Development and Validation of a Research-based Assessment: Reasoning about *P*-values and Statistical Significance

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Sharon Jacqueline Lane-Getaz

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Department of Educational Psychology: Quantitative Methods in Education
Statistics Education

Advisers
Joan B. Garfield, Ph.D.
Robert C. delMas, Ph.D.

June 2007

UNIVERSITY OF MINNESOTA

This is to verify that we have examined this copy of a doctoral thesis by

Sharon Jacqueline Lane-Getaz

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Joan B. Garfield, Ph.D.

Robert C. delMas, Ph.D.

Date

GRADUATE SCHOOL

ACKNOWLEDGMENTS

This adventure all began over tea and bagels at Joan Garfield's kitchen table when Joan encouraged me to apply to her new program in statistics education research at the University of Minnesota. Joan's influences on this dissertation are innumerable. In her class Becoming a Teacher of Statistics, we were asked to write a short paper on difficult topics in statistics. My topic was difficulties students have interpreting *P*-values and statistical significance. Joan encouraged me to pursue research on understanding the *P*-value in a subsequent course with her, Statistics Education Research Seminar. The extensive literature on misconceptions, misinterpretations, and misuse of the *P*-value became the basis for this project. Thank you, Joan, for your patience, high standards, and professional guidance throughout my graduate program.

I also thank my professor Michael Rodriguez who taught Survey Design, Sampling and Implementation and Principles of Educational and Psychological Measurement. Further thanks go to my committee members: Mark Davison, who taught me Advanced Measurement: Theory and Applications; Bob delMas, who taught me Regression and the General Linear Model; and Frances Lawrenz, who taught Qualitative Methods in Educational Psychology. I also owe thanks to Ernest Davenport, my teacher for Advanced Multiple Regression Analysis, Michael Harwell, who taught Advanced Research Methodologies in Education, and Jeff Long, who taught Longitudinal Analysis. Each of these professors added significant pieces to the puzzle that informed the completion of this project and gave me directions for future research.

My sincerest thanks are also extended to the five statistics education experts who helped craft the RPASS-1 instrument and the ten expert raters who provided invaluable feedback and advice on the RPASS test items. Interactions with each of you not only contributed to this project but also deepened my own conceptual understanding about the subtleties of this topic. I have to add a special "thank you" to Bob delMas who provided feedback on very early drafts, and Alice Bell, Paul Roback, Yukiko Maeda, Lija Greenseid, and Ranae Hanson who provided feedback on and proofreading of later, less painful drafts of these chapters.

Of course, I want to honor the love and patience of my partner Betsy and our daughter Audrey. They gave me the time and mental space to obsessively focus on completing this program and thesis but also knew to invade that space from time to time. They kept my feet firmly planted on the ground. I must also thank the first loves of my life, Thomas and Cordelia Lane, for instilling in me – and in my five siblings – a passion for life-long learning and a value for education.

ABSTRACT

This study developed the Reasoning about *P*-values and Statistical Significance (RPASS) scale and provided content- and some construct-related validity evidence. The RPASS scale was designed to assess conceptual understanding and misunderstanding of *P*-values and statistical significance and to facilitate future research about the effects of instructional approaches on this understanding.

During Phase I, a test blueprint was developed based on difficulties identified in the literature. RPASS-1 was piloted across four courses at the University of Minnesota, assessing five correct conceptions and 12 misconceptions (*N* = 333). In Phase II, learning goals were added to the blueprint from the ongoing literature review. Incorporating modifications from the blueprint, the pilot, and suggestions from five statistics education advisors produced RPASS-2.

During Phase III, RPASS-2 was administered. Feedback from two field tests and 13 student interviews (*n* = 61) produced a 25-item RPASS-3A. Ten experts from four colleges and universities rated RPASS-3A content and made modification suggestions. After individual meetings to review an interim RPASS-3B, all ten experts *agreed* or *strongly agreed* that the two subscales (correct conceptions and misconceptions) assessed the stated learning objectives or misconceptions. Deleting one redundant item produced RPASS-4.

In Phase IV, RPASS-4 was administered to students across five introductory courses at California Polytechnic State University, assessing 13 correct conceptions and 14 misconceptions (*N* = 224). On average, respondents answered 16 items correctly.

Results showed a higher mean proportion of correct responses for correct conception items versus misconception items. Statistical literacy items were the least difficult, and statistical thinking items were the most difficult.

RPASS-4 total score reliability was low (Cronbach's coefficient $\alpha = .42$, $N = 224$). Convergent and discriminant measurements were gathered in two courses to provide some evidence of construct-related validity ($n = 56$). Correcting validity coefficients for attenuation, RPASS-4 correlated moderately with the convergent and weakly with the discriminant measure.

In Phase V, a subsequent item analysis identified a 15-item subset of RPASS-4 items (designated RPASS-5) with estimated internal consistency reliability of $\alpha = .66$. RPASS-5 retained weak evidence of construct validity as obtained for RPASS-4. Inferences about respondents' understandings and misunderstandings were drawn from these 15 items.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Background

Noted statisticians and the guidelines for reform in statistics education suggest that the first course in statistics should emphasize “the conceptual meaning of ‘*P*-value,’ ‘confidence,’ and ‘statistical significance’”—fundamental inferential concepts (ASA, 2005; Cobb, 1992, 2000; Moore, 1997). Students in virtually every discipline encounter *P*-values when they read research articles in their field. Despite their common use, statistical significance test procedures and *P*-values have been misinterpreted by authors and misunderstood by students (see Nickerson, 2000). There is an apparent need to improve statistics instruction of this topic.

However, assessments are needed to measure what students do or do not know and to evaluate whether alternative instructional approaches improve understanding of statistical significance and, more specifically, the *P*-value. Both statistics and mathematics education professionals have called attention to a need for reliable and valid instruments to facilitate statistical and mathematics education research (e.g., Ben-Zvi & Garfield, 2004; Garfield, 2006; Shaughnessy, 1992).

