the null/alternative hypothesis, and how to conclude a hypothesis test procedure. *Specialized content knowledge* is the mathematical knowledge used in teaching, but not necessarily taught to students. An example of this in hypothesis testing is knowledge of advanced probability theory associated with theoretical sampling distributions, which includes knowing the relationship between calculus, probability, and statistics. This type of knowledge is beyond the scope of many introductory statistics students since calculus is not a typical pre-requisite for many introductory statistic courses. However, knowing the relationship between calculus and probability in

hypothesis testing are concepts statistics teachers should have learned in advanced statistics courses. *Knowledge at the mathematical horizon* is a teacher understanding the broader set of mathematical concepts with which a particular idea will connect. Students tend to learn in the moment since they are unaware what material is coming up in future lessons. Teachers, unlike students, should know the bigger picture and the importance of specific concepts. For example, a teacher should know that an understanding of simulations and sampling distributions will play a bigger role in developing students' understanding of statistical inference techniques.

Under the category of pedagogical content knowledge are also three sub-categories. *Knowledge of content and students* is the amalgamated knowledge that teachers possess about how students learn content. This includes knowledge of the developmental stages students go through to understand a concept and the kinds of misconceptions and conceptual hurdles student might encounter. Such knowledge requires both experience and knowledge of research on student learning that directly relates to the mathematical or statistical content being taught. An example of this might include knowing that theoretical sampling distributions might be an area that is beyond the scope of many students whose backgrounds are limited to only college algebra. *Knowledge of content and teaching* is knowledge focusing on the design of instruction. This includes knowing how to choose examples and representations to best guide student discussions toward mathematical ideas. An example of this include structuring task that uses the design simulations to develop an understanding of empirical sampling distributions to further develop their understanding of inferential ideas. *Knowledge of*

curriculum is an understanding on how to sequence and structure the development of a mathematical topic. For example, teachers have to be aware that students need to develop a deep understanding of probability and sampling distributions prior to teaching statistical inference (Hill et al, 2008).

One of the key issues some statistics education researchers (e.g. Groth, 2007; Burgess, 2006) have with the framework of Hill and colleagues (2008) is that it does not account for the differences between mathematics and statistics. Statisticians have argued that there are key differences between the field of statistics and mathematics. Rossman, Chance, and Medina (2006) outline 5 components they see as fundamental differences between mathematics and statistics: (1) the crucial role of context, (2) issues of measurement, (3) importance of data collection (4) variability of data and (5) importance of communication. The work of statisticians is steeped in context taken from the real world. An understanding of the context can lead a statistician to make decisions on the type of analysis and conclusions to solve a statistical problem. Data collection is also important because sampling from a population is difficult and requires statisticians to continually focus on issues of precision and accuracy of data collection (Gould & Ryan, 2013). The omnipresence of variability in data also allows the possibility of invalid conclusions. These differences were acknowledged by the statistics education research community who suggested a re-evaluation of statistical knowledge for teaching (e.g. Cobb & Moore, 1997; Groth, 2007; Rossman, Chance & Medina, 2006; Shaughnessy, 2007). As a result, several researchers have taken steps towards developing potential frameworks for *statistical knowledge for teaching* (SKT) that have either adapted or

extended frameworks originally focused on *mathematical knowledge for teaching* (e.g. Burgess, 2006 & 2008; Gonzalez, 2012; Groth, 2007; Noll, 2007; Sorto, 2006).

One such framework, introduced by Groth (2007), advocates for a modification of the Hill et al. (2008) framework to encompass notions of SKT. An influential piece in Groth's (2007) perspective of SKT was the framework for statistical problems presented in the GAISE report. In his framework, Groth describes the interaction between mathematical and nonmathematical ideas that occur when teaching statistics. The nonmathematical work encompasses understanding variation in data and making inferences based on the probabilistic behavior of data. Furthermore, statistical analysis also includes methods of generating appropriate research questions and developing proper data collection methods in order to create valid conclusions. Once statisticians have data, they can then apply appropriate statistical methods for analysis. Groth (2007) believes that a robust understanding of statistics requires an understanding of the mathematical and nonmathematical side of a statistical process. For teachers, this involves knowing how to set up appropriate statistical activities that not only encompass the importance of theoretical statistics involving mathematical theory (e.g. probability theory), but also non-mathematical methods (e.g. sampling techniques).

Groth's hypothetical framework for SKT (Figure 2) highlights the interrelationship between common content and specialized content knowledge that involve statistical concepts that are both mathematical and non-mathematical. Groth advocates that common and specialized content knowledge in SKT is still a growing area of research for the statistics education community. This has led others in the statistics

education community to focus more deeply on other aspects of SKT (e.g. Burgess, 2008; Gonzalez, 2012; Noll, 2007).

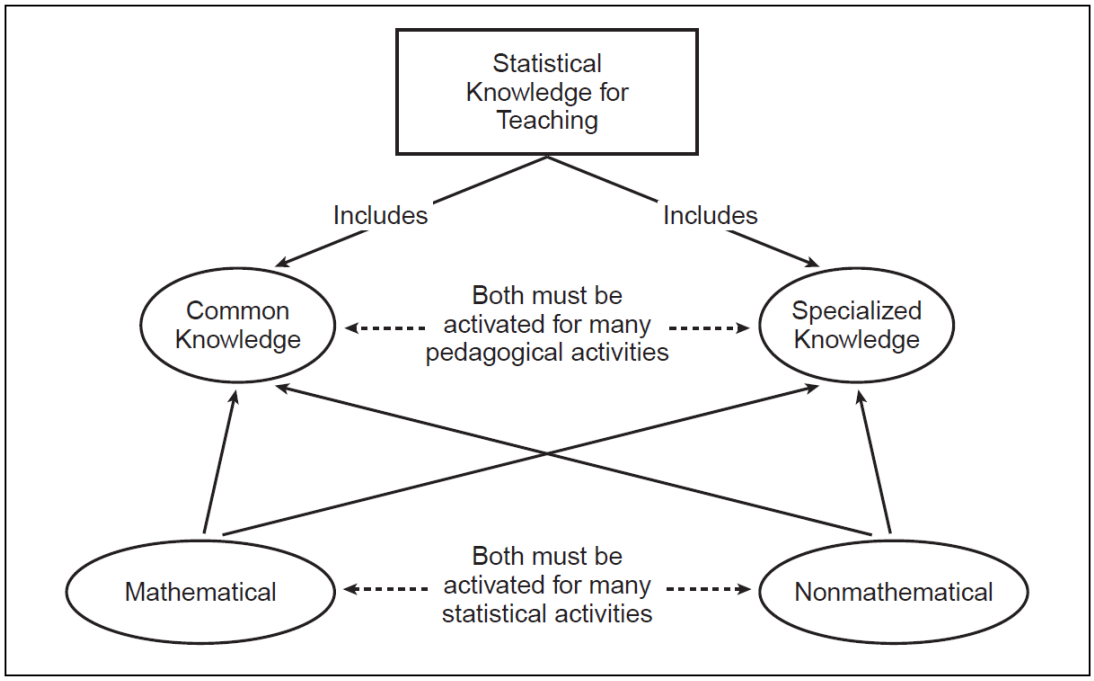


Figure 2. "Hypothesized structure of statistical knowledge for teaching (Groth, 2007, p. 429)."

Gonzalez (2012) work on SKT was also built on the work of Hill et al. (2008) and encompassing the notion of statistical literacy. In his work, Gonzalez expands on Hill et al.'s framework and describes how it might be applied to the concept of statistical literacy by focusing on four potential predictors of SKT: "(1) A model of SKT should be closely tied to a model of MKT; (2) Some knowledge components in the MKT model used must be redefined to meet the requirements for teaching statistics; (3) In order to conceptualize SKT, teachers' beliefs about statistics, teaching and learning must be considered; and (4) Tasks designed to elicit teachers' conceptions of variability would be helpful to provide indicators to measure SKT (Gonzalez, 2012, p. 4-5)." Through these indicators, researchers can help identify how teachers come to understand certain categories of SKT.

Using his expanded framework, Gonzalez (2012) then investigated how to analyze teachers' identification of concepts of variability. Figure 3 illustrates an example of an assessment tool designed by Gonzalez that expands the framework of Ball et al (2008) that builds on the connection of statistical literacy to assess topics of variability.

A: Indicators associated to Statistical Literacy (CCK):

1. Is the teacher able to give an appropriate and correct answer to the given task?
2. Does the teacher consistently identify and acknowledge variability and correctly interpret its meaning in the context of the given task?

B: Indicators associated to SCK:

1. Does the teacher show evidence of ability to determine the accuracy of common and non-standard arguments, methods and solutions that could be provided on a single question/task by students (especially while recognizing whether a student's answer is right or not)?
2. Does the teacher show evidence of ability to analyze right and wrong solutions that could be given by students, by providing explanations about what reasoning and/or mathematical/statistical steps likely produced such responses, and why, in a clear, accurate and appropriate way?

C: Indicators associated to HCK:

1. Does the teacher show evidence of having ability to identify whether a student comment or response is mathematically/statistically interesting or significant?
2. Is the teacher able to identify the mathematically/statistically significant notions that underlie and overlie the statistical ideas involved in the given task?

D: Indicators associated to KCS:

1. Is the teacher able to anticipate students' common responses, difficulties and misconceptions on the given task?
2. Does the teacher show evidence of knowing the most likely reasons for students' responses, misconceptions and difficulties in relation to the statistical ideas involved in the given task?

E: Indicators associated to KCT:

1. In design of teaching, does the teacher show evidence of knowing what tasks, activities and strategies could be used to set up a productive whole-class discussion aimed at developing students' deep understanding of the key statistical ideas involved in the given task, instead of focusing just in computation methods or general calculation techniques?
2. Does the teacher show evidence of knowing how to sequence such tasks, activities and strategies, in order to develop students' deep understanding of the key statistical ideas involved in the given task?

F: Indicators associated to KCC:

1. Does the teacher show evidence of knowing at what grade levels and content areas students are typically taught about the statistical ideas involved in the given task?
2. Does the designed lesson (or series of lessons) show evidence of teacher's understanding and support of the educational goals and the intentions of the official curriculum documents in relation to the teaching of the statistical contents present in the given problem, as well as statistics in general?

Figure 3. "Set of proposed indicators to assess SKT (Gonzalez, 2012, p. 5)."

A final example of a theoretical framework for SKT is introduced by Noll (2007) whose model for SKT also encompasses ideas of statistical literacy, statistical thinking, and statistical reasoning combined with Hill et al.'s (2008) categories of common content knowledge (CCK), specialized content knowledge (SCK), and knowledge of content and students (KCS) (Figure 4). In Noll's framework, she does not view common content knowledge and specialized content knowledge as mutually exclusive of knowledge of content and students. Instead, she emphasizes the important interaction between CCK and SCK. In particular, Noll emphasizes CCK as knowledge of a much wider spectrum whereas SCK is a specific category of CCK. For example, the knowledge needed to teach

hypothesis testing is very broad. For teachers, knowing the procedures to perform hypothesis testing is something that is taught to statistic students and is a form of CCK. Furthermore, knowing the relationship between the hypothesis testing methods and advanced probability is a form of SCK which is typically not taught to students.

Noll (2007) also emphasizes the relationship of CCK and SCK with KCS, highlighting that having a robust understanding (CCK and SCK) of a statistical concept can be useful in understanding student misconceptions and developing alternative pedagogical approaches to support student development. Using her framework, Noll (2006) investigated GTAs' statistical knowledge for teaching on the concepts of confidence intervals and variability. Through task-based interviews using hypothetical student work, she was able to assess GTAs' content knowledge (CCK and SCK) and their knowledge of content and students (KCS).

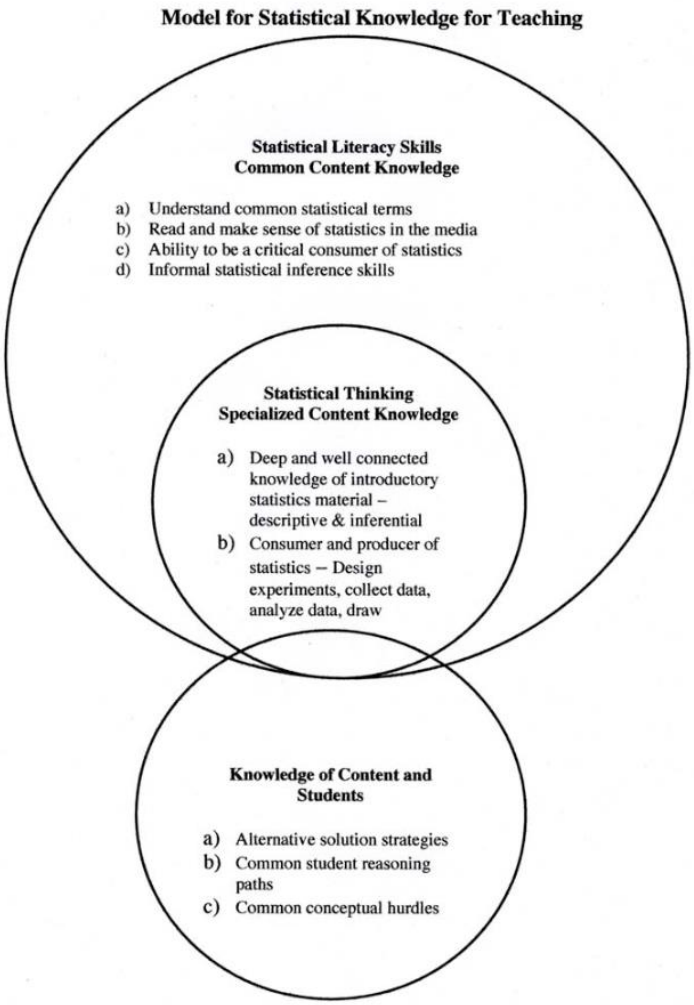


Figure 4. "Model for Statistical Knowledge for Teaching (Noll, 2007, p. 70)."

3.2. Research on Hypothesis Testing in the Classroom

Since the goal of this research is to analyze the statistical knowledge of CCIs and GTAs, it is appropriate to discuss the statistic education community's view of a robust understanding of hypothesis testing and probability appropriate for statistics teachers. Furthermore, it is also important to highlight the types of misconceptions that have arisen for both students and teachers in regard to probability in hypothesis testing and the current approaches by the research community to remedy these issues.

This section begins with background literature on hypothesis testing and the current pedagogical approach for teaching hypothesis testing in traditional statistics classroom. I then present research on student and teacher misconceptions discussed in the statistics education literature, current issues with the traditional approach to hypothesis testing, and potential causes for those misconceptions as highlighted by the statistics education literature. Finally, I discuss the current shifts in the pedagogical approaches discussed by the statistics education community regarding the instruction of hypothesis testing.

3.2.1. Traditional hypothesis testing. Theoretical hypothesis testing started in the early 1900s when statisticians Fisher, Pearson, and Gosset developed what they termed significance testing. *Significance testing* uses probability theory and formal logic to test claims by computing the likelihood of observed sample data. Significance is measured through the likelihood of observed sample data. The more *unusual* the sample, the more *significant* are the results of the test. Significance testing uses conditional reasoning from formal logic (i.e. *modus tollens*) by making a hypothesized assumption about a population, analyzing the probability of an observed sample in light of the assumption and then generating a conclusion (Cordani, 2010).

To analyze the probability of observed data Fisher used theories related to random sampling, probability distributions, and theoretical sampling distributions. A *theoretical sampling distribution* results from randomly sampling *all* possible samples of a fixed sample size from a population and then generating the statistic of interest from each sample to create a distribution of sample statistics. In significance testing the probability

of an observed sample statistic is found using statistical knowledge of an *implicit* theoretical sampling distribution generated by assuming an initial assumption about the population (i.e. null hypothesis) and then applying theories of probability distributions (e.g. normal distributions, Chi-squared, etc.). By using an observed sample statistic (i.e. test statistic) and probability theory, Fisher would compute the theoretical probability of getting a sample as extreme as the observed sample data assuming the null hypothesis and random sampling techniques. This probability is defined as the *p*-value. By using the *p*-value statisticians can decide whether the results of an observed sample can lead one to reject (of fail to reject) the null hypothesis. If the *p*-value is small, then there is strong evidence that the null hypothesis is invalid because the sample data illustrates the chances that something as extreme as the observed sample statistic is highly unusual given the null assumption. If the *p*-value is large, then there is weak evidence against the null hypothesis because the chance of getting a sample as extreme as the observed sample statistic is likely to happen. Since the *p*-value plays such a central role in the decision-making process of significance testing, practitioners of statistics should know how it is generated and what it represents.

To illustrate Fisher's process, consider the following example. Suppose a statistics student wanted to test whether the proportion of males in a college population was still 0.5 using an observed sample of 100 students that exhibited 70 males (Table 1).

Table 1. *Significance Testing Procedure for a College Proportion.*

Steps	**Explanation**
1. State a null hypothesis	Student states a null hypothesis about the population. **Example.** H_0: The proportion of male is 50%.

2. Compute the test statistic	Student uses the sample data to compute a test statistic using a formula specific for the hypothesis test. **Example.** Given a sample of 100 students with a proportion of 60% the test statistic for proportion can be computed using the formula to transform a sample proportion to standardized z-score: $z = \frac{0.7-0.5}{\sqrt{\frac{(0.5)(0.5)}{100}}} = 4$
3. Compute the *P*-value of the sample test statistic	Students use a probability distribution table or calculator to compute the probability of getting a statistic as extreme as the sample test statistic. **Example.** *P*-value = P(z > 4) ≈ 0.
4. Conclude the results of the hypothesis test.	Student uses the information in step 3 to make a decision about the null assumption. This conclusion is either to reject or fail to reject the null hypothesis based on the following criteria. Reject the null hypothesis if: *P*-value is small. Fail to reject the null hypothesis if: *P*-value is large. **Example.** *P*-value = 0 < 0.5. Thus we reject the null hypothesis.
5. Interpret the conclusion.	Student uses the result of step 4 to interpret the results of the hypothesis test. **Example.** Because we rejected the null hypothesis with a 5% level of significance, there is enough information to claim that the population is not equally distributed.

The central idea behind Fisher's approach is the importance of analyzing evidence based on probability of observed data (Rossman, 2008). The more unusual an observed sample is the more significant the evidence is against the null. An issue with Fisher's approach is that rejecting the null hypothesis did not mean one could accept an alternative hypothesis because the original work of Fisher made no claim of an alternative hypothesis (Clauser, 2008). Furthermore, Fisher did not select a fixed value to determine what would be considered an unusual *p*-value. As a result, choosing a value to quantify unusualness for the *p*-value is arbitrarily based on a statistician's choice.[3] This approach differed from the work of other statisticians of the era.

[3] Historically, the choice of 5% was also chosen arbitrarily by statisticians (Clauser, 2008).

Statisticians Jerzy Neyman and Egon Pearson (son of Karl Pearson) coined the concept of hypothesis testing which popularized the use of *null* and *alternative hypothesis*. Neyman-Pearson used much of Fisher's original work on significance testing (i.e. theoretical sampling distribution, probability and assuming a null hypothesis) with the exception that their method of hypothesis testing focused on minimizing errors in a testing process (Rossman, 2008). Significance testing uses two competing hypotheses (a null and alternative) to determine whether one should reject or fail to reject the null hypothesis. The Newman-Pearson's hypothesis test would then conclude with evidence for either the null or alternative hypothesis as the result of the test. In order to do this, a statistician would use theoretical sampling distributions and probability theory similar to Fisher's approach. Unlike Fisher's approach of analyzing a *p*-value, Neyman-Pearson's method focuses on minimizing the probability of committing errors in the hypothesis testing procedure. The errors are false negatives and/or false positives and are shown in Table 2 (Rossman, 2008). To illustrate, suppose the null assumption is true. Neyman-Pearson's method analyzes how to minimize the chances of incorrectly rejecting the null hypothesis. Rejecting the null hypothesis when it is true is defined as Type-I error. The term *level of significance* (α) is reserved for the probability of this error. Another potential error when concluding a hypothesis test is failing to reject the null hypothesis when it is false (i.e. the alternative hypothesis is true). This error is a Type-II error symbolically denoted as β.

Table 2. *Table for Type-I/II Errors in Hypothesis Testing*

	Validity of the Null Hypothesis	
Conclusion	**Null Hypothesis is true**	**Null Hypothesis is False**
Reject the Null Hypothesis	Incorrect Decision (Type-I Error)	Correct Decision
Fail to Reject the Null Hypothesis	Correct Decision	Incorrect Decision (Type-II Error)

The chance of not committing a Type-II error is defined as the *power of the test*, and is computed as 1-β. A common practice in introductory statistics is to focus on methods related to Type-I error but giving little emphasis on Type-II error. In classwork, the probability of committing a Type-I error can be expressed graphically through a sampling distribution and is termed the *critical region.* The value that marks the location of the critical region on the sampling distribution is termed the *critical value* (see Figure 5).

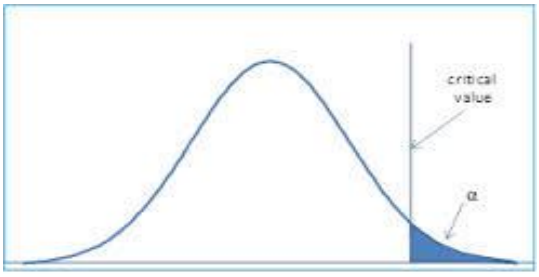


Figure 5. Theoretical sampling distribution with critical value and critical region.

The method of Neyman-Pearson is traditionally called the *Critical Region Approach* of hypothesis testing in introductory statistics classrooms to coincide with the procedure.[4] To illustrate the difference between Fisher's and Neyman-Pearson's method, Table 3 illustrates the same statistical problem as Table 1 using the Critical Region Approach.

[4] Introductory statistics textbooks also refer to this as the traditional method of hypothesis testing.

Table 3. *Critical Region Hypothesis Test for a Population Proportion*

Steps	**Explanation**
1. State null and alternative hypotheses	Student states a null hypothesis and alternative hypothesis about the population. **Example.** H_0: The proportion of male is 50%. H_A: The proportion of male is greater than 50%.
2. Compute the test statistic	Student uses the sample data to compute a test statistic using a formula specific for the hypothesis test. **Example.** Given a sample of 100 students with a proportion of 60% the test statistic for proportion can be computed using the formula to transform a sample proportion to standardized z-score: $z = \frac{0.7-0.5}{\sqrt{\frac{(0.5)(0.5)}{100}}} = 4$
3. Compute the Critical Value	Using a set level of significance, the critical value is found by looking at a table of different critical values based on the probability distribution of the test statistic. **Example.** Critical Value for 5% level of significance is 1.645 found on the standard normal distribution table.
4. Determine the likelihood of the sample test statistic.	Student checks whether the test statistic falls within the critical region set by the critical value computed in step 3. **Example.** z = 4 > 1.645 = critical value.
5. Conclude the results of the hypothesis test.	Student uses the information in step 4 to make a decision about the null assumption. This conclusion is either to reject or fail to reject the null hypothesis based on the following criteria. Reject the null hypothesis if: Test statistic falls in the critical region set by the critical value. Fail to reject the null hypothesis if: Test statistic does not fall in the critical region set by the critical value. **Example.** z = 4 > 1.645 = critical value. Thus we reject the null hypothesis.
6. Interpret the conclusion.	Student uses the result of step 5 to interpret the results of the hypothesis test. **Example.** Because we rejected the null hypothesis with a 5% level of significance, there is enough information to claim that there are more males in the college population.

One of the key benefits of this approach over Fisher's is that the process results in a conclusion whether a statistician rejects or fails to reject the null hypothesis. If a statistician rejects the null hypothesis then he/she has evidence for the alternative

hypothesis, and vice-versa. The process of Neyman-Pearson mainly focuses on the concept of 'errors' by implicitly analyzing the position of an observed sample data in a theoretical sampling distribution. The method as presented in Table 3 does not however show how to deal with Type-II error. In fact, detailed work regarding the computation and minimization of Type-II error is generally left for advanced statistics courses.

It is still common for statistics courses to teach the Neyman-Person method in modern statistics courses as the Critical Region approach.[5] The same cannot be said for Fisher whose method was updated around the 1940s to a hybrid method combining both approaches. The hybrid method is identified by statisticians as *null hypothesis significance testing* (NHST). NHST combines the ideas of error analysis and *p*-value and has been the standard approach to hypothesis testing instruction in statistical classrooms (Clauser, 2008; Stigler, 1999).

Prior to learning methods of hypothesis testing, most introductory statistics curriculum begins with chapters on descriptive statistics (e.g. mean, standard deviation, proportion, etc.) and probability distributions (e.g. binomial distribution, normal distribution, etc.) progressing to a chapter on (theoretical and empirical) sampling distributions. A culmination of these ideas leads to chapters on statistical inference (e.g., Bluman, 2012; Brase & Brase, 2012). A synopsis of traditional introductory statistics textbooks (e.g., Bluman, 2012; Brase & Brase, 2012) covers Critical Region and the NHST as standard curriculum. The process of NHST is broken into six steps: (1) state the null and alternative hypothesis, (2) compute the test statistic of an observed sample, (3) compute the *p*-value, (4) compare the *p*-value with the level of significance (5) conclude

[5] For an example, see Bluman (2012).

the results of the hypothesis test, and (6) interpret the results. Table 4 illustrates an example of NHST that tests for a greater proportion of males in a college population given that an observed sample of 100 students exhibited 70 males.

Table 4. *Null Hypothesis Significance Test for a Population Proportion.*

Steps	**Explanation**
1. State null and alternative hypotheses	Student states a null hypothesis and alternative hypothesis about the population. **Example.** H_0: The proportion of male is 50%. H_A: The proportion of male is greater than 50%.
2. Compute the test statistic	Student uses the sample data to compute a test statistic using a formula specific for the hypothesis test. **Example.** Given a sample of 100 students with a proportion of 60% the test statistic for proportion can be computed using the formula to transform a sample proportion to standardized z-score: $z = \frac{0.7-0.5}{\sqrt{\frac{(0.5)(0.5)}{100}}} = 4$
3. Compute the *P*-value	Students use a probability distribution table or calculator to compute the probability of the given test statistic. **Example.** *P*-value = P(z > 4) ≈ 0.
4. Determine the likelihood of the sample test statistic.	Student checks whether the *P*-value computed is lower than the level of significance. Alternatively, one can check whether the test statistic falls within the critical region set by the critical value computed in step 3. **Example.** *P*-value ≈ 0 < 0.05 = level of significance.
5. Conclude the results of the hypothesis test.	Student uses the information in step 4 to make a decision about the null assumption. This conclusion is either to reject or fail to reject the null hypothesis based on the following criteria. Reject the null hypothesis if: *P*-value < level of significance Fail to reject the null hypothesis if: *P*-value > level of significance **Example.** *P*-value = 0 < 0.5 = significance level. Thus we reject the null hypothesis.
6. Interpret the conclusion.	Student uses the result of step 5 to interpret the results of the hypothesis test. **Example.** Because we rejected the null hypothesis with a 5% level of significance there is enough information to claim that there are more males in the college population.

Students are instructed to use a NHST or Critical Region approach on various hypothesis test problems (e.g. multiple populations, means, standard deviations, etc.). Traditionally, lesson time is spent on learning how to compute a *test statistic*, *p*-value and *critical value* and determining which hypothesis test procedure to use based on the statistical problem (e.g. Bluman, 2012; Brase & Brase 2012). Knowledge of the underlying theoretical sampling distributions and the related probability distribution is briefly discussed, but deeper understanding of the underlying concepts of hypothesis testing is typically reserved for upper division statistics courses. Student achievement is then measured through assessments that evaluate correct computation and procedural fluency. For example, students must correctly identify the null/alternative hypothesis, compute the correct *p*-value, find the critical value and determine whether to reject/fail to reject the null hypothesis. Unfortunately, research reveals that traditional instruction and assessment of NHST has led to misconceptions for both students and teachers (Batanero, 2000; Batanero & Diaz, 2006; Castro Sotos et al, 2007 & 2009; Falk, 1986; Garfield & Ben-Zvi, 2008; Haller & Krauss, 2002; Liu, 2005; Thompson, Liu & Saldahna, 2007; Vallecillos, 2002; Vallecillos & Batanero, 1997).

3.2.2. Research on students' and teachers' understanding of NHST.

Developing a deep conceptual understanding of hypothesis testing is non-trivial for many students and teachers. Statistics education researchers have uncovered a variety of misconceptions regarding students' (e.g. Batanero, 2000; Batanero & Diaz, 2006; Castro Sotos et al, 2007; Castro Sotos et al, 2009; Falk, 1986; Garfield & Ben-Zvi, 2008; Haller & Krauss, 2002; Vallecillos & Batanero, 1997) and teachers' (e.g. Haller & Krauss,

2002; Heid, et al., 2005; Thompson, Liu, & Saldahna, 2007) understanding of NHST. For example, research by Vallecillos and Batanero (1997) revealed that students struggle identifying the null and alternative hypothesis when confronted with statistical problems where the null/alternative hypotheses are not clearly delineated. Another example is research regarding students' understanding of statistical power outlined by Nickerson (2000) who asserts that students tend to think Type-I and Type-II errors as the same and/or at times confuse the relationship between them. For example, students would interpret the probability of not committing a Type-I error as the probability of committing a Type-II.

Empirical research on teachers' understanding of hypothesis testing is seen in the work of Thompson, et al. (2007) who investigated eight secondary school teachers in a professional development seminar on statistical inference. The results of their research uncovered that even if teachers understood the logic of hypothesis testing, teachers had a tendency to misinterpret its use. For example, when teachers in the study were asked questions related to hypothesis tests, seven of the teachers applied methods that presumed they had access to the population. Thompson et al. noted that even if the teachers understood the logic of hypothesis testing, some fail to demonstrate understanding of its functionality.

The examples above illustrate just a few of the misconceptions found in the research literature on students' and teachers' understanding of hypothesis testing. Since the goal of this paper is to primarily focus on understanding of probability in the context of hypothesis testing, two important misconceptions are highlighted as areas in the

research students and teachers struggle with when understanding hypothesis testing: (1) misunderstanding the role of conditional probability in hypothesis testing concepts (e.g. p-value and level of significance) and (2) misconceptions regarding the role of sampling distributions in hypothesis testing.

3.2.2.1. The role of conditional probability in hypothesis testing. One of the common misconceptions for students and teachers is the inherent role of conditional probability in hypothesis testing (Garfield & Ben-Zvi, 2008). For example, simply interpreting the *p*-value as 'the probability of rejecting the null hypothesis' or 'the probability the null hypothesis is true' are cases where the *p*-value is not viewed as a conditional probability. Even if a student or teacher is able to recognize *p*-value and level of significance as conditional probabilities, some struggle identifying the assumption and conclusions of the conditional statement (Batanero, 2000).

The fact that Students and teachers struggling to understand the role of conditional probability in hypothesis testing is not surprising considering that research has shown that understanding theoretical conditional probability is quite difficult (Shaughnessy, 2007; Watson & Moritz, 2002). Unfortunately, the treatment of hypothesis testing in several introductory statistics textbooks uses methods based primarily on theoretical sampling distributions (e.g. Normal Distribution, Chi-Squared Distribution, Student-t Distribution, etc.) and conditional probability. Furthermore, the actual theory and reasoning behind the computation of the probability in many of these methods requires knowledge of calculus and advanced probability courses. This is generally not required for introductory statistics courses. As a result, almost all probability theory

related to hypothesis testing is implicit in methods presented in introductory statistics textbooks. To illustrate, traditional introductory statistics textbooks approach the computation of probability in hypothesis testing by transforming the observed sample to a statistic (e.g. *z*-score or *t*-score) utilizing the null assumption (e.g. Table 4). After transforming the observed sample to a statistic, the traditional approach is to symbolically represent the *p*-value as a standard probability statement. That is, choosing to represent the *p*-value symbolically as $P(data)$ rather than $P(data \mid hypothesis)$. The computation of probability is then done using either technology or statistical tables (e.g. standard normal distribution table). As a result, the role of conditional probability is implicit in the method which could potentially be misinterpreted by the student or teacher (Shaughnessy & Chance, 2005). Thus, it is no surprise that students and teachers struggle recognizing the *p*-value (or level of significance) as a conditional probability since it is not symbolically represented as one in standard curriculum.

There have been several empirical studies focused on students' misconception of the level of significance. For example, a common misconception regarding the level of significance is students swapping the assumption and conclusion of the conditional in the interpretation, that is, the level of significance is the probability that the null hypothesis is true once the decision to reject it has been taken (Batanero, 2000; Batanero & Diaz, 2006; Castro Sotos et al, 2009, Vallecillos, 2002). Examples of this can be found in work by Vallecillos (2002) who studied college students understanding of hypothesis testing. In her work, she surveyed 436 students and found that 53% were incorrectly interpreting the level of significance, stating that "A level of significance of 5% means that, on average, 5

out of every 100 times we reject the null hypothesis, we shall be wrong." Under the same study Vallecillos (2002) did follow-up interviews with seven individuals considered excellent statistics students. She found that majority of the students incorrectly interpreted the level of significance as 'the probability of the null hypothesis being true given that it has been rejected.' A similar phenomenon was seen in Castro Sotos et al. (2007) who believe the misconception of the level of significance could be linked to students having a deeper misconception inherited by probability (i.e. conditional probability). In their research, Castro Sotos et al. outline two other misconceptions they claim may be due to the conditional nature of a level of significance: (1) level of significance as the probability that one of the hypotheses is true; and (2) the level of significance is the probability of making a mistake.

The misconception of the conditional nature of probabilities in hypothesis testing can also be found in students' interpretation of the *p*-value. Batanero (2000) observed that students who are able to calculate a correct *p*-value and correctly conclude whether to reject or fail to reject the null hypothesis based on their calculated *p*-value struggled interpreting the *p*-value. A survey of college statistics courses by Hauler and Krauss (2002) found that out of 44 students, 30 believe that the *p*-value is the probability that you made a wrong decision given that you reject the null hypothesis. Garfield and Ben-Zvi (2008) have also identified several student misinterpretations of *p*-value in the research literature which include: (1) "The *p*-value is the probability that the null hypothesis is true; (2) The *p*-value is the probability that the null hypothesis is false; (3) A small *p*-value means the results have significance (statistical and practical significance

are not distinguished); (4) *p*-value indicates the size of an effect (e.g., strong evidence means big effects); (5) Large *p*-value means the null hypothesis is true, or provides evidence to support the null hypothesis; (6) If the *p*-value is small enough, the null hypothesis must be false" (p. 270).

Unfortunately, the misconceptions regarding the *p*-value are not confined to statistics students. Statistics education researchers have discovered teachers struggling with hypothesis testing (Haller & Krauss, 2002; Heid, et al., 2005; Thompson, Liu & Saldahna, 2007). Some researchers (e.g. Brewer, 1985; Gliner, Leech & Morgan, 2002; Haller & Krauss, 2002) even claim that textbooks may have caused some of the misconceptions, which are then exemplified by teachers who use such textbooks in their work. In an example presented by Haller and Kraus (2002), they illustrate one such book entitled *Introduction to Statistics for Psychology and Education* by Nunally (1975). In their example, they outline the following interpretations put forth by the author on the interpretation of NHST on the concept of *p*-value:

1. "the improbability of observed results being due to error"
2. "the probability that an observed difference is real"
3. "if the probability is low, the null hypothesis is improbable"
4. "the statistical confidence ... with odds of 95 out of 100 that the observed difference will hold up in investigations"
5. "the degree to which experimental results are taken 'seriously'
6. "the danger of accepting a statistical result as real when it is actually due only to error"

7. "the degree of faith that can be placed in the reality of the finding"
8. "sample mean actually differs from the population mean" (Nunally, p. 194-197).

Hauler and Krauss highlight that the author (Nunally) concludes these statements all define ways of describing the *p*-value even though they are all inherently wrong. In the same study, Haller and Krauss (2002) investigated statistics teachers' knowledge of the *p*-value. In their study, they investigated 30 university instructors, 39 scientific psychologists, and 44 psychology students to not only study the subjects' understanding of *p*-value, but also to compare whether misconceptions are shared amongst them. In their study, they provided the subjects with a survey shown in Figure 6.

Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means t-test and your result is ($t = 2.7$, d.f. $= 18$, $p = 0.01$). Please mark each of the statements below as "true" or "false". "False" means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

1) You have absolutely disproved the null hypothesis (that is, there is no difference between the population means). [ ] true / false [ ]

2) You have found the probability of the null hypothesis being true. [ ] true / false [ ]

3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means). [ ] true / false [ ]

4) You can deduce the probability of the experimental hypothesis being true. [ ] true / false [ ]

5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision. [ ] true / false [ ]

6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions. [ ] true / false [ ]

Figure 6. "Questionnaire on the interpretations of a p-value (Haller & Krauss, 2002, p. 5)."

In their work, they indicate that the correct response for all of these statements should all be false. A breakdown of subjects who made at least one mistake consisted of 80% statistics instructors, 100% of students and 89.7% of scientists. The greatest frequency of mistake across all three groups was found to be statement 5, which shares a similar definition to a Type-I error; illustrating that students, teachers, and researchers sometimes struggle differentiating between level of significance and *p*-value.

3.2.2.2. The role of sampling distributions in hypothesis testing. One of the most discussed topics in statistics education is students' and teachers' understanding of

sampling distributions and its role in statistical inference. Kahneman and Tversky (1982) conjectured that individuals tend to take a *singular* rather than a *distributional* perspective when making judgements about the probability of a single outcome. A *singular* perspective is one in which a person focuses on the causal system that produced the particular outcome and bases probability strictly on the propensities of the outcome. A *distributional* perspective relates the outcome to a sampling schema and views an individual outcome as a single instance of a set of similar cases where probability can be as estimated relative frequencies (Kahneman & Tversky, 1982; Saldahna & Thompson, 2002). Empirical evidence of this can be found in the work of Konold et al. (1993) who found that individuals tend to think that they can predict with *certainty* the outcome of an *individual* trial of an experiment. Konold et al. describe this as an *outcome approach,* where an individual tends to base predictions of uncertain outcomes on causal explanations instead of information gathered from repeating an experiment.

Even if students have grasped the concept of repeating a sampling process, misconceptions still arise regarding their understanding of a sampling distribution. Empirical research conducted by Saldahna and Thompson (2002) created an instructional sequence designed to support students' understanding of sampling distributions by connecting ideas of repeated selection, variability, and distribution in a course for secondary school students. In their research, they discovered that students had two different views of samples. The first is an *additive conception of sampling* where students view a sample as a subset of a population and multiple samples as multiple subsets. The second is a *multiplicative conception of sampling* where students viewed sample as "a

quasi-proportional mini-version of the sampled population, where the "quasi-proportional" image comes from anticipating a bounded variety of outcomes, were one to repeat the sampling process (Saldahna & Thompson, 2002, p. 6)." The researchers claim that having a multiplicative conception of sampling supports a deep understanding of statistical inference and is common among statisticians. In particular, it helps students explain *why* an outcome has a particular probability because it familiarizes them towards relating individual outcomes to a distribution of a class of similar outcomes.

The research regarding sampling distributions have led many researchers to believe that misconceptions in statistical inference arise because students and teachers lack the knowledge of sampling distribution and its role in statistical inference (e.g., Heid et al, 2005; Makar & Confrey, 2004 & 2005; Saldahna & Thompson, 2002; Thompson et al., 2007). Almost all statistical textbooks (e.g. Bluman, 2012; Brase & Brase, 2012) dedicate a few sections to (theoretical and empirical) sampling distributions, but focus on statistical inference methods that rely strictly on theoretical sampling distributions. Traditional approaches of hypothesis testing use theoretical sampling distributions even though research has shown that understanding a theoretical sampling distribution is difficult for students and teachers (e.g., Chance, delMas & Garfield, 2004; Lipson, 2002; Saldahna & Thompson, 2002; Thompson et al, 2004). Chance et al. (2004) writes that understanding sampling distributions is challenging because it requires an integration of several statistical concepts (e.g. random sampling, statistic, parameter, variability, probability distribution, etc.). Furthermore, Chance et al. states that those students who fail to develop a deep understanding of sampling distributions often develop mechanical

knowledge of statistical inference. This is troubling considering that sample distributions play a major role in conceptually understanding the relationship between statistical inference and probability.