A working group organized by the American Statistical Association—with funding from the National Science Foundation—met over three years to discuss how to use scientifically-based research methods in mathematics education research (Scheaffer & Smith, 2007). These 20 math educators and statisticians framed a five-component research program to improve research in mathematics education: generating ideas, framing the ideas in a research setting, examining research questions in small studies, generalizing results in larger studies, and extending results over time and institutions. One goal of this research program is to link and accumulate results across studies by using proven assessment instruments.

> Measures, especially scales, in mathematics [and statistical] education need to be transportable across studies. The development of appropriate measurement scales and adequate assessment of their properties (validity, reliability, and fairness) are critical. The key questions are: Can the essential background variables be measured? Are appropriate outcome measures and related measurement instruments already developed? (Scheaffer & Smith, 2007, p. 11)

A review of the research literature from the fields of statistics, statistical and mathematics education, psychology, and educational psychology reveals difficulties or misconceptions students may have understanding probability and statistics. Researchers have examined how people's prior intuitions, heuristics, and biases may impact their reasoning about problems in probability, data analysis, and descriptive statistics (e.g., Garfield & Ahlgren, 1988; Kahneman & Tversky, 1982; Konold, 1995). Further studies have been conducted to understand the impact of instruction on students' correct conceptions and misconceptions of probability and statistics (Chance, delMas, & Garfield, 2004; delMas & Bart, 1989; Fischbein & Gazit, 1983; Nisbett, Krantz, Jepson, & Kunda, 1993). However, without instruments that show evidence of making valid and reliable inferences about this topic, only a splintered and disconnected picture of how students understand and misunderstand inference can emerge.

A research instrument for assessing students' understanding of inference is needed to evaluate what students know and to document possible obstacles to correct understanding, including misconceptions, heuristics, competing strategies and prior intuitions. The 2002 National Research Council report suggests that awareness of misconceptions “permits teachers to interpret student comments more effectively and to create assessment items to test for evidence of them.” In light of the need for standardized instruments to link results across studies, it is important that research instruments be designed to facilitate the study of students’ correct conceptions and misconceptions of *P*-values and statistical significance across classrooms, disciplines, and studies.

In a summary of 25 years of research in statistics education, Garfield (2006) reiterated the need for research instruments in statistics education and described the role a measurement developer might take within a broader research program.

> [The] measurement developer’s [role] is to carefully define what the desired student outcomes are…to develop a new instrument, which may include items or modifications of previous instruments. The goal is to develop a high quality, valid and reliable instrument that will be useful beyond one particular study. (Garfield, 2006, p. 5)

The goal of this study is to develop an instrument for statistics education research that shows evidence of making valid inferences about students’ inferential understanding.

1.2 Description of the Study

This study involved the development and validation of the Reasoning about *P*-values and Statistical Significance (RPASS) scale. The RPASS was designed to support future research on students’ conceptual understanding and misunderstanding of *P*-values and statistical significance and the effects of instructional approaches on this understanding. The study unfolded in five phases.

During Phase I, a test blueprint was developed based on difficulties identified in the literature. RPASS-1 was piloted across four courses at the University of Minnesota, assessing five correct conceptions and 12 misconceptions (N = 333).

In Phase II, learning goals were added to the blueprint from the ongoing literature review. Incorporating modifications from the blueprint, the pilot, and suggestions from five statistics education advisors produced RPASS-2.

During Phase III, RPASS-2 was administered. Feedback from two field tests and 13 student interviews (n = 61) produced a 25-item RPASS-3A. Ten experts from four colleges and universities rated RPASS-3A content and made modification suggestions. After individual meetings to review an interim RPASS-3B, experts provided final validity ratings for the items and subscales based on stated learning objectives or misconceptions being assessed (RPASS-3C). Deleting one redundant item produced RPASS-4.

In Phase IV, RPASS-4 was administered to students across five introductory courses at California Polytechnic State University, assessing 13 correct conceptions and 14 misconceptions (N = 224). Convergent and discriminant validity evidence were gathered in two of the five courses (n = 56).

Phase V consisted of conducting a subsequent item analysis of the RPASS-4 responses to identify a subset of items (designated RPASS-5) that might produce a higher internal consistency reliability if administered to a new, large sample. Inferences about respondents' understandings and misunderstandings were drawn from these 15 items.

1.3 Structure of the Dissertation

The review of the literature (Chapter 2) includes a description of what P-values and statistical significance mean and what correct understanding of these concepts entails. The literature about what students do and do not know about inference is discussed, summarized, and critiqued. None of the instruments used in these studies are shown to be valid or to produce reliable scores. There emerges a need for a research instrument to assess all the difficulties and desired understandings about *P*-values and statistical significance. Existing research instruments in statistics education are reviewed. Reasons these assessments do not meet research needs for studying *P*-values and statistical significance are discussed. The literature review closes by framing the problem and research questions to be pursued in the research study.

The methods employed to develop the RPASS instrument and collect validity evidence are described in Chapter 3. RPASS scores, reliability, and validity evidence are reported in Chapter 4. Correlations of RPASS scores with similar and dissimilar instruments are reported as evidence of convergent and discriminant validity. Chapter 4 summarizes RPASS-4 results in three item groupings: (1) based on whether a correct conception or misconception was assessed; (2) based on the four content areas defined by the test blueprint; and (3) based on the three learning goals for a first statistics course: statistical literacy, statistical reasoning, and statistical thinking. Chapter 5 is a discussion of results from the 15 most reliable items based on the three item groupings summarized in Chapter 4 and by three levels of item difficulty. Limitations of the study and questions for instrument development and future research are also discussed.

CHAPTER 2: REVIEW OF THE LITERATURE

This chapter presents the literature reviewed to inform the development of a research instrument to assess students' conceptual understanding of *P*-values and statistical significance. The literature review begins with a discussion of what it means to understand *P*-values and statistical significance. Next, research studies about students' understanding and difficulties understanding this topic are summarized, reviewed, and critiqued. The difficulties identified in these studies frame the content for the Reasoning about *P*-values and Statistical Significance (RPASS) test blueprint and scale.