Empirical evidence of teachers struggling to understand the role sampling distributions play in statistical inference can be seen in the work of Makar and Confrey (2005). In their work, they studied secondary teachers at the end of a professional development sequence by giving tasks that had them compare two sampling distributions of data to make inferences about two populations. During the professional development, the teachers often used sampling distributions to investigate statistical concepts, but the interviews discovered that teachers still relied on descriptive statistics to compare distributions rather than inferential techniques. Some teachers who considered using a sampling distribution for the comparison tasks struggled to understand how it would be useful. This phenomenon was also observed in the work of Heid et al. (2005) who investigated eighteen prospective secondary teachers enrolled in a course designed to broaden understanding of statistical concepts. In their research, they discovered that participants could accurately describe and construct a sampling distribution but were inconsistent when trying to articulate distinctions between the distribution of a sample and a sampling distribution. As a result, the teachers in the study made deterministic statements rather than probabilistic inferences when reasoning through problems related to hypothesis testing.

One of the biggest issues surrounding probability in hypothesis testing is students and teachers not recognizing that samples should be viewed as part of a stochastic

process.[6] The goal of statistical inference is to analyze the probability of an observed sample as one potential sample in a set of possible samples drawn from a population. Liu and Thompson's (2004) research on probability and statistical inference with secondary school educations in a professional development program highlight that teachers' understanding of probability covers a large spectrum. They discovered that teachers in their study viewed probability in four ways:

1. "thinking that probability is a (subjective) judgment based on personal experiences;
2. thinking that probability is about predicting the state of a specific completed (or to-be- completed) event about which one does not know the actual result;
3. thinking that probability is about selecting one outcome from a set of possible outcomes;
4. thinking that probability is about imagining a collective of results all generated by a single-process that yields results that are more dense in some regions of possible values than in other regions, i.e., a stochastic conception of probability" (Liu & Thompson, 2004, p. 4).

According to Liu & Thompson, the representation of a probability question directly influences how teachers interpret a probability problem. In their work, they found that when a probability statement states a collection of people, teachers had a tendency to interpret probability as a group characteristic. When a probability question is stated as a single event, teachers were less likely to have images of a collection of similar events,

[6] A *stochastic process* is a collection of random variables indexed by a variable parameter (which is usually time).

illustrating that some teachers do not always view probability as a stochastic process. This is problematic considering that probability in hypothesis testing is conceptually based on a random sampling process and sampling distributions.

3.2.3. New pedagogical approaches to hypothesis testing. The previous sections highlight several misconceptions held by students and teachers regarding the traditional approach of hypothesis testing. Many of these misconceptions involve concepts of probability and/or a lack of understanding the role sampling distributions and random sampling play in hypothesis testing. This has led researchers to develop pedagogical approaches to hypothesis testing that emphasize the importance of simulations and sampling distributions in the understanding of hypothesis testing.

One of the key ideas highlighted by the statistics education literature is the importance of sampling distribution and simulations as a basis for understanding hypothesis testing (e.g., Chance, delMas & Garfield, 2004; Lipson, 2002; Saldahna & Thompson, 2002; Thompson et al, 2004). Teachers should be able to distinguish between a *theoretical* and *empirical* sampling distribution and know their roles in hypothesis testing. In traditional statistics textbooks, the theoretical sampling distribution plays a large role in the computation of probability in hypothesis testing. Teachers of statistics should understand that although the hypothesis test requires the use of well-known probability distributions (e.g. normal distribution) that is a direct result of an implicit sampling distribution. To determine the likelihood of an observed sample statistic requires teachers and students to understand that probability is a result of a stochastic process. The statistics education community has recommended pedagogical approaches

that emphasize empirical sampling distributions and simulations to improve the instruction of hypothesis testing (e.g., Chance, del Mas & Garfield, 2004; Chance, B. et al., 2007; Cordani, 2010; delMas, R., Garfield, J. & Chance, B., 1999; Diaz, 2010; Erickson, 2006; Garfield & Everson, 2009; Garfield & Zvi, 2008; Heid, 2005; Zieffler, Garfield, delMas & Reading, 2008; Zieffler, Garfield, et al., 2008).

In place of a traditional pedagogical approach of hypothesis testing, the statistics education community has focused on three interrelated strategies. First is emphasizing the importance of sampling distributions in the learning of hypothesis testing (Batanero, 2000; Lipson, 2003; Liu & Thompson, 2005; Heid, 2005; Makar & Confrey, 2005) arguing that sampling distributions form the basis for understanding the relationship between sample statistics and probability in a hypothesis testing argument. The second pedagogical strategy recommended by educators (e.g., Chance, del Mas & Garfield, 2004; Chance, B. et al., 2007; delMas, R., Garfield, J. & Chance, B., 1999; Dolor, 2013; Dolor & Noll, 2015; Erickson, 2006; Garfield & Everson, 2009; Garfield & Zvi, 2008; Heid, 2005; Weinberg, Wesner & Pfaff, 2011; Zieffler, Garfield, delMas & Reading, 2008; Zieffler, Garfield, et al., 2008) is that simulations can be an important pedagogical tool in teaching hypothesis testing because they provide a visual and/or physical experience of the sampling process to generate *empirical sampling distributions*. Unlike *theoretical sampling distributions, empirical sampling distributions* can be generated through computer simulations. Chance et al. (2004) recommends that simulations can build foundational ideas of sampling distributions by allowing students to explore how samples and sampling distributions behave with respect to a population. The strategies of

focusing on empirical sampling distributions and simulations lead towards the third strategy which is the teaching of hypothesis testing that stresses less on memorizing procedural ideas and more towards instructional approaches that focus on building students' *informal inferential reasoning* (Zieffler et al, 2008).

The definition of *informal inferential reasoning* (IRR) is based on the work of Zieffler et al. (2008) that they describe "as a way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples (p. 44)." Instead of starting with formal concepts (e.g. *z*-test or *t*-test), statistics educators (e.g., Anderson-Cook & Dorai-Raj, 2003; Chance et al., 2004; Chance & Rossman, 2006; Erickson, 2006; Garfield & Everson, 2009; Garfield et al., 2012; Lawton, 2009; Makar & Rubin, 2009; Pfannkuch, 2010; Pfannkuch, Forbes, Harraway, Budgett & Wild, 2013; Schneiter, 2008; Zieffler et al., 2008) have implemented new approaches where students begin with experiencing the sampling process by generating multiple samples (physically or electronically) to generate *empirical sampling distributions* as a pedagogical tool that is concrete for the student. Figure 7 illustrates the role of sampling distributions in a hypothesis testing process that student could build using computer simulations. Teachers could then generate lessons that connect ideas of probability in statistical inference using an *empirical sampling distribution*.

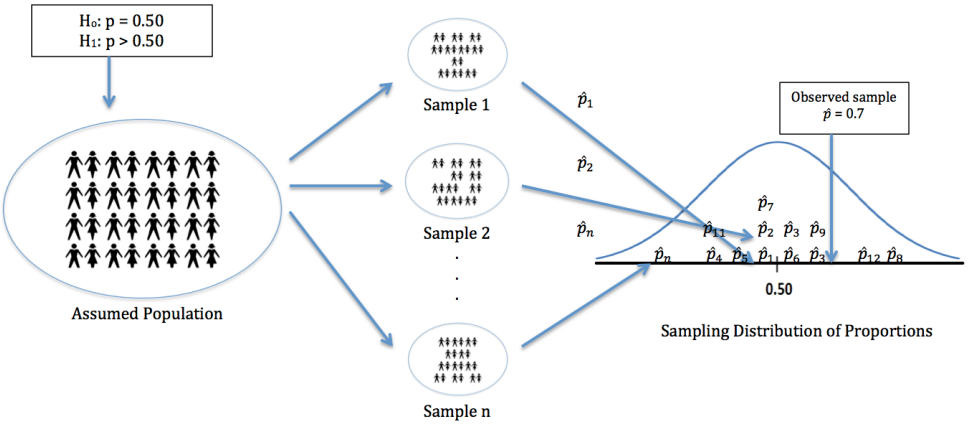


Figure 7. "Model of Hypothesis Testing with Empirical Sampling Distributions (Dolor & Noll, 2015, p. 64)."

Instead of relying on *theoretical sampling distributions*, students can use *empirical sampling distributions* to compute probability by observing and/or counting the number of statistics that fall within a given region (i.e. relative frequency) that can lead to concepts like the *p*-value and level of significance without using formal terminology. According to Zieffler et al, (2008) the reason students often find difficulty with formal statistical inference is because "they lack both experience with stochastic events that form the underpinnings of statistical inference and the experience of reasoning about these events" (p. 45).

With the statistics education community encouraging more simulation-based approaches for the teaching statistics, new curriculums have emerged acknowledging these new pedagogical approaches. Examples of this can be seen in the work of Weinberg et al. (2011) whose work highlights activities that build ideas of statistical inference through sampling activities that gradually introduces students to the mechanics of

hypothesis testing. Erickson (2006) suggests an approach where students can build formal concepts in hypothesis testing using simulation through five steps:

1) "Collect Data from the situation of interest – data that seem to reflect some phenomenon.
2) Design a *measure* of that phenomenon that you can calculate from the data. Ideally, this measure is a larger number if the phenomenon is strong and present, and small-even zero-when the phenomenon is absent. The value of this measure, using real data, is the *test statistic.*
3) Simulate the condition of the null hypothesis, and collect those data using repeated trials?
4) Compute the measure from simulated data, and repeat to build up a sampling distribution for the measure in question.
5) Compare the test statistic to the sampling distribution. The empirical *p*-value is the fraction of cases in the sampling distribution that are at least as extreme as the test statistic (Erickson, 2006, p. 3)."

According to Erickson (2006), this process supports students' development of concepts like the *test statistic* because it gives them the freedom to construct their own measures that make sense to them. Because the measure is by their own construction, simulations and sampling distribution could potentially become more accessible to students, which could lead to a deeper understanding of concepts such as a *p*-value. Substantial work can be seen from the research of Garfield, delMas, & Zieffler (2010 & 2012) who designed a new curriculum entitled Change Agents for Teaching and Learning Statistics

(CATALST) curriculum. In their curriculum, they introduce a simulation based approach to statistical inference where students must model statistical problems with respect to a specific context, repeatedly simulate data, and then use the resulting distribution to draw statistical inferences. By immersing students into the process of simulations, the class builds ideas on informal statistical reasoning that lead towards methods of formal statistical inference. The Comprehensive Assessment of Outcomes in Statistics (CAOS) test was used to assess learning outcomes for this course (delMas, Garfield, Oooms, & Chance, 2007).

There has been some empirical research regarding using simulation-based approaches in the statistics education community (Dolor & Noll, 2015; Erickson, 2006; Weinberg et al., 2010; Garfield, delMas, & Zieffler, 2012). For example, the work of Dolor and Noll (2015) implemented a curriculum similar to the model proposed by Erickson (2006) in a professional development course for secondary school statistics teachers. In their research, they developed activities using the methodology of guided reinvention to support teachers' reconstruction of formal concepts of hypothesis testing. Using the model of a goodness-of-fit hypothesis test as a basis for learning activities, the researchers had teachers generate their own measure to analyze the *unusualness* of samples generated from a population of four-color beans. A result of their research discovered that teachers who were able to generate their own measure were able to fluently discuss concepts related to sampling distributions in a hypothesis-testing context.

Currently, there is a need in the statistics education community to study the impact of a simulation-based approach of hypothesis on student learning. With the

statistics education community slowly transitioning to these new ideas, it is important for the research community to study the impact these pedagogical approaches on students' statistical understanding. It is also important to know whether statistics teachers possess the knowledge to teach using these new approaches.

4. THEORETICAL FRAMEWORK

The goal of this work is to study CCIs' and GTAs' knowledge of probability in the context of hypothesis testing. To accomplish this, it is important to discuss a theoretical model for analyzing the SKT that can be used to assess CCIs and GTAs. This section begins with a description of a theoretical framework that I use to analyze teachers' understanding. I then discuss the concept of a *p*-value by focusing on categories of understanding emphasized by the statistics education literature. During this process, I also discuss how the categories of understanding for a *p*-value relate to SKT.

4.1. A Theoretical Framework for Assessing SKT for p-value.

The literature review in the previous section on the various frameworks for statistics teachers' knowledge showed the overwhelming importance of statistical literacy, statistical thinking, and the framework of Hill et al. (2008) on the statistics education community's view of a statistics teacher. Noll's (2007) and Gonzalez's (2012) frameworks are of particular relevance to this research in regards to the importance of assessing statistics teachers' knowledge of probability related to hypothesis testing. This research hopes to expand the theoretical framework described by Noll (Figure 4) that highlights the strong connection between content knowledge (i.e. CCK and SCK) and its impact on pedagogical knowledge (i.e. KCS). Gonzalez's (2012) framework also highlight important indicators regarding the categories of CCK, SCK, and KCS that could be useful in determining a teacher's understanding of a statistical concept. I conjecture that by merging the frameworks of Noll and Gonzalez researchers can generate assessment tools to analyze teachers' knowledge of hypothesis testing concepts. Table 5

illustrates my proposed theoretical framework for assessing the categories of CCK, SCK, and KCS highlighted by Noll's framework merged with the indicators discussed by Gonzalez (2012) focusing on probability in hypothesis testing particularly related to the *p*-value.

Table 5. *Theoretical Framework for Analyzing Categories of SKT for the P-value.*

Category	**Indicators**
Common Content Knowledge	1. Is the teacher able to appropriately give a correct answer to questions related to *p*-value? 2. Does the teacher consistently identify and interpret concepts of a *p*-value in a statistical task?
Specialized Content Knowledge	1. Does the teacher have the ability to determine standard and non-standard approaches to arguments, methods, and solutions for statistical questions regarding a *p*-value? 2. Does the teacher have the ability to provide evidence to analyze correct and incorrect solutions given by students and provide justifiable explanations and reasoning for responses that are clear and accurate?
Knowledge of Content and Student	1. Is the teacher able to anticipate students' common responses and misconceptions for a statistical task related to the *p*-value? 2. Does the teacher show evidence of knowing the most likely reason for students' responses and misconceptions in regarding tasks related to *p*-value?

To illustrate the merging of the two frameworks (i.e. Noll and Gonzalez), I first focus on the category of CCK. Gonzalez (2012) describes a potential indicator for CCK occurs when a teacher can consistently identify, correctly compute, and interpret statistical concepts. Indicators of CCK for a *p*-value assess when a teacher can correctly and consistently answer questions related to a *p*-value that he/she would expect

introductory statistics students to answer. This includes knowing the definition of a *p*-value and how to compute one when presented with related statistical task. Indicators of SCK focus on content knowledge that is beyond what is typically taught in introductory statistics courses. Indicators of SCK include knowing alternative approaches to computing and interpreting a *p*-value which may include using an empirical sampling distribution as a way to compute and describe a *p*-value. The teacher must also have the knowledge to describe the reasoning behind statistical concepts using advanced statistical knowledge. This includes knowing the relationship between the *p*-value and calculus or describing why a *p*-value can be symbolized as a conditional probability. Indicators of KCS focus on anticipating and recognizing common student responses and misconceptions. This includes identifying common student misinterpretations of the *p*-value such and recognizing incorrect student usage of a *p*-value in a hypothesis-testing task.

It is important to note that categories of knowledge (CCK, SCK, and KCS) tend to overlap when analyzing a *p*-value. For example, recognizing and appropriately responding to an incorrect student interpretation of the *p*-value requires knowledge of KCS, SCK, and CCK since this requires knowledge of the correct interpretation and the ability to explain why an interpretation is incorrect. I theorize that even though all forms of SKT might be active when a teacher tries to recognize an incorrect interpretation, particular forms of knowledge can still be assessed using carefully designed tasks. For instance, the teacher might be able to correctly recognize a correct interpretation of a *p*-value (CCK), but he/she might not be able to explain why using concepts from advanced

statistics (SCK). Similarly, the teacher might recognize a common student misinterpretation of a *p*-value (KCS), but might have difficulty explaining why it is a misinterpretation (CCK and/or SCK).[7] Using my hypothesized framework and carefully designed assessment questions, I generate an initial model of CCIs and GTAs understanding of probability in hypothesis testing focusing on categories of CCK, SCK, and KCS.

4.2. Categorizing SKT for *P*-value.

In order to use the hypothesized framework to assess SKT for CCI and GTAs, it is important to discuss an understanding of probability in hypothesis testing. Since one of the research goals is to assess teachers' understanding of *p*-value, this section outlines important ideas on the understanding of the *p*-value. The work of Shaughnessy and Chance (2005) highlight questions commonly raised by statistics students and teachers on concepts related to the *p*-value. These include: (1) computing the *p*-value through sampling distributions (empirical or theoretical); (2) the interpretation of the *p*-value; (3) understanding *p*-values as tail probabilities and measures of extremes; (4) understanding the *p*-value as a conditional probability under the null hypothesis. I also add to this list the understanding of the *p*-value and its relationship to the level of significance. This includes knowing that *p*-values are compared to the level of significance to determine the conclusion of hypothesis testing as commonly taught in statistics textbooks (Bluman, 2012; Brase & Brase, 2012). Based on the work of Shaughnessy and Chance (2005) and research from the statistics education literature, Table 6 characterizes the categories of

[7] More detailed examples of CCK, SCK, and KCS will be described in the upcoming section where I discuss particular properties of the *p*-value.

the *p*-value that are the focus of my assessment. In these upcoming sections, I discuss how these categories relate to the proposed theoretical model in Table 5.

Table 6. *Categories of Understanding for the P-value.*

1. The definition (interpretation) of the *p*-value.
    - Understanding the assumptions and conclusions of the conditional probability for *p*-value.
    - Understanding *p*-values as measures of extremes results.

2. The symbolic representation of a p-value.
    - The standard representation of the *p*-value using the test statistic formula.
    - The representation of the *p*-value using the conditional statement.

3. The representation of a p-value in an empirical/theoretical sampling distribution.
    - The computation of the p-value.
    - The graphical representation of the p-value as one/two-tail probabilities.

4. The relationship of p-value with the level of significance.
    - The magnitude of the *p*-value in relation to the level of significance.
    - Understanding the relationship of the *p*-value with the level of significance in the hypothesis testing process.
    - Understanding the relationship between a result having statistical significance versus practical significance.

4.2.1. Category 1 – Definition of a p-value. The first category is the *p*-value's definition. In order to assess CCK for a *p*-value, understanding should focus on a teacher's ability to state (or identify) the *p*-value's definition. This includes correctly identifying and interpreting how the *p*-value relates to the null assumption and the observed sample in a hypothesis test problem. This also includes recognizing the relationship between the *p*-value definition and concepts of conditional probability. Furthermore, the teacher must understand that *p*-values refer to a set (or region) of sample statistics (not a single sample statistic) because it is defined as the probability of 'extremes.'

Assessing a teacher's SCK for a *p*-value can be determined by their understanding of a *p*-value's role in a stochastic process. For example, knowing that p-value measures extremes are important when relating *p*-value to the concept of a continuous random variables (e.g. sample means). Conceptually, the probability of a sample statistic for a continuous random variable is theoretically 0 because computation of probability is viewed as an area under a curve for a probability density function (i.e. integration). This is not the case for data whose underlying sampling distribution is a discrete random variable (e.g. counts). This is why statisticians focus on notions of extremes rather than the probability of a single sample statistic. Notions of extreme sample statistics also connect to an understanding of sampling distributions, which I discuss in the section that relates a *p*-value to empirical and theoretical sampling distributions.

To assess a teacher's KCS for a *p*-value definition means analyzing a teacher's ability to identify correct/incorrect student interpretation of the *p*-value. Common misinterpretations highlighted by the statistics education literature include viewing the *p*-value as the probability of a single event (i.e. the probability of the observed sample statistic), not recognizing the *p*-value as a conditional probability, and incorrectly identifying the assumption and conclusion in the interpretation. Useful tasks for this type of assessment may include giving a teacher set of student definitions and interpretations, analyzing how the teacher reacts to these statements.

4.2.2. Category 2 – Symbolic representation of a p-value. The second category is the symbolic representation of the *p*-value. Symbolically, the p-value can be represented either with or without a conditional statement. For example, in Table 3

(NHST for Population Proportions), the *p*-value is represented using a *z*-score and is given a symbolic representation of $P(z \geq 4)$. This representation of the *p*-value is the standard approach in introductory statistical textbooks commonly taught to students. This approach assumes the student and teacher knows the null assumption is an important component for computing the *z*-score and the overall *p*-value.[8] Since this representation is part of standard coursework, assessing CCK for statistics teachers can be determined by the teacher's ability to correctly identify this representation.

An alternative symbolical representation of the *p*-value is P(data | hypothesis). Reflecting on the example presented in Table 3, a teacher might choose to represent the *p*-value as $P(X \geq 0.7 \mid p_0 = 0.5)$ where X is a random variable and p_0 is the assumed population proportion. This representation explicitly shows the *p*-value as a conditional probability and identifies the role of the null hypothesis and observed sample in the conditional statement. This representation closely resembles the interpretation of the *p*-value as a conditional probability that assumes the null assumption and uses the sample statistics discussed in the statistical problem. The second representation requires a teacher to know the inherent role of conditional probability in the *p*-value's definition. The knowledge of the second representation is a component of SCK because discussion of the *p*-value as a conditional probability is typically discussed in advanced statistics courses. Furthermore, introductory statistic teachers uncommonly teach the second symbolic representation of the p-value. Therefore, tasks assessing SCK would focus on the ability of a teacher to recognize and relate both symbolic representations.

[8] This is typical in traditional hypothesis testing because computation of probability is done after the observed sample is transformed using a formula (e.g. *z*-score, *t*-score, etc.).

A common misconception with the second representation of a *p*-value is the inability to recognize the *p*-value as a conditional probability. Even if a teacher is able to represent a *p*-value as a conditional statement, students and teachers struggle identifying the assumption and conclusion of the conditional statement. Knowledge of these misconceptions aligns with KCS because they are related to student mistakes found throughout the statistics education literature. Other misconceptions common in introductory statistics related to the first representations is *p*-value as an equality symbol rather than an inequality (i.e. $P(Z = 4)$ rather than $P(Z \geq 4)$) or the inability to represent the *p*-value as a valid probability statement. Thus, assessment questions on CCK and SCK for the symbolic representation of a *p*-value should focus on the teacher's ability to correctly identify symbolic representations and discuss their relationships, while questions on KCS should assess the teacher's ability to recognize and describe incorrect symbolic representations of the *p*-value in a statistical task used by students.

4.2.3. Category 3 – Relationship of a p-value with sampling distributions. The third category is the relationship between the *p*-value and theoretical/empirical sampling distributions. The typical images of *p*-values are illustrated as either one-/two-tailed regions of a sampling distribution, which relate to regions of *unusualness* or *extremes.* CCK of this involves an understanding of the *p*-value under a theoretical sampling distribution. Figure 8 is an example of a one-tailed *p*-value region using a theoretical sampling distribution whose random variable is normally distributed (typical when performing hypothesis tests on means or proportions).[9] The *p*-value can then be computed using either statistical tables or with computers/calculators. Construction of

[9] Two-tailed *p*-values are figures where both tail regions in the sampling distribution are shaded.

these types of *p*-value images is standard practice in introductory statistics curriculum when conducting hypothesis tests tasks.

Another approach to computing the *p*-value region can be done using integration assuming the person knows the theoretical probability distribution corresponding to the statistic being used for the hypothesis tests problem. For the example, to compute the probability in Figure 8 one simply needs to integrate $\int_1^{\infty} f(x)\,dx$ where *f(x)* is the probability distribution function for a standard normal distribution $N(0,1)$. Computing the *p*-value using integration is typically discussed in upper division statistics courses since knowledge of the computation requires background in calculus. Furthermore, methods such as these requires knowing that the random variable being used to compute the *p*-value relates to a probability distributions function that has a closed form. This differs from introductory statistics students who only require introductory algebra as pre-requisites for the course. For these reasons, knowing this alternative form of computation aligns with SCK of teachers. Evidence of this type knowledge can be identified in a teacher by how he/she describes the relationship between calculus and statistics when discussing the *p*-values computation.

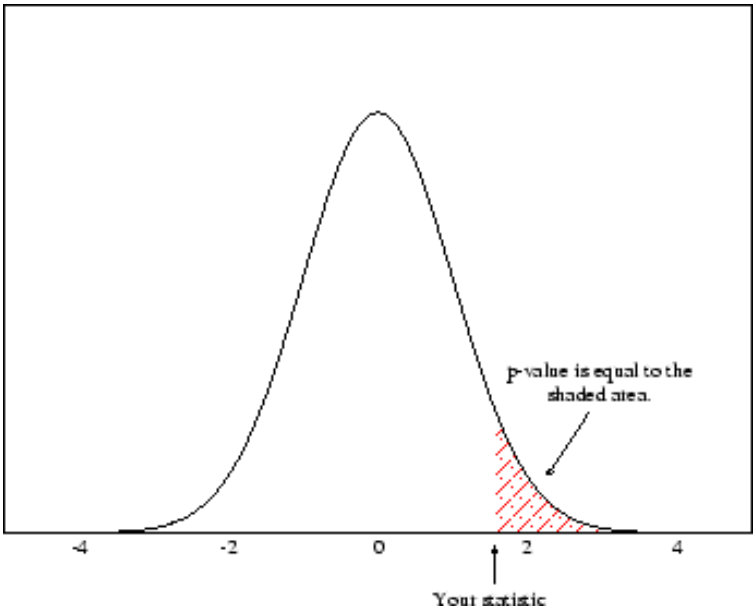


Figure 8. Image of p-value in a theoretical sampling distribution.

An empirical sampling distribution requires a teacher to understand that probability can be generated by repeatedly simulating statistics under a null assumption. For example, suppose one wanted to study the approval rating of the president and a survey was conducted from 30 people in the population. Figure 9 shows an example of an empirical sampling distribution for 100 samples with a sample size of 30 that was simulated from a population assuming a 50% approval rating for the president. The statistic being analyzed is a count of the number of "yes" votes in a sample of 30 people.

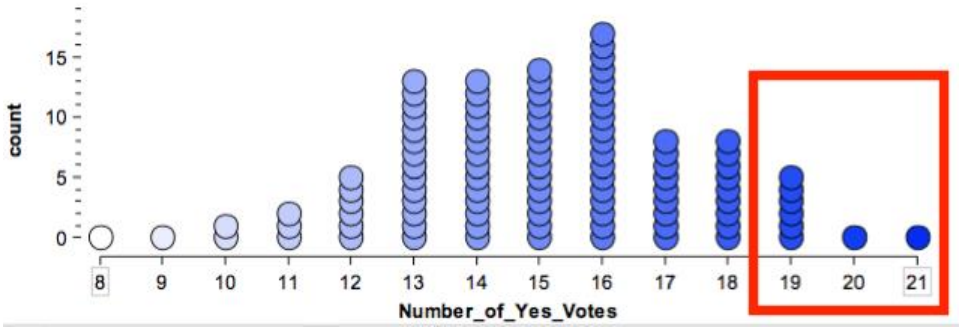


Figure 9. Empirical sampling distribution with population parameter p = 0.50.

Since the assumption was 50% of people approved of the president and a simulation was generated from a population with this assumption, it is expected that the highest frequency of points would fall around the count of 15 because of random variation. If the result of a sample survey was 19 out of 30 people approved of the president, then the *p*-value would be the region (i.e. set of values) consisting of values of 19 or more (for one-tail tests). The *p*-value would then be computed by counting the number of points that fell within that tailed-region using *relative frequencies*. In Figure 9, the *p*-value would be 8/100 or 8%. This understanding of a *p*-value illustrates SCK because it is an alternative approach than one presented in standard statistics texts. Identifying SCK would mean assessing a teacher ability to recognize and describe the relationship between these kinds of approaches and the standard approach in introductory statistics texts.

There is also important prerequisite knowledge for understanding both graphical representations of the sampling distributions shown above. The theoretical representation requires a deeper understanding of random variables, probability distributions, and sampling distributions. For example, teachers might recognize statistics like means and proportions have sampling distributions that relate to normal distributions. This is knowledge typically learned from prior statistics courses. Prerequisite knowledge for the empirical sampling distributions requires a person to know various ways to model data in order to simulate statistics for appropriate sampling distributions. This makes the empirical sampling distribution an approach more appealing as an instructional approach because is more accessible to students who may not have taken courses in statistics and

probability in the past. In regard to KCS, teachers must be mindful of correct/incorrect student computations using the empirical sampling distributions. For example, a student might incorrectly compute a *p*-value by only considering the number of observations that fell on the sample value and not the tail sections. In the example shown in Figure 9, this would be the student getting a *p*-value of 6% because the student only counted the observations at the outcome of 19. The teacher must then recognize that the student is not considering the important concept of "extremes" that is an important component of the *p*-value's definition. In order to recognize these misconceptions, the teacher must have CCK and SCK of the modeling process and KCS of misconceptions and hurdles students encounter when students attempt to use simulation-based approaches to hypothesis testing.

4.2.4. Category 4 – Relating a p-value to the level of significance. In this category, the goal is to focus on the magnitude of the *p*-value. A property of the *p*-value is its ability to measure the *unusualness* of an observed sample under the null assumption. Teachers should recognize that large *p*-values mean the probability that the observed sample occurred under the null assumption is large. In contrast, small *p*-values correspond to a sample that rarely occurs under the null assumption. This is illustrated in Category 3 where an observed sample gave a *p*-value of 8%. Statisticians decide whether the probability is small enough to warrant rejecting the null hypothesis.[10] This introduces the connection between statistical significance and hypothesis testing. Statistical significance of a hypothesis test occurs when there is strong evidence to reject the null hypothesis. A statistician does this in practice by observing small *p*-values which is

[10] In practice, most statisticians use a standard 5% or 1% as a cutoff measure of unusualness.

viewed as an observed sample showing strong evidence against the null hypothesis. Large *p*-values on the other hand do not show statistical significance because the observed sample is considered to be within an acceptable range of possible sample statistics under the null hypothesis.

A big concept that relates to statistical significance is the level of significance which statisticians use as a cutoff measure when analyzing the *p*-value. Figure 10 illustrates the relationship between the *p*-value and level of significance (α) in a theoretical sampling distribution for a one-tail hypothesis test.

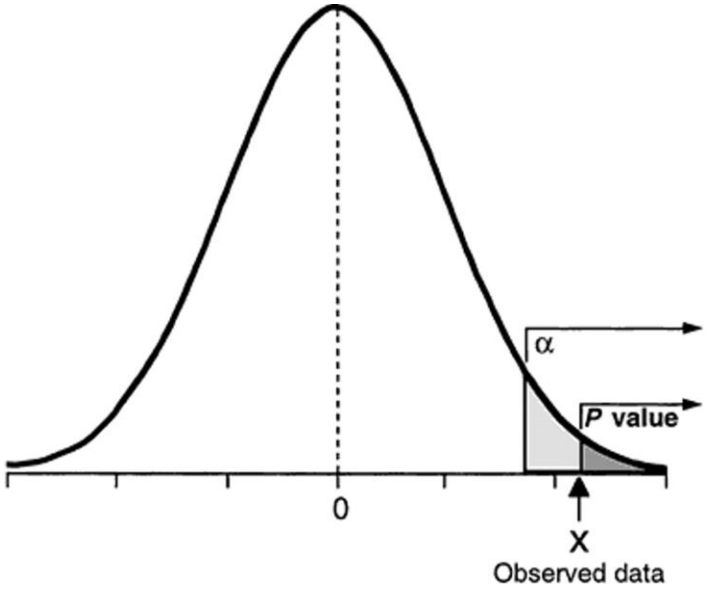


Figure 10. Relating a *p*-value with a level of significance in a sampling distribution.

Drawing the image shown in Figure 10 is common practice in introductory statistics to illustrate the relationship between the *p*-value and level of significance. Teachers should know in practice that one 'rejects the null hypothesis' when the *p*-value is less than the level of significance and 'fails to reject the null hypothesis' when the *p*-value is greater than the level of significance. This effectively links the concepts of level of significance

and the *p*-value with the notion of statistical significance. This is a form of CCK because this is knowledge shared between the teacher and student in introductory statistics as part of instruction. It is important to note that even if the topics of *p*-value and level of significance are connected procedurally, conceptually it is important for teachers to emphasize that a *p*-value does not have any direct relationship with the concepts of Type-I error. Delving deeper into the relationship between Type-I error and *p*-value leans further to SCK of a statistics teacher. For example, a teacher might have an understanding of statistical power and an understanding that Type-I error does not relate to a *p*-value. This type of knowledge is only emphasized in upper division statistical courses (Aberson et al., 2002).

In regard to KCS, teachers should recognize student misconceptions related to images like the one shown in Figure 10. For instance, introductory statistics students have a tendency to incorrectly label regions of the level of significance and *p*-value in their representations. Being able to recognize these errors and knowing alternative routes to solve these misconceptions relates to KCS. For example, a teacher might use empirical sampling distributions instead of theoretical sampling distributions as alternative approaches to help a student understand their misconception. Furthermore, teachers should be mindful of student misconceptions regarding the magnitude of the *p*-value and statistical significance. In particular, recognizing how the magnitude of the *p*-value is used to interpret the statistical significance of a result of a hypothesis testing problem.

5. METHODOLOGY

As highlighted in the previous sections, there is a need to study teachers' understanding of hypothesis testing. Since research has shown that students and teachers struggle with understanding probability in statistical inference, it is important that the statistics education community focus on studying the knowledge of statistics teachers. The overarching goal of this study is to analyze teachers' knowledge of probability in the context of hypothesis testing. In particular, the research questions proposed in this study are: What knowledge do CCIs and GTAs have about probability in the context of hypothesis testing? In particular,

1) How do CCIs and GTAs understand the concept of a *p*-value?
2) How do CCIs and GTAs understand the role of conditional probability in the context of hypothesis testing?
3) How do CCIs and GTAs understand formal/informal concepts of probability, simulations and sampling distributions when reasoning about tasks related to hypothesis testing?

To answer these questions this section outlines the methods used to gather and analyze the data on statistics teachers understanding of a *p*-value. The first section focuses on the data collection methods and assessment tools that were used. The second section focuses on the analytical process. The final section discusses issues of validity for this study.

5.1. Data Collection

To accomplish the research goals for this study, a mixed methods approach was applied to gather quantitative and qualitative data consisting of both survey data and clinical interviews. I first used a task-based web survey followed by a 60-minute interview with a small subset taken from the larger group of survey participants. The surveys provided me an opportunity to gain a general understanding of the population of GTAs and CCIs understanding and served as an opportunity to find volunteer interview participants. Section 5.1.1 outlines the details of the assessment survey and Section 5.1.2 describes the details of the interview protocol. Section 5.1.3 outlines information regarding the timeline and recruitment of the participants in the study. This section also includes demographic information for the survey and interview participants.

5.1.1. Assessment survey. To gather information on the general population of GTAs and CCIs, a survey was conducted focusing on participants' SKT on the concept of *p*-values. The goal of the survey design was to generate tasks related to *p*-value that highlights the information described in Table 6. A list of teachers and institutions were contacted with the help of members from American Mathematical Association of Two-Year Colleges (AMATYC) and CAUSE (Consortium for the Advancement of Undergraduate Statistics Education) to distribute the survey to potential candidates. For those universities and community colleges selected, the department chairs of the mathematics and statistics department were contacted regarding the research plans and goals of the study by e-mail. If the department chair agreed, he/she was requested to forward the e-mail containing a link to the survey to faculty in their department. Since the

survey was voluntary, there is possibility that the number of participants can vary dramatically. The survey was administered using the online survey tool *Qualtrics*. To provide incentive to take the survey, participants were given an opportunity to win one of four $50 Amazon gift cards raffled to the participants who completed the survey.

The survey consists of two main components (see Appendix B). The first component is a set of four content questions meant to assess the participant's SKT on the concepts of probability and hypothesis testing. The second component is a set of demographic questions meant to gather information on the gender, educational background, profession and teaching experience of the participants.

To ensure the survey items were appropriate, research was conducted on potential tasks from published sources in the statistics literature that highlighted areas of statistical literacy and statistical thinking involving the *p*-value. Since the goal of this research was to investigate teachers' understanding of *p*-value and simulation-based models of hypothesis testing, I focused primarily on source material that contained previously conducted research on these topics. During the synopsis of the research material, potential assessment items were chosen and a pilot survey was conducted in the Winter of 2015 originally containing six tasks. Preliminary findings showed that tasks taken from previous research material coincided with their original source material, but after discussions with experts in the statistics education field modifications were necessary in order to make the questions more applicable to better assess SKT.

Using the results of the pilot data and recommendations from experts in the statistics education community, a revised set of survey questions was generated. A total

of five modified content tasks was then created. As a final measure to ensure the quality of the survey questions, the current set of survey questions was distributed to a new set of experts in the statistics education community for a final review. After feedback from the experts, a final set of revisions was conducted prior to distribution of the survey resulting in a total of four survey tasks. A table describing the kind of SKT being assessed and the origin of each question can be found in Appendix A. In the upcoming sections, I discuss each task in detail and the SKT each question is meant to address.

5.1.1.1 Content Question 1 (The Drug Test Task). The Drug Test Task assesses the participant's understanding of the *p*-value's definition (Figure 11) in the context of a statistical problem by having them analyze verbal interpretations of the *p*-value. A verbal interpretation of the *p*-value describes the action of a person to use regular language (written or spoken) to interpret a *p*-value's meaning in the context of a statistical problem. This question is adopted from a CAOS test (delMas et al., 2007) question on the *p*-value's verbal interpretation. The phrasing of the task and the interpretations of Student A, B, and C are identical to the one presented in the CAOS. The interpretations of Student D, E, and F was added based on feedback from reviewers.

The task presents the participant with six potential interpretations of the *p*-value meant to analyze the participant's CCK of the *p*-value's interpretation in a statistical context. In introductory statistics, the *p*-value is taught as the probability of getting a statistic as extreme or more extreme as the observed statistic under the null assumption. Statistics teachers should know how to integrate the definition of a *p*-value in the context of a statistical task since this is a common question in statistics textbooks. Furthermore,

this question also assesses KCS because it can determine the participant's ability to recognize common student misinterpretations of *p*-value and prompts the participant to respond to these misinterpretations.

Based on the definition of a *p*-value, the only interpretation considered to be potentially valid is provided by Student A. It is important to note that unlike traditional interpretations, Student A replaces the words "as extreme or more extreme" with "as large or larger." The reason for this replacement is to highlight that the words "as extreme or more extreme" are sometimes difficult for students to conceptualize. Words such as *large* or *larger* are more colloquial. Using this replacement makes Student A's interpretation more representative of a student who is trying to conceptualize the meaning of *"as extreme or more extreme.*" The other five interpretations are common student misconceptions of the *p*-value that are highlighted in the statistics education literature. Unlike Student A's interpretation, which highlights the importance of the *p*-value as a conditional statement and focus on the observed sample, interpretations made by Student B, C, E, and F do not emphasize the element of conditional probability and make no mention of the relationship between the observed results and a null assumption. Student D's interpretation is a common misconception discussed in the work of Castro Sotos et al. (2009) that highlights a swapping of the assumption and conclusion in the conditional statement. Student B, C, D, E and F also make no mention of the *extremities* of results which is an important property of the *p*-value's definition.

A research article reports on a new drug test that is to be used to decrease vision loss. The article reports its results and a p-value of 0.04 in the analysis section. Below are six different student interpretations of the p-value.

Student A: *The probability of getting a result as large as or larger than the one in this study if the drug is actually not effective is 0.04.*

Is this student's interpretation of a p-value valid? ___ Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student B: *The probability that the drug is not effective is 0.04.*

Is this student's interpretation of a p-value valid? ___Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student C: *There is a 4% chance that the researcher made an error.*

Is this student's interpretation of a p-value valid? ___ Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student D: *If you were to repeat the study, there is a 4% chance of getting exactly the same result as the one in this study if the drug is actually not effective.*

Is this student's interpretation of a p-value valid? ___Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student E: *The probability that the drug is effective is 0.04.*

Is this student's interpretation of a p-value valid? ___ Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student F: *There is a 96% chance the drug is effective.*

Is this student's interpretation of a p-value valid? ___ Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Figure 11. Drug Test Task

5.1.1.2. Content Question 2 (The Graduate Student Task). The *Graduate Student Task* (Figure 12) assesses the participant's understanding of the magnitude (i.e. value) of the *p*-value and how it relates to statistical significance. This question focuses on the

participants' CCK by assessing an understanding of the *p*-value's magnitude commonly taught in introductory statistics courses. The question is originally from the CAOS test (delMas et al., 2007) that assesses whether a participant's ability to relate a p-value's magnitude with the statistical significance of a result. The task was mainly untouched from its original version with the exception of an additional prompt to have the participant explain their understanding of *p*-value and statistical significance.

The task was designed to assess the participant's procedural and conceptual understanding of the *p*-value in the hypothesis testing process. Procedurally, a *p*-value is compared to the level of significance to show statistical significance (which is set relatively small). Analyzing a procedural understanding of this question would focus on the participant's ability to recognize small *p*-values (e.g. less than 0.05 or 0.01) would result in a rejection of the null hypothesis and consequently showing statistical significance of a result. Analyzing a conceptual understanding would focus on a participant's understanding of the *p*-value as a measure of the likelihood of getting a sample as extreme as the observed sample under the null hypothesis. Having a small *p*-value means that the observed sample is extremely rare under the null hypothesis, this allows a statistician to favor a rejection of the null hypothesis. Having either understanding would result in the correct response being (b). Choices (a) and (c) relate to common student misconceptions of the *p*-value. Because (a) and (c) relate to student misconceptions, this question also focuses on the participants' KCS as identified in Table 5.

A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of p-value would she want to obtain?

a. A large p-value.
b. A small p-value.
c. The magnitude of a p-value has no impact on statistical significance.

Explain your understanding of the relationship between a *p*-value and statistical significance.

Figure 12. Graduate Student Task.

5.1.1.3. Content Question 3 (The Car Task). The *Car Task* highlights a symbolic understanding of the *p*-value (Figure 13). This is an original question inspired from the work of Shaughnessy and Chance (2005) and Castro Sotos et al. (2009). Symbolic representation describes the action of a person to use mathematical and/or statistical symbols to represent a *p*-value's meaning in the context of a statistical problem. The original pilot task consisted of a set of seven different symbolic representations of a *p*-value. Based on expert feedback, a decision was made to alter the list of choices to a smaller set and to give the participant more opportunity to explain their reasoning. After reviewing the pilot data, four interpretations were found to be the only ones chosen by participants. After feedback from experts, a decision was made to use six symbolic representations with the addition of prompting the participant for reasons for identifying a representation as invalid to allow a deeper understanding of the participants SKT. Furthermore, seeing how the teacher responds to the prompts allowed me the opportunity to assess the participants KCS since each prompt assesses how a teacher responds to student work.

The overall question evaluates the participant's ability to recognize the *p*-value as a conditional probability and/or a probability using a z-score transformation. Student A

takes a standard approach of transforming the raw data into a z-score that is common practice in the procedural computation of the *p*-value. The only exception to the representation is the use of a lowercase z, which implies a particular value after a transformation is done. Theoretically, an uppercase Z would be more appropriate to represent the use of a random variable. The use of a lowercase z is to emphasize a student-like approach to the transformation of a z-score. This is type of notation is more realistic to a student who is first learning to symbolize statistical concepts. Disregarding the exception of the lowercase z, Student A's representation is considered correct because it aligns with traditional statistical methods when doing a one-sample proportion test. Participants who consider this representation valid suggests a participant's ability to recognize a potentially correct representation of a *p*-value (i.e. CCK) and knowledge of a student's ability to give a potentially correct representation (i.e. KCS). The other correct response is Student E because it emphasizes the probability of getting a sample as extreme ore more extreme than the observed sample of $\hat{p} = 0.6$ assuming the population proportion is 0.50. This representation has a potential issue which is the use of $\hat{p}$ in place of a symbol for a random variable (e.g. X) that is more theoretically sound. This was done to illustrate a representation that more closely aligns with the statistical knowledge of introductory students. The symbol $\hat{p}$ is commonly used in introductory statistics course so it is more realistic that a student would use this symbol in place of a symbol for a random variable. With the exception of the usage of $\hat{p}$, Student E's representation is a way of symbolically representing *p*-value as a conditional probability that is uncommonly discussed in introductory statistics. For these reasons, recognizing E as a potentially

correct response highlights a SCK because it highlights a deeper understanding of a *p*-value that extends beyond an introductory statistics course. It also highlights KCS because it requires a teacher to assess the important features of the *p*-value that the student is trying to express symbolically.

The other symbolic representations reveal misconceptions of the *p*-value found in the research literature and were transformed into symbolic form. Student B's representations lacks a conditional form and incorrectly finds the probability of the observed sample. Student C does use a conditional probability, but incorrectly defines the events in the conditional probability. Rather than finding the probability of extreme values under the null assumption, the student's representation finds the probability of the null hypothesis assuming the observed sample data. Student D also uses a conditional probability statement, but also incorrectly defines the probability. As written, the student is defining the probability of the null assumption assuming the observed sample data, which is also incorrect. Student F is also an incorrect representation because it does not indicate the student is thinking of the *p*-value as a conditional probability although the student does make the correct use of the inequality notation to represent extreme values from the observed sample.

Research was conducted to determine whether people preferred hybrid over traditional gasoline powered vehicles. The article reports that their random sample showed that 60% of people preferred hybrids. The researchers gave a test statistic of z = 1.414 and a p-value of 0.08 using a right-tailed hypothesis test for proportions. Below are six different symbolic representations of the p-value given by introductory students when asked about this problem.

$$\textbf{Student A: } \boldsymbol{P(z \geq 1.414) = 0.08}$$

Is this student's symbolic representation of a p-value valid? If you think this question is invalid, explain why you think it is invalid.

Student B: $\boldsymbol{P(\hat{p} = 0.6) = 0.08}$

Is this student's symbolic representation of a p-value valid? If you think this question is invalid, explain why you think it is invalid.

Student C: $\boldsymbol{P(Reject\ the\ null\ hypothesis \mid \hat{p} = 0.6) = 0.08}$

Is this student's symbolic representation of a p-value valid? If you think this question is invalid, explain why you think it is invalid.

Student D: $\boldsymbol{P(p = 0.5 \mid \hat{p} = 0.6) = 0.08}$

Is this student's symbolic representation of a p-value correct? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student E: $\boldsymbol{P(\hat{p} \geq 0.6 \mid p = 0.5) = 0.08}$

Is this student's symbolic representation of a p-value valid? If you think this question is invalid, explain why you think it is invalid.

Student F: $\boldsymbol{P(\hat{p} \geq 0.6) = 0.08}$

Is this student's symbolic representation of a p-value valid? If you think this question is invalid, explain why you think it is invalid.

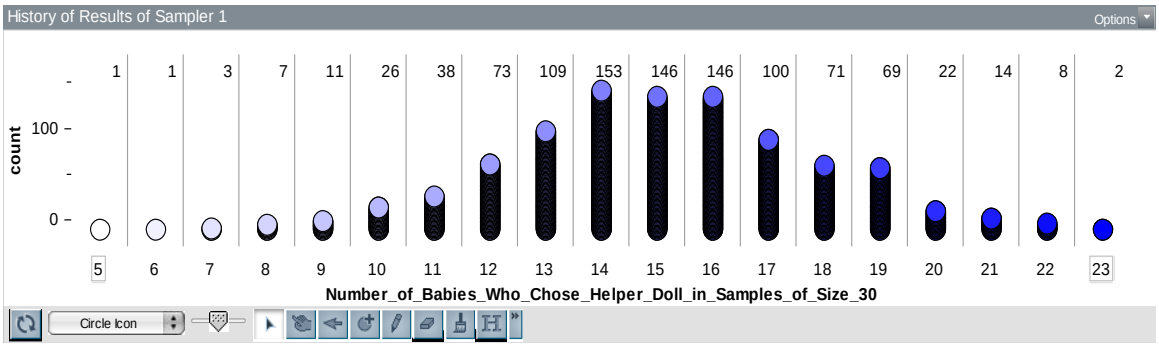
Figure 13. Car Task.

5.1.1.4. Content Question 4 (Helper-Hinderer Task). The final survey question is a modification of a problem that highlights a simulation-based approach of a one-tailed hypothesis test for a one-sample proportion (Figure 14). This question was inspired by the work of researchers (e.g. Chance et al, 2004 & 2007; delMas, et al, 1999; Erickson, 2006; Zieffler et al, 2008a & 2008b) from the CATALST curriculum that emphasize the importance of simulation and sampling distributions to understand concepts of statistical inference. The task is designed to assess the participant's understanding of a simulation-based approach to hypothesis testing. Furthermore, the task assesses the participant's understanding of the relationship between empirical sampling distributions and *p*-value.

The original question in the pilot research had the participant explain whether they agree/disagree with a student's approach of computing a *p*-value and level of significance using the empirical sampling distribution. The participant was then prompted to explain the different portions of the student work. A review of the pilot data revealed little about the participant's understanding of the simulation or how they thought about the model assumption. Based on reviews of the pilot data and expert feedback, the question was modified to allow a greater assessment of the participant's understanding of the modeling process and how they would use the model and compute a *p*-value to make a statistical decision.

A sociology study was conducted to determine whether babies are able to recognize the difference between good and bad. In one experiment, 30 six-month old babies were randomly selected. Each baby was shown two possible puppets to play with, a 'good' puppet that helped and a 'bad' puppet that hindered. 21 out of 30 babies showed a strong preference for the helper puppet over the hinderer. In order to determine if this result provides strong statistical evidence that babies really do have a preference for the 'good' or helper puppet, James, a statistics student, conducted the following test procedure:

- James gets a coin and flips the coin 30 times.
    - If the coin lands on the "heads", he records the baby as preferring the helper puppet.
    - If the coin lands on the "tails", he records the baby as preferring the hinderer puppet.
- James then used a computer simulation to repeat the previous step 1000 times.
- James then plots the distribution for the number of times a baby chooses the helper puppet from each of the 1000 samples of size 30. This is shown in the graph below.



i.) James' procedure is based on which assumption?

a. A baby is more likely to choose the helper puppet.
b. A baby is equally likely to choose either the helper or hinderer puppet.
c. A baby is more likely to choose the hinderer puppet.

Explain the reason for your choice.

ii.) Suppose James wanted to conduct a right-tailed hypothesis test using the simulated data.

- What would you estimate for the *p*-value?
- Explain how you found the *p*-value and interpret it in the context of James' research.

iii.) Based on your estimated *p*-value, what do you think should be James' conclusion?

a. There is statistically significant evidence that babies are more likely to choose helper puppets.
b. There is statistically significant evidence that babies are more likely to choose hinderer puppets.
c. There is statistically significant evidence that babies are equally likely to choose helper or hinderer puppets.

Explain the reason for your choice.

Figure 14. Helper-Hinderer Task.

In this question, I present the work of a student, James, who uses a simulation-based approach to perform a one-tailed hypothesis test for a single proportion on data conducted about a baby's preferences to helper and hinderer puppets. James begins by creating an empirical sampling distribution through a sequence of coin flips. A sample consists of 30 coin flips, which is repeated 1000 times. By using the model of a coin flip to represent a baby's choice of the puppet, James is assuming that a baby has no preference for a helper or hinderer puppet. The purpose of the empirical sampling distributions created by James can be used to determine the likelihood of the sample result compared to what would be expected to happen by chance based on James' assumption that there was no preference.

In order to assess a participant's understanding of the simulation-based approach, sub-questions were asked about James' approach. The first question assesses whether the participant is able to recognize the null assumption. As mentioned above, the correct response is (c) because of James' use of the coin to simulate a baby's response. This question also assesses whether a person can correctly identify the null hypothesis in a simulation-based approach to hypothesis testing. Choice (a) and (b) illustrate potential alternative hypothesis for the simulation-based model.

The second question assesses the participant's ability to compute a *p*-value using the empirical sampling distribution. Because the observed sample yielded 21 out of 30 babies preferring the helper puppet and since the definition of the *p*-value is a measure of extremes, a correctly computed *p*-value would result in a value of 2.24% by adding the relative frequency for the outcomes of 21, 22, and 23 (see Figure 15).

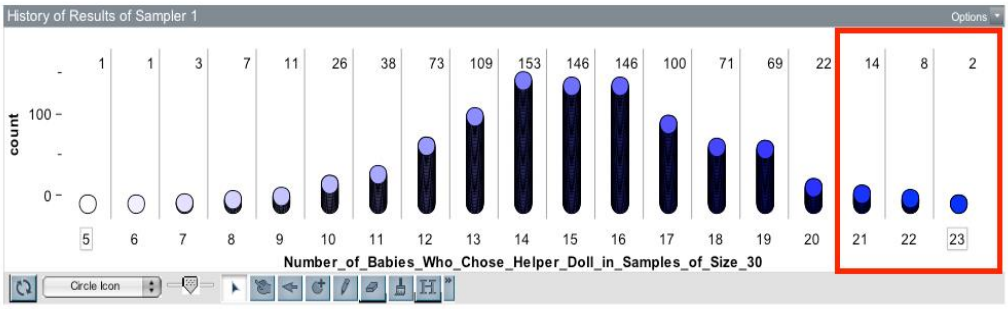


Figure 15. Empirical sampling distribution with a correct p-value region marked.

I hypothesize that a common misconception regarding this problem would be participants only computing a *p*-value of 14/1000. This is due to the participant neglecting the property that *p*-value includes all extremes. As a result, they would only look at the relative frequency for the outcome of 21 (see Figure 16). There might also be participants

who are unable to compute the *p*-value because they fail understand how to relate the empirical sampling distributions to a *p*-value resorting to more theoretical methods such as using a binomial distribution or a standard one proportion test as typically done in the introductory statistics book.

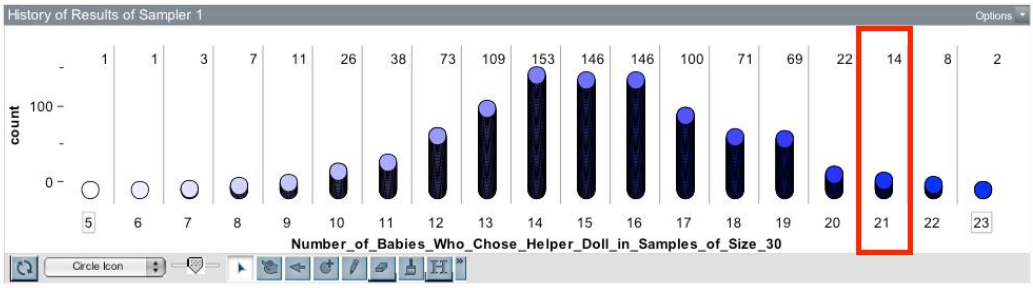


Figure 16. Empirical sampling distribution with only the observed value marked.

The third question will assess the participant's understanding of the overall process to generate a conclusion about James' reasoning. I hypothesize that individual's knowledgeable of this process should recognize this question as the final steps of a hypothesis test where a statistician must either reject/fail to reject the null hypothesis and claim statistical significance of the results. Since the previous question has the participant compute the *p*-value, I hypothesize that a correct conception would lead the participant to emphasize how the *p*-value is relatively small, showing there is evidence to suggest James' reasoning is correct. This is assuming one uses a 5% level of significance. Alternatively, if a participant chooses a 1% level of significance, then the comparison of the level of significance to the *p*-value would refute James' reasoning. An individual who does not possess a correct conception would struggle discussing these types reasoning and may just focus on sample results.

This task touches on all three categories of SKT. This task assesses a participant's SCK of a *p*-value because this is not a typical approach to hypothesis testing for traditional introductory statistics classes, but is one statistics educators have used to understand the concept of a *p*-value. Therefore, this task can be useful to determine if a teacher is knowledgeable about simulation-based approaches and to assess whether the teacher knows how it can be a useful strategy to support student development (i.e. KCS). As more of the introductory statistics curriculum moves towards a simulation-based approach to hypothesis testing, this form of knowledge may eventually shift to one that is common in more statistics classrooms making this task useful in accessing CCK.

5.1.2. Interview. Following the survey, participants were given the opportunity to participate in clinical interviews. The goal of the interviews was to further explore teachers' SKT based on their responses to the survey and clarify any explanations. The interview responses were used to triangulate participant responses with their survey to determine if there were any inconsistencies. Participants were compensated with a $25 Amazon gift card for their interview time.

To gather interview participants, each survey participant was prompted at the end of the survey asking whether they would potentially like to be interview participants. Those participants who agreed were then categorized based on their profession (CCI or GTA). Participants were then chosen from the two categories based on their responses to the survey questions. The primary goal was to gather individuals with various levels of statistical knowledge on the *p*-value. In order to do this, an initial review of their survey data was conducted to see what kind of responses were provided by the participants. The

responses for each question was carefully evaluated by myself and a secondary researcher to determine whether a participant's understanding of a *p*-value was distinct from other participants. The analysis included assessing whether the participant's response was correct/incorrect and if explanations showed different levels of understanding of the *p*-value. After an initial selection was generated, several participants were contacted electronically to determine if they were available for follow-up interviews. A final count of seven interviews were conducted (3 CCIs and 4 GTAs). The interviews were conducted on-site or through video conference calls using Skype. Each video was recorded and transcribed for analysis.

5.1.2.1. Follow-up survey questions. The goal of the interview was to gather additional qualitative data to generate a snapshot of a participant's understanding. Each survey task had a standard list of follow-up interview questions meant to further investigate a participant's survey response. A list of the follow-up interview questions can be found in APPENDIX C. In addition, certain interview questions have been designed to cater to specific student response from the survey. The *Helper-Hinderer Task* have questions where the student will be asked to compute a probability for the *p*-value. Since participants might have several approaches to the problem, alternative solutions introduced as 'hypothetical' student responses will be shown to challenge the participant's understanding. To help the participant explain their reasoning, tools such as a calculator, coins, computer simulations, etc. will also be provided during the interview.

5.1.3. Survey and interview participants. To gather the participants for the survey and interview necessary to complete the study 50 institutions (30 universities and

20 community colleges) were contacted via their department chair to pass along the survey tool to their respective faculty. In addition, support from CAUSE (Consortium for the Advancement of Undergraduate Statistics Education) helped in the distribution of the survey to a larger population through a newsletter.[11]

55 participants from various teaching and educational backgrounds completed the survey. The demographic data of the participants was divided into two main categories: Graduate Students and Instructors. A total of 30 graduate students and 25 instructors took part in the survey. The category of graduate students included those individuals who classified themselves as either a graduate teaching assistant (GTA), a graduate research (GRA) assistant or both. The category of instructors included those participants who classified themselves as community college instructors (CCI), four-year university instructors (FYI) or both. The choice to separate the participants into these two categories was to highlight that unlike instructors, a graduate student's primary role is a student while an instructor's primary role is a teacher. Furthermore, graduate students potentially switch between roles of graduate teaching assistants and graduate research assistants at their institutions. Instructors have the potential to be employed at multiple institutions. For example, adjunct instructors might teach at a community college and a four-year university.

Table 7 shows the demographic information of the graduate students. A total of 22 male graduate students and 8 female graduate students participated in the survey. The data also shows that from the 30 graduate students, 22 participants classified themselves

[11] Since exact information is not provided by CAUSE regarding their information database, it is unknown how many institutions were contacted.

as graduate teaching assistants, 5 participants classified themselves as graduate research assistants and 3 participants classified themselves as both a graduate teaching and research assistant. Regarding academic levels, 21 of the 30 graduate students classified themselves as doctoral candidates, 6 graduate students were masters degree candidates and 3 graduate students classified themselves as having a masters degree but did not mention if they were either masters or doctoral candidates. Regarding academic focus, 17 of the 30 of the graduate students classified themselves as focusing on a degree in statistics, 5 focused on mathematical sciences, 1 mathematics education, and 4 specialized in a major besides mathematics and statistics. There were also a set of graduate students with multiple majors: 4 of the graduate students majored in mathematics and statistics, 1 graduate student was a triple major in mathematics, mathematics education, and statistics education and 1 graduate student majored in mathematics, statistics and an outside major. Finally, regarding teaching responsibilities almost 26 of the 30 graduate students were currently teaching classes at four-year universities and 4 graduate students had no teaching responsibilities at the time of the survey.

Table 7. *Demographics for the Graduate Student Participants.*

Demographic Information	**Frequency**
Total Graduate Students	30
Gender	
• Male	22
• Female	8
Academic Position	
• Graduate Teaching Assistant (GTA)	22

• Graduate Research Assistant (GRA)	5
• Graduate Teaching and Research Assistant	3
Academic Level	
• Current Masters Student	6
• Master's Degree	3
• Current Doctoral Student	21
Academic Degree	
• Mathematical Science	5
• Mathematics Education	1
• Statistics	17
• Statistics Education	0
• Mathematical Science and Statistics	4
• Mathematical Science, Mathematics Education, Statistics Education	1
• Mathematical Science, Statistics, Other	1
• Other	1
Teaching Responsibilities	
• Four-Year University	26
• Community College	0
• None	4

Table 8 shows the demographic information of those participants categorized as faculty instructors at their institutions. A total of 14 male and 11 female instructors completed the survey. Furthermore, 15 were community college instructors, 9 were four-year instructors, and 1 instructor who taught both at community college and a four-year university. In regards to academic level, 16 instructors had a masters degree, 8 had a doctoral degree, and 1 was a current doctoral candidate. The instructors also had a variety of academic backgrounds: 6 of the instructors had a degree in mathematical sciences, 6

had a degree in statistics, 1 had a degree in mathematics education. There was also a set of instructors with a double major. One instructor had a degree in mathematics and statistics, three instructors had a degree in mathematical science and mathematics education, one instructor had a degree in statistics and statistics education, one instructor had a degree in mathematics and an outside major.

Table 8. *Demographics for the Instructors.*

Demographic Information	**Frequency**
Total Instructors	25
Gender	
• Male	14
• Female	11
Academic Position	
• Community College Instructors (CCI)	15
• Four-Year Instructors (FYI)	9
• Community College and Four-Year Instructors	1
Academic Level	
• Master's Degree	16
• Doctoral Degree	8
• Current Doctoral Student	1
Academic Degree	
• Mathematical Science	6
• Mathematics Education	1
• Statistics	6
• Statistics Education	0
• Mathematical Science and Statistics	1
• Mathematical Science and Mathematics Education	3
• Statistics and Statistics Education	1
• Mathematical Science and Other	1

• Mathematical Science, Mathematics Education, and Statistics Education	1
• Other	5
Teaching Responsibilities	
• Community College Instructor	15
• Four-Year University Instructor	9
• Community College and Four-Year University	1

After conducting an initial analysis on the survey data, seven candidates participated in follow-up interviews to help further assess their understanding of the *p*-value. As specified in the methodology, these seven candidates were chosen based on their willingness to participate, teaching background (Graduate Students and Instructors) and the survey responses provided by the participant. While these seven participants did share some similarities in their understanding of the p-value from a preliminary analysis of their survey, they also provided a variety of survey responses that demonstrated various ways teachers might understand concepts related to the *p*-value. A summary of the interview participants' demographics is presented in Table 9.

Table 9. *Demographics of the Interview Participants.*

Name	Demographics
Tod	Graduate Teaching Assistant Current Doctoral Student in Mathematics Science
Angie	Graduate Teaching Assistant Current Doctoral Student in Statistics
Sean	Graduate Teaching Assistant Current Doctoral Student in Statistics
Carol	Graduate Teaching Assistant Current Masters Student in Mathematics and Mathematics Education
Jane	Community College Instructor and Four-Year University Instructor Masters Degree in Mathematics Education
Phil	Community College Instructor Masters Degree in Mathematical Science, Statistics Education, and Mathematics

	Education
James	Community College Instructor Masters Degree in Mathematical Science

5.2. Analysis

In this section, I describe the analysis that was conducted. First, I will describe the methods used to analyze the survey data outlining the methods used to organize and categorize the types of themes that occurred during the study. Second, I will then discuss the methods used to analyze the themes found in the interview data.

5.2.1. Survey analysis. The initial plan for the survey analysis was to gain quantitative and qualitative results on the SKT of the population of GTAs and CCIs. Each assessment question will be summarized individually and a running count of the individual responses was accumulated to generate a summary of the overall data.

Since the survey tasks also included tasks that prompted qualitative responses, a deeper analysis of a participant's understanding can be determined through their written work. To analyze the qualitative data from the surveys, the qualitative research method of *thematic analysis* was used to identify types of understanding that might be missed from the multiple-choice portion of the survey tasks. Thematic analysis is a widely used research tool for analyzing qualitative data such as written statements or interview data (Creswell, 2007; Guest, Macqueen, & Namey 2012). The defining features of thematic analysis include: (1) identifying key themes in text which are translated into codes which are then aggregated into a coding scheme, (2) use techniques in addition to theme identification and data reduction technique and (3) can be used to build a theoretical model or find solutions to real-world problems. Furthermore, Guest, Macqueen, and

Namey (2011) highlight that the strengths of thematic analysis include: (1) well suited for large data sets, (2) good for team research, (3) can be used to study topics other than individual experiences, and (4) addition of quantitative techniques adds analytic breath.

Using thematic analysis, the qualitative response for each survey question was used to determine the CCK, SCK, and KCS of the participant. Regardless if the participant chooses an incorrect response in the multiple-choice portion, their qualitative response was still analyzed for content to determine emerging themes in the data. To illustrate, sample responses from pilot data for the *Graduate Student Task* include:

- **"**The p-value would be the likelihood that the particular sample would be collected given a null hypothesis it true. A small probability would indicate statistically significant results, or more specifically, that the sample collected is not consistent with the null hypothesis and thus the null hypothesis should be rejected."
- "A small P value suggests that the sample provides enough evidence that one can reject the null hypothesis for the entire population."

The first response illustrates a participant's understanding of the *p*-value as a conditional probability. Evidence for this is indicated by the participant's explicitness in describing the null assumption and its relation to the observed sample. The first response also indicates understanding of the relationship between the magnitude of the *p*-value, the sample data, and the null hypothesis. This is illustrated by their response of statistical significance connecting to notions of unusual sample data and consistency with the null

hypothesis. This links to important ideas of the relationship between the likelihood of the observed sample and how that provides evidence against a null hypothesis.

The second response indicates a form of procedural knowledge of the *p*-value. In their response, I hypothesize that the participant is aware of the process of rejecting the null hypothesis for a small *p*-value based on the observed sample. The second response does not indicate whether the participant's sees the *p*-value as a conditional probability or how he/she views the concepts of likelihood as the first response. It only focuses on how the *p*-value connects to the decision-making process of hypothesis testing.

Based on these two sample responses, an initial analysis would show that both responses would be categorized differently for the type of knowledge illustrated from both participants. This categorization helps in distinguishing types of SCK that exists amongst participants. The first response would be coded to a category related to SCK. Evidence in the first response is illustrated in their discussion of the conditional nature of the *p*-value and recognizing the importance of the null hypothesis in describing the *p*-value. The creation of the coding scheme is also compared with responses from interview responses. A more detailed illustration of the coding scheme process is described in the interview analysis.

5.2.2. Interview analysis. Thematic analysis also played a significant role in analyzing the interview responses. Using thematic analysis, potential themes were found within the data. These themes were used to generate a model of statistical knowledge of teachers of various aspects of the *p*-value. The process of generating these themes were the result of four phases of analysis using the survey and interview data: (1) post-survey

and post-interview analysis, (2) transcription analysis, (3) coding scheme, (4) chronicling emerging themes.

5.2.2.1. Post-survey and post-interview analysis. After surveys were conducted and analyzed, participants who were willing to participate in an interview had their survey data pre-screened prior to the interview. An initial summary of each interviewee's survey was generated to create an initial image of the SKT of the interviewee using the framework of Table 5 as a guide. After the interview, a post interview summary was generated to highlight key ideas discussed during the interview. The pre-interview and post-interview summaries were then compared amongst the interviewees to highlight any discrepancies in the participant's understanding.

5.2.2.2. Transcription analysis. After post-interview analysis was completed, each interview was transcribed verbatim to highlight important ideas and statements regarding the participant's knowledge of the *p*-value. The interview transcription was also used to determine areas of miscommunication and potential areas of scrutiny in the data that might be caused by the interviewer or interview protocol. After transcription, a summary was then generated on a subject's understanding. The transcription summary was then compared to summaries from the initial phases to generate a model of the individual's understanding.

5.2.2.3. Coding scheme. In order to generate a sufficient coding scheme for the qualitative data, a Test-Retest Method was used. An initial coding scheme was generated in the first pass of the then went through a cyclic process until a reasonable measure of accuracy was achieved. The coding scheme focused on categorizing emerging themes in

the data on ideas related to CCK, SCK, and CKS. After an initial coding scheme was been generated, survey responses, videos, and transcripts were re-analyzed to find sources of evidence that either confirm or refute the initial categories. After evidence was collected, one additional researcher was trained the use of the coding scheme. After training, the researcher was given all the data, which was then blindly coded using the pre-generated coding scheme. Codes were then compared between researchers to determine if refinement was needed. Whenever there was inconsistency between coders, discussions were done regarding the inconsistency between the codes. Based on the discussion, the coding scheme was modified to help generate consistency between coders. New sets of segments were then recoded with this process continuing until the percentage agreement of at least 80%.

5.2.2.4. Chronicling emerging themes. After coding the data, a final reflection was conducted with the purpose of chronicling emerging themes through a case study of each of the interview participants. An overall analysis was then conducted to determine areas of potential implausibility that might result from the process of collecting the survey and interview data. For example, there is a possibility that a participant might misinterpret survey/interview questions. In order to check for such inconsistencies, results across the different subjects were analyzed to determine if the misinterpretation was seen across subjects or specifically to one participant. Deeper analysis was then conducted to determine if the misinterpretation of a question was due to the task design or knowledge of the participant.

5.3. Validity

A main concern with this study is to warrant the validity of the data collection and analysis. To ensure valid results, the overall design of the study contained layers of validity checks. As discussed in the previous questions, a survey tool was first generated using questions based on published sources in the statistics literature that have been peer-reviewed by experts in the statistics education community. Four of the survey items were taken directly from assessment items previously published while two original questions were developed from research in the statistics education community. Furthermore, the current survey was the result of modification after a pilot study was previously conducted in winter 2015 and suggestions from experts in the statistics education community. To further increase the validity of the survey questions, the current set of survey questions was distributed to experts in the statistics education community for a final review and modifications prior to distribution to the target population. This will help in generating a strong representation of the teacher population of GTAs and CCIs. The interview questions help to triangulate the data with the survey responses. Having multiple sources of data ensures that any coding schemes generated will have multiple sources of evidence to clarify any misleading information and validate any claims.

The final step to ensure validity is to focus on the analysis of the data process. There were several stages of analysis involving quantitative and qualitative analysis. Multiple stages of qualitative analysis ensure that any results have been properly documented and analyzed. Having an additional researcher will also support the validity of any generated coding scheme.

6. RESULTS

The goal of this research is to study teachers' understanding of the *p*-value. To do this, survey and interview data was collected from graduate students and instructors to assess their understanding of *p*-value through a series of survey and interview questions designed around concepts related to the *p*-value. These concepts include an understanding of a verbal interpretation of *p*-value, the meaning of the *p*-value's magnitude, a symbolic representation of *p*-value, and a *p*-value's role in a simulation approach to hypothesis testing.

After analyzing the survey and interview responses, three major themes emerged from the data that will be the focus of this chapter. The first major theme focused on how participants understood the magnitude of the *p*-value. This theme emerged from the participants' responses to the *Graduate Student Task* where they discussed the relationship between a *p*-value's magnitude and a statistically significant result. The second theme focused on participants' understanding of the role sampling distributions and simulations play in their understanding of the *p*-value. The *Helper-Hinder Task* was an important component in analyzing a participant's understanding of simulations and empirical sampling distributions because it allowed me to assess their ability to correctly compute a *p*-value using an empirical sampling distributions. The third and final theme emerged from analyzing how teachers viewed the role of conditional reasoning in their understanding of the *p*-value. Portions of the survey and interview questions gave participants opportunities to discuss the relationship and usage of the *p*-value as a

conditional probability. Here I present the results of the *Drug Task* and *Car Task* that helped elicit two ways of describing the p-value: verbally and symbolically.

6.1. Teachers' Understanding of the *p*-value's Magnitude.

A major component of a robust understanding of a *p*-value means knowing how a *p*-value's magnitude relates to statistical significance. A procedural understanding of the *p*-value magnitude includes knowing how to generate conclusions to a hypothesis test by comparing the *p*-value with the level of significance. A conceptual understanding of the *p*-value's magnitude includes knowing that it is a probability that measures the *likelihood* of an observed sample occurring under the null hypothesis. Viewing the *p*-value as a measure of *likelihood* is useful when describing the evidential strength of a researcher's conclusion because it relates the notions of *unusualness* or *rarity* of the observed sample under the null hypothesis. Being able to coordinate the relationship between these two ideas is part of a robust understanding of a *p*-value's magnitude.

The procedural understanding of the *p*-value is prominently taught in introductory statistics courses when covering hypothesis testing, so I expected almost all participants to recognize the need for small *p*-values to claim the results of an experiment are statistically significant. It was harder to predict what type of thinking would emerge as participants discussed the *p*-value's magnitude and its relationship to statistical significance. The data reveals a mix of participants who shared different understandings on the magnitude of the *p*-value.

Participants' understanding of the *p*-value's magnitude was prevalent in their discussion of the *Graduate Student Task* that was designed to assess how participants'

view the relationship between statistical significance and the magnitude of the *p*-value (i.e. Figure 17). If you recall, the *Graduate Student Task* presents asks a participant to choose what magnitude *p*-value is needed if a graduate student wants to show the results of an experiment are statistically significant.

A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of p-value would she want to obtain?

a. A large p-value.

b. A small p-value.

c. The magnitude of a p-value has no impact on statistical significance.

Explain your understanding of the relationship between a *p*-value and statistical significance in the context of this problem.

Figure 17. Graduate Student Task

An overview of the survey and interview data from the *Graduate Student Task* revealed four primary ways the participants understood the relationship between the *p*-value's magnitude and statistical significance. The first was some participants had a conceptual understanding of the *p*-value's magnitude where they viewed the *p*-value as a measure of *likelihood* of an observed sample under a null assumption. The second was some participants held a procedural understanding of the *p*-value's magnitude where they focused primarily on the comparison of the *p*-value with a fixed level of significance. The third was some participants emphasized the importance of choosing an appropriate level of significance based on practical significance and statistical power and how it might relate to a *p*-value's magnitude. The fourth is a misconception of the *p*-value where participants described the *p*-value as the probability or likelihood the null hypothesis is

true. Furthermore, some participants showed an ability to overlap these different ways of thinking.

To summarize the results of the *Graduate Student Task*, this section is broken into two parts. The first section is a summary of the survey responses from the 55 participants. Here I present a general view on how participants understood the relationship between a *p*-value's magnitude and a statistically significant result. The second section summarizes the interview responses given by the seven participants as they discussed the relationship between statistical significance and *p*-value. The final section will summarize the survey and interview data to describe the important connections between the data to help paint a picture of teachers' understanding of a *p*-value's magnitude.

6.1.1. Survey results of the Graduate Student Task. As previously mentioned, the *Graduate Student Task* is a two-part task consisting of a multiple-choice and a short answer portion. The results of the survey responses for this task are broken into two sections. The first section summarizes the multiple-choice portion of the task where participants were asked what magnitude of a *p*-value is desired to have a statistically significant result of an experiment. The second section of the survey results is a summary of the participants' short answer portions where they described the relationship between statistical significance and *p*-value. It is in the results of the short answer portion where I present the various categories of thinking on a *p*-value's magnitude that emerged from the participant responses.

6.1.1.1. Quantitative results of the Graduate Student Task. In the *Graduate Student Task*, participants were asked to select one of three choices for the magnitude of

the *p*-value needed for a graduate student to conclude the results of an experiment are statistically significant. The three choices were: *small p-value, large p-value,* and *statistical significance does not relate to the p-value* (the correct response being a small *p*-value). Table 10 lists the three possible choices and the number of participants that chose each response broken down by demographic information.

Table 10. *Results for the Multiple-Choice Portion of Graduate Student Task.*

	GTA	**GRA**	**GTA & GRA**	**CCI**	**FYI**	**CCI & FYI**	**Total**
Large p-value	0	0	0	0	0	0	0
Small p-value	21	5	3	14	9	1	53
Statistical Significance does not relate to p-value	1	0	0	1	0	0	2
Total	22	5	3	15	9	1	55

The results in Table 10 show that majority of participants (53 out of a total of 55) correctly identified that a small *p*-value is necessary. Only two participants incorrectly choose that statistical significance is not related to a *p*-value. This suggests a strong majority of the participants possess the CCK that small *p*-values are desired when trying to show significant results. The two participants who selected the incorrect choice of '*statistical significance is not related to p-value'* both had backgrounds in mathematical science and experience teaching elementary statistics courses in their respective institutions.

The results of the multiple-choice portion of the task might suggest that the background of the participants did not seem to play a major factor in a participant's ability to make the correct choice considering that the 53 participants who did select the

correct response originated from various professional or educational backgrounds. It is worth noting that the two participants whose educational background was in mathematical sciences were the only individuals who gave an incorrect response might suggests that there could be a potential gap for those teachers without a formal educational background in statistics. Furthermore, these two participants also identified themselves as having taught elementary statistics courses in the past but still gave an incorrect response, suggesting that even some practicing teachers of statistics could potentially lack knowledge on the magnitude of the *p*-value.

6.1.1.2. Qualitative results for the Graduate Student Task. The results in the previous section indicate that most participants could correctly identify that small *p*-values are needed to have statistically significant results. While majority of the participants selected the correct response to the multiple-choice portion, there were differences in the qualitative response when participants were asked to discuss the relationship between the *p*-value and statistical significance. This section is a summary of qualitative responses and the identifiable categories based on the analysis of the various explanations given by the participants.

Category *SSHP (Statistical Significance and High P-Value)* was a result of the two participants who chose the incorrect response in the multiple-choice section (i.e. statistical significance does not relate to *p*-value). Their explanations followed a similar theme where they describe that choosing the level of significance governs whether the results of an experiment are statistically significant regardless of the *p*-values magnitude. This can be seen from the responses given by the two participants in Table 11.

Table 11. *Examples of SSHP.*

Participant Information	**Explanation**
GTA Masters Candidate in Mathematical Science	We choose a completely arbitrary "significance level" (generally a "small" number, i.e. 0.05) and then we say a statistical experiment is "statistically significant" if our computed p-value is less than our significance level. If we chose a stupid significance level (say 0.99), then a "large" p-value like p=0.9 is statistically significant.
CCI Masters Degree in Mathematical Science	It would depend on the level of significance chosen for this problem.

These two excerpts illustrate a procedural understanding of the relationship between the level of significance and the *p*-value. This is very explicit in the first excerpt given by the GTA who uses the example of a significance level of 0.99 to describe how a research can have statistically significant results with a *p*-value as large as 0.9. This is true in theory since the procedure for showing a statistically significant result is based on checking whether the *p*-value is smaller than the level of significance. However, the choice of a significance level is not done arbitrarily. Statisticians are genuinely cautious with choosing a high level of significance because choosing appropriate significance levels is engrained in theories of statistical power and practical significance. Furthermore, choosing a large significance level does not make sense conceptually. The *p*-value measures the likelihood of an observed sample under a null hypothesis (i.e. the smaller the *p*-value, the greater the evidence to refute the null hypothesis). Setting a large level of significance allows samples that are not unusual to be acceptable evidence to refute the null hypothesis. Because of these responses, Category SSHP was designated to encompass those who gave incorrect responses in the multiple-choice portion based

purely on the idea that the statistical significance of a result is defined by the choice of the level of significance regardless if the level of significance is large or small.