The next section describes the need for a valid and reliable research instrument to assess these difficulties. The fourth section of the literature review evaluates existing research instruments in statistics education that assess students' general statistics understanding. The content and development of these research instruments are described and critiqued to inform the development of the RPASS. Issues with assessing and interpreting results from these kinds of assessments are discussed. Combining the research about students' understandings and difficulties with what has been learned in the development of statistics education research instruments, questions are posed to guide the development and validation of the RPASS.

2.1 What It Means to Understand *P*-values and Statistical Significance

2.1.1 Defining *P*-values and Statistical Significance

In order to develop and validate a research instrument to assess reasoning about *P*-values and statistical significance, one must define what *P*-values and statistical significance mean and what constitutes a correct understanding. The textbook presentation of statistical significance testing is often a hybrid of the Fisher and Neyman-Pearson approaches (Huberty, 1993).

2.1.1.1 Defining the *P*-value

Noted textbook authors use three basic elements in their definitions of the *P*-value: (1) assuming the null hypothesis is true, (2) stating the *P*-value is a probability, and (3) assessing the likelihood, rareness, or consistency of obtaining a sample statistic as or more extreme than the statistic of an observed sample. The terminology and emphases reflect a mixture of the Neyman-Pearson fixed-*a* method and Fisher's *P*-value method for statistical significance testing. Even though significance test procedures followed the Neyman-Pearson approach with null and alternative hypotheses, interpretations tended to be more Fisherian (Huberty, 1993; Wainer & Robinson, 2003). *P*-value definitions varied in their emphasis and terminology employed. Some defined the *P*-value in relation to test statistics.

> The *P*-value (also sometimes called the *observed significance level*) is a measure of inconsistency between the hypothesized value for a population characteristic and the observed sample. It is the probability, assuming that H_o is true, of obtaining a test statistic value at least as inconsistent with H_o as what actually resulted. (Devore & Peck, 2005, p. 419)

> The probability, computed assuming H_o is true, that the test statistics would take a value as extreme or more extreme than that actually observed is called the *P*-value of the test. The smaller the *P*-value, the stronger the evidence against H_o provided by the data. (Moore, 2004, p. 346)

> The *p*-value is computed by assuming the null hypothesis is true and then determining the probability of a result as extreme (or more extreme) as the

> observed test statistic in the direction of the alternative hypothesis. (Utts & Heckard, 2004, p. 360)

The definitions written by others suggest a more Fisherian approach as they do not mention test statistics at all.

> The *P*-value is the probability of getting a result at least as extreme as the observed result, in the direction conjectured by the researchers. (Chance & Rossman, 2006, p. 78)

> The ultimate goal of the calculation is to obtain a *P*-value—the probability that the observed statistic value (or even more extreme value) could occur if the null model were correct. (De Veaux, Velleman, & Bock, 2006, p. 454)

> The *P*-value for a test is the probability of seeing a result from a random sample that is as extreme as or more extreme than the one you got from your random sample *if the null hypothesis is true*. (Watkins, Scheaffer, & Cobb, 2004, p. 445)

2.1.1.2 Defining Statistical Significance

Similarly, some statistical significance definitions reflect more of a Fisherian or Neyman-Pearson approach. Devore and Peck (2005) and Utts and Heckard (2004) present a Neyman-Pearson definition of statistical significance. Furthermore, Utts and Heckard explain why the *P*-value approach would be emphasized throughout most of their textbook.

> When the value of the test statistic leads to the rejection of H_o, it is customary to say that the result is statistically significant at the chosen level α. (Devore & Peck, 2005, p. 436)

> If the test statistic is in the rejection region, conclude that the result is *statistically significant* and reject the null hypothesis. Otherwise, do not reject the null hypothesis. (Utts & Heckard, 2004, p. 448)

> It is always true that if the *p*-value ≤ α, then the test statistic falls into the rejection region, and vice versa. However, the *p*-value method…allows us to determine what decision would be made for every possible value of α. ...For that reason, and because statistical software programs and research journals generally report *p*-values, we will continue to emphasize the *p*-value approach. (Utts & Heckard, 2004, p. 449)

Some authors' definitions of statistical significance reflect a Fisherian perspective.

> This central issue of statistical significance remains exactly the same…: How often would such an extreme difference between the groups arise purely from the randomization process, even if there were truly no difference between the groups? When this probability (again called a *p-value*) is very small, we declare the group difference to be *statistically significant*. (Chance & Rossman, 2006, p. 153)

> We can define a 'rare event' arbitrarily by setting a threshold for our *P*-value. If the *P*-value falls below that point, we'll reject the null hypothesis. We call such results statistically significant. (DeVeaux et al., 2006, p. 475)

> If the *P*-value is as small as or smaller than α, we say that the data are statistically significant at level α. (Moore, 2004, p. 349)

> A sample proportion is said to be statistically significant if it isn't a reasonably likely outcome when the proposed standard is true. (Watkins et al., 2004, p. 435)

Utts and Heckard (2004) add an "advanced technical note" on Bayesian statistics (p. 360). The *P*-value is presented as the conditional probability $P(\mathrm{A} \mid \mathrm{B})$, where A is the observed data or test statistic or one more extreme and B is the null hypothesis being true.

However, to compute the conditional probability *P*(B | A), one needs to compute *P*(A and B) and, therefore, would need to know the probability that the null hypothesis is true, the very information that is sought.

> …In fact, the null hypothesis is either true or not, so even considering *P*(B) only makes sense if we think of it as a personal probability representing our belief that the null hypothesis is true. There is a branch of statistics, called *Bayesian statistics*, that utilizes the approach of assessing *P*(B) and then combining it with the data to find an updated *P*(B). (Utts & Heckard, 2004, pp. 360-361)

Thus, students may be exposed to a variety of perspectives and approaches to this topic in textbooks and articles that are all consistent with each other, and each one presents a valid perspective that is worth knowing.

2.1.2 Importance of *P*-values and Statistical Significance in Learning Statistics

The understanding and interpretation of the *P*-value is an important topic in the introductory course (ASA, 2005; Moore, 1997). For students learning statistics, attaining a conceptual understanding of the *P*-value and statistical significance opens the door to a wide array of statistical procedures that utilize this inferential logic, including testing coefficients in simple and multiple regression and longitudinal analysis and testing group equivalence or membership in one- and two-way analysis of variance (ANOVA), analysis of covariance (ANCOVA), and multivariate ANOVA (MANCOVA). Hagen (1997) summarizes the importance of the null hypothesis significance test (NHST):

> At its simplest level, the NHST logic is used to evaluate the significance of a two-variable correlation or a difference between two groups. With more complex inferential methods, the same logic is used. …It is unlikely that we will ever be able to divorce ourselves from that logic even if someday we decide that we want

to. (Hagen, 1997, p. 22)

2.1.3 Controversy about the Use of *P*-values in Significance Testing

The significance testing procedure has been criticized for its misuse and misinterpretation. Controversy about significance testing procedures began shortly after Fisher, Neyman, and Pearson popularized this frequentist approach and has continued to draw criticism (Berkson, 1942; Cohen, 1994; Falk, 1998; Kirk, 1996; Kline, 2004; Menon, 1993; Nickerson, 2000; Rozeboom, 1960; Wainer & Robinson, 2003).

2.1.3.1 Criticisms of the Use of *P*-values in Significance Testing

Four common criticisms of the use of *P*-values in significance testing are discussed to characterize the controversy (Cohen, 1994; Kirk, 1996; Kline, 2004). First is the claim that the *P*-value does not tell researchers what they really want to know (Kline, 2004). What researchers want to know is the probability the null hypothesis is true, given the data observed $P(H_o \mid \text{Data})$ (Cohen, 1994). The probability $P(H_o \mid \text{Data})$ could be computed using Bayes' theorem, as discussed by Utts and Heckard (2004).

A second criticism of significance tests "arises from a misapplication of deductive syllogistic reasoning" when interpreting the *P*-value (Cohen, 1994, p. 998). If the *P*-value is less than say .05, then one misinterpretation is that there is a .05 probability that the null hypothesis is true; the null hypothesis is deemed improbable (Cohen, 1994). Cohen argues that (1) by misapplying Boolean logic in probabilistic problems, the logic becomes invalid, and (2) only Bayesians assign probabilities to the truth or falsity of hypotheses.

A third criticism of the procedure is that the null hypothesis is always false when it is written as a nil hypothesis (i.e., comparing a parameter to zero) (Cohen, 1994).

Cohen argues that everything is related in soft psychology. One may find a statistically significant correlation when just a weak relationship exists. Cohen's criticism is that researchers tend to focus on the *P*-value rather than on the strength or weakness of a relationship. Cohen observed that researchers report these weak, statistically significant relationships as if they were important or practically significant.

A fourth criticism of significance testing is that it reduces a statistical study to a simplistic dichotomy – to reject or fail to reject the null hypothesis – based on comparing the *P*-value to an arbitrary threshold, like $\alpha = .05$ (Kirk, 1996; Nickerson, 2000). Fisher made a similar criticism of the Neyman-Pearson decision procedure (see Hubbard & Bayarri, 2003). Fisher suggested his *P*-value approach was not intended to be a "one shot" reject or fail to reject decision. He suggested that a study would be designed so that if it were repeated, the replication would rarely fail to reject the null hypothesis.

> In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result. (Fisher, 1951, p. 14)

Thus, some criticisms of the *P*-value apparently stem from whether one ascribes to a frequentist or subjectivist (Bayesian) approach to probability (see Bennett, 1998). Other criticisms echo disagreements between the Fisher and Neyman-Pearson approaches. Some criticisms are not launched at the procedures *per se* but at the misuse and misinterpretation of results from these procedures.

2.1.3.2 Debating the Use of *P*-values in Research Reports

Concerns over the misuse and misinterpretation of *P*-values and significance testing have prompted editors of psychology and education journals to call for abandoning its use (see Daniel, 1998). Entire journal issues (e.g., *Experimental Education, 61,* 1993; *Research in Schools, 5,* 1998) have been devoted to debating alternatives or complements to reporting *P*-values as the sole indicator of significant results.

Despite the call for abandoning its use, the 1999 APA Task Force on Statistical Inference stopped short of banning *P*-values from research reports. The APA recommendations suggested complementing *P*-values with other statistical measures. These statistical measures and procedures included power analysis, sample sizes, and effect sizes (see Capraro, 2004; Kirk, 1996; Thompson, 1998), confidence interval estimation (Cohen, 1994; Cumming & Finch, 2005; Good & Hardin, 2003), bootstrap studies (Thompson, 1993), and replication studies (Cohen, 1994).

Understanding the *P*-value and basic significance test procedures provides the foundation for more advanced procedures encountered in a second course in statistics, including applying the general linear model, testing coefficients in multiple regression, testing models for goodness of fit, and more advanced statistical procedures like linear mixed methods and hierarchical linear models. It is, therefore, important for statistics educators to define what introductory statistics students should understand about this topic to prepare students for reading and understanding the research literature in their fields of study.

2.1.4 Learning Goals for *P*-values and Statistical Significance

One way to define what students should understand is to categorize learning outcomes in a taxonomy. One that is used in statistics education is statistical literacy, reasoning, and thinking (see Ben-Zvi & Garfield, 2004; Chance, 2002; delMas, 2002; Garfield, 2002; Rumsey, 2002). Ben-Zvi & Garfield describe statistical literacy as the ability to organize and work with data, tables, and representations of data. Statistical literacy includes "understanding basic concepts, vocabulary, and symbols, and understanding probability as a measure of uncertainty" (Ben-Zvi & Garfield, 2004, p. 7). Garfield (2002) describes statistical reasoning as interpreting statistical information, connecting concepts, and being able to explain statistical processes. Statistical reasoning skill "combines ideas of data and chance" (Garfield, 2002, p. 1). Statistical thinking is described as understanding the "bigger picture" processes involved in conducting statistical investigations (Lane-Getaz, 2006; Pfannkuch & Wild, 2004). The "statistical thinker is able to move beyond what is taught in the course, to spontaneously question and investigate the issues and data involved in a specific context" (Chance, p. 4). In his 2002 commentary, delMas depicts statistical reasoning and thinking as intersecting learning goals embedded within statistical literacy. He describes how a particular concept or topic might have learning goals for statistical literacy, reasoning, and thinking.