The qualitative responses of the 53 participants who correctly made the choice of the "small *p*-value" were also analyzed for common themes. Four additional categories were identified as notable differences in understanding amongst the 53 participants based on their survey responses.

Category SSLP (Statistical Significance and Low *P*-value) describes those participants whose responses focused on the comparison of the *p*-values with the level of significance, but emphasize using small values for the level of significance. This is illustrated in the following excerpts shown in Table 12.

Table 12. *Examples of SSLP.*

Participant Information	**Explanation**
FYI in Statistics	Assuming she has a small level of significance (alpha) a statistically significant result would mean a smaller p-value than the alpha.
GTA. Masters Degree in Mathematical Science	The smaller the p-value, the higher the statistical significance. Typically, you set a cutoff value for significance before obtaining results. For example, alpha = 0.05. If the p-value is less than 0.05, you say that it is significant at the 5% level, even if the p-value you obtained is much smaller.

These responses highlight a procedural understanding of the relationship between the *p*-value and the level of significance that originates from knowing the hypothesis testing procedure. It is important to note that these excerpts share a very similar way of thinking as those in SSHP since they both describe a comparison of the level of significance to the *p*-value. Unlike the participants in SSHP who made incorrect choice in the multiple-

choice portion, the emphasis on using a small level of significance is crucial. Using a small level of significance such as 1% or 5% is common practice in introductory statistics courses. This is explicit in the second excerpt where the participant describes "typically" using a 5% level of significance. Unfortunately, those participants who fell in this category gave no further evidence why the level of significance must be small beyond simply stating a small level of significance. There were participants who did give a reason for choosing a small level of significance. These participants made connections to concepts of likelihood, statistical power, and practical significance. These participants will be described in future categories.

Category SSL (Statistical Significance and Likelihood) describes participants who chose a *small p-value* and gave a qualitative response describing how the *p*-value is a measure of the *likelihood* of an observed sample under a null hypothesis. This is shown in the following excerpts from the survey shown in Table 13.

Table 13. *Examples of SSL.*

Participant Information	**Explanation**
GTA in Stat	The smaller the p-value, the less likely you are to have randomly observed the data under the null hypothesis. Thus, with a small p-value, the null hypothesis becomes unlikely, and is rejected. We call data significant if it leads us to reject the null hypothesis, in favor of the alternative.
GRA in Math Ed.	The smaller the p-value, the less likely that the null hypothesis is true. A smaller p-value indicates that with the given null hypothesis, it is very unlikely that the null was true to get the result. The smaller the p-value, the stronger the evidence that the null isn't true.
GTA in M.S.	P-value measures the probability of obtaining the test statistic in the extreme direction suggested by the alternative hypotheses under the null. Therefore, a small P-value indicates that given the null the chance of getting the test statistic getting extreme values suggested by alternative is highly unlikely.

Notice that in all the given excerpts, we see a consistent way of thinking that focuses on how the *p*-value is an indicator to see how much an observed sample agrees with a null hypothesis. The main theme in these excerpts is the concept of *likelihood.* Unlike the previous categories where the participants only mention a comparison with the level of significance, these excerpts describe the *p*-value as a measure of evidence against the null hypothesis by viewing it as the likelihood of the observed sample. This type of understanding shifts from a standard procedural comparison as those in Category SSHP and SSLP, to a conceptual understanding of the relationship between the *p*-value, the null hypothesis, and the observed results of an experiment. This type of understanding helps highlight a flaw in participants' reasoning who fell in Category SSHP because it gives a practical reason why we do not choose a large level of significance. Setting a large level of significance (e.g. 99%) is contradictory to the concept of using the sample data as evidence to reach a statistically significant conclusion because it allows *almost all* samples to force a rejection of the null hypothesis. It is also important to note the type of thinking expressed by participants in this category coincide with the original concept of the *p*-value intended by Fisher when he originally designed hypothesis testing to show experimental results were statistically significant; whose design of hypothesis testing never relied on comparison of a level of significance but an understanding of the *p*-value's meaning.

Category SSD (Statistical Significance relates to Decision) described those participants who chose a small *p*-value in the multiple-choice portion, but discussed how the *p*-value is describing the probability of making the decision to reject the null

hypothesis. A sample of two explanations provided by participants is illustrated in Table 14.

Table 14. *Examples of SSD*.

Participant Information	**Explanation**
GTA doctoral candidate in Mathematical Science	The smaller the P-value the more likely the Null Hypothesis is rejected.
GTA doctoral candidate in Mathematics Education	The smaller the p-value, the less likely that the null hypothesis is true. A smaller p-value indicates that with the given null hypothesis, it is very unlikely that the null was true to get the result. The smaller the p-value, the stronger the evidence that the null isn't true.

In both excerpts, participants highlight how the *p*-value is a way to measure the probability of making a decision about the null/alternative hypothesis. In the first excerpt, the participant notes that the smaller the *p*-value, the more likely the null hypothesis is rejected. I hypothesize that the participant's thinking focuses on the relationship between rejecting the null hypothesis when the *p*-value is small that comes because of the procedure of hypothesis testing. While rejecting the null hypothesis is common for small *p*-values, the *p*-value does not measure the probability of making a decision of rejecting the null hypothesis. It is simply a consequence of the decision-making process of hypothesis testing. The second participant's explanations mention that the smaller the *p*-value, the less likely that the null hypothesis is true. This excerpt similarly connects the *p*-value as a probability of the null hypothesis. This is once again not true since the *p*-value only measures the behavior of a sample under the null hypothesis and not the

probability of the null hypothesis. It is important to note that the *p*-value only gives us evidence against the null hypothesis under random chance, which is sometimes mistaken as the likelihood the null hypothesis.

Category SSPPS (Statistical Significance and Power and/or Practical Significance) was the final category that described those participants who chose a *small p-value* in the multiple-choice portion, but discussed how the magnitude of the *p*-value and level of significance is related to theories of statistical power and/or practical significance. A sample of two explanations provided by participants is illustrated in Table 15.

Table 15. *Examples of SSPPS.*

Participant Information	**Explanation**
CC with Masters in Mathematical Science	Statistical significance depends on comparing a P-value to a chosen significance level. When the P-value is less than the significance level, results are deemed "statistically significance". The practical significance of research results is another thing entirely. The trick is choosing that significance level---not a simple task.
FYI with Doctorate in Other Field	Even though the results might not be of practical significance (i.e., the effect size is small), a p-value lower than the researcher's chosen alpha-level would provide evidence of statistical significance.

While participants in Category SSL and 4 describe the *p*-value as measures of likelihood and evidence against the null assumption, the participants in these excerpts discuss notions of "practical significance" and/or relate to notions of statistical power when discussing connections between *p*-values and a level of significance. Choosing an appropriate level of significance is an important idea done by statisticians when

performing statistical experiments. This understanding extends beyond a typical procedural view of simply comparing the *p*-value with an arbitrary level of significance, but a deep theoretical understanding for why it is inappropriate to choose a big level of significance. This way of thinking also provides a reason why the arguments made by participants in Category SSHP are considered incomplete because choosing a big level of significance is counterintuitive to the ideas of looking for results that would provide evidence against a null assumption. It also gives a theoretical reason for the thinking displayed in Category SSLP where participants simply mentioned choosing a small level of significance. Furthermore, Category SSPPS provides a deeper theoretical reason for the responses expressed in SSL whose explanations for wanting a small *p*-value relied on a surface level description of the *likelihood* of a *p*-value. Category SSPPS responses is more robust way of thinking procedurally and conceptually about the relationship of *p*-values and level of significance using theoretical ideas steeped in the theories of statistical power. This level of understanding tends to be under emphasized in many introductory statistics classes, but is one valued as important to actual statisticians.

In addition to the categories, there were participants who fell into more than two categories. For example, there were participants whose explanations illustrated a combination of Categories SSLP and SSL. These participants chose the correct answer to the multiple-choice portion of the *Graduate Student Task*, explained how the procedure of the *p*-value being compared with the level of significance, and then mentioned the *p*-value as a measure of likelihood of the observed sample under the null assumption. A sample of one such excerpt of this type includes the one shown in Table 16.

Table 16. *Examples of Hybrid Thinking of the Magnitude of the P-value.*

Participant Information	**Explanation**
GTA and Master's Candidate in Statistics	"Statistically significant" means that the p-value is less than the level of significance, which is set at the beginning. Also, the definition of p-value is the probability of getting a statistic as extreme as or more extreme than what we observed, if the null is true. If we assume the null is true (or assume a statement of status quo or no difference) and then we get results that are very different from what the null claims, then this would be good evidence that the null is not true. In other words, if we calculate a statistic that is unlikely if the null actually is true, then we know that the probability of getting a statistic at least as extreme as what we got is very low. Thus, the p-value is low, and we would reject the claim, or the null hypothesis.

In this excerpt, we see the participant mention both a comparison of the level of significance with the *p*-value in the first sentence, but also the conceptual relationship of the *p*-value as a measure of an observed sample's likelihood under the null hypothesis. This shows a hybrid understanding of the *p*-value that connects the procedure of comparing the *p*-value with the level of significance with a conceptual understanding for choosing a small level of significance. I view this as an integration of a procedural and conceptual understanding of a *p*-value's magnitude. Statisticians choose a small level of significance in the first place is because researchers want to set a small cutoff value for samples they consider to be *unusual enough* under the null assumption. If a *p*-value falls below this cutoff value of unusualness, then there is a clear indication that an observed sample is behaving very differently from a null assumption and results in significant results.

Table 17. *Results of Categories for the Magnitude of the P-value.*

	GTA	GRA	GTA & GRA	CCI	FYI	CCI & FYI	Total
Category SSHP – Statistical Significance Relates to High P-Value	1	0	0	1	0	0	2
Category SSLP – Statistical Significance Relates to Low P-Value	5	0	0	3	6	0	16
Category SSD – Statistical Significance relates to Decision	4	1	0	0	1	0	6
Category SSL – Statistical Significance relates to Likelihood	5	0	1	7	1	1	15
Category SSPPS – Statistical Significance relates to Power and/or Practical Significance	0	0	0	0	0	0	0
Hybrid Category of SSL & SSD	1	1	0	1	0	0	3
Hybrid Category of SSLP & SSL	1	0	1	0	0	0	2
Hybrid Category of SSL & SSPPS	0	1	0	0	0	0	1
Hybrid Category of SSLP & SSPPS	0	1	0	1	1	0	3
Responses that could not be categorized	5	1	1	2	0	0	9
Total	22	5	3	15	9	1	55

After going through the coding process of Test-Retest, the qualitative results for this showed an inter-reliability rate of approximately 81% based on the discussed categories. Table 17 displays the results of the coding process for the total count of the categorizations describing how the participants viewed the relationship between a *p*-value's magnitude and statistical significance. The table also highlights whether participants either fell in a single category of combinations of categories. The results of the categorization showed a diverse understanding amongst the participants. Even though

53 out of the 55 participants all agreed on the correct magnitude of a *p*-value needed to claim a statistically significant conclusion, the results in Table 17 show that not all participants think of the *p*-value's magnitude in the same manner.

A total of 39 participants' explanations only incorporated a single way of explaining a small *p*-value. The primary explanations utilized either using SSLP or SSL as a common theme in the relationship between a *p*-value and statistical significance. Seven of the four-year university instructors fell into SSLP, which is the more procedural view of *p*-value. This contrasts with the seven community college instructors who fell in SSL which is a more conceptual view of the *p*-value. The graduate students were about equally spread amongst the categories of (SSLP, SSD, and SSL). Furthermore, none of the participants noted simply discussing statistical power (SSPPS) in their explanations. When they did mention ideas related to SSPPS, it was in relation to other concepts such as level of significance (SSLP or SSL).

A total of nine participants' explanations incorporated hybrid ways of relating *p*-value to statistical significance. Three of the participants discussed concepts related to SSL, but whose explanations began suggesting the misconception that the *p*-value can measure the probability of the null/alternative hypothesis (SSD). Two participants made connections between choosing a small level of significance (SSLP) and how it relates to likelihood (SSL). This is surprising considering that being able to coordinate procedural understanding and conceptual understanding of the *p*-values magnitude is a crucial idea in understanding a *p*-value. When relating to concepts of statistical power and/or practical significance (SSPPS), one participant made a connection to likelihood (SLP). Three

participants chose to relate their understanding of choosing a small level of significance (SSLP) to the importance of statistical power and/or practical significance (SSPPS). This is not surprising considering that discussion of statistical power and practical significance is deeply connected to notions of choosing an appropriate level of significance and the practicality of a statistical claim. Therefore, it makes sense that none of the participants would simply discuss statistical power or practical significance without some relationship to the procedure of choosing a level of significance or reflecting on the context of making a statistical claim.

Finally, nine of the participants gave either no explanation or responses that did not reveal useful information regarding understanding of the *p*-value's magnitude. The largest majority that fell in this category was graduate teaching assistants. Unfortunately, it is still difficult to make larger claims about the GTAs (or CCIS) overall understanding of the magnitude of the *p*-value. When some GTAs were given an opportunity to further articulate their explanation during the interviews, two GTAs provided even deeper understandings of the *p*-value than otherwise categorized in the survey responses data while others only articulated an understanding that even further provided evidence that they only thought of the *p*-value at a procedural level (SSLP). This indicates potential room for improvement in the demographic of graduate teaching assistants in developing the ideas of magnitude of a *p*-value and statistical significance.

6.1.2. Interview Results of the Graduate Student Task. The variety of qualitative responses provided by the seven interviewees on the relationship between statistical significance and the *p*-value was one of several reasons for choosing them.

Table 18 is a summary of the seven interviews, their survey responses, and the categories of understanding discussed in the previous section. An initial look at the categorization showed that amongst the seven participants, some shared overlapping categories. Sean's and Phil's survey responses were coded as SSLP, which related to how statistical significance was a result of comparing a *p*-value with a fixed level of significance. Angie's and Jane's survey responses were coded as SSL, how the *p*-value helps measure the unusualness of a sample in relation to a claimed null hypothesis. James' response was coded as Category SSPPS whose response noted the importance of practical significance and the difficulty of choosing an appropriate level of significance. Tod's explanation fell into the misconception of the *p*-value categorized as SSD. Lastly, Carol's survey explanation did not provide useful information.

Table 18. *Responses for the Interview Participants on the Graduate Student Task.*

Participants	**Responses and Categorization**
Tod	*The smaller the P-value the more likely the Null Hypothesis is rejected. (SSD)*
Angie	*Small p-value indicates small probability of given results under the null. This indicates a significant different from expectation, and thus a statistically significant result with which we can reject the null. (SSL)*
Sean	*You set up the level of significance beforehand and it your p-value is smaller you have shown it true.* *(SSLP)*
Carol	*The smaller the p-value, the better the statistical significance! (Uncategorized Response)*
Jane	*The p-value is describing the probability of getting results as extreme or more extreme as what is found in the study. If this value is large, it is saying that the results are not that extreme under the assumption of the null hypothesis. The smaller the p-value, the more statistically significant the results are because it is describing a very small probability of it occurring under the null hypothesis.* *(SSL)*

Phil	*"Small" is relative - it's only "small" when compared to the significance level. So, if this student wants to "show that the results of an experiment are statistically significant", she wants a P value that comes in under alpha (assuming that she has controlled for any extraneous variables and set up correct hypotheses).* *(SSLP)*
James	*Statistical significance depends on comparing a P-value to a chosen significance level. When the P-value is less than the significance level, results are deemed "statistically significance". The practical significance of research results is another thing entirely. The trick is choosing that significance level---not a simple task.* *(SSLP & SSPPS)*

After analyzing the interview responses, it became apparent that some of the interviewees had a more robust understanding of the relationship between the *p*-value and statistical significance than was initially revealed in their survey data. The categories that resulted from the analysis of the survey responses were still useful in analyzing the interviewees understanding of the magnitude of the *p*-value and provided a useful foundation when analyzing how the interviewees understood the relationship between the *p*-values magnitude and a statistically significant result.

Analysis of interviews provided greater evidence for the categorization initially generated in the survey regarding ways the participants viewed the *p*-value's magnitude both procedurally and conceptually. First, I present excerpts from the interview showing evidence of participants who only focused on the procedural comparison of the *p*-value with the level of significance (SSLP). Second, I present excerpts from the interview showing evidence of participants who primarily focused on the conceptual understanding of the *p*-value as a measure of likelihood (SSL). Finally, I present excerpts from the

interview showing evidence of the hybrid thinking where we see participants discussing how the procedure of comparing the *p*-value with the level of significance relates to concepts of likelihood, practical significance, and statistical power.

6.1.2.1. Evidence of procedural understanding of a p-value's magnitude. Carol was a Masters student in mathematical science and mathematics education with less than a year of experience in teaching statistics. An initial categorization of her survey response did not show a clear evidence how she connected statistical significance with the *p*-value. When given the opportunity to discuss further her response, regarding why she chose a small *p*-value and how it might relate to her thinking of statistical significance, she relied heavily on the procedure of comparison with the level of significance as her response. For example, in the excerpt below we see how Carol describes her thought process when she was asked about the relationship between a small *p*-value and the results of the hypothesis test.

Interviewer: What does that tell me about my null hypothesis?

Carol: That you are going to reject it. Right because well it all depends on what your level of significance is. I think students have to interpret themselves. Because if we are testing the p-value against an alpha of 0.05. And the student gets an p-value of like 0.048 and they are super close together and you could either reject or not reject it because they are so close together and I think there it is up to the student to make like, to make the decision to actually think about the procedure and think about the data itself and then make his or her decision.

Interviewer: So here is one question. So, let's say you are doing a hypothesis test and you just got the p-value and you didn't even compare with the level of significance. Like I said, I got my p-value of 0.01 or 0.001. Can I just claim that my null is false? Can I do that just by looking at the p-value without actually comparing it with the level of significance?

Carol: I think so yeah.

Interviewer: In what way?

Carol: I think that is because you are conditioned to compare p-value with 0.05 or 0.01. And anything lower than 0.01 we are going to safely assume that we are just going to reject it.

Interviewer: Do you always have to make that comparison?

Carol: Yes.

In Carol's explanation, I wish to highlight how she discusses components of the standard practice of hypothesis testing done in introductory statistics courses; the procedure of rejecting the null hypothesis when the *p*-value is smaller than the level of significance. Furthermore, when questioned whether it is necessary to even make the comparison of the *p*-value with the level of significance even with a small *p*-value she responds with the idea of being 'conditioned' to compare it with a 0.05 or 0.01 but never giving a reason how those 0.05 or 0.01 are chosen or discussing the *p*-values as a measure of likelihood. Her use of the 'conditioned' is also indicative of memorizing a procedure that has been a pitfall of many introductory statistics students for simply using any standard level of significance without thinking about the consequences. This illustrates

greater evidence for Category 2 style thinking that was illustrated by many of the responses found in the survey.

6.1.2.2. Evidence of conceptual understanding of a p-value's magnitude. Jane was an instructor in CC and FUI with a Masters in Mathematics Education who at the time was also teaching introductory statistics courses. In her initial survey response, Jane's responses put her in Category 3. When Jane was asked during the interview for further clarification of her survey response, she seemed to push strongly for the idea on how *p*-value is a measure of likelihood of the observed sample under the null assumption. An example of this was seen in the interview excerpt below.

Interviewer: How do you relate the p-value and statistically significance. So, if a student asks you how are they related?

Jane: So, the p-value can give us this kind of how extreme it (the sample) is under the assumption and if it (p-value) is giving a very small value then it (the sample) is very extreme which makes it statistically significant.

Interviewer: Okay. Last one for this one. If a student asked you does the actual size of the p-value being an important thing or should I just not care?

Jane: If you are trying to determine statistical significance then the size is very important.

Interviewer: And from what I am seeing from your picture is that one.

Jane: It has to be out there (points to the tail of a normal distribution) in order...it has to be out from the tail to provide statistical significance.

In the exchange above, we see Jane focus solely on the importance of the tail sections of a distribution and emphasizes that small *p*-values lead to statistical significant results because it shows how extreme a sample is a null assumption. This thinking does not rely on a comparison of the *p*-value with the level of significance, but the importance of the likelihood of the observe sample. Furthermore, Jane's response to how she would respond to a student differs from that of Carol above who focused on the comparison. Jane's responses relied on a visual model of a distribution and the idea of where the sample would fall in a distribution. The thinking displayed by Jane in the above excerpt gives stronger evidence for the type of thinking displayed by participants in Category 3.

6.1.2.3. Evidence of hybrid understanding of a p-value's magnitude. The last four individuals the interview (Phil, Angie, Sean, and James) each gave various categories of survey responses, but during the interview the importance of choosing an appropriate level of significance was eventually brought up by all four of these participants. Initially, Phil and Sean both fell into Category 2 in their survey response, Angie was in Category 3, and only James was in Category 5. Analysis of these four interviewees showed that at the end, all four of them displayed Category 5 thinking.

Each interviewee could discuss ideas related to the likelihood of the observed sample to the null hypothesis, mention comparison of the level of significance with the *p*-value, and discuss ideas on how to choose an appropriate level of significance. The interviewee, Phil, illustrated an example of this as shown in the following excerpt.

Phil: P-value kind of tells you how far once side or which one you should lean to. Like I should be leaning towards H_0 because the p-value is huge. What does

that mean? There is a lot of agreement between the data I saw and the null hypothesis I think I was going to see. A lot of agreement means I believe in the null. That kind of a that kind of a flow charty way.

Interviewer: So, if a student were to ask you, does the size of the p-value, is that an important thing or is it just a number to me? How would you respond to that?

Phil: Well I think the answer to both those questions is yes, it is just a number. The way research is done these...again Fisher never intended p-values the way they are being used in research. When he proposed them, he literally proposed that's the definition and that was it. They are being used unfortunately I just did stats for an OSU student about a year ago and she gave me a bunch of numbers and she's like can you find statistical significance with this? She didn't bother telling me alpha equals 5%, she simply says...this is a person who doesn't understand statistics at all, but she knows enough that she will not get published unless I can somehow manipulate this data so that less than 5%. You're laughing because you get the joke. It's completely ridiculous. I did a non-parametric test and I'm like look 20 different associations and I found 2 if I do them pairwise that find significance. Here is why I think you shouldn't publish this. I feel like I am p-hacking with your data. I am basically throwing stuff at a wall and seeing what sticks and two of them manage to 5% or less come out. You can't find more? Like do you understand what you are asking me to do? You're asking me to find potentially associations just to get you

> small p-values. I tell these students these stories all the time because she is not the first one to have this happen. To ask me to do this. So the small p-value is...think about it this way. By the time you've gotten to the crunching of the numbers in your research experiment, 99.9999% of the hard work is already done. You've collected data that is hopefully pointing at what it is you're trying to measure. You have established guidelines and benchmarks that hopefully, when you put the line in the sand and alpha equal whatever it is, is being meaningful in a practical and statistically significant way.

In this excerpt, we see Phil discussing an experience he had with a student related to the differences between comparing the level of significance and the meaning of the *p*-value as a measure of *unusualness*. In his explanation, he describes how the student is simply focused on the procedural comparison of the *p*-value with the level of significance and in his rebuttal also discusses that this is something you should not be doing in practice because it does not agree with the work of Fisher who original meant for the *p*-value as a measure of the *unusualness* for a observed sample against the null assumption. His ability to relate the two types of understanding shows strong evidence of combining SSLP and SSL. This way of thinking also influences his pedagogical view because it gives him an ability to assess how a student is thinking of the relationship between the *p*-value and statistical significance incorrectly. He remarks that just trying to get a *p*-value below a level of significance is not the only important part of the hypothesis test but thinking about what that *p*-value means is just as vital. He uses the metaphor of putting a line in

the sand for choosing a level of significance based on *guidelines* and *benchmarks* which shows evidence of SSPPS thinking.

Evidence of Phil being in SSPPS was mainly overshadowed by his focus on the idea that *p*-values helps provide evidence against the null assumption. It was only apparent his thinking was in Category 5 when he was asked to discuss what he think is meant by statistical significance.

> **Phil**: Statistically significantly is different from the other one. So the first thing to back up a little bit more, your statistically significance...one of the first things I have students understand in the first couple of weeks we are doing confidence intervals is the difference between statistical significances and practical significance. Because that is huge, because you can get statistically significant results by increasing your sample size without having real practical effect. And we can convince ourselves pretty easily that until we're experts in our fields, we won't know what practical significance is and we can discuss that all day long with our colleagues. So statistical significance essentially means that you know...when something is statistically changed or different. We have gotten data that have moved far enough away, quote, unquote, far enough away from some null assumption that you flag it as something as changed or it hasn't gotten far enough away.

In the excerpt, we see Phil discussing the difference between statistical and practical significance but still focus on the importance of the data showing a difference from a null assumption. He also continues to discuss ideas that relate to the difference between

statistical significance and practical significance when he makes comments about setting appropriate sample sizes.

As mentioned, similar ways of thinking were expressed by the other interviewees. For example, Sean, makes the following statement when discussing the importance of the choosing the level of significance.

Sean: What were the p-values for experiments where you used a larger sample size? What were the p-values for a smaller sample size? Because it was like does it mean anything? That's where you have the alpha. The alpha almost carries like information from previous studies and from the field itself. So it is saying this is the statistical significance for this field. You might have a 0.01 for a psychology experiment. You might have a 0.01 for an engineering experiment.

If you might recall, Sean's background is a doctoral candidate in statistics who also has experience doing statistical research. In his discussion, he focuses on two important ideas. One is the importance of choosing appropriate samples sizes and the importance of prior research when choosing an appropriate level of significance. These are all important aspects of the theoretical concept of statistical power that is something not discussed by all three other interviewees (i.e. Phil, James, and Angie). Another interviewee, Angie, who also fell in this category brought up a similar idea of the importance of choosing an appropriate level of significance by discussing two different contexts where choosing an arbitrary the level of significance could have severe consequences.

Angie: And where you put that obviously depends on a lot of things. Umm...how important it is you are correct in your assumption that it is for example that the

> alternative is true. So in medicine for example, we wouldn't want to use a 5% p-value and say you know we tested this drug we found we found with 5% significance that it helps people. That's not going to quite cut in in the medical community, but if you are just doing a study on will this installing a sidewalk here do students want a sidewalk installed through this walkway. Something really small and you say yes more students want one here than want the garden that is in place then 5% would cut it. So, it is really just are the results you see errant enough from your assumption that you are willing to conclude the alternative.

Like Sean, Angie is a doctoral candidate in statistics with a background in statistical research. In her explanation, we see an elaboration on the importance of context in choosing the level of significance that relate to the severity of a result. In her work, she focuses on the differences between concluding the results of a survey on sidewalks versus the results of a medical experiment. In her description, she views the results of hypothesis testing on medical experiment being a much bigger consequence. This focus on the importance of practical significance of the *p*-value and how it is important to consider what the results might mean if we just arbitrarily set a level of significance without much thought.

6.1.3. Summary of the Graduate Student Task. The Graduate Student Task provided evidence of ways teachers might understand the role of the *p*-value in making statistically significant claims. The initial survey provided a strong foundation for analyzing how we might come to categorize ways teachers view the magnitude of the *p*-value and statistical significance. The interviews provided even greater evidence for these

categories by showing us how teachers come to understand the procedural and conceptual concepts of a *p*-value's magnitude. The categories themselves can also be beneficial for future researchers because it shows a hierarchical understanding a *p*-value's magnitude that applies to both teachers and students. SSHP itself shows a base level understanding where the focus is simply on comparing a *p*-value with a level of significance without justification for choosing a level of significance. SSLP is the next step because it relies on fixed values for level of significance (e.g. 1% or 5%) to help in making that comparison, but it still does not give meaning to the *p*-value. SSL starts to give meaning to the *p*-value as a measure of likelihood but extra understanding is necessary to connect it to SSLP. Unfortunately, just understanding SSL and SSLP is very limited because it's reason for choosing a level of significance is simply based on a personal view of *unusualness.* In other words, some researchers might consider 1% to be an acceptable level of significance because it is already a small cutoff measure of unusualness, but researchers might see it as not unusual enough. SSPPS is where most statisticians and we hope teachers would be because it provides the most robust understanding of the relationship between *p*-values and level of significance. The reason is because SSPPS thinking about statistical power and practical significance helps researchers considers what might be appropriate cutoff measures of *unusualness* through a set of fixed method and guidelines that have a theoretical foundation.

The interview data does show some promising results because it illustrated that even those participants who fell into lower categories in the survey data eventually showed SSPPS thinking during the interview. Unfortunately, this was not true for all the

interviewees. Of the seven, only four showed evidence of SSPPS. This data shows that there is a still a need for professional development in the concept of a *p*-value's magnitude.

6.2. Computing *p*-values with Samplings Distributions.

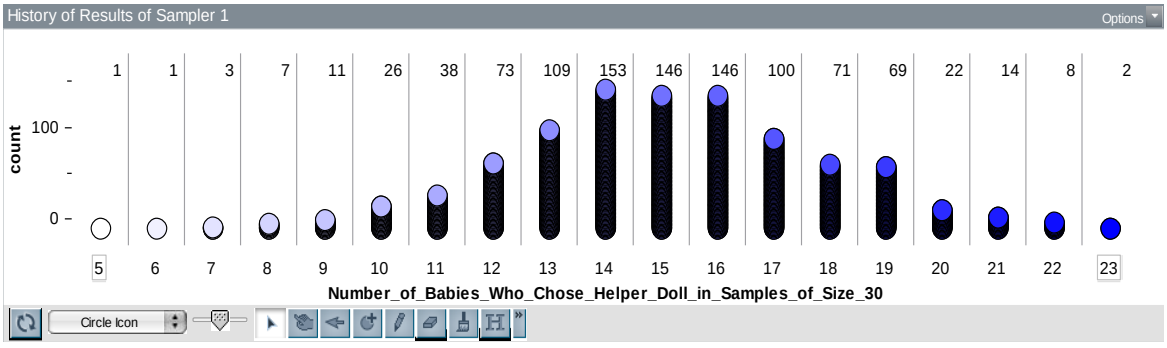
A component of a robust understanding for a *p*-value includes knowing how it relates to sampling distributions and simulations. Traditionally, hypothesis testing in introductory statistics classes rarely use simulation approaches to hypothesis testing. As discussed in the literature, traditional courses rely on methods that shroud deep theoretical ideas of statistical inference. The simulation approach to hypothesis testing removes some of the theoretical complexity through the creation of (computer) models that generate empirical sampling distributions. The computation of the *p*-value then becomes more accessible to students because it can be found using relative frequencies. This makes it easier to assess ways an individual might compute a *p*-value. This motivated the investigation of the second theme that analyzes participants' understanding of the relationship between *p*-values, sampling distributions, and simulations.

The main purpose of this section is to present results on the participants' understanding of the *p*-value computation when using a simulation approach to hypothesis testing. The primary task used to analyze how participants understand the relationship between *p*-value, simulations, and sampling distribution was the *Helper-Hinderer Task* (Figure 18). The task was developed to assess participants' knowledge by presenting a hypothetical student solution to a statistical problem portrayed as a simulation approach to a one-proportion hypothesis test. I primarily focus on data

resulting from the second question where the participants were prompted to find an approximate *p*-value and to explain their computation. The survey results showed seven categorical approaches where participants correctly (and incorrectly) computed the *p*-value using a relative frequency or utilized theoretical probability. I also present interview data that further elaborates why participants chose to do specific computational approaches. The interview contained noteworthy moments where the participants made important connections between their understanding of the simulation models and how it potentially affects their pedagogical thinking.

A sociology study was conducted to determine whether babies are able to recognize the difference between good and bad. In one experiment, 30 six-month old babies were randomly selected. Each baby was shown two possible puppets to play with, a 'good' puppet that helped and a 'bad' puppet that hindered. 21 out of 30 babies showed a strong preference for the helper puppet over the hinderer. In order to determine if this result provides strong statistical evidence that babies really do have a preference for the 'good' or helper puppet, James, a statistics student, conducted the following test procedure:

- James gets a coin and flips the coin 30 times.
  - If the coin lands on the "heads", he records the baby as preferring the helper puppet.
  - If the coin lands on the "tails", he records the baby as preferring the hinderer puppet.
- James then used a computer simulation to repeat the previous step 1000 times.
- James then plots the distribution for the number of times a baby chooses the helper puppet from each of the 1000 samples of size 30. This is shown in the graph below.



i.) James' procedure is based on which assumption?

a. A baby is more likely to choose the helper puppet.

b. A baby is equally likely to choose either the helper or hinderer puppet.
c. A baby is more likely to choose the hinderer puppet.

Explain the reason for your choice.

ii.) Suppose James wanted to conduct a right-tailed hypothesis test using the simulated data.

- What would you estimate for the *p*-value?
- Explain how you found the *p*-value and interpret it in the context of James' research.

iii.) Based on your estimated *p*-value, what do you think should be James' conclusion?

a. There is statistically significant evidence that babies are more likely to choose helper puppets.
b. There is statistically significant evidence that babies are more likely to choose hinderer puppets.
c. There is statistically significant evidence that babies are equally likely to choose helper or hinderer puppets.

Explain the reason for your choice.

Figure 18. Helper-Hinder Task

6.2.1. Survey Results: Computing p-values with sampling distributions. The second question of the *Helper-Hinderer Task* was designed to asses a participant's ability to compute an approximate *p*-value using an empirical sampling distribution generated via a simulation. Participants were directly asked to find an approximate *p*-value and then prompted to explain their computation. The correct *p*-value was expected to be 0.024. Table 19 shows the results of the number participants who correctly computed the *p*-value, those who computed an alternative correct *p*-value, and those who computed an incorrect value broken down their demographically.

Table 19. *Results of P-value Computation for Helper-Hinderer Task.*

	GTA	GRA	GTA & GRA	CCI	FYI	CCI & FYI	Total
Correct Computation	15	5	2	11	6	0	39
Incorrect Computation	6	0	1	4	2	0	13
Correct Computation using alternative methods	1	0	0	0	1	1	3
Total	22	5	3	15	9	1	55

Table 19 shows that 39 of the 55 participants correctly computed the expected *p*-value of 0.024 using the empirical sampling distribution. There were also 16 of the 55 participants who gave responses other than 0.024.[12] Of the 16 participants, three gave a computation of a *p*-value that could be considered correct. These computations included those participants who correctly used a binomial distribution, normal approximations, or a one proportion hypothesis test. While the computation of the *p*-value by these participants did not use the empirical sampling distribution as instructed in the task, they still correctly computed the *p*-value using standard statistical procedures. As a result, they were counted differently than those participants who computed a *p*-value that were clearly incorrect. Those participants who were assessed as giving incorrect *p*-values used approaches that did not correspond to either a correct theoretical or empirical approaches to the *p*-value's computation.

While majority of the participants gave the expected response for the *p*-value, the fact that approximately 29% (16 out of 55) of the participants gave varied responses shows a large percentage of participants that did not understand the purpose of the empirical sampling distribution provided. These individuals included nine of the graduate students and seven of the instructors. These initial results shows a potential gap in the knowledge of the participants regarding their understanding of a simulation approach to hypothesis testing.

After reviewing the results of the various computations, the responses provided by the participants' explanations were categorized. Seven categories of computational

[12] The responses given by these 16 people will be explained later in this section.

approaches were identified from the survey. Two categories were identified for the 39 participants who gave the correct *p*-value of 0.024 and five additional categories were identified for those participants who gave incorrect computations. These categories will be explained in further detail below.

Category CRF (Computation with Relative Frequency) describes those participants who that only described a computation where they counted the number of observations in the empirical sampling distributions that occurred for the outcomes of 21 and higher. Sample excerpts from the survey include those seen in Table 20.

Table 20. *Examples of CRF.*

Participant Information	**Explanation**
FYI Doctorate in Statistics	Counted the number of replicates that were 21 or more and divided by 1000.
GTA Doctoral Candidate in Statistics	By looking at the graph, you can see that 24 of the experiments had at least 21 babies preferring the helper.
GTA Doctoral Candidate in Mathematics	It is the area passed the expected results. So (14+8+2)/1000.

The excerpts given in Table 20 shows a process done by participants where they first noticed the observed sample gave a proportion of 21/30. They proceeded to then count the frequency of occurrences for 21 or greater in the empirical sampling distribution which amounts to (14+8+2)/1000 = 24/1000 = 0.024. An image of the empirical sampling distribution from the *Helper-Hinderer Task* with the marked *p*-value region is shown in Figure 19 below. The approach described by these participants aligns correctly with the expected computation of a *p*-value by computing the relative frequency from the

empirical sampling distribution using the definition of the *p*-value as "the probability of getting results as extreme or more extreme as the observed sample statistic."

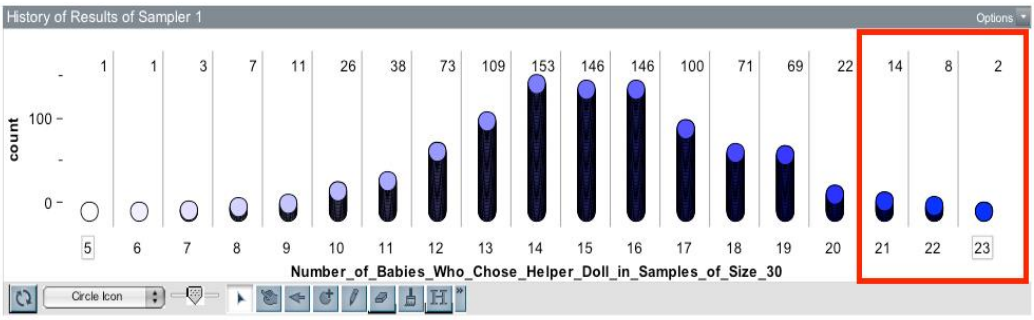


Figure 19. Empirical sampling distribution with the correct *p*-value region.

Category CRFANH (Computed with Relative Frequency Assuming Null Hypothesis) describes those participants whose explanations encompassed the ideas of a correct computation of the *p*-value using relative frequency with an additional emphasis of the null assumption. Unlike those participants in Category CRP who only described how they are counting the frequency of occurrences, these participants included in their explanations a reference to a null hypothesis. Sample excerpts for this category include those seen in Table 21.

Table 21. *Examples of CRFANH.*

Participant Information	**Explanation**
GTA Doctoral Candidate in Mathematics	The plot shows the (approximate) distribution of "number of babies who chose helper dolls in a sample of size 30" under the assumption that babies are equally likely to choose either the helper or hinderer puppet. The p-value is then the probability of obtaining a result as extreme as the one we did (i.e. 21 or more babies out of 30) under that assumption. The plot shows that 21/30 occurred 14 times, 22/30 occurred 8 times, and 23/30 occurred 2 times (and presumably 24 and above did not occur at all!), so the p-value is (14+8+2)/1000, since there were

	1000 samples from the distribution.
FYI Doctorate in Statistics	(14+8+2)/1000 (number of times 21 or higher observed divided by total number of trials) / Assuming that a baby is equally likely to choose either the helper or the hinderer puppet, the probability of 21/30 babies choosing the helper is 0.024.

In both excerpts, we see the participants refer to the null hypothesis of the problem (i.e. a baby is equally likely to choose the helper or hinderer puppet). As I alluded to in the literature, understanding the role of the null hypothesis in the computation of the *p*-value is problematic for many students because they focus primarily on the computation and not on the overall process of hypothesis testing. The fact that these participants emphasized the importance of the null hypothesis in their response shows evidence they value the connection between the computation of the *p*-value and the null hypothesis.