To inform development of the test blueprint, there is a need to describe what constitutes evidence of a correct and full understanding of *P*-values and statistical significance. Based on input from statistics education professionals, the National Science Foundation-funded website Tools for Teaching and Assessing Statistical Inference (http://www.tc.umn.edu/~delma001/stat_tools/) provides a list of instructional outcomes for statistical inference compiled from experts in statistics education (Garfield, delMas, & Chance, 2005).

Instructional outcomes for statistical inference from the tools website are discussed in terms of statistical literacy, reasoning, and thinking goals. Students who have statistical literacy related to *P*-values and statistical significance would understand that a *P*-value assesses the role played by chance variation (Garfield et al., 2005). These students understand that small *P*-values mean the observed event is rare in the null distribution. Extending Ben-Zvi and Garfield's (2004) definition of statistical literacy to this topic suggests students are able to recognize formal definitions, symbols, and graphical representations of *P*-values and statistical significance.

Extending Ben-Zvi and Garfield's (2004) definition of statistical reasoning to this topic also suggests that students should be able to interpret results, make statistical comparisons, and make connections between different inferential concepts. Students who reason statistically would understand how *P*-values, significance levels, and confidence intervals are related; for example, a 95% confidence interval for a mean will reject any value of the null hypothesis that is outside the interval at a two-tailed 5% significance level. The student who exhibits proper reasoning about *P*-values understands how sample size relates to the magnitude of the *P*-value. The student would understand that large sample sizes lead to statistical significance even when there are small, unimportant differences from a practical perspective (Garfield et al., 2005).

Students who think statistically should be aware of *P*-values and statistical significance in the broader context of statistical investigations. A statistical thinker understands why *P*-values should be complemented with confidence intervals (see de Veaux et al., 2006). They understand that a *P*-value indicates that there may be an effect but not how large an effect there is. Statistical thinking includes being aware that the *P*-value is limited by the quality of the sample being tested. Deviations from the expected (error) may be related to issues with sampling or study design. These students would understand why there is nothing sacred about setting the significance level at .05 (Garfield et al., 2005).

2.2 Literature about Understanding *P*-values and Statistical Significance

A review of the recent literature to investigate people's understanding and misunderstanding of statistical significance, and more explicitly the *P*-value, yielded 10 empirical studies. Seven were observational studies investigating inferential understanding. Three were empirical studies investigating instructional interventions to improve inferential understanding. The summary of this section identifies 14 difficulties identified in this literature and categorizes these difficulties in four groupings to inform development of the preliminary test blueprint for the RPASS.

2.2.1 Observational Studies Related to Understanding *P*-values and Statistical Significance

Six of the seven observational studies reviewed were quantitative, using surveys or questionnaires to collect data. Vallecillos-Jimenez and Holmes (1994) investigated whether introductory statistics students believed statistically significant results prove the truth or falsity of hypotheses. They had the largest sample of the studies cited with 436 participants. The voluntary sample included students in the following areas: 93 psychology students, 75 business students, 70 civil engineering students, 63 computer science students, 43 pedagogy students, 61 medicine students, and 31 mathematics students. Students were asked to justify their responses to 4 of the 20 true-false items in the survey. Nearly a third of the responses (32.6%) reflected a belief that significance tests definitively *prove* the truth or falsity of the null or the alternative hypothesis. The true-false items analyzed in the paper read, "A statistical test of hypotheses correctly carried out establishes the truth of one of the two hypotheses, either the null or the alternative one" (Vallecillos-Jimenez & Holmes, 1994, p. 4). Seventy-seven percent of the respondents provided a justification for this item. Of these, 89.8% of the justifications suggested a misinterpretation of the item wording.

Wilkerson and Olson (1997) surveyed 52 University of Idaho graduate students with a six-item questionnaire. Twenty respondents were pursing a Ph.D., 14 an Ed.D., and 16 a master's degree. Two did not respond to this question. The study investigated whether interpretations of significance tests reflected an understanding of the relationship between treatment effects, sample size, and Type-I and Type-II error. One item read that if two studies reported the same *P*-value, "the study with the smaller *n* provides better evidence of treatment effect" (Wilkerson & Olson, 1997, p. 629). Only 1 of 52 participants answered this item correctly. Results suggest the impact of sample size on treatment effects was not well understood. However, these early career researchers did show evidence of understanding the influence of sample size on statistical significance.

Even though Oakes' 1986 study is not recent, it is discussed to provide historical context for the Falk and Greenbaum (1995) and Haller and Krauss (2002) replications. Oakes surveyed 70 academic psychologists from the United States. His instrument consisted of six true-false items associated with a brief problem scenario. The problem scenario reported statistically significant results. All six statements were incorrect interpretations of statistical significance. The items confused the small *P*-value with the probability of the truth or falsity of the hypotheses, as Type I error, and as suggesting that observed results were reliable. More than 40 of the 70 psychologists selected "true" for half of the false statements. Only three respondents correctly marked all six statements as false. Sixty percent interpreted statistically significance to mean the experiment conducted was *reliable*.

For part 2 of the exercise, the psychologists were to select the interpretation of the *P*-value they typically use from the list. If their interpretation was not on the list, they could write their own interpretation. Eight of the 70 respondents wrote out a correct interpretation. Only two of these respondents correctly interpreted statistical significance for both parts of the exercise. Oakes speculated that if the correct interpretation were included on the first part of the survey, more respondents might have answered correctly.

Falk and Greenbaum (1995) gathered intermediate level students' interpretations of items similar to those used by Oakes. This study added a correct option from which to choose. The 53 participants were students of psychology at the Hebrew University of Jerusalem. Respondents had completed two statistics courses. In a previous course, the students read Bakan's (1966) paper warning readers of the common difficulties with interpreting statistically significant results. Seven of the 53 respondents (13.2%) chose the correct option that all the other statements were false.

Haller and Krauss' (2002) replicated Oakes' study across six German universities (*n* = 113). The voluntary sample included methodology instructors (*n* = 30), scientists from non-methods courses (*n* = 39), and psychology students (*n* = 44). Results showed that 80% of the methodology instructors, 89.7% of non-methods instructors, and 100% of the psychology students selected at least one of the misinterpretations of the *P*-value as true (Haller & Krauss, 2002). Considering the results from Oakes (1986), Falk and Greenbaum (1995), and Haller and Krauss, misinterpreting the *P*-value as the probability of the truth or falsity of hypotheses appears to be common and persistent.

Mittag and Thompson (2000) studied perceptions of significance testing among members of the American Educational Research Association (AERA). A mail survey was sent to a stratified random sample of 1,037 AERA members with a 21.7% response rate. About half of the respondents were male (49.3%), and most had earned a Ph.D. (83.6%). The sample appeared to be representative of the AERA population with respect to the postal locations and AERA organizational divisions (Mittag & Thompson, 2000). The 29 items were rated on a scale from 1-*agree* to 5-*disagree*, and were analyzed in nine clusters. While the focus of the study was broader than *P*-values, five topics were directly related to interpretations of the *P*-value and statistical significance, including sample size impact, reliability, practical significance, Type I error confusion, and confusion of whether *P*-values are indicative of sample or population characteristics.

First, most AERA respondents correctly interpreted the influence of sample size on statistical significance. Second, respondents were neutral as to whether significant results were evidence of reliability. The results for three items suggested the respondents harbored misconceptions. Many interpreted non-significant findings as unimportant, confused *P*-values with Type I error, and believed *P*-values test the probability of results occurring in the sample, rather than the probability of results occurring in the population (Mittag & Thompson, 2000).

Williams (1999) employed both qualitative and quantitative methods to examine 18 introductory students' conceptual and procedural understanding of statistical significance. Students were asked to "talk aloud" while completing three open-ended tasks: one concept-mapping task and two tasks solving hypothesis-testing problems. Students were also engaged in semi-structured follow-up interviews to further probe their understanding. Conceptual understanding of "significance" was explored in the concept-mapping task. Students arranged concept labels on a page and were required to provide the links connecting the concepts. Procedural knowledge was investigated by providing students with summary data to hand calculate two statistical significance tests, a large one-sample *z*-test, and a small two-sample *t*-test.

Students had difficulty linking the concepts in the mapping task. Nine of 18 students completed the large sample *z*-test problem. Only two students completed the small two-sample *t*-test problem. Thus, interpretation of results depended mainly on follow-up interview data. Students' interview responses were classified based on whether they expressed correct, partially correct, or incorrect conceptions of four concepts:

hypotheses, significance level, *P*-value, and significance. There were 17 correct interview comments concerning major ideas of hypotheses and 13 comments correctly describing the significance level. Hypotheses and significance level seemed to be the best understood of the concepts.

Student responses included no correct descriptions of the *P*-value and only one correct description of statistical significance. There were, however, eleven comments that correctly tied *P*-values to other concepts. Williams attributed the poor responses to students' difficulties with statistical language as opposed to evidence of misconceptions. Williams (1999) noted one misconception – unique to her study – that these students believed that the "*P*-value is always low" (p. 559).

Williams' study of introductory students did not identify one potential difficulty believed to confuse introductory students as described by Batanero (2000). Batanero suggested that introductory students misapply the Boolean logic of the converse in some of their statistical analyses, switching the hypothesis and conclusion of a conditional. Similar discussions appear in the work of Falk and Greenbaum (1995).

2.2.2 Empirical Studies Investigating Instructional Interventions to Improve Inferential Understanding

Three lines of empirical research have explored instructional interventions to improve introductory students' inferential understanding (Collins & Mittag, 2005; Lipson, Kokonis, & Francis, 2003; Saldanha & Thompson, 2002, 2006). Two studies are related to how understanding sampling variation may facilitate developing inferential understanding. Lipson and colleagues focused their efforts on how students perceive what is encountered in a computer simulation and how to design a computer intervention to improve inferential understanding. Saldanha and Thompson have focused their efforts on the development of activities along a learning trajectory to improve understanding of sampling variation and thereby improve inferential understanding. Collins and Mittag investigated the impact of statistical graphing calculators on students' understanding.

Lipson and colleagues (2003) engaged students in a simulation activity constructed to introduce the notion of the *P*-value but "without formalizing the concept" (p. 3). They engaged students in a problem context in which sample data collected by a journalist suggest that the Australian postal service achieves 88% on-time deliveries. The post office claimed 96% on-time delivery. Students were to resolve the cognitive conflict between the sample statistic and the expectation based on the population parameter (the post office claim). All subjects in the study acknowledged that the more *likely* explanation was that the post office claim was inflated. However, students used absolute terms to explain the difference rather than statistical explanations, such as "the Australia Post lied" (Lipson et al., 2003, p. 8). Based on this research, Lipson and colleagues (2003) suggested a model to describe the development of students' inferential understanding using simulation software to introduce inferential concepts. The model included four developmental stages:

1. Recognizing computer representations (e.g., for population, sample, and sampling distribution)
2. Integrating concepts of population, sample, and sampling distribution
3. Dealing with the contradiction that a sample may deviate from the hypothesized population
4. Explaining results from a statistical perspective

Students who did integrate concepts of populations, sampling, and sampling distributions did not embrace the contradiction between observed samples and what was expected based on population parameters (Lipson et al., 2003). Students seemed to miss the concept of sampling variability, the objective of the activity. Directions for future research focus on simplifying the software interface and integrating sound with the images to lessen students' cognitive load (Lipson, Kokonis, & Francis, 2006).