While the majority of the participants correctly computed *p*-value using the empirical sampling distribution, 16 participants gave *p*-value's that differed from the expected computation of 0.024. Table 22 summaries the different computations of the *p*-value by the 16 participants.

Table 22. *Alternative Computations of the P-value for the Helper-Hinderer Task.*

Incorrect Computations of the p-value	**Frequency**
p-value = 0.09	1
p-value = 0.05	2
p-value = 0.04	1
p-value = 0.03	1
p-value = 0.01	2
p-value = 0.01426	1
p-value < 0.05	3
p-value < 0.025	1
p-value > 0.05	1
p-value = 1/(2^30)*[30nCr16 + 30nCr17 + ... + 30nCr30]	1
p-value = find the z (because it's a proportion)	1
No Computation Provided	1

One participant gave no computation and no explanation on how he/she would compute a *p*-value. This participant was a community college instructor with a mathematical science background. Another participant simply stated to find a 'z' but did not provide any computed value. This was a graduate teaching assistant in statistics. The 14 other participants who did give a description gave various approaches to the computation. Even though Table 22 shows a variety of computed *p*-values the responses for their computations provided by these participants showed areas of commonality. A total of five categories were identified as ways these participants approached the computation of the *p*-value.

Category CIOS (Computation Ignores Observed Sample) identifies the approach done by two participants that gave an incorrect computation of *p*-value = 0.01. One was a graduate teaching assistant in statistics and the other was a four-year university instructor who majored in a field outside mathematics and statistics. These participants' description of their computational method showed a misconception regarding the use of the empirical sampling distribution in computing the *p*-value. This is seen in their explanations shown in Table 23.

Table 23. *Examples of CIOS.*

Participant Information	**Computation and Explanation**
GTA Master's Degree Candidate in Statistics	*Computation: p*-value = 0.01. *Explanation:* It is bootstrapping method. one side test. P(T>21|p=0.5) = 10/1000=0.01.
FYI Doctorate in Other	*Computation*: *p*-value = 0.01 *Explanation*: 10/1000 is the approx. area cut off by 21.

In these two excerpts, we see the participants computing a *p*-value by only considering all values *greater than* 21. An image of the region being computed by the participants is shown in Figure 20. This is made explicit in the explanation provided in the first excerpt where the participant uses the ">" symbol in his/her conditional statement rather than "≥." The second excerpt is more challenging because the participant only mentions that it is the area cut off by 21. I hypothesize that even though the participant did note the value of the observed sample statistic (i.e. 21), the participant does not include the frequency of the observed sample statistic in their overall count and only considers occurrences more extreme than the observed sample statistic because a total of only 10 frequency points occur past 21. The methods described by these two participants incorrectly compute the *p*-values because it disregards the observations that occur at 21, which is an important component of the *p*-value's definition. In other words, this category describes a misconception where the computation of the *p*-value does not include the probability of the observed sample outcome. Interestingly, this type of misconception is not problematic in the traditional hypothesis testing that primarily use sampling distributions that rely on continuous random variables because computing a *p*-value using the ">" instead of the "≥" in the continuous case would yield similar results. This only becomes problematic in empirical sampling distributions because relative frequency is used in the computation.

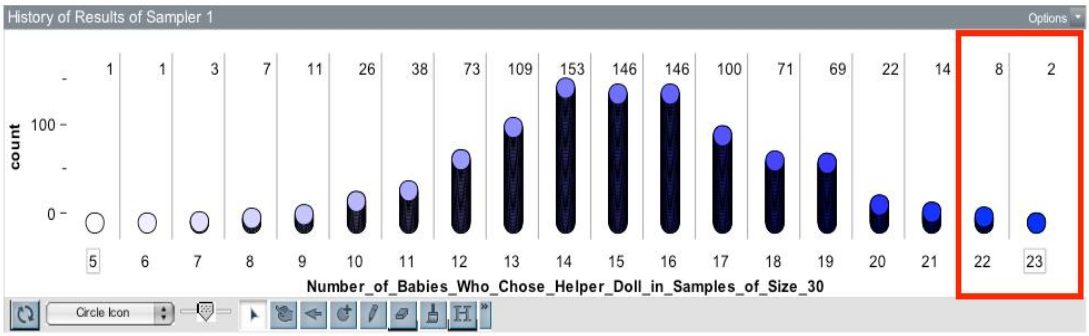


Figure 20. Empirical sampling distribution illustrating CIOS.

Category CUR (Computation using Regions) describes three participants who computed a *p*-value of 0.09. In the participant's work, the computation of the *p*-value was incorrect because of a misconception regarding their usage of the empirical sampling distribution. This is seen in the following excerpt in Table 24.

Table 24. *Examples of CUR.*

Participant Information	**Computation and Explanation**
GTA Master's Candidate in Other	*Computation*: *p*-value = 0.09 *Explanation*: I just took the heights corresponding to 16,17,18,19,20 and added them up (which was 408) which I, then, subtracted from 500 (total of the right half). / 0.09 is still not a very small p-value, it is not a strong evidence in the favor of alternative hypothesis which would be the validity of the claim that the babies do in fact differentiate between good and bad.
CC Doctorate in Other	*Computation: p-value* = 0.05 *Explanation:* 24/500 observations at 21 or above.

In first excerpt, we see the participant is counting all the values to the right of the center (i.e. 15) of the empirical sampling distribution which he/she notes is 408 and then subtracting that from 500. This would produce a value of 92, if divided by 1000 would

yield about 0.09, which is noted in the participant's explanation. The major flaw is that the participant assumes exactly 500 simulated statistics fell to the right of the value 15 in the empirical sampling distribution. This is reminiscent of an introductory statistics student's approach to computing probability by subtracting known areas under a curve. I hypothesize that he/she chose to subtract those values of 16 up to 20 to imitate an introductory statistics student performing the computation: $P(X \geq 21) = 0.5 - P(16 \leq X \leq 20)$ when finding probability using a normal probability distribution. The reason the participant's mentions the value of 500 is because the participant thought the region to the right of the center (15) is equal to 0.5. By subtracting the probability for the outcomes of 16 to 20 from 0.5 would leave a set of tail outcomes that fall to the right 20. Unfortunately, this approach gives an incorrect result because the empirical sampling distribution is not necessarily a symmetrical around the center of 15. It also complicates the problem since one can easily compute the probability by counting the frequency of outcomes from 21 and higher. In the second excerpt, we see the participant also use a denominator of 500 in their computational approach, which also assumes a symmetrical empirical sampling distribution. The approach taken by these participants is interesting because how they connect methods in introductory statistics (i.e. finding probability using statistical tables) to an empirical sampling distribution.

Category CLS (Computation with the Level of Significance) describes four participants whose description of their computation was based on the level of significance. Excerpts of these types include the excerpts in Table 25.

Table 25. *Examples of CLS.*

Participant Information	**Computation and Explanation**
GTA Masters Candidate in Mathematical Science	*Computation: p*-value > 0.05 *Explanation:* The proportion of the toss at a given significant level.
GTA Masters Candidate in Mathematical Science and Mathematics Education	*Computation: p*-value < 0.05 *Explanation:* In order for James' procedure to be statistically significant, he needs a small p-value.
CC with Masters Degree in Statistics	*Computation: p*-value < 0.025 *Explanation:* level of significance = .05, split into .025 in each tail.

In each of these responses we notice that each of the participants would mention significance level or note a standard level of significance value (e.g. 0.05, 0.01, etc.). Each of these participants' *p*-value primarily used inequalities to represent the value for their *p*-value. I hypothesize that these responses stem from an understanding of the procedure of comparing a *p*-value with the level of significance. The first two excerpts stated the *p*-value to be related to a level of significance of 0.05. The third excerpt gives a *p*-value < 0.025 that corresponds to a two-sided hypothesis testing where the statistical significance of 0.05 is divided by two. I hypothesize that these participants struggled finding the utility and purpose of the empirical sampling distribution so they relied on their understanding of the hypothesis testing step where you compare the level of significance with a *p*-value. Their inability to compute a correct *p*-value using the empirical sampling distributions shows a gap in their understanding of the relationship between *p*-values and simulations.

Category CTP (Computation using Theoretical Probability) describes those participants who used theoretical probability to compute the *p*-value. Excerpts of these are shown in Table 26.

Table 26. *Examples of CTP.*

Participant Information	**Computation and Explanation**
GTA Doctoral Candidate in Statistics and Other	*Computation: p*-value = 0.04 *Explanation:* Under the null hypothesis, the number of helper puppets chosen is distributed binomially with parameters n=30 and p=1/2. Binomial with n=30 can be approximated as normal with mean np=15, and variance np(1-p)=30/4=7.5. He can convert the baby data to be approximately standard normal by subtracting 15 from 21, and then dividing by sqrt(7.5) [a number between 2 and 3], yielding a number slightly larger than 2. So, with 1.96 being the 95% extreme probability number, our p value is slightly smaller than 0.05.
GTA/GRA Masters Candidate in Statistics	*Computation: p*-value = 1/(2^30)*[30nCr16 + 30nCr17 + ... + 30nCr30] *Explanation:* We could interpret the p-value as being the probability that James' obtained 16 or more heads. The calculation is simply a binomial coefficient.

In these participants' explanations, we see participants use their understanding of a binomial distribution to compute the *p*-value. In the first excerpt, we see the participant use his/her understanding of a normal approximation to a binomial to compute a *p*-value = 0.04. This participant gave a *p*-value of 0.04 as the previous answer. The second participant also notices the problem is a binomial probability problem, but uses a formula to compute the probability. While the second participant does not provide an exact *p*-value, he/she does give a formula written as: 1/(2^30)*[30nCr16 + 30nCr17 + ... + 30nCr30]. Both approaches provide an appropriate theoretical probability, but ignore the empirical data. We do not know if these participants simply prefer theoretical

calculations to be exact in this type of scenario and/or if they do not understand how to work with the empirical sampling distributions.

Category CHTM (Computation using Hypothesis Testing Method) describes three participants who used a formal hypothesis test method to compute the probability of the *p*-value. This is seen in the following sample excerpts in Table 27.

Table 27. *Examples of CHTM.*

Participant Information	**Computation and Explanation**
FYI with Masters Degree in Other	*Computation: p*-value – 0.03 *Explanation:* ran a X2 test with df=1.
CC/FYI with Masters in Mathematics Education	*Computation: p*-value = 0.01426 *Explanation.* With a null hypothesis that the population proportion is .5 and an alternative that the proportion is greater than .5, we can find a z-score by taking the difference between the sample proportion of .7 and null proportion of .5 and dividing by the standard error of sqrt(.5*.5/30) and with the z-score got the associated area to the right to represent the p-value.

The first excerpt we see one participant who ignores the empirical sampling distributions and chooses to find the *p*-value using a chi-squared test to generate a *p*-value = 0.03. The second participant also ignores the empirical sampling distribution and performs a proportion test to generate a *p*-value = 0.01426. Both these excerpts represent sufficient approaches if one were to conduct a formal hypothesis testing for the original statistical problem James is trying to solve, but it disregards the purpose of using the empirical sampling distribution generated by James. While participants in this category do use theoretical probability models like those in Category CTD, the distinction is the explicit usage of the participants to perform a formal hypothesis test.

The responses for the computation of the *p*-value showed a wide variety of computations for the *p*-value. After going through the coding process of Test-Retest, the qualitative results for this showed an inter-reliability rate of approximately 91% based on the discussed codes. Table 28 shows the final count for the number of participants that fell within each category along with their demographic information.

Table 28. *Summary of Categories for the Computation of the P-value.*

	GTA	**GRA**	**GTA & GRA**	**CCI**	**FYI**	**CCI & FYI**	**Total**
Category CRF – Computation using Relative Frequency	8	1	0	4	2	0	15
Category CRFNH – Computation using Relative Frequency assuming Null Hypothesis	7	4	2	7	4	0	24
Category CIOS – Computation Ignores Observed Sample	1	0	0	0	1	0	2
Category CUR – Computation using regions	1	0	0	1	1	0	3
Category CLS – Computation using level of significance	3	0	0	2	0	0	5
Category CTP – Computation using Theoretical Probability	1	0	1	0	0	0	2
Category CHTM – Computation using hypothesis testing methods	1	0	0	0	1	1	3
No Response	0	0	0	1	0	0	1
Total	22	5	3	15	9	1	55

While a large percentage of the graduate student participants were mainly in Category CRF and CRFANH (a correct computation of the *p*-value), there were still a small number who fell in the categories for incorrect computations. What is interesting from the results is those participants who used theoretical probability models to compute the *p*-value were mainly instructors even when told to use the simulated data to compute a *p*-

value. Furthermore, the only participant who did not provide a response was from a community college instructor whose background was mathematical science. At first glance, the survey results indicate a potential gap regarding teachers' understanding of a simulation approach to hypothesis testing, specifically in their ability to compute *p*-values using empirical sampling distributions.

6.2.2. Interview Data: Computing p-values with sampling distributions.

After reviewing the data from the *Helper-Hinderer Task,* one of the follow-up interview tasks focused on having participants elaborating on their computation of the *p*-value and to discuss their opinion of using an empirical sampling distribution to perform the computation. Prior to the interview, the seven interviewees gave different responses to the computation of the *p*-value. A summary of their responses and categorization for their responses is shown in Table 29. Only two of the seven (Tod and Phil) gave a correct computation of the *p*-value in their initial survey response. Tod fell into category CRF and Phil fell into category CRFANH. The other responses that deviated from the value of 0.024. James initially did not provide a computation claiming he could not read the graph, but he did provide an explanation that showed evidence he could do a correct computation. Angie gave an explanation that showed evidence she could do a correct computation of the *p*-value but gave a computation based on theoretical probability (i.e. CTP). Carol gave an explanation that hovered around an idea that she was comparing a level of significance categorizing her as CLP. Jane explained using a method similar to a traditional one-proportion test, which categorized her as a CHTM.

Table 29. *Categories of Interviewee Responses for the Computation of the P-value.*

Participant	**Computation and Explanation**
Tod	*Computation of P-Value:* p=0.024 *Explanation:* It is the area passed the expected results. So 14+8+2/1000 (CRF)
Carol	*Computation of P-Value:* p<.05 *Explanation:* In order for James' procedure to be statistically significant, he needs a small p-value. (CLP)
Sean	*Computation of P-Value:* Find the z (because it's a proportion) *Explanation:* Find the area from |z| to infinity (CHTM)
Angie	*Computation of P-Value:* P(n>=21 | p=.5) from a binomial distribution. *Explanation:* 8 years of statistics education; I didn't find it specifically, but we could either use a normal approximation or calculate it explicitly depending on the level of the course involved. (Or use the 2.4% the simulation provided) (CTP)
Jane	*Computation of P-Value:* 0.01426 *Explanation:* With a null hypothesis that the population proportion is .5 and an alternative that the proportion is greater than .5, we can find a z-score by taking the difference between the sample proportion of .7 and null proportion of .5 and dividing by the standard error of sqrt(.5*.5/30) and with the z-score got the associated area to the right to represent the p-value. (CHTM)
Phil	*Computation of P-Value:* 0.024 *Explanation:* If I'm to assume that there were no samples in which more than 23 of the 30 babies chose the helper doll, then a good estimate would be (14 + 8 + 2)/ 1000, or about 2.4%. (CRF)
James	*Computation of P-Value:* The graph's difficult to read, but considering black areas below the 21, 22, and 23 circles, I'd guess less than 5%. *Explanation:* To calculate the P-value, I'd need the counts from James's simulation. Otherwise, I'd estimate as above. / The P-value means that when James assumed there's no preference and simulated 30 babies making the choice 1000 times, he observed the sample proportion (21/30 babies prefer helper) in less than 5% of the simulations. (CLS)

In this section, I highlight the reasoning of the participants that emerged from the interviews. The initial categorization of the survey might suggest that many of the interviewees did not know how to use the empirical sampling to find the *p*-values. Results of the interviews uncovered that six of the seven interviewees could compute the *p*-value using the empirical sampling distribution with Carol being the only exception. What was interesting was those interviewees whose survey data focused on theoretical ideas had a very strong opinion for the use of theoretical approaches to computing the *p*-value. The first opinion was a preference to the theoretical model over the simulation model when computing a *p*-value. The second opinion was based on an understanding that either a theoretical or empirical approach to hypothesis testing would be appropriate.

Considering that statistics education community is pushing towards the use of modeling and simulation approaches to deepen understanding of hypothesis testing, it was vital to investigate a participant's ability to compute the *p*-value using an empirical sampling distribution and whether they saw value in using simulations. In this section, I highlight three primary themes that emerged as the interviewees discussed their opinions of computing a *p*-value using an empirical sampling distribution. The first theme highlights what happens when a participant inability to comprehend the simulation model hampered their ability to compute an appropriate *p*-value using the empirical sampling distribution. The second theme highlights that even though some participants understood how to compute a *p*-value using an empirical sampling distributions, they believed that theoretical methods were more appropriate. The third theme highlights a set of

participants who knew how to use the empirical sampling distributions to compute a *p*-value and felt that either approach (theoretical or empirical) is equally appropriate.

6.2.2.1. Struggling with empirical sampling distributions. As mentioned, several of the survey participants could compute a *p*-value correctly using the empirical sampling distribution presented in the *Helper-Hinderer Task.* One particular excerption was Carol who gave a computation categorized as CLP. In her initial survey response, she wrote that the p-value < 0.05 because, "James needs a small *p*-value in order to show statistical significance." Analysis of the Carol's data showed evidence that she knew procedures of hypothesis testing, but struggled discussing relationships that tied ideas of sampling distributions to *p*-values. This became apparent when she discussed her understanding of the simulation model of hypothesis testing presented in the *Helper-Hinderer Task.* Below is an excerpt from the interview showing an exchange where Carol is asked whether she has seen a simulation approach to hypothesis testing prior to the study.

Carol: I have never seen anybody do this. I mean I can see how they…I never seen anybody do this and thought that was kind of clever but I would take that data with a grain of salt.

Interviewer: Why would you say that?

Carol: Because it's not...it's not what he is actually trying to test. Like he is just using this experiment to say something about another experiment. And in my head I'm just like why don't you just conduct the original experiment to begin with.

Interviewer: Okay. What do you think...when you say conduct the original experiment, what do you mean by that?

Carol: Like here they want to see like they see if there is a good baby or bad baby or something, a good puppet or bad puppet. So, these have to do with like humans and there are so many variables going on with like a baby. Like where they grow up in. Like I don't know, a strict Asian family right? Or like a super traditional family where they going to think a lot of things are bad. But if you were to like take baby who is straight out of Compton and that baby was exposed to many things that are bad, like their environmental factors you have to put into play where this is just super mechanic. Like this simulation is robotic.

Based on the response given by Carol, I hypothesize that the relationship between simulations and empirical sampling distributions was not a part of her overall understanding of hypothesis testing. In addition, the context of the problem itself seems gives her pause in the experiment's legitimacy and her trust in the simulation. This is expressed in her initial response about James' simulation being 'clever but take the data with a grain of salt' shows a reluctance to the modeling of the problem. Her statement where she mentions that the method done by James "is not what he is trying to test" shows that she does not recognize James' method as a valid way of conducting hypothesis test. This follow-up response shows a further lack of experience with empirical sampling distributions and simulation. She instead focuses on describing potential biases with the modeling process. Even when asked about the computation of

the *p*-value, she insisted on discussing a procedural approach she learned from prior statistical work as show in the excerpt below.

> **Interviewer:** Alright. Do you think you can actually use this (James' work) information to get the p-value? Like his distribution he has right there?
>
> **Carol:** Just by looking at this I think I could.
>
> **Interviewer:** How would you?
>
> **Carol:** I would just have to find $\bar{x}$ the standard deviation.
>
> **Interviewer:** Alright.
>
> **Carol:** And from there I could calculate the test statistic and the p-value.
>
> **Interviewer:** Is that the only way you think it is possible to do that? Not saying isn't another way, I'm just wondering in your opinion is that the only way to get it.
>
> **Carol:** In my experience that is the only way to get it unless you throw it into your computing system and let that do that.

The type of understanding displayed by Carol shows that she places a lot of value in the procedural method of hypothesis testing. Even when asked about the empirical distribution, Carol mentions the sample mean and standard deviation which is not necessary when using a simulation approach to hypothesis testing. This shows a lack of understanding would makes it difficult to determine whether a person's ability to see the pedagogical value of a simulation approach. This was not the case for the two other interviewees whose knowledge of the simulation approach was adequate, but whose opinion still showed a preference to a theoretical approach.

6.2.2.2. Preferring theoretical over empirical probability. One such interviewee was Jane whose original computation from the survey was through a z-score

approximation (i.e. CHTM). In her survey work, she wrote down that the *p*-value for the *Helper-Hinderer Task* is 0.01426 which she computed using a traditional one-proportion hypothesis test. This would be considered a correct procedural way of doing a hypothesis test for a single proportion that is traditionally taught in introductory statistics, but ignores the empirical sampling distribution produced by the student.

During the interview, Jane displayed an understanding of sampling distributions that was evident when she discussed ideas of repeated sampling and her description of the empirical sampling distribution in the *Helper-Hinderer Task.* Further evidence was in her ability to discuss a correct computation of the *p*-value in the *Helper-Hinderer Task* when she correctly articulated the computation of the *p*-value of 0.024 using the empirical sampling distribution. What is interesting was her pedagogical view of the empirical sampling distribution versus using her procedural methods. When asked about the validity of using a simulation and computing a *p*-value using the empirical sampling, her response seemed to prioritize the traditional approach as shown in the excerpt below.

> **Interviewer:** So if this was the question you asked your students and some student came up to you and said I learned hypothesis testing by just doing a simulation and the student did it this way. Let's say a student came to class and they are expected to know hypothesis testing and one student did it this way (formula) and one student did it this way (simulation), would you find it valid or both valid?
>
> **Jane:** I think in the end I would consider them both valid but I would almost enforce the preference for this route (z-score formula) because of accuracy.

Interviewer: Okay. Alright.

Jane: Just because if they keep doing it (simulation), it could lead to errors with the hypothesis test.

Interviewer: So you prefer this one (formula)?

Jane: Yeah.

Interviewer: Is it because the error is off by a 1%?

Jane: Yeah I guess the error is throwing me.

In the above except we see Jane's aversion to the simulation model. In her description of the simulation model, she discusses potential underlying 'error' in the computation because the simulation model is approximating the *p*-value. Jane reasoning is based on her comparison with her theoretical approach versus her empirical approach that resulted in a difference of 1%. What makes Jane's argument interesting is that the approach she believes to be more accurate is also an approximation method.[13] It was not clear whether she realized that her method of doing hypothesis testing using proportions relied on an approximation of a normal distribution, but it seems that traditional methods prioritize her choice in how she would evaluate student work.

Another interviewee who shared a similar opinion as Jane was Angie. Angie also preferred a theoretical approach to hypothesis testing over a simulation model. In her original survey response, she wrote a conditional statement and noted using a binomial distribution. Angie then gave a follow-up explanation where she mentioned that the *p*-value is 2.4% if you use the empirical sampling distribution. She showed an understanding of sampling distributions in the overall scheme of hypothesis testing. Just

[13] The method used by Jane is based on an approximation method for a binomial distribution.

like Jane, Angie mentions in her interview that she would prefer a theoretical approach over an empirical approach to hypothesis as illustrated in the excerpt below.

> **Angie:** Yeah I said there is two approaches here. One you can actually calculate specifically from a binomial distribution and say you know it is going to be exactly this probability of getting exactly n of 21 or more. Or you could do a normal approximation which is what I would teach my students to do and 21 is a little small for a normal approximation but 21 worked pretty well for a normal approximation anyway so I think it would be fine. Umm...and lastly or you can use the 2.4 simulation that he found. That would be one other approach. So there is a couple different ways. I would probably, personally I would teach the normal approximation because we don't get into the binomial distribution and I would probably prefer something theoretical over simulation as well.
>
> **Interviewer:** Why would you say that?
>
> **Angie:** I mean simulation works if that is all you can use, but if you have theory that's probably more sound than simulation unless you have some reason to believe your assumptions are wrong but if you're simulating under the same assumptions your theory is just as sound.

Angie's explanations once again highlight a preference for theoretical model to compute a *p*-value. In her explanation, Angie expresses the importance of getting a "specific" more exact way of computing the *p*-value through a binomial distribution model. Unlike Jane who discusses a normal approximation model and computes it using a z-score,

Angie's background as a doctoral student in statistics seems to play into her understanding on alternative approaches to finding a *p*-value. She expresses being more comfortable with a theoretical model and would only use simulations if necessary. The thinking displayed by Angie is consistent with statisticians who rely on simulations models when the theoretical approaches cannot be appropriately determined. This seems to play into her pedagogical choice where she notes preferring to teach using a normal approximation rather than a simulation approach.

6.2.2.3. Favoring simulations. In contrast to the interviewees above, three interviewees (Sean, James, and Phil) felt strongly about their opinion on the simulation model. One interviewee was Sean who was a doctoral student in statistics. During the interview, Sean showed a strong understanding of empirical sampling distributions and simulations. Several times during the interview, Sean would mention his experience working with simulations and conducting bootstrapping models. When Sean was questioned regarding the validity of the simulation model, he seemed comfortable in its applicability to answer a statistical question related to hypothesis testing. This is illustrated in the following exchange when Sean was asked about the validity of the simulation model.

> **Sean:** When I say mimic is resampling...so there was three things we talked about. Resampling from the population. Pull a sample of 50 from the population a 1000 times. That's not really bootstrapping. The other thing I've done is...the first one was in context to the null. Bootstrapping is taking your sample, sampling from it with replacement equally size samples and get

statistics from each sample. This seems to be...I forget your question...this seems to be pulling from...the idea of almost creating a distribution in kind of a very low tech way. A reference distribution under the null in a very low tech way.

Interviewer: Okay, but are you still kind of okay with the fact how he did it?

Sean: Yeah, he is just basically assuming that his coin...well there is an additional assumption that his coin is fair. But as long as his coin is fair, it is a perfectly legitimate way of doing it.

In the excerpt, we see Sean drawn upon his experience with simulations helped him articulate how the process done by the student is like the procedure taught in traditional statistics courses. He makes a mention of common ideas from introductory statistics courses such as creating a probability distribution. When probed further on his understanding of the model, Sean continues to describe notions of repeated sampling to confirm the validity of the model. When Sean was later asked about how he would compute the *p*-value using the empirical sampling distribution, he quickly just used the empirical sampling distribution as a way of doing it as show in the excerpt below.

Sean: Well 21 out of 30 babies so and then you say okay the number of babies that chose the helper doll out of a sample size of 30 using a coin. So, he got 14 and he got 8 and he got 2. 21, 22, 23 and on. This is basically saying 21 to infinity but we are limited because this is a very discrete way of doing it plus it is a discrete sample.

Sean having experience and a strong understanding of simulation approaches proved to be useful when assessing James' model to not only find the correct *p*-value, but also discusses important ways the empirical sampling distribution is related to theoretical ideas. For instance, in the excerpt above we see Sean relate to ideas of discrete sampling and how *p*-values measure tail regions that extend to infinity.

Phil also shared a similar opinion to regarding the validity of the simulation model. Just like Sean, Phil computed a correct *p*-value using the empirical sampling distribution. Phil also mentions several times how he has done work with simulation models in the classroom. While Phil shares a similar opinion as Sean regarding the models validity, his explanation regarding his opinion on the model touches on a pedagogical reasoning why such models are useful for classroom the classroom.

Interviewer: What do you think about number 4 (*Helper-Hinderer Task)*? My first question, but I think you've answered already it. You've probably done simulation models for hypothesis tests.

Phil: I personally think this should be way we should teach it on day one. This is my personal opinion and I think Allan (Rossman) agrees with me, but there is no reason to force parametric test if you can simulate using non-parametric methods like this. These are...kids get these. Students get none...they understand flipping coins and rolling dice. They get these. Dice are used in MTH105. I am totally down where you can model it either...I like using manipulatives too. You obviously said, James gets a coin and flips the coin 30 times.

Interviewer: Yeah.

Phil: So, you set it up as flipping a coin, but yeah computer simulations are huge so. It's funny you're question number 1, I had to look at it two or three times. It can't be that easy. I'm like so many of the questions you've asked at this point were really like, they forced me to look really carefully at what the student have written and I expected them all to be at that level well duh, equally likely from the coin and then I realized that is not obvious from anybody. You know they flip coins all the time, but they forget the context of what a coin flip is. It is an unbiased equally likely outcome. There you go, there is your part b.

Unlike the previous interviewees who favored a theoretical approach, Phil was very focused on the pedagogical value of the empirical sampling distribution. Throughout the interview, Phil showed strong evidence of understanding both traditional and simulation-based approaches to hypothesis testing. His explanation of the simulation model and its validity relates to how a teacher would discuss *p*-values to an introductory statistics class that might have very little understanding of theoretical probability. In his explanation, he suggests a preference to these "non-parametric" methods over the "forced parametric" because it is something his students would be able to understand. The connection between his understanding of both approaches and his experience as a teacher was shown particularly in this excerpt when he suggests a simulation approaches things students would likely understand. This shows an example of a teacher whose mixture of different forms of statistical knowledge come together to guide his pedagogical practices.

6.2.3. Summary: Computation of the p-value using sampling distributions.

This section outlines some important ideas related to the relationship between *p*-values,

simulations and sampling distributions. The survey results show that there is still inconsistency with the way teachers view the role of simulations to conduct hypothesis testing. With only 39 out of 55 correctly computing a *p*-value even when provided an explicit sampling distribution shows that there are still gaps in the knowledge of our current statistics teachers. Furthermore, those who even relied on theoretical models of computing the *p*-value would struggle computing the *p*-value incorrectly. This lack of knowledge on the relationship between simulations and hypothesis testing also caused further differences of opinions when discussing the validity of simulation approaches.

In the interview data, we see that participants who have never learned a simulation approach to hypothesis testing struggle computing *p*-values using empirical sampling distributions. Because of this lack of experience working with simulation models, we see an effect on teachers' opinions on its usefulness in the classroom and how it might support student development. Of all the interviewee participants, three questioned the simulation model and still preferred to stick to theoretical approaches. The two participants who did see the value of a simulation-approach was based on actual experience working with simulations both in statistical practice and in teaching. One teacher had a very strong belief that it could serve a greater purpose as a pedagogical tool in developing students understanding of *p*-values and hypothesis testing in general. Unfortunately, because of the minimal data gathered from the interviews on teachers' opinions of a simulation approach to hypothesis testing, it is unknown how much of the general population of statistics teachers share a similar view of simulations.

6.3. Teachers' Understanding of Conditional Probably and the *p*-value.

An important component of a robust understanding of a *p*-value includes knowing it is a conditional probability structured around a relationship of the observed results and null hypothesis. Recall from the literature review that prior research indicated students overlook and/or misinterpret the conditional relationship of the *p*-value. In addition, prior research shows that often teachers have similar misconceptions as their students. Because understanding that the *p*-value is a conditional probability is important, data was analyzed to see what themes would emerge from the data when participants discussed their understanding of verbal and symbolic representations of the *p*-value.

There were two primary instances where participants were pressed to discuss the relationship between a p-value and conditional probability. The first instance was the *Drug Task,* which assessed participants' notions of verbal interpretations of the *p*-value. The second instance was the *Car Task,* which assessed participant's notions of symbolic representations of the *p*-value.

To illustrate the participants' understanding of the relationship between a *p*-value and conditional reasoning, this section will examine data from the survey and the interview that highlight concepts related to a verbal and symbolic representation of the *p*-value. I then provide dialogue from the interviews that highlight how teachers view the relationship between a verbal and symbolic representation of the *p*-value.

6.3.1. Survey results: verbal interpretations of the p-value. The definition of the *p*-value is defined as 'the probability of getting a result as extreme or more extreme as the observed result assuming the null hypothesis.' This is a standard way for students and

teachers to verbally interpret a *p*-value in many introductory courses (and sometimes in statistical practice). The *Drug Task* was designed to incorporate the interpretation into a given context (Figure 21).

A research article reports on a new drug test that is to be used to decrease vision loss. The article reports its results and a p-value of 0.04 in the analysis section. Below are six different student interpretations of the p-value.

Student A: *The probability of getting a result as large as or larger than the one in this study if the drug is actually not effective is 0.04.*

Is this student's interpretation of a p-value valid? ___ Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student B: *The probability that the drug is not effective is 0.04.*

Is this student's interpretation of a p-value valid? ___Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student C: *There is a 4% chance that the researcher made an error.*

Is this student's interpretation of a p-value valid? ___ Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student D: *If you were to repeat the study, there is a 4% chance of getting exactly the same result as the one in this study if the drug is actually not effective.*

Is this student's interpretation of a p-value valid? ___Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student E: *The probability that the drug is effective is 0.04.*

Is this student's interpretation of a p-value valid? ___ Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student F: *There is a 96% chance the drug is effective.*

Is this student's interpretation of a p-value valid? ___ Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Figure 21. Drug Test Task.

The task uses the context of a drug experiment on vision loss as a context for interpreting the *p*-value. Through this context, the participant is asked to identify whether six hypothetical student interpretations of the *p*-value are valid/invalid. A standard approach of interpreting the *p*-value for this context is, "the probability of getting a result as extreme or more extreme as the observed results assuming the drug is not effective." Student A provides the closest example to the kind of verbal interpretation that aligns with the *p*-value's interpretation. The difference between the standard interpretation mentioned and the one presented by Student A is the usage of the words 'large or larger' in place of 'extreme or more extreme.' The other five interpretations were all designed to be invalid interpretations. Student D does mention the importance of conditioning on the null hypothesis, but the interpretation fails to describe the important property that *p*-values measure the probability of getting results "as extreme or more extreme as the observed sample result." Student Interpretation B, C, E, and F are all written in a way that (among other properties of the *p-value*) do not highlight the important idea that the *p*-value is a probability computed assuming a null hypothesis. Because the focus of this section is to assess a participant's ability to recognize the importance of conditioning on the null hypothesis (i.e. the drug is not effective), it was important to assess whether they

thought Student A's interpretation was valid because it is an interpretation most teachers consider to be valid. Second, it was important to note participant responses to Student D's interpretation because it was correctly conditioning on null assumptions, but is syntactical incorrect because of the phrase "the same result" in place of "as extreme or more extreme as the observed result." Finally, it was necessary to analyze whether the lack of conditional reasoning in the other student interpretations (B, C, E, and F) would raise an issue for the participants.

Since the primary focus of this section is to analyze whether participants noted the importance of conditional reasoning in a *p*-value's verbal interpretation, it was therefore important that analysis on responses for Student A and D were handled differently than Students B, C, E, and F. Since Student Interpretations A and D were both written as verbal interpretations that correctly conditioned on the null hypothesis, it was important to determine if a participant gave an appropriate explanation if they found the interpretation to be invalid. Examples of Student A's interpretation that could be considered appropriate can be seen in Table 30. These explanations were categorized as accurate feedback (AF) if the participant criticized the syntactic use of the word "large or larger" as misleading. Some participants preferred the use of the word "low or lower" or "extreme or more extreme" because they felt better aligned with the problem's context. Participants such as these noted that because the context refers to a decrease in vision loss, using the word "lower" would be considered acceptable. Using the word "extreme" would also be considered accurate because it is similar to the standard definition of a *p*-value. This can seen in the explanations provided by the CCI in Table 30 who notes that

the word "Larger" to be misleading. Inaccurate feedback (IAF) include those explanations that were viewed as either not helpful or incorrect ways to assess the statement because it does not provide further connections to ideas of a *p*-value. An example of IAF can be seen in the explanation provided by the FYI in Table 30 who gives very little information regarding the importance of conditional reasoning of a *p*-value.

Table 30. *Accurate and Inaccurate Feedback for Student Interpretation A.*

Category of Feedback	**Participant Explanation**
Accurate Feedback	CCI (Masters in Math/Stat/Stat Ed): "Larger" is misleading - I think they would have meant "more extreme". Although, with a simple tweak, this one's correct: "The probability of getting these results (or even more extremely lowered results) than the ones in this study, if the drug is actually not effective, is 0.04."
Inaccurate Feedback	FYI (Ph.D. in Statistics): This seems to be the interpretation of a one tailed test.

Examples of explanations that categorized as AF and IAF for Student D is shown in Table 31. As mentioned, the main error with Student D's interpretation was the use of the phrase "exactly the same result" which does not account for the fact that *p*-values are measures of tail regions. Examples of accurate feedback for those interpretations can be seen in one of the GTA responses in Table 31 who notes the incorrect phrase, "exactly the same result" as the error and notes that *p*-values are a measure of area. Examples of IAF encompass explanations that did not notice this key issue with the interpretation. This can also be seen in the sample explanation provided by the other GTA in Table 31 who makes no mention of the error described.

Table 31. *Accurate and Inaccurate Feedback for Student Interpretation D.*

Category of Feedback	**Participant Explanation**
Accurate Feedback	GTA (M.S. in Other): p-value does not focus on "exactly the same result", it marks the area under the curve beyond that particular z-value.
Inaccurate Feedback	GTA (Doctoral Candidate in Mathematics): I've never heard this interpretation, but it's not correct since it disagrees with the valid interpretation

Table 32 summarizes the results for the responses for Student Interpretation A and D. Based on the data collected, a total of 43 out of 55 the participants identified Interpretation A as a valid verbal interpretation of the *p*-value. For the 12 participants who considered Interpretation A as invalid, four participants gave appropriate feedback that focused on the syntactical error of the interpretation noting a preference for "small or smaller" or "extreme or more extreme" but still acknowledged the interpretation as "mostly valid". Eight of the participants did not note the syntactic error, but gave erroneous responses or provided no explanations. For Interpretation D, six of the participants noted the interpretation as valid, while 49 noted the interpretation as invalid. Of the 49 participants, 32 participants gave accurate feedback noting the error of using the word "exactly the same" while 17 gave inaccurate feedback or provided no explanation.

Table 32. *Results for Student Verbal Interpretations A and D.*

	GTA	**GRA**	**GTA & GRA**	**CCI**	**FYI**	**CCI & FYI**	**Total**
STUDENT VERBAL INTERPRETATION A RESULTS							
Valid	19	5	3	8	7	1	43
Invalid w/ Accurate Feedback (AF)	1	0	0	3	0	0	4

Invalid w/ Inaccurate Feedback (IAF)	2	0	0	4	2	0	8
Total	22	5	3	15	9	1	55
STUDENT VERBAL INTERPRETATION D RESULTS							
Valid	2	0	0	2	2	0	6
Invalid w/ Accurate Feedback (AF)	11	5	2	10	3	1	32
Invalid w/ Inaccurate Feedback (IAF)	9	0	1	3	4	0	17
Total	22	5	3	15	9	1	55

As mentioned, the coding for the explanations provided for Student Interpretations B, C, E, and F were handled differently than those of Student A and D. A primary misconception with these four verbal interpretations (B, C, E, and F) was a lack of conditional reasoning in the student's work. While these four interpretations were designed to highlight other issues regarding the *p*-value, a primary issue that should have been noted by the participants was the lack of the conditional reasoning. Thus, explanations of the participants were categorized based on whether the participant noted the lack of conditional reasoning. Table 33 shows examples of sample excerpts for Student B.

Table 33. *Conditional and Non-Conditional Feedback for Student Interpretation B.*

Categories	**Participant Explanation**
Conditional Feedback	GTA (Doctoral Candidate in Statistics): p-values measure the probability of finding a result such as you got given the null is true
Non-Conditional Feedback	GTA (Masters Candidate in Math/Math Ed): The p-value is the measure of evidence against the null, not that the null is true.

Conditional Feedback (CF) was used to categorize those participants whose explanation noted aspects of conditional reasoning. This can be seen in the first

participant's explanation in Table 33 where they note that the *p*-value is a measure of a result given a null being true. This example shows evidence that the participant is criticizing a lack of conditional reasoning in the student interpretation.