Describing their research, Saldanha and Thompson (2006) contend that "a conception of sampling that entails images of the repetitive sampling process, the bounded variability among sampling outcomes, and the fuzzy similarity between sample and population… is a powerful and enabling instructional endpoint" to develop conceptions of inference (p. 2). The teaching experiment modified the lessons in order to facilitate students' construction of inferential concepts. Their instructional trajectory engaged students in a series of activities to facilitate understanding sampling variation (Saldanha, 2004). The researchers believe a Multiplicative Conception of Sampling (MCS) – interconnecting repeated random sampling, variability of sample statistics, and the resultant distribution – may be linked to understanding inference (Saldanha & Thompson, 2002, 2006). Some evidence has not supported this belief (Saldanha, 2004).

In the third empirical study reviewed, Collins and Mittag (2005) investigated the impact of using inference-capable graphing calculators on inferential understanding. A treatment group (*n* = 22) received graphing calculators capable of inferential statistics and the no treatment group (*n* = 47) used graphing calculators without this capability. There was no randomization of subjects to treatment, but some control measures were employed. Both groups had the same instructor. To the degree possible, the curriculum did not differ between groups, with the exception of the calculators used. Final examination scores were the dependent variable, and scores on prior tests were used as statistical controls. No significant or practical differences were found between the two groups when controlling for performance on prior tests. The final exam items and answers were appended to the article. The small sample size and confounding factors of uncontrolled classroom effects were noted as limitations of the study.

2.2.3 Summary and Critique of Research on Understanding and Misunderstandings of Inference

2.2.3.1 Summary and Critique of Observational Studies

A summary of the methods used in the seven observational studies appears in Appendix A, Table A1. The respondent sample sizes in the observational studies were sufficient ($n \geq 50$). Vallecillos-Jimenez and Holmes (1994) and Haller and Kraus (2002) did sample across different student disciplines and institutions. However, only in the Mittag and Thompson (2000) survey was a random sampling process employed to allow generalization to the AERA member population. The remaining samples are subject to selection bias.

The number of items used in most of these studies was insufficient (six or fewer items). However, Vallecillos-Jimenez and Holmes (1994) conducted one of the better studies in terms of numbers of items. Their 20-item questionnaire would potentially have more reliability than the instruments being used in most of the other survey studies cited. Nevertheless, based on the follow-up interviews conducted, Vallecillos-Jimenez and

Holmes suggest that item wording confounded the results obtained. Thus, measurement error appears to threaten the internal validity of this study.

Wilkerson and Olson (1997) studied a nonrandom sample of 52 graduate students enrolled in summer courses. The researchers administered a five-minute, one-page questionnaire with six items. The small number of items limits content validity. The researchers recommend that the study be replicated to find if results might be generalized. However, the items were not included in the article to facilitate assessing the item content or to potentially replicate the study.

Replicating studies across different populations can strengthen generalizability of results. Falk and Greenbaum (1995) replicated Oakes' (1986) results with 53 intermediate level psychology students at The Hebrew University of Jerusalem. The Haller and Krauss (2002) study further validated results with 113 German students and professors. While there was no formal validation study, Oakes debriefed his respondents, providing some substantive evidence of the validity of the results he obtained. Even though none of these studies employed a random sample, the replications of the study across diverse groups help to substantiate the validity of Oakes' results.

The most valid of these studies is the Mittag and Thompson (2000) survey of AERA members' perceptions of statistical significance. In addition to using a stratified random sample that allowed generalization of results to the target population (AERA members), the larger number of items in the instrument (29) facilitated sampling more content than any of the other studies reviewed. Second, the items were analyzed in nine clusters, providing greater insight into the respondents' understanding. A limitation of the study is that the respondents' inferential understanding may have been biased. Those who responded to the mail-back survey were likely to have a stronger statistical background than those who did not return the survey.

In Williams' (1999) investigation of students' conceptual and procedural understanding of statistical significance, students' concept maps were only partially constructed, and few students completed the procedural calculations for the large or small sample hypothesis tests. Therefore, conclusions were limited to the interview results. The student interviews were confounded by the students' poor statistical literacy. Students had difficulty correctly describing what they did or did not know. Even though these results cannot be generalized, Williams' findings suggest the difficulty of constructing items to assess students' apparently fragile knowledge of *P*-values and statistical significance testing.

2.2.3.2 Summary and critique of empirical studies

A summary of the methods used in the three empirical studies investigating instructional interventions on inferential understanding appears in Appendix A, Table A2. The empirical lines of research being pursued by Saldanha and Thompson (2006), Lipson et al. (2003), and Collins and Mittag (2005) might benefit from having a standardized instrument to assess the impact of these instructional interventions on students' understanding of inferential concepts. The teaching experiments use a qualitative methodology aimed to improve instructional approaches specific to the study. This study was not designed to produce generalizable results.

Collins and Mittag's (2005) comparative study was further confounded by limitations related to comparing existing groups. The lack of randomization of treatments to groups introduced potential treatment-settings interference, treatment infidelity, and selection bias. There is insufficient evidence to suggest that any of the instructional interventions studied were effective.

2.2.3.3 Summary and critique of instruments used to assess students' reasoning about *P*-values and statistical significance

The quality of the surveys and questionnaires used in the research reviewed limits the inferences that might be drawn from them. Four of the surveys used six or fewer items, suggesting poor coverage of the content domain. Only two instruments used a potentially reliable number of items ($\geq$ 20 items) (Vallecillos-Jimenez & Holmes, 1994; Mittag & Thompson, 2000).

Some evidence of content validity was provided by five of these studies by including the instrument used in the article (Collins & Mittag, 2005; Falk & Greenbaum, 1995; Haller & Krauss, 2002; Mittag & Thompson, 2000; Oakes, 1986). Inclusion of the items allows readers to informally assess item content. By providing the items used, the researchers also facilitated the replication of their studies. Furthermore, items could also be harvested for other studies.