Non-Conditional Feedback (NCF) was used to categorized those participants who explanations made no mention of aspects conditional reasoning. Instead, the participant focuses on other parts of the interpretation they valued as incorrect. The can be seen in the second participant's explanation in Table 33 who focuses primarily on notions of evidence against the null. It is completely unknown however whether the participants who provided similar explanations understands that *p*-values are conditional probabilities, but it does show that the primary feedback they chose to provide was not on conditional reasoning. For consistency, a similar categorization of conditional/non-conditional feedback was used for the Student Interpretations C, E, and F since all three verbal interpretations lacked conditional reasoning.

Table 34 summarizes the survey results of the *Drug Task* for interpretations B, C, D, and F and whether they gave notions of conditional reasoning. For Interpretation B, four participants marked it as valid. For those 51 participants who marked the interpretation as invalid, 23 participants noted the lack of assuming some null hypothesis in the interpretation and 28 participants focused on other concepts related to the *p*-value unrelated to the conditional reasoning. For Interpretation C, four participants marked it as a valid interpretation. For the 51 participants who marked it as invalid, only nine mentioned the lack of conditional reasoning or assuming a null hypothesis. For Interpretation E, all the participants marked the interpretation as invalid. Only 17 of the

participants noted the lack of a conditional reasoning while 38 noted other aspects of the interpretation related to the *p*-value. Finally, for Interpretation F, four participants marked it as valid and 51 marked it as invalid. Furthermore, 13 of the 51 who marked it as invalid noted the lack of conditional reasoning in the interpretation.

Table 34. *Results for Student Verbal Interpretations B, C, E and F.*

	GTA	**GRA**	**GTA & GRA**	**CCI**	**FYI**	**CCI & FYI**	**Total**
STUDENT VERBAL INTERPRETATION B RESULTS							
Valid	2	0	0	1	1	0	4
Invalid w/ Conditional Feedback (CF)	6	4	2	6	4	1	23
Invalid w/ Non-Conditional Feedback (NCF)	14	1	1	8	4	0	28
Total	22	5	3	15	9	1	55
STUDENT VERBAL INTERPRETATION C RESULTS							
Valid	1	0	0	1	2	0	4
Invalid w/ Conditional Feedback (CF)	5	1	1	1	1	0	9
Invalid w/ Non-Conditional Feedback (NCF)	16	4	2	13	6	1	42
Total	22	5	3	15	9	1	55
STUDENT VERBAL INTERPRETATION E RESULTS							
Valid	0	0	0	0	0	0	0
Invalid w/ Conditional Feedback (CF)	7	3	0	4	2	1	17
Invalid w/ Non-Conditional Feedback (NCF)	15	2	3	11	7	0	38
Total	22	5	3	15	9	1	55
STUDENT VERBAL INTERPRETATION F RESULTS							
Valid	2	0	0	2	0	0	4
Invalid w/ Conditional Feedback (CF)	4	1	0	6	1	1	13
Invalid w/	16	4	3	7	8	0	38

Non-Conditional Feedback (NCF)							
Total	22	5	3	15	9	1	55

An initial summary of the survey data from the *Drug Task* shows that several participants could correctly identify the correct verbal interpretation of the *p*-value. A large number of participants showed the ability to correctly identify valid/invalid interpretations of the *p*-value. The inconsistency came in the explanations provided by the participants regarding why interpretations were invalid. Looking at the results of verbal interpretations provided by Students B, C, E, and F, we see that the number of participants who explicitly focused on the lack of the conditional reasoning varied quite a bit. Instead of consistently focusing on the lack of conditional reasoning, several participants chose to focus on other aspects of the verbal interpretation they felt was problematic. There is possibility that some participants simply chose to give feedback for an invalid interpretation of a *p*-value based what they felt was a more pressing issue in the interpretation. Evidence for this can be seen when analyzing feedback for interpretations (B, C, E, and F) as a whole. Thirty participants noted a lack of conditional reasoning in at least one of the four interpretations. Twenty-five participants made no mention of conditional reasoning when they were asked to provide why a verbal interpretation was invalid. This shows some evidence that some of the participants did not understand that *p*-value was a conditional probability and/or found that giving feedback on the conditional reasoning of the *p*-value was not a primary criticism for verbal interpretations given by students.

6.3.2. Survey results: symbolic representations of the p-value. A symbolic representation of the *p*-value is not often discussed as a major component in introductory

statistics courses even though it is an important concept in theoretical statistics. Since the *p*-value is defined as a conditional probability, knowing how to symbolically represent it using a conditional probability statement should be part of a teacher's specialized knowledge. This thinking formed the basis for the *Car Task* (Figure 22) that assesses whether participants are able to recognize correct symbolic representations of the *p*-value.

Research was conducted to determine whether people preferred hybrid over traditional gasoline powered vehicles. The article reports that their random sample showed that 60% of people preferred hybrids. The researchers gave a test statistic of z = 1.414 and a p-value of 0.08 using a right-tailed hypothesis test for proportions. Below are six different symbolic representations of the p-value given by introductory students when asked about this problem.

Student A: $\boldsymbol{P(z \geq 1.414) = 0.08}$

Is this student's symbolic representation of a p-value valid? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student B: $\boldsymbol{P(\hat{p} = 0.6) = 0.08}$

Is this student's symbolic representation of a p-value valid? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student C: $\boldsymbol{P(Reject\ the\ null\ hypothesis \mid \hat{p} = 0.6) = 0.08}$

Is this student's symbolic representation of a p-value valid? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student D: $\boldsymbol{P(p = 0.5 \mid \hat{p} = 0.6) = 0.08}$

Is this student's symbolic representation of a p-value correct? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student E: $\boldsymbol{P(\hat{p} \geq 0.6 \mid p = 0.5) = 0.08}$

Is this student's symbolic representation of a p-value valid? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student F: $\boldsymbol{P(\hat{p} \geq 0.6) = 0.08}$

Is this student's symbolic representation of a p-value valid? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Figure 22. Car Task.

The *Car Task* uses a context of a study on hybrid and gasoline-powered vehicles to determine whether participants can identify valid symbolic representations of the *p*-value. Like the *Drug Task*, the participant is presented with six hypothetical student symbolic representations of the *p*-value. Two representations (A and E) were designed to be correct. The first representation was by Student A who uses a z-score representation of the *p*-value that relies on an *implicit conditional* that is the result of the computational transformation to a *z*-score. The second was Student E who uses an *explicit conditional* probability statement to write out the *p*-value. The other representations were written to illustrate incorrect ways of writing a *p*-value symbolically. Focusing strictly on concepts related to the conditional nature of a *p*-value, Student B and F did not use either a z-score transformation or a conditional probability statement. Student C and D uses a conditional notation but incorrectly conditions on the observed sample instead of the null assumption.

Since the focus of this section is on the importance of the conditional reasoning, I first present the results of the two valid representations of the *p*-value (A and E) from the *Car Task*. I then focus on the results of Student B and F, whose representation was designed to assess whether participants found the lack of a conditional statement problematic. Finally, I present the results of Student C and D, which contained (among other things) an incorrect conditional statement by assuming the observed sample instead of the null assumption.

Because this task also analyzes connections between a *p*-value and conditional probability, a similar type of analysis used in the *Drug Task* was reapplied for this task. A big challenge that occurred during analysis was that a number of the participants took issue to the use of specific notation used by students. In particular, the use of symbols (i.e. $z, \hat{p}, \text{and } p$) in the different symbolic representations in the hypothetical student work became a main focal point of their criticism. This made it challenging to categorize themes related to conditional probability in several of the explanation. As a result, a coding strategy similar to the one used for the *Drug Task* was applied in an attempt to code areas where conditional reasoning popped up in the participants work.[14]

Student B and F whose symbolic representation did not use a conditional probability or imply some form of conditional reasoning was analyzed on the participant's ability to recognize the lack of conditioning reasoning in the symbolic form. For those participants who found representations B and F as invalid, a similar approach of noting whether the participant provided explanations that align with conditional feedback (CF) and non-conditional feedback (NCF) was applied for these representations as those seen in the *Drug Task*. Since Student Interpretation A and E were designed to represent correct symbolic representations of the *p*-value, explanations from participants who marked these representations invalid were categorized using a similar approach of appropriate feedback (AF) and inappropriate feedback (IAF) to see if their explanations focused on syntax. Student C and D both used an incorrect conditional representation of the p-value. Since these representations explicit conditionals, a similar treatment was

[14] A more in-depth discussion of the issues surrounding the semantic issues will be addressed in the section on limitations of the research.

applied to the categorization of the explanations to see whether the participant provided appropriate/inappropriate feedback regarding the type of things they felt was incorrect about representation.

Results of the categorization for the various participants can be found in Table 35. The data in the table shows that most of the participants could correctly identify the valid representations of the *p*-value in the *Car Task*. For Student A's symbolic representation (z-score representation of the *p*-value), 44 participants identified it as a valid representation. For Student E (the conditional version of the *p*-value), 39 participants marked the representation as valid. Closer inspection of the survey data revealed 9 out of the original 44 participants who said Student A was valid marked Student E as invalid. On the other hand, 7 of the 39 participants who selected Student E (conditional version of the *p*-value) as valid marked Student A as invalid. These numbers show slight inconsistencies in the number of participants who validated implicit versus explicit conditional version of *p*-value. This shows some evidence that some teachers do not see a direct connection between a standard and non-standard form of writing a *p*-value symbolically.

For the symbolic representations that did not use a conditional probability, 10 participants noted that Student B and 23 participants noted Student F lacked a conditional representation. Furthermore, 44 students did not provide feedback that noted conditional reasoning for Student B and 23 did not provide feedback that noted conditional reasoning for Student F. Instead, the participants focused on other areas that would make the representation invalid (e.g. problems with equality, finding the wrong probability, etc.)

Table 35. *Results for the Car Task on the Symbolic Representations of the P-value.*

	GTA	GRA	GTA & GRA	CCI	FYI	CCI & FYI	Total
STUDENT SYMBOLIC REPRESENTATION A RESULTS							
Valid	17	5	1	13	7	1	44
Invalid w/ Appropriate Feedback (AF)	2	0	0	0	0	0	2
Invalid w/ Inappropriate Feedback (IAF)	3	0	2	2	2	0	9
Total	22	5	3	15	9	1	55
STUDENT SYMBOLIC REPRESENTATION B RESULTS							
Valid	1	0	0	0	0	0	1
Invalid w/ Conditional Feedback (CF)	2	2	1	2	3	0	10
Invalid w/ Non-Conditional Feedback (NCF)	19	3	2	13	6	0	44
Total	22	5	3	15	9	1	55
STUDENT SYMBOLIC REPRESENTATION C RESULTS							
Valid	2	0	0	0	0	0	2
Invalid w/ Conditional Feedback (CF)	2	0	1	5	3	0	11
Invalid w/ Non-Conditional Feedback (NCF)	18	5	2	10	6	1	42
Total	22	5	3	15	9	1	55
STUDENT SYMBOLIC REPRESENTATION D RESULTS							
Valid	2	0	0	0	0	0	2
Invalid w/ Appropriate Feedback (AF)	7	2	2	11	5	1	28
Invalid w/ Inappropriate Feedback (IAF)	13	3	1	4	4	0	25
Total	22	5	3	15	9	1	55
STUDENT SYMBOLIC REPRESENTATION E RESULTS							
Valid	14	3	2	12	7	1	39
Invalid w/ Appropriate Feedback (AF)	2	1	0	1	0	0	4
Invalid w/	6	1	1	2	2	0	12

Inappropriate Feedback (IAF)							
Total	22	5	3	15	9	1	55
STUDENT SYMBOLIC REPRESENTATION F RESULTS							
Valid	5	0	0	4	2	0	11
Invalid w/ Conditional Feedback (CF)	5	4	2	8	4	0	23
Invalid w/ Non-Conditional Feedback (NCF)	12	1	1	3	3	1	21
Total	22	5	3	15	9	1	55

6.3.3. Interview results of the p-value conditional nature. The results of the survey data showed that not all the participants could correctly identify verbal and symbolic interpretations of the *p*-value. In order to investigate more deeply about teachers reasoning around a conditional probability, seven interviewees were chosen particularly because they showed a variety of responses to the *Drug* and *Car Task*. A summary of the survey responses for the seven interviewees can be found in Table 36 and 37.

Table 36. *Results of the Drug Task for the Interviewees.*

Participant Responses to Drug Task	**Response**
Tod	
▪ Student Verbal Interpretation A	VALID
▪ Student Verbal Interpretation B	VALID
▪ Student Verbal Interpretation C	VALID
▪ Student Verbal Interpretation D	INVALID (AF)
▪ Student Verbal Interpretation E	INVALID (AF)
▪ Student Verbal Interpretation F	VALID
Angie	
▪ Student Verbal Interpretation A	VALID
▪ Student Verbal Interpretation B	INVALID (NCF)
▪ Student Verbal Interpretation C	INVALID (IAF)
▪ Student Verbal Interpretation D	INVALID (IAF)
▪ Student Verbal Interpretation E	INVALID (AF)
▪ Student Verbal Interpretation F	INVALID (NCF)
Sean	
▪ Student Verbal Interpretation A	VALID
▪ Student Verbal Interpretation B	INVALID (CF)
▪ Student Verbal Interpretation C	INVALID (AF)
▪ Student Verbal Interpretation D	VALID
▪ Student Verbal Interpretation E	INVALID (AF)

▪ Student Verbal Interpretation F	INVALID (CF)
Carol	
▪ Student Verbal Interpretation A	VALID
▪ Student Verbal Interpretation B	INVALID (NCF)
▪ Student Verbal Interpretation C	INVALID (IAF)
▪ Student Verbal Interpretation D	INVALID (IAF)
▪ Student Verbal Interpretation E	INVALID (IAF)
▪ Student Verbal Interpretation F	INVALID (NCF)
Jane	
▪ Student Verbal Interpretation A	VALID
▪ Student Verbal Interpretation B	INVALID (CF)
▪ Student Verbal Interpretation C	INVALID (IAF)
▪ Student Verbal Interpretation D	INVALID (IAF)
▪ Student Verbal Interpretation E	INVALID (AF)
▪ Student Verbal Interpretation F	INVALID (NCF)
Phil	
▪ Student Verbal Interpretation A	INVALID (AF)
▪ Student Verbal Interpretation B	INVALID (CF)
▪ Student Verbal Interpretation C	INVALID (IAF)
▪ Student Verbal Interpretation D	INVALID (IAF)
▪ Student Verbal Interpretation E	INVALID (AF)
▪ Student Verbal Interpretation F	INVALID (CF)
James	
▪ Student Verbal Interpretation A	INVALID (AF)
▪ Student Verbal Interpretation B	INVALID (CF)
▪ Student Verbal Interpretation C	INVALID (IAF)
▪ Student Verbal Interpretation D	INVALID (IAF)
▪ Student Verbal Interpretation E	INVALID (IAF)
▪ Student Verbal Interpretation F	INVALID (NCF)

Table 36 indicates that the interviewees gave different results on their impressions for verbal interpretation of the *p*-value. Almost all the interviewees marked Student A's verbal interpretation as valid with the exception of Phil and James. Even though Phil and James marked Student A's verbal interpretation as invalid, they provided appropriate feedback that they understood the interpretation of the *p*-value and gave an explanation that commented on the context of the interpretation. Two of the interviewees incorrectly marked other interpretations as valid. Tod marked four (i.e. A, B, C, and F) of the six verbal interpretations as valid which leads me to believe that he possessed minimal understanding of a *p*-value's verbal interpretation. Sean was another who marked Student

D's verbal interpretation as valid, but whose explanations for invalid interpretations contained evidence of conditional reasoning. Jane, Phil, and James marked all the other invalid interpretations correctly and made some note in their explanations that they were thinking about conditional probability in at least one of the invalid interpretations. The only exception was Angie and Carol who did not make any mention of concepts related to conditional reasoning for verbal interpretations they considered invalid.

Table 37. *Results of the Car Task for the Interviewees.*

Participant Responses to Car Task	**Response**
Tod	
▪ Student Interpretation A	VALID
▪ Student Interpretation B	VALID
▪ Student Interpretation C	VALID
▪ Student Interpretation D	INVALID (NCF)
▪ Student Interpretation E	VALID
▪ Student Interpretation F	VALID
Angie	
▪ Student Interpretation A	VALID
▪ Student Interpretation B	INVALID (NCF)
▪ Student Interpretation C	INVALID (NCF)
▪ Student Interpretation D	INVALID (CF)
▪ Student Interpretation E	VALID
▪ Student Interpretation F	INVALID (CF)
Sean	
▪ Symbolic Representation A	INVALID (AF)
▪ Symbolic Representation B	INVALID (NCF)
▪ Symbolic Representation C	INVALID (NCF)
▪ Symbolic Representation D	INVALID (CF)
▪ Symbolic Representation E	VALID
▪ Symbolic Representation F	INVALID (CF)
Carol	
▪ Symbolic Representation A	VALID
▪ Symbolic Representation B	INVALID (NCF)
▪ Symbolic Representation C	INVALID (NCF)
▪ Symbolic Representation D	INVALID (NCF)
▪ Symbolic Representation E	INVALID (IAF)
▪ Symbolic Representation F	INVALID (NCF)
Jane	
▪ Symbolic Representation A	VALID
▪ Symbolic Representation B	INVALID (NCF)
▪ Symbolic Representation C	INVALID (NCF)
▪ Symbolic Representation D	INVALID (CF)

▪ Symbolic Representation E	VALID
▪ Symbolic Representation F	INVALID (NCF)
Phil	
▪ Symbolic Representation A	VALID
▪ Symbolic Representation B	INVALID (CF)
▪ Symbolic Representation C	INVALID (CF)
▪ Symbolic Representation D	INVALID (CF)
▪ Symbolic Representation E	VALID
▪ Symbolic Representation F	INVALID (NCF)
James	
▪ Symbolic Representation A	VALID
▪ Symbolic Representation B	INVALID (NCF)
▪ Symbolic Representation C	INVALID
▪ Symbolic Representation D	INVALID
▪ Symbolic Representation E	VALID
▪ Symbolic Representation F	INVALID (CF)

Table 37 also showed varied responses by the interviewees regarding their understanding of the symbolic representations of the *p*-value. In the symbolic representation, Tod was the most inconsistent who marked five of the six symbolic representations as valid. This once again showed that he struggled in understanding notions of *p*-value. Carol was another participant who incorrectly marked representation E (explicit conditional statement) as invalid. She also made no mention of conditional reasoning in her explanations. Tod and Carol in particular struggled the most with the symbolic and verbal interpretations in general even during the interviews, which is potentially due to their difficulty understanding of the *p*-value in general. The other five interviewees all marked the same symbolic representations as valid (i.e. A and E) and the rest as invalid. Furthermore, these five interviewees noted issues with conditional reasoning in at least one of their invalid explanations.

After reviewing their interview and survey data, a key theme that emerged was how the interviewees were coordinating their understanding of the verbal interpretation

with the symbolic representations of the *p*-value. The verbal interpretations of the *p*-value are standard practice in introductory statistics classes, but a symbolic representation is rarely discussed in detail. Surprisingly, the symbolic representation of a *p*-value is prominent in upper-division statistics classes when discussing theories of hypothesis testing. Being current (or future) teachers of statistics, the participants in this study are individuals we expect to have a deeper knowledge of statistics and probability than average introductory statistic students. Therefore, part of a teacher's content knowledge should be the ability to articulate how verbal and symbolic representations of a p-value are related. This section will highlight areas in the interview data where we see how teachers' think about the relationship between the verbal and symbolic representations. First, I present data showing two interviewees (Carol and Tod) who struggled with their understanding of verbal and symbolic representations of the *p*-value. Second, I present data showing the work of two interviewees (Angie and Phil) that illustrated strong understanding of the relationship between verbal and symbolic representations of the *p*-value. Third, I will present data where two participants (Sean and James) had conflicted views on the use of a symbolic representation in the classroom. Lastly, I will present data showing one particular interviewee (Jane) who illustrated a cognitive shift in her understanding of the *p*-value's conditional reasoning that resulted from her comparing verbal and symbolic representations.

6.3.3.1. Struggling with verbal and symbolic representations of a p-value.

During the interviewers, two participants' work on the verbal and symbolic representations showed they struggled with the understanding of a *p*-value. Both these

interviewees had a strong background in mathematics, but minimal work in statistics and/or teaching statistics. Carol was a masters student in mathematics and mathematics education and Tod was a doctoral student in mathematics. During the interview, it was quite clear that they struggled describing the true meaning behind a verbal interpretation where most of their responses focused on a reiteration of a definition or their understanding about a hypothesis testing procedure. When asked to discuss symbolic representations, they struggled further due to their inability to understand the meaning behind the *p*-value's conditional reasoning, conditional notation, and/or ideas relations to sampling distributions.

The first interviewee we will look at is Carol who showed a strong preference to a procedural approach to a *p*-value. When first asked about her definition of the *p*-value, she provided the following response below.

> **Interviewer:** So, the first question before we get into the survey, can you describe for me your definition of a p-value?
>
> **Carol:** It's like the probability that...it's like the amount of evidence...per verbatim?
>
> **Interviewer:** Yeah sure. I just want to know what is your answer. If you think it is a textbook answer that's fine too.
>
> **Carol:** I think it's a textbook answer. I don't really know what it means, but I know that it's like the amount of evidence against the null or in favor of the alternative. Something like that.

In her response, we see that Carol is very hesitant about her understanding of the *p*-value's definition. She even explicitly exclaims that she does not know what it means, but

knows it is measuring evidence against the null. While she is correct in saying that the *p*-value does provide evidence against the null, she does not provide further information beyond her understanding of the procedure.

Tod gave a similar response when asked about his definition of the *p*-value. When prompted on how he would tell a student what a *p*-value is, he struggled with the overall definition as seen in the excerpt below.

> **Interviewer:** So, assume I am a student and I ask you what is a p-value Tod? Can you tell me what it is?
>
> **Tod:** So, the p-value is the probability that null hypothesis can be rejected...no.
>
> **Interviewer:** No that's fine. Go ahead. Take your time.
>
> **Tod:** I'm trying to think. The p-value is...reject...the probability that the null hypothesis is true right? I guess you're not supposed to answer that.
>
> **Interviewer:** I'm just wondering what do you think.
>
> **Tod:** I'm trying to remember now. It's the probability that...
>
> **Interviewer:** If you want to use some of this...
>
> **Tod:** The probability that the null hypothesis is true.
>
> **Interviewer:** So, you are saying it is the probability that the null hypothesis is true?
>
> **Tod:** Right.

Here we see Tod's understanding of the *p*-value as the probability of the null hypothesis being true instead of thinking about it as the probability of the observed sample under the null hypothesis. This is a common misconception found in the research and was seen in the data found in the *Graduate Student Task.* This misconception becomes a major issue for him and when asked to respond to other verbal representations he would fall back on

the idea that *p*-value's measure the probability of the null hypothesis, which becomes a major crux to his understanding.

As Tod and Angie progress to the symbolic notation, it started to become clear that conditional reasoning of the *p*-value and conditional probability (in general) was a major hurdle in their understanding of the *p*-value. Angie and Tod both encountered different struggles when discussing the symbolic notation. Angie for example focused very much on her procedural understanding of the *p*-value. This can be seen when she discussed the first symbolic representation (Student A) that uses a *z*-score notation.

Interviewer: So why do you say that's (Student A) valid?

Carol: Well okay, we are comparing our...yeah the p-value is a probability. It is an area under the curve. And so when you are trying to find the probability of this guy under here, z = 1.414. I am going to assume that this is the test statistic.

Interviewer: Yeah.

Carol: So, this is the test statistic and this symbolic representation here is asking us to find this guy right here. This guy right here 1.414. Is asking us to find the probability that we get a value larger than our test statistic right here. And that's I guess are the links.

Here we see Carol discuss the symbolic representation of Student A and how her validation focuses on specific concepts related to hypothesis testing procedures. She uses the words like tests statistic and area under the curve, which is reminiscent of ideas related to conducting a hypothesis test procedure where you find a test statistic and mark

p-value regions. While she says Student A is valid, she never remarks about the importance of conditioning on the null hypothesis but only on getting value's larger than the test statistic. It was later revealed in the interview while discussing Student E's symbolic representation (explicit conditional) that she struggled understanding conditional reasoning.

Interviewer: So, this one (Student E) you say is invalid?

Carol: Yeah.

Interviewer: And it is the same reasoning because it is not?

Carol: So, I am seeing p and p hats here and I am thinking like there is no p distribution that I know of and in order for us to calculate the p-value we have to know the distribution and that is how we got the test statistic and none of the...that's how we get our p-value. And none of the or...we have to know our distribution and that is how we get our test statistic and our p-value. But none of these have any indication what the distribution is because we are comparing this guy right here (points at Student E), we are comparing the probability that our population proportion is 1/2 given that our sample is 0.6. Like we are asking for a probability of a probability, but we're not asking anything about the distribution or the data. It is giving us something about the data and it is giving us something about population but we're not, it doesn't look like we are going to use this to make an inference about anything.

Interviewer: So, that is the same reason to both of these (Student E and F)? So, this one (Student F) doesn't have a conditional though?

Carol: Right, but we don't have a $\hat{p}$ distribution like…

Interviewer: So, when you see the z, what does that tell you about the distribution?

Carol: That it's standard normal.

The excerpt above shows Carol struggling with her understanding of the underlying sampling distribution of proportions and how it relates to the observed sample and the population proportion. She notes that the notation is difficult for her because it gives information about the population proportion and the sample proportion, but says nothing about a distribution or the data. This shows evidence that she does not recognize an underlying sampling distribution of sample proportions under the null hypothesis. When she makes a reference to notations like Student A (i.e. z-score representation) she knows that distribution is standard normal, but I conjecture that she still does not see the underlying sampling distribution that is the foundation of the hypothesis testing process. Furthermore, she also mentions how she is unaware how she is going to make an inference using the information that is given. This shows a struggle she has with the role conditional reasoning of the *p*-value and sampling distributions plays in a hypothesis test procedure.

Prior to asking Tod about the explicit conditional representation for the *p*-value (i.e. Student E), it was revealed that he had no knowledge of the conditional notation as seen in the excerpt below when he discussed Student C's symbolic representation.

Interviewer: How are you seeing this symbol down the middle (line in the conditional)?

Tod: This one (points at the line in the conditional statement)?

Interviewer: Yeah, the little line down the middle.

Tod: Oh, reject the null hypothesis *such that* p is...so they are saying random sample showed 60 so I assume the preference level is 60.

When asked about the meaning behind the line in the conditional notation, Tod described it to mean "such that" notation. This is in contrast to the meaning of the bar in conditional probability, which is supposed to symbolize the break between the assumption and event whose probability is being measured. His confusion is a direct result of his knowledge of mathematical set notation where bars represent "such that" in set theory. Probing further revealed that he had done very little work with conditional probability in general. This is not surprising considering his background is pure mathematics. This resulted in him struggling further with the rest of the task itself because it was his first time thinking about the conditional probability notation.

Because Carol and Tod's understanding of *p*-value was so limited, it was very difficult for them to further discuss how this would apply to student learning that went beyond a restating of a definition of the *p*-value's procedural use. The work of Carol and Tod illustrate the difficulties of new instructors first learning notions of conditional probability and *p*-value. This is in part due to missing pieces of their statistical knowledge. The work of these two participants are contrasting examples to those done by the interviewees, Angie and Phil, who in the next section show a robust understanding of conditional reasoning, symbolic notation, and *p*-value.

6.3.3.2. Relating verbal and symbolic representations of a p-value. Two of the interviewees showed a strong initial understanding of the conditional nature of the *p*-

value. The first was Angie, a GTA and doctoral candidate in statistics, who expressed a clear understanding that a *p*-value measures the probabilistic relationship between an observed result and the null hypothesis. Angie showed a strong understanding of the importance of the conditional nature in the literary interpretation of the *p*-value. She correctly identified the valid interpretations of the *p*-value and provided correct explanations for invalid interpretations. During the interview when Angie was asked to clarify how she viewed the relationship between a *p*-value and conditional probability, she gave the following response.

> **Angie:** I would say here, first off, we are trying to test under the assumption that the drug is not effective. And we don't actually have a probability for that, that's the assumption we have going in. So, we are not interested in assessing the likelihood of that statement. Then we want to do from there, under the assumption that the drug has no effect, do we see the results that we got are consistent with that belief or are they more extreme and probably in the direction of drug effectiveness that might indicate that our initial assumption was incorrect. So, we want to find evidence of our initial assumption. We see here that the results we got under our initial assumptions would only happen about 4% of the time, so there is two situations. One where in the extreme case, 4% results that could happen if the drug weren't effective or two the drug is actually effective. How effective has to be determined. That drug effectiveness is the reason we got this extreme results.

In Angie's excerpt above, we see her using her own interpretation of the *p*-value in the context of the *Drug Task*. In her explanations, she focuses on the *p*-value as a measure of the relationship between the observed results and the null hypothesis (e.g., the drug is not effective) through colloquial terms. For example, she used words like *belief* to describe the concept of assuming the null hypothesis in a hypothesis testing process. She also used *consistency* as a way to describe how researchers measure the behavior of observed data under a null hypothesis.

The second interviewee, Phil, also showed a strong understanding of the conditional nature of the *p*-value. During the interview, Phil was particularly adamant that the process of hypothesis testing is built around the idea of assuming the null hypothesis and computation of the *p*-value is a measure of the relationship between the observed result and a null hypothesis. Phil, like Angie, focused on how the *p*-value itself is a measure of the *consistency* between the observed sample and the null hypothesis as seen in the excerpt below.

> **Phil:** Oh yeah. It's the way...I've wrestled with the wording to make it less 'mathy' and more intuitive for years. And one of the ways, and I might go back to this one too. I treat is as an agreement between the observed and null hypothesis. The less agreement the more difference. The more agreement, the more you can believe the null. I am using terms that are not statistically correct. Cause you use believe the null or...but the idea is that you have a small p-value, that means that there is a little bit of agreement or not much agreement with what you observed and what you thought you should have

> observed under the hypothesis. So, you have to have both. That's why I got so excited in your survey whenever I saw a conditional probability. Because that is the whole point, it is conditional probability. Given this, what's the chance this happens.

In Phil's excerpt, we see how he elaborates more on the importance of the conditional reasoning of a *p*-value and his attempt to make the definition "less mathy." Unlike Angie, Phil, used words like *agreement* and *difference* to describe how the *p*-value is measuring the relationship between what was observed and the null hypothesis. While the words used by Phil are different from Angie they both, in essence, describe in their own words what a *p*-value is meant to measure. Unlike Angie, Phil has a long background of teaching introductory statistics in community college with a background in mathematics and mathematics education. As he explains his interpretation of the *p*-value, we see more of a push towards a non-standard interpretation of a *p*-value that he believes would be more applicable to students. This reveals his pedagogical practice and knowledge of student content that is a result of teaching statistics for several years.

Analysis of both Phil and Angie's interviews showed a strong understanding of the symbolic representation of the *p*-value. Both correctly identified the invalid and valid symbolic representations in the *Car Task.* They also gave thoughtful responses for the invalid symbolic representations of the *p*-value by focusing on the lack and/or misuse of the conditional statements. This was more apparent in the interview as they discussed the relationship between the symbolic and verbal representations of the *p*-value. For

example, below is an excerpt showing Angie's discussion of why she believed Student A and Student E were both valid representations of the *p*-value.

Interviewer: How would you write it?

Angie: So unfortunately in the course I teach, because it is so low level we don't use this type of notation very often. I had one instructor who used it and we didn't use it much in the context of p-values. We only use it in the context of talking about really basic probability. You have probability A here and probability B. What is the probability of A and B. Stuff like that so we never got into this notation. If I were to be very thorough I would say E probably. A is also valid because it is addressing the values that it is looking at the standardized values other than some other value. So, it is inherent in that saying conditioned on some values. E is the best here the probability of getting this extreme results of 60%. Yeah the probability this extreme results under the assumption of our null hypothesis that the proportion is actually 50% is 0.08. So, that would definitely be the best one there. I would say is not bad because by calculating the z value they are kind of stating that you know the proper proportion is 50% and we are 1.4 standard deviations above that.

From Angie's discussion, we gleam deeper insight into why she thought Student A and E symbolic representations were both valid. For Student E, Angie believed the symbolic representation was valid and linked the symbolic representation to her verbal interpretation of the *p*-value in this context. For Student A, Angie focused on the implicit assumption of the null hypothesis in the computation of the z-score.

Phil shared a similar kind of argument as Angie when describing the validity of Student A's symbolic representation as shown in the excerpt below.

Phil: So, if you go back to our original definition of a p-value, the chance of data or more extreme under a null hypothesis. Under a null hypothesis, then if you set it up intuitively and hopefully with a big ole graph on the board or a graph on their notebooks, then you can kind of work your way through these and see which ones tie to it. Like that number...the first one there. Student A's choice. The chance that z is bigger than or equal to 1.414. So, nowhere in that statement is there a conditional probability. There is no condition of a given null hypothesis. The thing about it is, is I wouldn't have that statement by itself on the board. That would be right next to a curve with 0 in the middle.

Interviewer: Alright

Phil: And then under that, you'd have what corresponds to 0. What corresponds to 0...this is a...hold on let me read the context again. Well actually, it would be 0 difference because they talk about hybrid over gasoline powered. You'd have the z-score axis and then under that you'd have the contextual axis.

Interviewer: Alright.

Phil: Literally, that is enough. That is enough because you've got the context in the curve. The sampling distribution showing the null distribution of the results. There is no difference in preference between these two things and then your data landed at 1.414 and more extreme is taken care off by the greater than sign.

In Phil's discussion, we see him discussing properties of the normal distribution. He also expresses how, in the classroom, he would not simply just use a z-score symbol without first making it explicit that the *z*-score is a consequence of assuming a null hypothesis. He does this by first describing the null hypothesis in the context. He then describes that in spite of the fact that Student A's use of a z-score does not show a conditional statement, the representation has an underlying null hypothesis being implicitly assumed by also discussing how he would accompany the notation with a graphical representation on the board. Phil's approaches the question once more pedagogically by relating his discussion to how we would present the material in a classroom.

In regards to Student E, Phil also showed a clear understanding on how a verbal interpretation relates to a symbolic representation of the *p*-value. The excerpt below illustrates Phil mapping the two representations to show the connections.

> **Phil:** I like E because number one, the condition, the assumption is in the right place.
>
> **Interviewer:** Alright.
>
> **Phil:** 0.5 is the assumption. It should be after the little vertical slash. And the p-hat being the 0.6 should be in front of the slash if you are going pure symbols. It should be greater than or equal to not just got equal to. You are getting away from center (the middle of the distribution).

The transcripts from of Angie and Phil highlight a very robust understanding of the conditional nature of the *p*-value through their description of the verbal and symbolic representations. Even though the both showed a strong understanding of the symbolic

representation and saw the value of representing a *p*-value using a conditional probability statement, their pedagogical vies of these ideas is noteworthy. Phil expressed a lot of ideas related to his ways of teachings these ideas while Angie herself focused largely on the assessment of the statements. Furthermore, both interviewees liked the symbolic representation of the *p*-value as a conditional statement, but had very different views of its use in the classroom.

One question that came up during the interview was whether Angie felt using a conditional probability to represent a *p*-value in her classroom was something she would advocate. The excerpt below illustrates the kinds of pedagogical choice she would make if she were teaching a symbolic representation in the classroom.

Interviewer: Would you be surprised if they (students) actually put E up there?

Angie: If I were actually teaching this (Student E) notation I would probably stress conditional notation. So, I would be quite happy with them putting E but would be quite content with F (no conditional statement).

Interviewer: …in E you stress this conditional statement but in A you don't.

Angie: Yeah.

Interviewer: Can't you just say condition on the fact that since you are using a normal distribution, your center is 0.

Angie: Yeah that is what I was stating in A. So, the difference in F and A, you need this assumption to get to that point that the probability z is greater than whatever is 0.08. But with F, if you just presented me with this probability statement, the probability that $\hat{p}$ is greater than 0.6, I actually can't assess that

so I actually don't know what that is equal to. Even with more information, knowing this other stuff, if I don't know the value p I can't assess that. If somebody gave me a response from Student A without everything right of the equal sign, I could say oh that's equal 0.08. That's just a normal distribution.

Interviewer: Okay so here it is kind of similar to a question I gave you in the very first. What would you think would be a necessary and sufficient way that you want your students to write the p-value symbolically?

Angie: Symbolically I would actually say something close to what student C was doing. Conditioning on the null hypothesis. Probability of $\hat{p}$ greater than equal to 0.6 conditioned on just H_0. That would be sufficient for me.

In the above exchange, Angie seems to take some considerations on the different ways she wants her students to write a *p*-value symbolically. While she favored the conditional probability as her overall choice (Student E), she felt it was necessary to allow several freedoms for her students. One such freedom is the necessity of the conditional form of the *p*-value. When probed about why she felt Student A's representation is appropriate even though it is not written as a conditional statement, she focuses on the importance of an assumption and how it is necessary to compute a probability. She then uses this reasoning to express how she finds issues with Student F as problematic because it is not clear what assumption is being made in order to find the probability. When finally asked about what she thought would be a reasonable expectation for her students, she still expressed that a conditional statement was what she desired but with the added freedom that a verbalization of a null assumption would be appropriate rather than strict symbols.

Phil strongly favored Student E's approach on the use of the conditional probability statement. In spite of this, Phil also seemed very hesitant on the use of symbols. This is seen in the exchange below when asked about his general view of a symbolic representation for the *p*-value.

> **Interviewer:** So, if you were to give me just your general abstract view of what the symbolic representation should be?
>
> **Phil:** So, I do class-by-class depending on how comfortable they seem with...like my classes will change term to term depending on my audience. I don't teach the exact same thing every term.

In the excerpt above, we see how Phil comments that deciding whether to use a symbolic representation depended highly on the audience of the class. The fact that he has the knowledge of his students and his understanding of symbolic representations allowed him to take more liberties in making pedagogical choices. So, decisions whether teaching specific symbolic representations versus sticking with verbal representation in the classroom was based his personal view of the needs of his students.

Phil and Angie represent examples of two participants who showed a strong understanding of the verbal and symbolic representations of the *p*-value. Their ability to relate the different representations illustrates a specialized kind of knowledge of the *p*-value that is not common amongst introductory statistics students, but is one that is appropriate for teachers of statistics. Furthermore, their understanding and ability to relate the two representations also revealed ways it might support the type of pedagogical decisions they choose to make in the classroom.

6.3.3.3. Conflicting views of the symbolic representations of a p-value. In addition to Phil and Angie, two notable participants were Sean and James. Sean and James showed a strong understanding of the conditional nature of the *p*-value during their discussions of the interview.

The interviewee Sean felt that a symbolic representation was something that should be written in the classroom. Sean was one of the participants focused deeply on how *p*-values measured the *likelihood* of a sample in relation to a null assumption. Sean's discussion of the symbolic representation showed a deeper understanding of his conditional reasoning particularly when discussing why conditioning on the null hypothesis was important as shown in the excerpt below.

Interviewer: So, this whole idea, if they were to represent it correctly you would want to think of it like a conditional statement?

Sean: Because I mean the probability that without the context it doesn't mean anything. I mean the probability that there is 3 oranges in your garbage can is kind of like that.

Interviewer: Okay

Sean: I mean if you said the probability that there is 3 oranges in your garbage can given that you live in Florida.

Interviewer: So here is a question I was asking to people especially if they taught statistics. We don't usually write it as a conditional probability, but it seems to me from your understanding it should be written as a conditional.

Sean: Yes.

In the excerpt above, we see Sean discussing why it is important to state the null assumption. He uses the context of oranges in the garbage. In his work, he describes the event of finding oranges in a garbage can to illustrate the importance of conditional probability by describing how the location of the garbage can affect the probability of the certain events.

Sean is also a notable case because he felt strongly that Student A's symbolic representation (i.e. *z*-score version of a *p*-value) was invalid. There was evidence that Sean had a strong procedural understanding of hypothesis testing so it was surprising that he found it to be invalid considering Student A's symbolic representation is a common representation of the *p*-value in introductory statistics classes. When probed further regarding why he thought it was invalid, Sean provided the following reasoning shown below.

Sean: Just that he (Student A) isn't making an assumption that the null is true.

Interviewer: Okay.

Sean: That is my sticking point.

Interviewer: So, some people would consider this (Student A) to be valid. Just out of curiosity.

Sean: I can see it is because coming from just a strictly math stat point of view, the probability that z is greater than that on a normal distribution, a standard normal distribution, is 0.08. I can see how that can be valid.

Interviewer: So, one of the things I think this is typically how we would write it in like intro stat course. So, I was wondering if you think...so you said it was invalid but now you are saying it is valid?

Sean: I can see how it is valid, but then again I am at the sticking point of the conditional probability.

Here we see Sean requiring that any symbolic representation of a *p*-value must require a conditional statement. This type of view almost hinders his understanding that when applying z-scores, there is an implicit null assumption being applied into the computation. While he states that he does see the potential for Student A's symbolic representation, he feels quite strongly that due to his understanding of the conditional nature of the *p*-value just using a *z*-score representation of a *p*-value feels lacking. This is different from the ideas discussed by Angie and Phil who recognized Student A's symbolic representation as acceptable because they potentially have a deeper understanding of the meaning behind the transformation. Here we can see how a strong a need for a conditional statement can prove to be detrimental to one's own understanding.

In contrast to Sean, James had a different opinion on the symbolic representation of the *p*-value. During the interview, James demonstrated in his survey and interview that he understood that the *p*-value is a conditional probability. When asked about regarding his pedagogical opinion on the usefulness of a symbolic representation in the classroom showed a distinct dislike of its usage in the classroom as illustrated in the excerpt below.

Interviewer: So, I know like as far is the symbolic cases, to see a student write this one (Student E). Do you think that's good?

James: I wouldn't think that. I would think they wrote it (the symbolic notation).

Interviewer: Really?

James: I mean when student write down symbolic things in math classes it doesn't tell me anything about what they got.

Interviewer: Okay. But as far as a teacher is concerned do you think they should understand what that actually means?

James: Yes. But, I don't know how to check that other than asking them one-on-one or having them write down in sentences what it actually means and they complain about me. You made a sentence. This is a math class.

In the excerpt, we see James explain how symbolic notation might prove to be unfruitful in assessing a student's understanding of a *p*-value. He discusses how a symbolic notation itself could potentially be something students memorized, but may not conceptually understand. This illustrates he uses his knowledge of students as a way to justify whether a symbolic representation might not be appropriate in the classroom.

These two conflicting views on the usage of a symbolic form of the *p*-value in the classroom illustrate different beliefs amongst statistics teachers. Sean's demonstrates a person with a strong understanding of a *p*-value who values the importance on formalization as an important idea in developing understanding of particular concept. This potentially stems from his background as a doctoral statistics student where notation plays an important of his understanding. James on the other hand is a community college instructor with greater experience teaching introductory statistics. His knowledge of

students and what he believes is accessible and necessary knowledge guides his pedagogical practices

6.3.3.4. Developing an understanding of conditional reasoning. During analysis of the interview data, one notable interviewee was Jane who showed a distinct shift in her thinking of the conditional nature of the *p*-value. Jane was a statistics teacher who showed a strong understanding of the verbal interpretation of the *p*-value. She had a background in teaching statistics and was currently teaching at a community college and four-year university with a Masters in Mathematics Education. The survey data showed she had a strong understanding of conditional reasoning and was able to correctly identify correct verbal and symbolic representations of the *p*-value. When first asked about her definition of the *p*-value during the interview she gave the following response,

> **Jane:** Okay so the p-value would be the under this null hypothesis we have this behavior of a curve. Then we get a sample we have a certain statistic that comes out and the p-value describes how extreme that value is under that assumption. So, the p-value represents this area under the curve being extreme or more extreme than that sample value.

The response given by Jane coincides with a standard interpretation of a *p*-value that we would expect from students in introductory statistics courses. Her use of words like *extreme, statistic,* and *null hypothesis* tend to be key things teachers look for when assessing a student's verbal interpretation of a *p*-value. When asked about the symbolic representation during the interview, there was a notable shift in the conversation where she discussed how her understanding of the verbal interpretation related to a symbolic

representation of a *p*-value. This can be seen in the excerpt below where she discusses her thoughts on Student E's representation.

> **Jane:** Yeah. I liked E a lot in that it has idea that it has the null in there. The null assumption that we are given this population proportion of 0.5 and then what is the probability of having a sample greater than or equal to 0.6. Like with the curve that maps kind of what I did there but built under 0.5 being our population out here at 0.6. The probability of being to the right of there.

Here we see Jane give a thorough description of the symbolic representation for Student E and how it relates to a verbal interpretation of a *p*-value. She also linked the symbolic representation to how she imagines the *p*-value is represented in a distribution where she imagines 0.5 as the center, 0.6 as the location of the observed sample, and the tail region as the *p*-value. While she did like the use of a conditional statement of Student E, that did not hinder her understanding and assessment of Student A's work as seen below.

> **Jane:** And A, although it (Student A symbolic representation) doesn't mention the null at all it kind of is able to separate from it by moving to the standard normal curve which I like in that.

Here we see Jane discussing the implicit null assumption that is necessary in order to transform the observed sample proportion to the standard normal curve, similar to explanations offered by Phil and Angie. This is quite different from the work of Sean above who almost felt the need for a conditional statement.

The interesting moment came when Jane asked if she thinks the *p*-value should be written as a conditional statement, Jane makes the following remarks.

> **Jane:** Yeah like...I was going to say I can see doing it with the 244 students but they also don't see conditional probability at all in 244 so it feels even more disconnected. I like the conditional probability way of writing it. I can see actually getting the students to do it.
>
> **Interviewer:** Okay.
>
> **Jane:** Yeah with F, I can't tell that if students wrote it that way they do understand what it represents. Probably in the back of their heads, there is this (points at the conditional probability) they just aren't representing it in that.
>
> **Interviewer:** So, that leads to my question. Do you think there is a relationship between how we view p-values and its notion as a conditional probability? Like is it a conditional probability or is it not a conditional probability?
>
> **Jane:** Yeah, I would say it is a conditional probability. I never thought of it that way, but whenever we talk about a *p*-value we do have this given assumption of the null being true.

In the last statement, we see Jane mention that prior to the study she never viewed the *p*-value as a conditional statement. As she ponders the ideas of a symbolic representation, she recalls the ideas of the verbal interpretation. Through the comparison of the symbolic representations with a verbal interpretation of a *p*-value, Jane discovers that *p*-values are conditional probabilities. This becomes a moment of clarity regarding the conditional nature of the *p*-value. This emergence of new knowledge required her to

coordinate the meaning of a *p*-value with the concept of conditional probability. What makes this particularly interesting is that through Jane's task of comparing different representations, we see her develop a new type of understanding of the *p*-value she did not originally possess. We also see a how this new understanding affects her pedagogical decisions in the classroom in her first statement where she mentions that she likes the ideas of using conditional probability to write a *p*-value in her teaching practice. This is primarily in light of her new understanding of a *p*-value. This is in contrast to the pedagogical view of James who believes there is no value in understanding the relationship between the symbolic and verbal interpretations of *p*-value. Furthermore, we also see important pedagogical ideas discussed by Jane who discusses how she might assess students who simply write a *p*-value symbolically without a conditional statement (i.e. Student F). She notes that it is hard to assess what exactly the student is thinking, showing the value of formalizing the *p*-value symbolically as a conditional statement.

6.3.4. Summary: Conditional probability and *p*-values. The need to understanding the *p*-value as a conditional probability is crucial in order to develop a robust understanding of the *p*-value. This is particularly important for teachers because it can help determine the type of pedagogical decisions they make in the classroom. The results in this section show that understanding both a verbal and symbolic representation of the *p*-value can result in several pitfalls even when focusing solely on the *p*-value's conditional nature.

One of the important results of the analysis indicate that it is difficult to assess whether teachers or students really understand the conditional nature of the *p*-value solely

through their ability to identify the correct verbal and symbolic interpretation. The interviews revealed that some teachers who might appear to understand a verbal interpretation of the *p*-value might not be able to transfer that same type of knowledge to a symbolic representation. Participants in the survey appeared to have an understanding of the *p*-value's interpretation by their ability to identify the correct statement, but some explanations were very limited and appeared to almost be a rephrasing of the definition. It was only through their discussion of the verbal and symbolic representation together during the interviews that I was able to thoroughly see participants make direct connections between the concept of the *p*-value and conditional probability. Carol and Tod for example had a lot of misconceptions that became more visible during their discussions of notational usage. Jane in particular clearly knew the verbal interpretation, but later revealed in her interview to have missing areas of content knowledge when she was asked to relate verbal and symbolic representations.

Just studying a person's ability to choose the correct interpretation of the *p*-value may not be as different as them simply memorizing a concept without understanding it's underlying meaning. There is however, some value in having individuals compare multiple representations of statistical concepts like the *p*-value. Allowing students and teachers the opportunity to analyze the relationships between the symbolic and the verbal can lead to new avenues for learning. I also find that by studying both (verbal and symbolic) side-by-side can be useful in revealing a person's understanding of a *p*-value and how it might inform their teaching practices.

7. DISCUSSIONS

The overarching goal of this research is to analyze the SKT of teachers on the topic of probability in the context of hypothesis testing. In particular, the main research question of this study is: What knowledge do CCIs and GTAs have about probability in the context of hypothesis testing? In particular,

1) How do CCIs and GTAs understand the concept of a *p*-value?
2) How do CCIs and GTAs understand the role of conditional probability in the context of hypothesis testing?
3) How do CCIs and GTAs understand formal/informal concepts of probability, simulations and sampling distributions when reasoning about tasks related to hypothesis testing?

To answer these questions, I began with a review of the research on SKT. I then presented literature from the statistics education community on teachers' and students' understanding of probability in hypothesis testing and new pedagogical approaches in the statistics education literature that researchers believe could improve student understanding of hypothesis testing. In the designing of this study, the theoretical frameworks of Ball et al. (2005), Noll (2007), and Groth (2007) were paramount in creating a potential framework to assess the SKT of teachers' understanding of probability. Using a research by Gonzalez (2012) a set of assessment questions was developed to assess CCK, SCK, and KCS through a framework for analyzing SKT in regard to the *p*-value (Table 38). This helped in the creation of survey and interview tools to study the knowledge of GTAs and CCIs with the goal of generating an initial model of

teachers' understanding of probability in the context of hypothesis testing focused on the concept of a *p*-value.

Table 38. *Theoretical Framework for Analyzing Categories of SKT for the P-value.*

Category	Indicators
Common Content Knowledge	1. Is the teacher able to appropriately give a correct answer to questions related to *p*-value? 2. Does the teacher consistently identify and interpret concepts of a *p*-value in a statistical task?
Specialized Content Knowledge	1. Does the teacher have the ability to determine standard and non-standard approaches to arguments, methods, and solutions for statistical questions regarding a *p*-value? 2. Does the teacher have the ability to provide evidence to analyze correct and incorrect solutions given by students and provide justifiable explanations and reasoning for responses that are clear and accurate?
Knowledge of Content and Student	1. Is the teacher able to anticipate students' common responses and misconceptions for a statistical task related to the *p*-value? 2. Does the teacher show evidence of knowing the most likely reason for students' responses and misconceptions in regarding tasks related to *p*-value?

The results show that by analyzing SKT of GTAs and CCIs on the *p*-value provides a deeper understanding of teachers' knowledge of probability in the context of hypothesis testing. It also provides the research community with information that can lead to greater implications for statistics education. In this chapter, I first begin with a section discussing the three compelling themes that resulted from the data in regards to the SKT of graduate students and instructors: (1) the magnitude of the *p*-value, (2) the computation of a *p*-value using empirical sampling distributions, and (3) the relationship

between *p*-value and conditional probability. I then discuss further implications of the results of this research in regard to statistics education curriculum in the various grade levels, professional development, and future research. I will then end with a section on limitations of the research.

7.1. SKT of Probability on the concept of the *p*-value.

The data gathered in this study showed that when participants are faced with solving statistical tasks on the *p*-value, they needed to access and coordinate various aspects of SKT (i.e. CCK, SCK, and KCS). For example, when assessing a student's statistical work participants must be knowledgeable of the various ways students view a problem (i.e. KCS). The teacher must compare the students' work with the common knowledge shared in the statistical classroom between teachers and students (i.e. CCK). At times, being able to understand a student's work might extend beyond methods taught in the classroom and require the teacher to access more specialized knowledge learned outside the content he/she is required to teach (i.e. SCK). Being able to see the pedagogical values of alternative approaches to statistical inference requires knowing the limitations of students' knowledge (i.e. KCS). An important part of this research is to highlight how SKT underlies graduate students and instructors understanding of probability in the context of hypothesis testing. To provide concrete examples, I discuss in the sections that follow areas where the participants accessed and related notions of CCK, SCK, and KCS with respect to the various statistical tasks provided in this study.

7.1.1. Teachers' understanding of p-value's magnitude. Investigating teachers' understanding of a *p*-value's magnitude showed various ways graduate students and

instructors understood the magnitude of the *p*-value. Data from the *Graduate Student Task* showed that several participants understand the procedural usage of the *p*-value's magnitude to show a statistically significant result. In particular, most graduate students and instructors understand how the size of a *p*-value can be used to claim a statistically significant result. Even participants who noted that a *p*-value did not relate to statistical significance still provided explanations that showed they understood procedurally why a small *p*-value relates to statistically significant results. Knowledge of the *p*-value's magnitude to make statistically significant claims aligns with CCK of a *p*-value because it encompasses shared (or basic) knowledge that every statistically educated citizen should have after completing an introductory statistics course.

Aside from knowing that small *p*-values are needed for a statistically significant result, four primary ways of thinking were discovered that categorized how teachers deeply understood the relationship between a small (or large) *p*-value and statistical significance that touched on elements of CCK and SCK. In the sections that follow, I will first discuss three types of categories of thinking that I believe highlight important and correct connections between CCK and SCK of the *p*-value's magnitude. These three ways of thinking (SSHP/SSLP, SSL, and SSPPS) highlight important ideas that encompass a robust way of thinking about a *p*-value's magnitude. I then discuss category SSD, which I believe is an important misconception that should be highlighted as an area of consideration by the statistics education regarding a *p*-value's magnitude.

7.1.1.1. Connecting CCK and SCK of a p-value's magnitude. Aside from knowing that small *p*-values are needed for a statistically significant result, three primary

ways of thinking were discovered that categorized how teachers should understand the relationship between a small (or large) *p*-value and statistical significance that touched on elements of CCK and SCK. The first is based primarily on a procedural understanding of the *p*-value being compared to the level of significance (i.e. Categories SSHP and SSLP). The second is some graduate students and teachers view the *p*-value as a measure of likelihood of the observed sample under the null assumption (i.e. Category SSL). The third is some graduate students and teachers believe that when analyzing the magnitude of the *p*-value, one must also consider the factors of statistical power and practical significance (i.e. Category SSPPS). I consider these three ways of thinking to be crucial components for developing a robust understanding of a *p*-value's magnitude for teachers (and students) and can be viewed as examples of SCK. A number of participants expressed only one way of explaining about the relationship between the *p*-value and level of significance. Still, in order to do this requires them to access forms of SCK that focuses on a deeper understanding of the procedure of hypothesis testing. In particular, those who categorized as SSPPS used notions of statistical power and/or practical significance, which is a primary example of SCK regarding a *p*-value's magnitude because it allows teachers to provide a necessary and sufficient explanation why specific choices of a level of significance might not be appropriate.

Being able to explain the connection between a *p*-value's magnitude and statistical significance requires coordination between CCK and SCK. This is seen as participants tried to relate their multiple-choice responses to their understanding of the *p*-value's magnitude. Most participants focused mainly on a single category of thinking

about a *p*-value's magnitude. In particular, most of the participants focused on comparing the *p*-value with a level of significance (i.e. SSHP and SSLP) or discussing likelihood (i.e. SSL). Even during the interview, some participants only extended their thinking to a single category. Only few participants illustrated hybrid ways of thinking. For example, some participants were able to relate how a small level of significance and small *p*-value (SSLP) related to ideas of likelihood of an observed data (SLP). Some were even able to extend their thinking to explain why the comparison of a *p*-value to a level of significance (SSLP) relates to concepts of statistical power and/or practical significance (SSPPS).

The one category in particular that came up the least was how *p*-values and statistically significant claims relate to statistical power and/or practical significance (i.e. SSPPS). This is common considering that discussing statistical power and practical significance is not a primary focus of many introductory statistics courses. It should however be considered as part of a teacher's SCK because it provides a deeper theoretical reason for choosing an appropriate level of significance that go beyond the standard values (i.e.1% or 5%) used in the statistics textbook. It is also a type of understanding members of the statistics community feel is undervalued even though it has great importance in statistical practice (Aberson et al., 2002).

A major concern for the statistics education community is that the data shows very few of the participants possessed all three categories of thinking or had the ability to properly coordinate how they related. This shows a potential gap in the statistical knowledge of teachers regarding the magnitude of the *p*-value that we as a community

should address. As researchers, being knowledgeable of these three categories of thinking can be useful in the professional development of SKT regarding a *p*-value's magnitude for statistics instructors in general.

7.1.1.2. A misconception of the p-value's magnitude. A way of thinking about the magnitude of the *p*-value that I wish to highlight is the misconception that *p*-values measure the truthfulness of a null hypothesis (i.e. SSD). Category SSD is a concept that should be viewed with great importance because it is a common misconception held by many students. The data shows that even graduate students and instructors who might have a strong procedural and conceptual understanding of the *p*-value have the potential to interpret the *p*-value as a way to measure the probability of the null hypothesis.

As defined, the *p*-value only measures the probability of the getting results as extreme or more extreme than the observed result assuming the null hypothesis. Saying the *p*-value is the likelihood of the null hypothesis implies that we know the probability of the null hypothesis (which is not true). It is important for teachers (and students) to understand that hypothesis testing is a decision-making process structured around a random process. It is through this decision-making process that we are only able to show evidence against a null hypothesis based strictly on the probability of an observed sample conditioned on a null hypothesis. This brings into light an important distinction between evidence versus likelihood, which could potentially be a source of many misconceptions.

A common metaphor for a hypothesis test is a judicial trial where we show evidence to prove whether a person is innocent or guilty in a trial. The greater the evidence against the individual, we tend to think the greater likelihood the person is

guilty. Since hypothesis testing is structured around a random process built on ideas of conditional probability, statisticians try to make similar claims but with a caution that our evidence is strictly on the probability of the observed sample structured around a decision-making process. We see similar misconceptions when discussing confidence intervals where research has shown that people misinterpret that a 95% confidence interval means the probability the population parameter is contained inside the interval is 95% instead of relating the probability to the randomization process. Category SSD is something that I believe is worth investigating because it connects further to the importance of conditional reasoning and what it means to understand the decision-making process of hypothesis testing. Understanding this type of thinking extends to notions of SCK because it requires a deeper level of understanding of conditional reasoning and how it fits into the overall structure of a hypothesis test. Training teachers to know how to respond to this misconception is also relevant in regard to developing KCS. Knowing that current and prospective teachers of statistics have difficulty understanding this subtle misconception means there is a greater chance introductory statistics students who have even less background in statistics will struggle with this common misconception.

7.1.2. Teachers' understanding of p-value's computation. The data from the *Helper-Hinderer Task* has shown that when presented with a simulation-based approach to hypothesis testing, the computation of a *p*-value (and explanation of the computation) is not consistent amongst the current population of graduate students and instructors. This is not surprising considering that simulation-based approaches are rarely discussed in

introductory statistics courses. In the sections that follow, I wish to highlight three main themes that I believe are important ideas when considering how probability concepts like a *p*-value are computed when teachers are asked to use empirical sampling distributions to compute probability. The first theme centers on those who are able to compute probability of a *p*-value correctly using empirical sampling distributions. The second theme centers on misconceptions of a *p*-value as a direct result of using empirical sampling distributions. The third is a focus on computing *p*-values that favors theoretical over empirical probability methods.

7.1.2.1. Correct computations with empirical sampling distributions. Although many of the participants were able to compute the *p*-value correctly using a relative frequency, only a small number emphasize the importance of a null hypothesis. The concept that the *p*-value is computed assuming a null hypothesis is an important idea that research has shown to be under emphasized because most students focus primarily on the procedure of computing and not the underlying meaning behind the computation. In this study, a similar phenomenon was seen where only a few participants emphasized the null hypothesis when they explained their computation. For teachers, it is important to consistently emphasize the relationship between the *p*-value and the null hypothesis because it is a critical part of understanding the role of conditional probability plays in the computation in both theoretical and simulation based approaches.

The data also shows that not all teachers are able to proficiently compute the *p*-value correctly using an empirical sampling distribution. Those graduate students and instructors who were unable to compute the *p*-value correctly either computed the relative

frequency incorrectly, used theoretical probability methods, or applied methods based on traditional hypothesis testing. This highlights an inconsistency in the type of SCK held by some of the participants regarding alternative approaches to hypothesis testing. These inconsistencies are important areas of consideration because they are potential areas of student misconceptions that we as a community should be aware of as we move towards more simulation-based approaches to statistical inference.

7.1.2.2. Misconceptions when using empirical sampling distributions. Those participants who incorrectly computed the *p*-value using relative frequency fell into two categories: CUA and CIOS. Recall, the CUA category is when a participant would apply methods typically used in traditional statistics to find probability in empirical sampling distributions. This category of CUA fails particularly because simulations do not always generate a perfect distribution. Those participants who fell in this category clearly noted the frequency of the data, but did not to attend to the fact that simulated data does not guarantee "perfectly-shaped" sampling distributions that are common in theoretical sampling distributions. This shows a potential misconception teachers (and students) might encounter as they transition between computing *p*-values from a theoretical to an empirical sampling distribution. It also shows a gap between CCK and SCK in a small sample of graduate students and instructors.

The category of CIOS shows an inconsistency by teachers whose relative frequency computation of the *p*-value ignores the observed sample. In other words, their computation of the *p*-value is a strict inequality, which does not correspond to the meaning of the *p*-value. As mentioned, this method of computation would not be

problematic in traditional introductory statistics because most approaches use a continuous probability distribution so the computation of the *p*-value using strict inequality does not affect the final value. This however is not the case of empirical sampling distributions whose outcomes are discrete. This shows another potential hurdle that might occur when transitioning teachers (and students) towards computing *p*-value's using empirical sampling distributions. I want to emphasize that using a strict inequality is a misconception that statistics teachers (and students) need to be conscious of because it is inconsistent with the true definition of a *p*-value.[15] This is type of misconception can only be visible in the calculation of a *p*-value when computing a probability using a discrete or empirical probability distribution. Traditional textbooks rarely (if at all) teach hypothesis testing using discrete probability distributions, but with the rise of simulations and empirical sampling distributions as a new standard to hypothesis testing, statistics educators need to be aware of this potential misconception.

One possible misconception that I was expecting to see, but did not occur in the study, is a participant who only computed the probability of the observed sample outcome. Although this did not show up in the research data, it is one misconception that should also be part of a statistics teachers' KCS. The misconceptions mentioned here can be easily discussed using empirical sampling distributions with students because it does not require the use of theoretical probability ideas that is usually beyond the scope of many introductory statistics students. Knowing these types of misconceptions and how they can be addressed using simulations shows the value of understanding a simulation-

[15] This misconception is a lack of noticing *p*-values are defined using "as extreme or more extreme."

based approach to hypothesis testing which consequently makes it a valuable part of a statistics teachers' SKT.

7.1.2.3. Theoretical over empirical methods. In addition to participants who use relative frequencies to compute a *p*-value, the data revealed that some participants struggle understanding the purpose of the empirical sampling distribution that was provided to them. Some participants relied heavily on traditional methods of finding a *p*-value even when prompted to use the empirical sampling distribution to compute an approximate *p*-value. When prompted regarding which approach they would prefer to use in teaching, only a few participants understood the relationship between a theoretical and empirical approach to hypothesis testing. Even amongst those who could coordinate the two ways of performing hypothesis testing, some did not recognize the value of a simulation approach to hypothesis testing as pedagogically useful. Participants who held this type of view clearly showed the necessary CCK to perform hypothesis testing in introductory statistics courses, but lacked some of understanding of the simulation-based approach (i.e. SCK) and how it might be useful in helping students development (i.e. KCS). Instead, some participants focused mainly on the existence of approximation errors in the simulation even though their suggested theoretical method would also lead to similar errors.

Those who did see the value of the empirical sampling distributions clearly viewed its importance in the teaching of hypothesis testing. This view was most common amongst those participants who had a strong background in simulation and possessed abundant experience teaching introductory statistics. This shows an amalgamation

between SCK and KCS because it shows a strong example of a teacher who understands what a simulation-approach to hypothesis testing is meant to exhibit and how it might be useful in developing knowledge that is accessible by introductory statistics students.

7.1.3. Teachers' understanding of conditional probability. One of the big ideas in hypothesis testing is that all probability being computed in a hypothesis test is a result of conditional probability. One particular probability being computed is the *p*-value that is based on the relationship between the observed result and the null assumption. Understanding this idea is key because it clarifies the role of the *p*-value in the decision-making process and why the *p*-value is a way to measure evidence against the null hypothesis.

A major goal of this dissertation is analyzing how graduate students and instructors see the role of conditional probability in the context of hypothesis testing. This was done using the *p*-value as the main probability concept. Through the *Drug Task* and *Car Task*, the participants were asked to assess various verbal and symbolic representations of the *p*-value. In order to make appropriate judgments whether a specific verbal and/or symbolic representation was valid also required the participant to coordinate their knowledge of CCK, SCK, and KCS. To illustrate this, in the sections that follow I first discuss the results of the verbal interpretations. Second, I discuss the results of the symbolic representations of the *p*-value. Finally, I discuss the importance of relating verbal and symbolic representations in the practice of teaching *p*-value.

7.1.3.1. Verbal interpretations of the p-value. Being able to write an appropriate verbal interpretation can be viewed as CCK considering it should be part of a person's

everyday statistical practice after completing an introductory statistics course. In terms of the data, several of the participants could correctly identify a standard verbal interpretation of the *p*-value within a statistical context. Furthermore, a number of participants were able to correctly identify incorrect verbal interpretations.

The most interesting data came primarily from the explanations of participants when they provided explanations for interpretations they considered invalid. Most explanations offered by participants focused more on aspects of a student's interpretation that sometimes had very little connection with a *p*-value's conditional reasoning. This is interesting considering that the conditional nature of the *p*-value should be a primary focus of any interpretation of the *p*-value. This could be a potential weakness of the survey design, or that teachers themselves sometimes do not value the importance of conditional reasoning when verbally interpreting a *p*-value. For this reason, I have some reservations regarding these results. I say this because, some participants were able to correctly mark all verbal interpretations as valid/invalid but would then not make a single mention regarding the lack of conditional probability in those they marked as invalid. One example is the interviewee (Carol) who was able to identify the correct verbal interpretation but struggled explaining the conditional relationship. Carol in particular had some experience seeing the definition of the *p*-value during her work as a statistics recitation leader, but spent very little time deeply thinking about how a *p*-value is a conditional probability. This is primarily due to fact that her background was in mathematics and mathematics education with very minimal experience taking and/or teaching statistics courses. It is possible that only identifying a correct verbal

interpretation will not be the most adequate way to assess a participant's understanding of a *p*-value as a conditional probability since a valid interpretation itself can be easily memorized.

7.1.3.2. Symbolic representations of the p-value. The participants' work on the symbolic representations of the *p*-value provided useful information on teachers' understanding conditional probability because it forced participants to view a *p*-value in a lens that is not typical in standard practice. While the verbal interpretations of the *p*-value can be viewed as CCK in the introductory statistics classroom, a symbolic representation is primarily SCK used by statisticians.

By assessing whether a participant had the ability to identify correct symbolic representations requires them to consider the importance of symbols and formalization of statistical concepts. The resulting data showed that there was less consistency amongst the participants that marked symbolic representations as valid/invalid.[16] In regard to conditional probability, many of the participants identified with the standard use of a *z* statistic as a way to represent a *p*-value. This is not unusual considering that using a *z*-score is standard practice in introductory statistics where the conditional probability is implicit in the computation. When presented with a representation where the conditional probability of *p*-value was made explicit, the data shows a large number of participants who felt the representation were valid. However, when directly comparing the results for the implicit conditional and explicit conditional directly, there revealed to be a slight discrepancy in number of participants who found both representations valid. Some

[16] It is important to note that a major hurdle was in syntax used for some of the representations, which caused a few participants to focus on specific symbols. This will be discussed further in the section on future research.

participants were unfamiliar with the use of a conditional probability to represent a *p*-value or unfamiliar with the notation for conditional probability. Those participants who struggled with the explicit form were particularly those with a background in pure mathematics. This makes sense considering that they might not have much experience working with conditional probability and using an explicit conditional form of a *p*-value is not commonly introduced except in upper level statistics courses.

When analyzing explanations for symbolic representations that participants viewed as invalid, many participants chose to focus on other aspects of the representation that once again had very little to do with discussing conditional probability. For example, many participants took to noting specific usage of notations like the $\hat{p}$ to be a primary focus of their disagreement with certain symbolic representation and paid little attention to the fact that some symbolic representations lacked a proper conditional statement. While the design of the task itself was meant to illustrate a realistic view of a student with limited ability to write a symbolic notation, this was enough of a distractor to cause most participants to ignore the important component which was the representation being presented lacked a conditional notation. This was very similar to what was seen when participants gave explanations to invalid verbal interpretations where they chose to provide feedback that did not highlight a lack of conditional reasoning in the student's work. The fact that this phenomenon occurred in both the verbal and symbolic could potentially be due to the task design of the survey or that we as an education community do not place a significant emphasis on that fact that conditional probability is an important aspect of a *p*-value even though it is explicit in its definition.

7.1.3.3. Reasons for relating verbal and symbolic representations. While the results of participants' ability to assess verbal and symbolic representations provide some useful background information regarding how some graduate students and instructors might be able to assess student work, an important result from analyzing the data came from the interviews where the participants made direct connection between the verbal and symbolic representation of *p*-values. Those that could make direct connections showed the ability to coordinate what could be considered informal statistical concepts with formal statistical ideas. In this case, the verbal interpretations might be viewed as an informal approach to the *p*-value because it is a written (or spoken) description of the *p*-value's interpretation in the context of a statistical task but is a form of CCK that is part of introductory statistics. The symbolic representation can be thought of as formal considering that it uses statistical symbols and notations that align with formal statistical practice. For example, the use of probability notation (or conditional probability notation) is a formalization of the verbal form of finding the probability of a given event. The coordination of these ideas as seen in the data is non-trivial, but expected of teachers of statistics and should be considers as an area of SCK considering that advance representations of probability (or conditional probability) is sometimes beyond the understanding of introductory statistics students.

An important concept that arose from the interview data that I wish to highlight was the pedagogical value of discussing the relationship between verbal and symbolic representations, which was shown to be enlightening to some of the participants. A major design concern with the *Car Task* on symbolic representations of the *p*-value is that

having an explicit conditional statement of a *p*-value present as one of the possible hypothetical student representations (i.e. Student E) might be too revealing. One participant (Jane) noted that during the interview the moment she saw the conditional probability form of the *p*-value, she immediately had to rethink her understanding the *p*-value. I hypothesize that the cause of this shift in thinking was due to Jane observing this very notation that caused her to reconstruct the meaning behind the notation. In particular, by seeing the symbolized form and rethinking how it relates to the verbal interpretation was a useful avenue in constructing new understanding of the *p*-value.

In traditional hypothesis testing, the use of a conditional probability to represent the *p*-value is rarely done. It is therefore not surprising that teachers who might have a limited statistics background might not realize a *p*-value can be written as a conditional probability statement. Jane for example, was a participant who at first illustrated a strong understanding of the *p*-value's verbal interpretation during the interview. She also had a strong background in the concept of conditional probability notation. By providing her an opportunity to discuss the two representations allowed her to connect her CCK of a *p*-value's verbal interpretation and conditional probability to construct a deeper understanding behind the relationship between a *p*-value and conditional probability that did not exist. This can almost be seen as a development of a new kind of SCK that is not commonly discussed in the introductory statistics classroom. While the other participants did not show such an explicit realization of the conditional nature of the *p*-value as Jane, I would suspect there is a possibility other participants might have had a similar realization.

7.2. Implications and Future Research

This research presented here highlights important ideas focused on the statistical knowledge for teaching for the college level, but I believe the findings can lead to further implications for statistics education in general. In this section, I will highlight areas of implications for this research and areas for future research that I feel directly connect with the results of this dissertation.

7.2.1. Implications for undergraduate education statistics curriculum. The research presented here shows that there is a need by researchers and educators to consider the restructuring of the current way statistics is taught. Two big topics that have been widely discussed in the statistics education community under the umbrella of statistical inference is transitioning into a statistics curriculum that uses simulations as a centerpiece in the teaching of statistical inference and relating the importance of informal and formal statistical inference. In this section, I will first discuss the implications of using simulations as a basis for teaching statistical inference. I then discuss how we as statistics education community should consider the importance of informal and formal statistical inference in the classroom. I then end with an additional section discussing the role of conditional probability in the teaching of statistical inference.

7.2.1.1. Simulations and statistical inference in the classroom. At present, in order to be proficient in the computation of a *p*-value in traditional statistics text, teachers and students must know how to correctly compute a test statistic (*z, t,* etc.) and use the appropriate probability distribution table to generate a probability. While teachers and students might be able to procedurally compute the *p*-value using traditional methods,

research has shown that many of them fail to understand the deeper theoretical ideas surrounding these procedures (e.g. Batanero, 2000; Batanero & Diaz, 2006; Castro Sotos et al, 2007; Castro Sotos et al, 2009; Falk, 1986; Garfield & Ben-Zvi, 2008; Haller & Krauss, 2002; Vallecillos & Batanero, 1997, Haller & Krauss, 2002; Heid, et al., 2005; Thompson, Liu, & Saldahna, 2007). This is no surprise for many introductory statistics students who learn these procedures, but are ill prepared due to a lack of pre-requisite knowledge (i.e. calculus and advanced probability) to develop a robust understanding. Currently, many institutions continue to teach introductory statistics courses using these traditional procedures. Cobb (2007) describes these as archaic methods considering the advancement of educational technology and the accessibility of computers in the modern classroom.

The statistics education community has advocated for the use of simulations in the teaching of statistics for years because it allows students to develop core ideas of statistical inference built upon concepts that are within the scope of a typical introductory statistics students. A significant hurdle in developing a robust understanding of hypothesis testing is the sampling distribution. Traditional hypothesis testing uses theoretical sampling distributions to compute a *p*-value, but research has shown that students struggle understanding sampling distributions and their overall role in hypothesis testing. By using a simulation-approach to hypothesis testing, empirical sampling distributions take center stage in the hypothesis testing process. Rather than using recipes structured around theoretical computations, students can create models to simulate data to generate empirical sampling distributions to then compute a *p*-value using relative

frequencies. The computation of probability concepts in statistical inference like the *p*-value using an empirical sampling distribution should be a starting point for understanding probability since it can be easily accessible by a larger population of students. This helps alleviate some of the theoretical baggage in the hypothesis testing process by making the computation of probability easily accessible to students. This allows teachers to better assess whether students understand core concepts of hypothesis testing.

Understanding a simulation-based approach to hypothesis testing is currently viewed as SCK for teachers. I say this because simulation-based approaches are rarely introduced as standard practice in introductory statistics and are viewed as alternative approaches to performing hypothesis tests by statisticians. One of the big weaknesses in the traditional curriculum is that most textbooks use theoretical sampling distributions even though empirical sampling distributions are more accessible especially in the computation of probability for many students. As statistics education push for more simulation-based approaches to statistical inference, it is vital that teachers of statistics are proficient in the use and instruction of the simulation based approach to hypothesis testing. With the current shift in the curriculum outlined by American Statistical Association (ASA) in the GAISE report (2012), soon simulation-based approaches will begin to become the norm of introductory statistics classrooms. As result, simulation-based approaches to statistical inference will soon align with CCK for many statistics teachers.

Furthermore, if we as a community continue to value teaching traditional methods of hypothesis testing in the classrooms then more time needs to be spent discussing the similarities and differences between theoretical and simulation-based approaches of statistical inference. In this research, one of the big ideas that came out was that teachers felt a simulation-based approach to computing a *p*-value (or hypothesis testing in general) was not sufficient because it could give rise to approximation error. In particular, some participants felt *p*-values were better approximated using theoretical approaches like a binomial approximation instead of simulations because of the underlying approximation error in the technique. This was an interesting discovery that was seen in a number of participants who felt that a theoretical approach for a one-proportion was more valid even though it also contained problems of approximation errors. Having discussions between theoretical and simulation-based approaches to statistical inference can be useful in filling gaps in the SKT of current and future statistics teachers. It can also be useful discussions in introductory statistics classrooms that currently (or planning) to teach traditional methods. This can help in developing deeper meaning behind methods that at present are a mystery for most introductory statistics students.

In addition to understanding how to compute probability concepts (e.g. *p*-value) using empirical sampling distributions, it is also important that professional development focuses on training teachers in the use of a simulation-based approach to hypothesis testing in the introductory statistics classroom. Not only will they learn how to compute probability concepts correctly using empirical sampling distributions, but they will be able to develop a better understanding of the purpose of the empirical sampling

distribution in practice. During the interviews, some of the participants expressed great value in a simulation-based approach mainly because they have experience using such methods. Those who did not have large experience using these types of methods appeared to favor a traditional approach and did not seem clear how it might affect student development. It is therefore important to support the professional development of teachers on the understanding of simulation-based approaches and how it can be useful in the classroom.

A natural step would be to focus on the research and design of curriculum that helps teachers (and students) better understand concepts of probability through simulations. Current curriculums that use a simulation-based approach to statistical inference have seen much success in the work of Garfield, delMas, and Zieffler (2012) in their CATALAST curriculum. Similar results can be seen in empirical research with teachers as seen in work conducted by Dolor & Noll (2015) who used a curriculum based on a simulation-based approach to hypothesis testing to support in-service and pre-service teachers gain a deeper understanding of the role test statistics and sampling distributions play in statistical inference. Work such as these show the potential of using a simulation-based approach in practice and how it can lead to deeper understanding of hypothesis testing by both students and teachers.

7.2.1.2. Informal and formal statistical inference in the classroom. One of many areas discussed in the statistics education literature is developing teachers' and students' informal inferential reasoning (Zieffler et al, 2008; Makar & Rubin, 2009). I believe it is important to focus on how students and teachers construct relationships between informal

and formal. In particular, introducing classroom activities that build on the constructing and comparing of verbal and symbolic representations of the *p*-values is something I see as a fruitful part of learning about probability concepts in statistical inference. In this research, we see how individuals can construct new knowledge through a formalization of a *p*-value.

Based on this research, I believe helping student and teachers see connections between a verbal and symbolic representations have significance in the learning of statistical concepts. The data shows there was value in having teachers talk about the symbolic representation of the *p*-value because it enlightened their understanding that it was a conditional probability. A potential tweak to such formalizations tasks is to have participants symbolize their verbal interpretations as a first phase prior to showing possible initial understandings of the concept and their ability to formalize. This will allow a better view of a teacher's initial understandings of a symbolic representation and provide more robust evidence of developing knowledge as those seen in the interviewee Jane.

Research has shown that the transition from informal to formal mathematics and statistics is non-trivial. Students struggle with this transition when they need to write formulas to represent their statistical knowledge. This research shows that there is value that can be found through the discussion of formal and informal ideas because it allows an individual to rethink their understanding of a particular concept.

7.2.1.3. Conditional probability and statistical inference. Conditional probability is a topic commonly taught in introductory statistics, but is given a minor role in many

introductory statistics classrooms. The work of Watson & Moritz (2002), highlight that conditional probability is a concept that many students struggle with even though statisticians view it as a major component of statistical understanding. This is troubling considering that understanding all aspects of probability is necessary to deeply understand statistical inference.

In the research shown here, we see that some graduate students and instructors struggle understanding conditional probability in the context of hypothesis testing even though all probability is conditional probability. This was common amongst those participants who did not possess a strong statistical background. This is not surprising considering that discussing the direct connections between conditional probability and statistical inference is primarily done in upper-divisions statistical courses. In upper-division statistics, the defining and writing of hypothesis testing concepts like the *p*-value or level of significance are done using formal conditional probability notation. This unfortunately is inconsistent when we teach introductory statistics to the larger student population where the discussion of the relationship between conditional probability and statistical inference is almost non-existent. In fact, most undergraduate statistics textbooks (e.g. Brase & Brase, 2012) only spend a few sections discussing conditional probability before moving on to statistical inference without ever making connections between the two concepts.

As a result, I believe a restructuring of our current curriculum needs to emphasize a deeper connection between the concepts found in statistical inference and conditional probability. Even though statistics instructors on all levels use similar definitions for *p*-

value and level of significance, the fact that we as a community underemphasize such connections in the introductory statistics classrooms might be a reason for several of the misconceptions we might be currently seeing.

7.2.2. Implications for the preK-12 classrooms. The research presented here focuses primarily on the SKT of teachers in the undergraduate level, but I believe this information can also be useful information for the work currently being done in the preK-12 classrooms. In particular, this research can be useful when considering current research on potential learning trajectories for students of statistics as they progress towards college.

One of the underlying ideas that have been growing in the preK-12 curriculum is a restructuring of the curriculum to focus more on developing students' statistical reasoning and thinking. A report put forth by the American Statistical Association (ASA) entitled *Guidelines for Assessing Instruction in Statistics Education: A preK-12 Framework* (Franklin et al., 2007) detail a potential learning trajectory that focuses on developing the statistical understanding of students prior to reaching college. The report outlines the importance of developing students' ability through levels of statistical maturation as they gain a deeper understanding of the statistical problem-solving process that includes notions of questioning, conjecturing, data analysis, investigation, and inference.

At present, the GAISE report for the preK-12 highlights a start in the reform of traditional curriculum that closely relates to the work presented here. In particular, the report itself stresses that classrooms should spend more time using simulations and

empirical sampling distributions as pedagogical tools to develop students' understanding of statistical inference. The report also highlights the appropriate times to introduce formal statistical concepts and methods that I believe is an important component especially if students wish to transition to collegiate level statistics where formalization is still a primary component of many statistics classrooms. I believe the work done here provide useful information that applies particularly to the recommendations put forth in GAISE and highlight potential areas where students might struggle as they come to develop their statistical knowledge in these areas.

Furthermore, in order for students to properly develop their statistical knowledge, it is important that teachers themselves have received proper training and preparation in these grade levels. A report released by the American Statistics Association (ASA) entitled the Statistics Education of Teachers (SET) (Franklin et al., 2015) highlight recommendations for the teachers in preK-12 grade level and the kinds of SKT needed to be well-prepared teachers of statistics. One particular recommendation highlighted by SET states the following,

> "Prospective teachers need to learn statistics in a way that enable them to develop deep conceptual understanding of the statistics they teach. The statistical content knowledge needed by teachers at all levels is substantial, yet quite different from that typically addressed in most college-level introductory statistics courses. Prospective teachers need to understand the statistical investigative process and particular statistical techniques/methods so they can help diverse group of students understand this process as a coherent, reasoned activity. Teachers of

> statistics must also be able to communicate an appreciation of the usefulness and power of statistical thinking. Thus, coursework for prospective teachers should allow them to examine the statistics they will teach in depth and from a teacher's perspective (p. 5)."

This recommendation points to the need for a reevaluation of the professional development programs for all levels to focus on developing deeper SKT. This includes requiring teachers to understand all aspects of statistical inference like simulation-based approaches that is starting to become more standard in middle and high-school curriculum. This is no surprise considering that SET report is based upon the recommendations put forth by GAISE.

With standard curriculum currently pushing towards the important of statistical inference and deeper understanding of the use of simulation-based approaches in the classroom, this research in particular can become useful in how we train preK-12 grade teachers so that they have the necessary knowledge to teach students as they progress towards college. Furthermore, the tasks seen in this research can be potential professional development tools that can be used for training and assessing many middle and high-school teachers especially if they plan to teach more advance statistical ideas in the classroom.

7.2.3. Implications for professional development for graduate students and instructors. The results of the data illustrate a potential need for more research regarding professional development of graduate students and instructors in the college level. The survey and interview data show that not all our current and future introductory statistics

teachers are consistent in how they view an understanding of the *p*-value or hypothesis testing in general.

The work of Speer, Gutman, & Murphy (2010) mention the educational community should consider the importance of professional development of our graduate teaching assistants. As our educational community transitions in the next few years, more graduate teaching and research assistants will become the future faculty for many college institutions. If the primary goal of graduate programs is focused on developing graduate students to become professionals in their future careers, then this should include helping those who plan to move towards careers in academia.

A major concern is that several graduate programs focused primarily on developing the mathematical or statistical content knowledge of their students through coursework, but little opportunity is spent developing their pedagogical content knowledge. This research highlights that graduate teaching assistants with minimal teaching and/or research experience were unable to highlight important aspects of SKT that focused on the importance of pedagogy. This is not surprising considering that for many graduate teaching assistants, graduate programs are their first introduction into teaching. It might be worth conjecturing that most graduate students in statistics might have sufficient training in their statistical content knowledge, but a potential lack of pedagogical training. Therefore, I believe we as a community might need to re-evaluate our current graduate programs to also emphasize the importance of developing the pedagogical content knowledge of graduate teaching assistants if we as a community want them be successful educators and future colleagues in the university level.

Mesa et al. (2012) mentions that there is a lack of research regarding teachers in community colleges. This research is a useful launching point in assessing the statistical knowledge of these teachers. With the student population of community college continually growing, it is important that we continue to focus on this community of educators that teach a large portion of introductory statistics courses.

The research presented here shows there is a vast amount of information we can still learn from the instructors of community colleges in regard to their SKT. The community college instructors who were interviewed in this study showed diverse levels of understanding regarding content and pedagogy. This is quite common considering the diversity of instructors in these institutions. This brings into light the question of professional development for these individuals. Unlike graduate students who are expected to continually take coursework to improve their statistical and mathematical content knowledge, it is not required for community college instructors to take further coursework in the fields they teach. Most professional development is primarily due to their work on outside projects, conferences, and self-study which is done on an individual basis. Unfortunately, this makes it difficult to track how different instructors develop their SKT in both content and pedagogy. The results of this research show that there is some evidence that wealth of teaching experience by community college instructors does contribute to their pedagogical content knowledge. Unfortunately, there is very little information provided in this research to make a stronger claim that experience and knowledge correlate strongly for this population. Still, this brings into question what kind

of professional development is necessary for these individuals who are already considered working professionals in their field.

The data presented here shows a potential gap in the statistical knowledge of teaching for graduate students and instructors in four-year universities and community colleges. This brings into mind the importance of accountability regarding our undergraduate institutions in the selection and preparation of the individuals who teach our introductory statistics classrooms. In order to know what is necessary, it might be useful as an educational community to consider standards necessary to teach introductory statistics courses in the college level. To do this requires further research into the knowledge of undergraduate introductory statistics teachers. This knowledge can then help administrators plan professional development programs so that current and future statistics teachers are better prepared to teach introductory statistics courses. It is vital that those who are given the opportunity to teach several of our current and future students are properly trained if we wish as an education community wish to maintain a high-level of educational standards.

7.2.4. Implications for future research. There are areas of future research that I believe can benefit from the work done in this research. In this section, I highlight potential avenues of future research that directly connect with the work presented here.

I believe more work can be done exploring concepts related to the *p*-value. This study shows that understanding the definition of the *p*-value is non-trivial even for those who teach it. I suggest future researchers to expand this research through the creation of more refined assessment tools. The interview results in this research showed diverse

levels of thinking that one might not see through survey tools. Future plans for this research should focus on continued work in the interviewing of current and prospective statistics teachers through tasks that might better assess teachers' SKT of a *p*-value. For example, finding better ways to assess formal and informal representations of the *p*-value as those suggested here might be useful areas of improvement. Also, we need to give more opportunities for teachers (and students) to solve statistical problems by having them walk through modeling tasks using simulations that include the computation of the *p*-value. This will can give us information about their modeling process and how probability concepts fit in their overall thinking of statistical inference. Furthermore, a particular concept of research that shows potential for future research is teachers' understanding of the meaning behind the phrase "as extreme or more extreme" in the definition of the *p*-value. In this study, there were areas in the participants' work that showed they struggled in conceptualizing the meaning behind the concept of "extreme" values. Evidence of this can be found in the computation of the *p*-value using empirical sampling distributions where some of the participants did not include the observed sample value when they computed the relative frequency of the *p*-value. Furthermore, during the discussion of the conditional probability, some participants recognized the importance of talking about extreme values and how this impacts the need to symbolically write a *p*-value. I only mention a few ideas related to the *p*-value, but I believe the concept is still a wide-open of area that needs further investigation.

This research only touches on a few concepts related to probability in hypothesis testing. In particular, this research focuses primarily on the concept of a *p*-value, but

minimally on the concept of level of significance. Therefore, a prospective extension of this research is to create assessment tools of SKT focusing on level of significance and statistical power since these concepts also fall into the umbrella of probability in hypothesis testing and encompass concepts of conditional probability. As highlighted in the literature review, level of significance is a concept of hypothesis testing that has proven to be difficult for students and teachers. Furthermore, statistical power has also proved to be a challenging area of study because understanding statistical power requires a strong understanding of the process of hypothesis testing and the concepts related to it (Aberson et al., 2002). In order to assess an understanding of statistical power requires researchers to first assess whether the individual even has a basic understanding of hypothesis testing concepts like the research being conducted here.

Finally, I also feel there are avenues to pursue research regarding the connections between teachers' understanding of statistical concepts and how it relates to the pedagogical choices they make in the classroom. The work of Schoenfield (2011) highlights that the decisions made by mathematics teachers can be seen as a function of their knowledge, goals, and beliefs. In this research, we see how teachers' statistical knowledge played a significant factor in the type of pedagogical choices graduate students and instructors make regarding concepts of the *p*-value. In particular, how the participants felt about the usage of symbolic notation of the *p*-value and simulation-based approaches to hypothesis testing in the teaching of students. Thus, I believe this research is a good launch point to investigate how the work of Schoenfield applies to the types of

decisions current and future teachers of statistics make in the classroom based on their SKT.

7.3. Limitations of the Research

While the data itself provided a lot of fruitful information on graduate students and instructors' understanding of the *p*-value, there are limitations to the kinds of implications that can be made by this study. One of the major limitations of this research is the overall sample size of the data. Considering the overall population of graduate students and instructors, a sample of 55 individuals may not be large enough to make deeper implications about the results. Furthermore, the sample itself was not random, but through convenience. It is likely that many of the participants who chose to complete the survey represent those individuals who have a deeper understanding of statistical ideas and are therefore more willing to participate in the study.

A second area that might limit further implications of this research is due to the survey design. During the course of the interviews, all the participants provided more in depth explanations on their understanding of the *p*-value when given the opportunity. During the survey, many participants limited their explanations to a single sentence or short excerpts. As a result, the information provided in the survey only showed a small piece of a participant's understanding of the *p*-value. This created limitations to the kind of implications we can say about teachers in general because I was only able to categorize the kinds of understandings based solely on snippets of a participant's explanations.

Lastly, there is need to improve the tasks themselves. The *Car Task* in particular needs further modification. One major issue encountered in the results was that usage of specific notation caused some of the participants to disregard key aspects of the problem. As mentioned in the methodology, the task was designed to highlight hypothetical student work on symbolic notation of the *p*-value. The choice to use specific notations in the six hypothetical student representations was done to make the student work feel more authentic, but also hindered the analysis. The data itself showed that specific notations caused some participants to focus primarily on semantic issues rather than greater concepts like conditional probability of the *p*-value. This made categorization of participant's understanding challenging and minimally informative as originally expected. A redesign of the task could be useful in helping better assess teachers' understanding of *p*-value's verbal and symbolic representation of the *p*-value.

8. CONCLUSIONS

I believe this research will be useful in filling a gap in the statistics education literature of SKT on aspects probability in the context hypothesis testing particularly in the realm of GTAs and CCIs. One of key proponents of this research that differs from other current assessment tools in the literature is the importance of analyzing a teachers' understanding of a distinct concept of probability (i.e. *p*-value) that I believe needs continued research.

Probability itself is a subject that we as an educational community have come to realize is a difficult concept for many students and by combining it with concepts like statistical inference only increases the cognitive load for introductory statistics students. Unfortunately, probability plays an important role in the understanding of statistical inference. Makar & Rubin (2009) outline three important facets of statistical inference: probabilistic understanding, generalization, and recognizing the importance of data. This research highlights the probabilistic understanding of statistical inference particular to the concept of a *p*-value. Based on the results of the study, I propose to future educators to continue to re-evaluate aspects of SKT of teachers for all levels of teachers in the realm of probability and statistical inference. This research shows that there is still potential for a growing body of research that has yet to be explored. In the previous section, I offer a few suggestions that extend beyond the concept of the *p*-value that I believe could be great avenues for future research. Being able to deeply understand teachers' and students' knowledge of probability is especially important considering the drastic reform now being conducted in the statistics education curriculum throughout all education levels.

By assessing the SKT of GTAs and CCIs, I was able to provide a framework outlining ways to assess teachers' understanding of probability specific to the *p*-value. I believe assessing SKT of teachers should always be an important responsibility of all institutions if they wish to provide quality education for their students. Garfield and Ben-Zvi (2008) mention that very few courses focus on the preparation of teachers of statistics at any educational level. In their work, they highlight that graduate courses for future statistics teachers need: (1) emphasize statistical literacy and develop statistical thinking; (2) use real data; (3) stress conceptual understanding rather than mere knowledge of procedures; (4) foster active learning in the classroom; (5) use technology for developing understanding and analyzing data; and (6) use assessments to evaluate the student learning. This research focuses on the assessment aspect highlighted by Garfield and Ben-Zvi, which I conjecture will be useful for designers of professional development curriculum for elementary, secondary, and collegiate teachers.

By knowing the baseline SKT that teachers possess on concepts of hypothesis testing, curriculum designers can develop material to support current and upcoming statistics teachers. For example, one of the key features of this study is to not only analyze teachers' statistical content knowledge, but also touch on areas of pedagogical knowledge by assessing the teachers' knowledge of student misconceptions and cognitive limitations when learning about hypothesis testing. By knowing this knowledge, researchers and policy-makers will be better informed in making decisions when choosing and designing curriculum to strengthen teachers' knowledge in statistics classrooms that align with current guidelines such as those prescribed in the GAISE

report (Aliaga et al. 2012). This is especially crucial in the professional development of instructors in the collegiate level where many researchers feel there is a need for empirical research on mathematical and statistical knowledge of teachers (Speer et al. 2010; Mesa et al., 2014).

I also recommend we continue research into the population of graduate teaching assistants and community college instructors. In this research, I highlight only a small glance into the SKT of these two populations through their work on concepts of probability and statistical inference. By analyzing how graduate students and instructors discuss the *p*-value, I hope to give researchers a lens into the current understandings and misconceptions held by statistics teachers in the college level. Understanding the SKT of teachers in the collegiate level is a body of research that has been primarily overlooked by both the statistics and mathematics education community even though college instructors play a significant role in educating our overall population.

With the statistics education community pushing for greater reform of the statistics curriculum, it is important that teachers are properly trained in new pedagogical approaches. I believe this research is needed considering the important role GTAs and CCIs play in the development of undergraduate students' statistical knowledge. I envisioned that this research is a stepping stone towards improving professional development by setting a standard in the types of knowledge we expect in statistics teachers in all levels of education.

REFERENCES

Aberson, C. L., Berger, D. E., Healy, M. R., & Romero, V. L. (2002). An Interactive Tutorial for Teaching Statistics Power. *Journal of Statistics Education*, *10*(3).

Aliaga, M., Cuff, C., Garfield, J., Lock, R., Utts, J., & Witmer, J. (2010). *Guidelines for Assessment and Instruction in Statistics Education College Report*.

Anderson-Cook, C. & Dorai-Raj, S. (2003). Making the Concepts of Power and Sample Size Relevant and Accessible to Students in Introductory Statistics Courses using Applets. *Journal for Statistics Education*, *11*(3).

Ball, D. L., Hill, H. C., & Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *American Educator*, *29*(1), 14–17, 20–22, 43–46.

Batanero, C. (2000). Controversies around the Role of Statistical Tests in Experimental Research. *Mathematical Thinking and Learning*, *2*(1–2), 75–98.

Batanero, C., & Diaz, C. (2006). Methodological and Didactical Controversies around Statistical Inference. Presented at the Proceedings of 38th Conference of the French Statistical Conference, Paris.

Blair, R., Kirkman, E. E., & Maxwell, J. (2013). *Statistical Abstract of Undergraduate Programs in the Mathematical Sciences in the United States: Fall 2010 CBMS Survey*. American Mathematical Society.

Bluman, A. G. (2012). *Elementary statistics: A step by step approach* (8th ed.). New York, NY: McGraw Hill.

Brase, C. H., & Brase, C. P. (2012). *Understandable Statistics* (10th ed.). Boston, MA:

Brooks/Cole Cengage Learning.

Brewer, J. K. (1985). Behavioral Statistics Textbooks: Source of Myths and Misconceptions? *Journal of Statistics Education*, *10*(3), 252–268.

Burgess, T. (2006). A Framework for Examining Teacher Knowledge as Used in Action while Teaching Statistics. Presented at the ICOTS-7.

Castro Sotos, A. E., Vanhoof, S., den Noortgate, W. V., & Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests? *Journal of Statistics Education*, *17*(2).

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2007). The Role of Technology in Improving Student Learning of Statistics. *Technology Innovations in Statistics Education*, *1*(1).

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about Sampling Distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 295–323). Netherlands: Kluwer Academic Publishers.

Clauser, B. E. (2008). War, Enmity, and Statistical Tables. *Chance*, *21*(4), 6–11.

Cobb, G. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum. *Technology Innovations in Statistics Education*, *1*(1).

Cobb, P., & Moore, D. S. (1997). Mathematics, Statistics, and Teaching. *The American Mathematical Monthly*, *104*(9), 801–823.

Conference Board of the Mathematical Sciences. (2010). Statistical Abstract of Undergraduate Programs in Mathematical Sciences in the United States. AMS.

Cordani, L. K. (2010). The very beginning of a class on inference: Classical vs. Bayesian. In *Invited Paper Referred*. Brazil.

DeChenne, S. E. (2010). *Learning to Teach Effectively: Science, Technology, Engineering, and Mathematics Graduate Teaching Assistants' Teaching Self-Efficacy* (Dissertation). Oregon State University.

delMas, R., Garfield, J., & Chance, B. (1999). A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning. *Journal of Statistics Education, 7*(3).

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

Diamond, R. M., & Gray, P. (1987). *National Study of Teaching Assistants* (Research/Technical). Syracuse University, Center for Instructional Development.

Dolor, J. (2013). Developing hypothetical learning trajectories for teachers' developing knowledge of the test statistic in hypothesis testing. Presented at the Proceedings of the 16th Annual Conference on Research in Undergraduate Mathematics Education, Denver, Colorado.

Dolor, J., & Noll, J. (2015). Developing Teachers' Understanding of Hypothesis Test Concepts Using Guided Reinvention. *Statistics Education Research Journal*, *14*(1).

Erickson, T. (2006). Using simulation to learn about inference. Presented at the ICOTS-7, Brazil.

Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning*, *9*(1), 83–96.

Franklin, C., Bargagllottl, A. E., Case, C. A., Kader, G. D., Scheaffer, R. L., & Spangler, D. A. (2015). *Statistics Education of Teachers* (p. 88). American Statistical Association. Retrieved from http://www.amstat.org/asa/files/pdfs/EDU-SET.pdf

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. Alexandria, VA: American Statistical Association. Retrieved from http://www.amstat.org/education/gaise/

Garfield, J., & Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning: Connecting Research and Practice*. Springer Netherlands.

Garfield, J., delMas, R., & Zieffler, A. (2010). Developing Tertiary-Level Students' Statistical Thinking Through The Use of Model-Eliciting Activities. In *ICOTS8 Invited Paper*.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM - The International Journal on Mathematics Education*, *44*(7), 883–898.

Garfield, J., & Everson, M. (2009). Preparing Teachers of Statistics: A Graduate Course for Future Teachers. *Journal for Statistics Education*, *17*(2).

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *Journal of Experimental Education*, *71*(1), 83–92.

Godino, J. D., Batanero, C., & Font, V. (2007). The Onto-Semiotic Approach to Research in Mathematics Education. *ZDM - The International Journal on Mathematics Education*, *39*(1–2), 127–135.

Golde, C. M., & Dore, T. M. (2001). *At Cross Purposes: What experiences of doctoral students reveal about doctoral education.* Philadelphia, PA: A report prepared for The Pew Charitable Trusts.

Gonzalez, O. (2012). A Framework for Assessing Statistical Knowledge for Teaching Based on the Identification of Conceptions on Variability held by Teachers. Presented at the International Conference on Mathematical Education, Seoul, Korea.

Gould, R., & Ryan, C. (2013). *Introductory Statistics: Exploring the World Through Data*. Boston, MA: Pearson.

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

Guest, G., MacQueen, K. M., & Namey, E. E. (2012). *Applied Thematic Analysis*. Thousand Oaks, California: Sage Publications.

Haller, H., & Krauss, S. (2002). Misinterpretations of Significance: A Problem Students Share with Their Teachers. *Methods of Psychological Research Online*, *7*(1).

Heid, M. K., Perkingson, D., Peters, S. A., & Fratto, C. L. (2005). Making and Managing Distinctions - The Case of Sampling Distributions. In *Proceedings of the 27TH Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Vicksburg, VA:Virginia Tech. Retrieved

from http://citation.allacademic.com/meta/p18727_index.html

Hill, H., Ball, D. L., & Schilling, S. (2008). Unpacking "pedagogical content knowledge": Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372–400.

Hill, H. C., Sleep, L., Lewis, J. M., & Ball, D. L. (2007). Assessing Teachers' Mathematical Knowledge. In F. K. Lester (Ed.), *Second Handbook of Research in Mathematics and Learning*. Charlotte, NC: Information Age Publishing.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuition. *Cognition*, *11*, 123–141.

Konold, C., Pollatsek, A., Well, A., Lohmeier, J., & Lipson, A. (1993). Inconsistencies in Students' Reasoning about Probability. *Journal for Research in Mathematics Education*, *24*(5), 392–414.

Kvatinsky, T., & Even, R. (2002). Framework for Teacher Knowledge and Understanding about Probability. Presented at the ICOTS6.

Lawton, L. (2009). An Exercise for Illustrating the Logic of Hypothesis Testing. *Journal of Statistics Education*, *17*(2).

Lipson, K. (2003). The Role of the Sampling Distribution in Understanding Statistical Inference. *Mathematics Education Research Journal*, *15*(3), 270–287.

Liu, Y. (2005, August). *Teachers' Understandings of Probability and Statistical Inference and their Implications for Professional Development* (Dissertation). Vanderbilt University, Nashville, Tennessee.

Liu, Y., & Thompson, P. (2004). Teachers' personal and pedagogical understanding of

probability and statistical inference. In *Proceedings of the Twenty-sixth Annual Meeting of the International Group for the Psychology of Mathematics Education*. Toronto, Canada.

Liu, Y., & Thompson, P. (2005). Teachers' Understandings of Hypothesis Testing. In *Proceedings of the Twenty-seventh Annual Meeting of the International Group for the Psychology of Mathematics Education*. Vicksburg, VA:Virginia Tech.

Liu, Y., & Thompson, P. W. (2009). Teachers' understandings of proto-hypothesis testing. *Pedagogies*, *4*(1), 126–138.

Luft, J. A., Kurdzuek, J. P., Roehrig, G. H., & Turner, J. (2004). Growing a garden without water: Graduate teaching assistants in introductory science laboratories at a doctoral/research university. *Journal of Research in Science Teaching*, *41*(3), 211–233.

Lutzer, D., Rodi, S. B., Kirkman, E. E., & Maxwell, J. (2007). *Statistical Abstract of Undergraduate Programs in the Mathematical Sciences in the United States* (CBMS SURVEY/2005). AMS. Retrieved from http://www.ams.org/profession/data/cbms-survey/cbms2005

Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 353–373). Dordrecht: Kluwer Academic Publishers.

Makar, K., & Confrey, J. (2005). "Variation-talk": Articulating meaning in statistics. *Statistics Education Research Journal*, *4*(1), 27–54.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105.

Mesa, V. (2012). Achievement goal orientation of community college mathematics students and the misalignment of instructors' perceptions. *Community College Review*, *40*(1), 46–74.

Mesa, V., Sitomer, A., Strom, A., & Yonatta, M. (2012). Research Commentary: Moving from anecdote to evidence: The need for a research agenda in community college mathematics education. Presented at the Research in Undergraduate Mathematics Education, Portland, OR.

Mesa, V., Wladis, C., & Watkins, L. (2014). Research problems in community college mathematics education: Testing the boundaries of K-12 research. *Journal for Research in Mathematics Education*, *45*, 173–193.

Metz, M. L. (2010). Using GAISE and NCTM Standards as Frameworks for Teaching Probability and Statistics to Pre-Service Elementary and Middle School Mathematics Teachers. *Journal of Statistics Education*, *18*(3).

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). Common core state standards for mathematical practice. National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C.

Nickerson, R. S. (2000). Null Hypothesis Testing: A Review of an Old and Continuing Controversy. *Psychological Methods*, *5*, 241–301.

Noll, J. (2007). *Graduate Teaching Assistants' Statistical Knowledge for Teaching*

(Doctoral Dissertation). Portland State University, Portland, OR.

Noll, J. (2008). An Investigation of Graduate Teaching Assistants' Statistical Knowledge for Teaching. Presented at the Conference on Research in Undergraduate Mathematics Education, San Diego, CA.

Nunally, J. C. (1975). *Introduction to statistics for psychology and education*. New York: McGraw Hill.

Pfannkuch, M. (2010). Inferential Reasoning: Learning to "Make a Call" in Practice. In *ICOTS8 Invited Paper Refereed*.

Pfannkuch, M. (2011). The Role of Context in Developing Informal Statistical Inferential Reasoning: A Classroom Study. *Mathematical Thinking and Learning*, *13*, 27–46.

Pfannkuch, M., Forbes, S., Harraway, J., Budgett, S., & Wild, C. J. (2013). "Bootstrapping" students' understanding of statistical inference. Retrieved from www.tlri.org.nz

Rossman, A. (2008). Reasoning About Informal Statistical Inference: One Statisticians' View. *Statistics Education Research Journal*, *7*(2), 5–19.

Rossman, A., Chance, B., & Medina, E. (2006). Some important Comparisons between Statistics and Mathematics, and Why Teachers Should care. In *Thinking and Reasoning with Data and Chance*. Reston, VA: NCTM.

Saldanha, L., & Thompson, P. (2002). Conceptions of Sample and their Relationship to Statistical Inference. *Educational Studies in Mathematics*, *51*(3), 257–270.

Schneiter, K. (2008). Two Applets for Teaching Hypothesis Testing. *Journal of Statistics Education*, *16*(3).

Schoenfeld, A. (2011). Toward Professional Development for Teachers Grounded in the Theory of Decision Making. *ZDM - The International Journal on Mathematics Education*, *43*, 457–469.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. Lester (Ed.), *Handbook of research design in mathematics and science education* (pp. 307–334). Mahwah, NJ: Lawrence Erlbaum Associates.

Shaughnessy, J. M., & Chance, B. (2005). *Statistical Questions from the Classroom*. Reston, VA: National Council of Teachers of Mathematics.

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

Sorto, M. A. (2006). Identifying Content knowledge for Teaching Statistics. Presented at the ICOTS-7.

Speer, N., Gutman, T., & Murphy, T. J. (2010). Mathematics Teaching Assistant Preparation and Development. *College Teaching*, *53*(2), 75–80.

Speer, N., Smith III, J. P., & Horvath, A. (2010). Collegiate mathematics teaching: An unexamined practice. *The Journal of Mathematical Behavior*, *29*, 99–114.

Stigler, S. (1999). *Statistics on the Table*. Cambridge: Harvard University Press.

Tarr, J., & Lannin, J. K. (2005). Using research-based knowledge of students' thinking in conditional probability and independence to inform instruction. In *Exploring probability in school: Challenges for teaching and learning* (pp. 215–238). New York: Springer.

Thompson, P., Liu, Y., & Saldanha, L. (2007). Intricacies of Statistical Inference and

Teachers' Understandings of Them. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (1st ed., pp. 207–231). Mahwah, NJ: Psychology Press.

Thompson, P., Saldanha, L., & Liu, Y. (2004). Why statistical inference is hard to understand. Presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Vallecillos, A. (2002). Empirical evidence about understanding of the level of significance concept in hypothesis testing by university students. *Themes in Education*, *3*(2), 183–198.

Vallecillos, A., & Batanero, C. (1997). Activated concepts in the statistical hypothesis contrast and their understanding by university students. *Reserchers En Didactique Des Mathematiques*, *17*(1), 29–48.

Watson, J. M., & Moritz, J. B. (2002). School Students' reasoning about conjunction and conditional events. *International Journal for Mathematics Education in Science and Technology*, *33*(1), 59–84.

Weinberg, A., Wiesner, E., & Pfaff, T. (2010). Using Informal inferential reasoning to develop formal concepts: Analyzing and activity. *Journal of Statistics Education*, *18*(2).

Wild, C. J., & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, *67*(3), 223–248.

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What Does Research Suggest About the Teaching and Learning of Introductory Statistics at the College Level? A Review of the Literature. *Journal of Statistics*

Education, *16*(2), 1–25.

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

APPENDIX A: TASK DESCRIPTIONS

Assessment Question	Question Goals and Outcomes
Drug Test Task	- Modified version of the question from the CAOS test (delMas et al., 2007) to have participant respond to hypothetical student interpretations. - Assesses whether a participant is able to interpret the meaning of the p-value in-context. - Assesses whether a participant is able to recognize the conditional nature of the *p*-value. - Participant should recognize that p-value measures the probability of getting extreme samples using the observed sample as the point of reference, thus making Student A's interpretation the only valid statement. - Assess KCS by analyzing how a participant might responds to student correct/incorrect interpretations through qualitative responses. - Participant should recognize Student A as the only correct interpretation.
Graduate Student Task	- Modified version of the question from the CAOS test (delMas et al., 2007). - Assess whether participant is able to interpret the magnitude of the p-value in a hypothesis test task. - Participant should understand that a small p-value provides statistical significant results. - Assesses how a participant views the relationship between statistical significance and *p*-value through a qualitative response.
Car Task	- Original question inspired by the work Castro Sotos et al. (2009) and Shaughnessy and Chance (2005). - Assesses whether the participant is able to recognize symbolic representation of a *p*-value. - Determines whether participant can recognize student common misconceptions of a *p*-value's representation related to conditional probability. - Participant should recognize that Student A and Student F are both correct symbolic representations.

Approval Poll Task	- Original question inspired by the work of Chance et al. (2004 & 2007) and the CATALST curriculum. - Assesses whether a participant is able to recognize the null hypothesis in a simulation approach to hypothesis testing. - Assesses whether a participant is able to recognize the role of the sampling distribution in a hypothesis test. - Assesses whether a participant is able to correctly compute a *p*-value using an empirical sampling distribution. - Assesses whether a participant is able to correctly conclude the results of a simulation-based approach to hypothesis testing.

APPENDIX B: SURVEY QUESTIONS

1. A research article reports on a new drug test that is to be used to decrease vision loss. The article reports that the sample results gave a p-value of 0.04 in the analysis section. Below are six different student interpretations of the p-value.

Student A: *The probability of getting a result as large as or larger than the one in this study if the drug is actually not effective is 0.04.*

Is this student's interpretation of a p-value valid? ___ Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student B: *The probability that the drug is not effective is 0.04.*

Is this student's interpretation of a p-value valid? ___Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student C: *There is a 4% chance that the researcher made an error.*

Is this student's interpretation of a p-value valid? ___ Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student D: *If you were to repeat the study, there is a 4% chance of getting exactly the same result as the one in this study if the drug is actually not effective.*

Is this student's interpretation of a p-value valid? ___Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student E: *The probability that the drug is effective is 0.04.*

Is this student's interpretation of a p-value valid? ___ Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student F: *There is a 96% chance the drug is effective.*

Is this student's interpretation of a p-value valid? ___ Valid ___ Invalid

If you think this question is invalid, explain why you think it is invalid.

2. A graduate student is designing a research study. She is hoping to show that the results of an experiment are statistically significant. What type of p-value would she want to obtain?

a. A large p-value.
b. A small p-value.
c. The magnitude of a p-value has no impact on statistical significance.

Explain your understanding of the relationship between a *p*-value and statistical significance in the context of this problem.

3. A research was conducted to determine whether people preferred hybrid over traditional gasoline powered vehicles. The article reports that their random sample showed that 60% of people preferred hybrids that yielded a test statistic of z = 1.414 and a p-value of 0.08 using a right-tailed hypothesis test for proportions. Below are six different symbolic representations of the p-value given by students.

Student A: $P(z \geq 1.414) = 0.08$

Is this student's symbolic representation of a p-value valid? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student B: $P(\hat{p} = 0.6) = 0.08$

Is this student's symbolic representation of a p-value valid? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student C: $P(Reject\ the\ null\ hypothesis \mid \hat{p} = 0.6) = 0.08$

Is this student's symbolic representation of a p-value valid? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student D: $P(p = 0.5 \mid \hat{p} = 0.6) = 0.08$

Is this student's symbolic representation of a p-value correct? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

Student E: $P(\hat{p} \geq 0.6 \mid p = 0.5) = 0.08$

Is this student's symbolic representation of a p-value valid? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

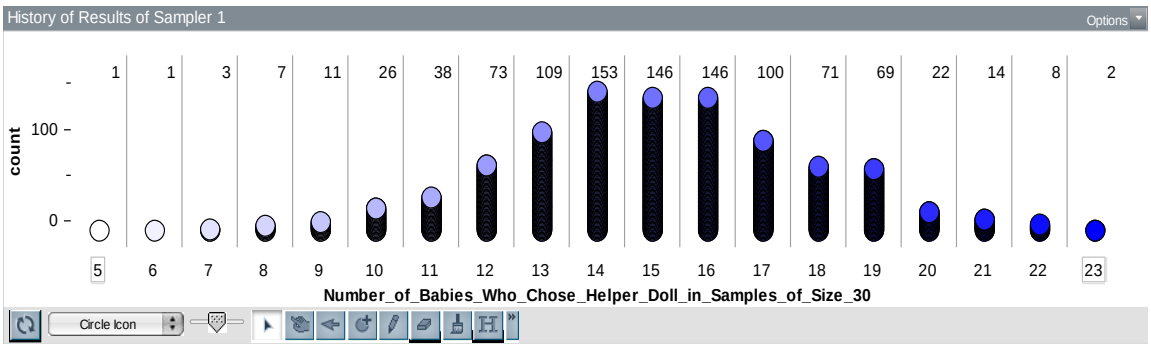
Student F: $P(\hat{p} \geq 0.6) = 0.08$

Is this student's symbolic representation of a p-value valid? ___ Valid___ Invalid

If you think this question is invalid, explain why you think it is invalid.

4. A sociology study was conducted to determine whether babies are able to recognize the difference between good and bad. In one experiment, 30 six-month old babies were randomly selected. Each baby was shown two possible puppets to play with, a 'good' puppet that helped and a 'bad' puppet that hindered. 21 out of 30 babies showed a strong preference for the helper puppet over the hinderer. In order to determine if this result provides strong statistical evidence that babies really do have a preference for the 'good' or helper puppet, James, a statistics student, conducted the following test procedure:

- James gets a coin and flips the coin 30 times.
  - If the coin lands on the "heads", he records the baby as preferring the helper puppet.
  - If the coin lands on the "tails", he records the baby as preferring the hinderer puppet.
- James then used a computer simulation to repeat the previous step 1000 times.
- James then plots the distribution for the number of times a baby chooses the helper puppet from each of the 1000 samples of size 30. This is shown in the graph below.



i.) James' procedure is based on which assumption?

a. A baby is more likely to choose the helper puppet.
b. A baby is equally likely to choose either the helper or hinderer puppet.
c. A baby is more likely to choose the hinderer puppet.

Explain the reason for your choice.

ii.) Suppose James wanted to conduct a right-tailed hypothesis test using the simulated data.

- What would you estimate for the *p*-value?
- Explain how you found the *p*-value and interpret it in the context of James' research.

iii.) Based on your estimated *p*-value, what do you think should be James' conclusion?

a. There is statistically significant evidence that babies are more likely to choose helper puppets.
b. There is statistically significant evidence that babies are more likely to choose hinderer puppets.
c. There is statistically significant evidence that babies are equally likely to choose helper or hinderer puppets.

Explain the reason for your choice.

5. Gender: Male Female

6. What is your highest academic level and degree? (Circle all the apply)

a. Bachelor's Degree.
b. Current Master's Student
c. Master's Degree
d. Current Ph.D. Student
e. Ph. D.

7. What academic level are you currently (or planning) to teach? (Circle all the apply)

a. Four-year College/University
b. Community College
c. High-School
d. Elementary School

8. How many years have you been a mathematics/statistics teacher?

9. How many different kinds of statistical courses have you taken (including those you are presently taking)? If you answered at least one, what were those courses?

10. How many times have you taught a mathematics/statistics course (including those you are currently teaching)?

11. Based on your response to 10, how many of those times have you taught introductory statistics course?

12. Have you taught other kinds of statistics courses besides introductory statistics? If yes, what were these other courses?

APPENDIX C: INTERVIEW QUESTIONS

Follow-up Question to Drug Test Task:

- Can you describe in your own words the definition of the *p*-value?
- What do think the *p*-value means in this context?
- Can you explain your reason for agreeing/disagreeing with each of these statements?
- Suppose a student wrote (a) as their definition of a *p*-value. What would your response be to that student?
- Suppose a student does not mention the notion of extreme values in their definition. What would your response be to that student?

Follow-up Question to Graduate Student Task:

- Can you describe in your own words the definition of statistical significance?
- What do you think statistical significance means in the context of this question?
- Can you explain in your own words how a p-value is related to statistical significance?
- Suppose a student asks you about the importance of the size of the *p*-value? How would you respond be to that student?

Follow-up Questions to the Car Task:

- Suppose a student in your class were to ask for a symbolic representation of a *p*-value. How you would respond?
- Have participant discuss each representation and explain in their own words why they agree/disagree.
- Have participant describe in their own words each of the symbolic representation in the context of the problem.
- Suppose a student asked if there was an alternative way to symbolically write a *p*-value for this problem. How would you respond?

Follow-up Questions to Helper-Hinderer Task:

- Question 1: Walk through each assumption and discuss why they are valid/invalid assumptions.
- Question 2: Walk through each interpretation and discuss why they are not valid/invalid choices.
- Question 3: Have the participant discuss their method of computing a *p*-value for the given task?
    - Discuss alternative ways of computing the *p*-value with the participant and have them respond to each.
- Question 4: Discuss participant's conclusion for the given task.
    - If the participant does not mention the use of *p*-value from Question 3, first have them discuss their reasoning then introduce the idea of using the *p*-value to make a conclusion.
    - If the participant does not mention using a level of significance, introduce the idea of level of significance into the problem and ask how they might be use it for the given task.
    - Prompt the participant to find a critical value for a 1% (or 5%) level of significance if they have not already done so and question how they might use it to complete the given task.