A limitation to the empirical studies is that no standardized assessment instrument was used to evaluate the curricular improvements. No quantitative assessments were used in the studies conducted by Saldanha and Thompson (2006) or Lipson et al. (2003). Classroom specific assessments (rather than standardized assessments) were used in the Collins and Mittag (2005) study for the response variable and the statistical controls.

However, statistics education researchers have shown that scores from classroom assessments do not correlate with scores obtained from research assessments (Garfield, 1998; Garfield & Chance, 2000; Liu, 1998; Tempelaar, Gijselaers, & Schim van der Loeff, 2006).

2.2.3.4 Summary of what these studies reveal about inferential understanding

These empirical studies suggest that difficulties understanding *P*-values and statistical significance may be common and persistent. Some difficulties may be tied to the subjects under study. The basic literacy of statistical significance was difficult for introductory statistics students to attain (Batanero, 2000; Williams, 1999). Introductory, intermediate-level, and some graduate students struggled with the influence of sample size on statistical significance and differentiating treatment effects from statistically significant results (Haller & Krauss, 2002; Hubbard & Bayarri, 2003; Williams, 1999; Wilkerson & Olson, 1997). Some descriptions of inferential logic were confusing for statistics students and instructors (Haller & Krauss, 2002; Falk & Greenbaum, 1995; Oakes, 1986). Similarly, experienced researchers had some difficulties with recognizing misinterpretations of the *P*-value as the probability of the truth or falsity of hypotheses (Haller & Krauss, 2002; Oakes, 1986).

However, there are problems with drawing inferences from the results of these studies. Most of the studies used an insufficient number of items to sample across the content domain. There was no validation that the items used measured the intended content. Furthermore, none of the studies reported reliability of scores for the instrument used. Only one study employed a random sampling process to allow some generalization to a larger population than those studied (Mittag & Thompson, 2000).

Despite widespread concern with how people reason about statistical inference, none of the measures targeted to assess how students understand and misunderstand *P*-values and statistical significance reported evidence of score reliability or validity of item content. Results suggest that the proposed difficulties cited in these studies should be consolidated into one instrument and reliability and validity evidence gathered to facilitate future research on this topic.

2.2.4 Categorization of Difficulties Understanding *P*-values and Statistical Significance

The 14 difficulties that people seem to have with understanding and interpreting *P*-values and statistical significance were identified in the literature and sorted into four categories in Table 2.1. At the simplest level is misunderstanding statistical significance terminology and basic concepts (B-1, B-2). Next is confusing relationships between inferential concepts (R-1 to R-5). The third group focuses on misapplying the logic of statistical inference (L-1 to L-3). The last group includes misinterpreting the *P*-value as the probability of the truth or falsity of hypotheses (H-1 to H-4).

Table 2.1

Categorization of Difficulties Understanding P-values and Statistical Significance

Category	Difficulties	Reference
	Misunderstanding Basic terminology and concepts	
B-1	Confusion about the basic language and concepts of significance testing	Batanero, 2000 Williams, 1999
B-2	Believing the *P*-value is always low	Williams, 1999
	Confusing Relationships between inferential concepts	
R-1	Confusion between test statistics and *P*-values	Williams, 1999
R-2	Confusion between samples and populations	Mittag & Thompson, 2000 Saldanha & Thompson, 2006 Lipson et al., 2003
R-3	Confusion between significance level, α and Type I error rate and the *P*-value	Haller & Krauss, 2002 Hubbard & Bayarri, 2003 Mittag & Thompson, 2000 Williams, 1999
R-4	Believing *P*-value is independent of sample size	Mittag & Thompson, 2000 Nickerson, 2000 Wilkerson & Olson, 1997
R-5	Believing reliability is *1* – *P*-value	Haller & Krauss, 2002 Mittag & Thompson, 2000 Nickerson, 2000 Oakes, 1986 Daniel, 1998
	Misapplying the Logic of statistical inference	
L-1	Misusing the Boolean logic of inverse ($a \rightarrow b$ confused with not-$b \rightarrow$ not-a) to interpret a hypothesis test (confusion of the inverse)	Batanero, 2000 Falk & Greenbaum, 1995 Oakes, 1986 Nickerson, 2000
L-2	Misusing the Boolean logic of converse ($a \rightarrow b$ replaced with $b \rightarrow a$) (confusion of the converse)	Batanero, 2000
L-3	Thinking the *P*-value is the probability chance *caused* results observed	Daniel, 1998 Kline, 2004

Table 2.1 (continued)

Category	Difficulties	Reference
	Misinterpreting the *P*-value as the probability of the truth or falsity of Hypotheses	
H-1	Interpreting *P*-value as the probability the alternative hypothesis *(H_a)* is true	Falk & Greenbaum, 1995 Haller & Krauss, 2002 Nickerson, 2000 Oakes, 1986
H-2	Interpreting *P*-value as the probability that accepting the alternative hypothesis *(H_a)* is false	Falk & Greenbaum, 1995 Haller & Krauss, 2002 Williams, 1998, 1999
H-3	Interpreting *P*-value as the probability the null hypothesis *(H_0)* is true	Falk & Greenbaum, 1995 Haller & Krauss, 2002 Oakes, 1986
H-4	Interpreting *P*-value as the probability the null hypothesis *(H_0)* is false	Falk & Greenbaum, 1995 Haller & Krauss, 2002 Nickerson, 2000 Oakes, 1986

Note. Each of the difficulty categories are linked to one or more RPASS items later in this paper.

2.3 Need for a Research Instrument to Assess Reasoning about *P*-values and Statistical Significance

A research instrument is needed that that assesses each of the 14 difficulties cited in Table 2.1. This list of difficulties was, therefore, used to design the RPASS test blueprint. During the design and development of a research instrument, it is important to assess the reliability of scores and the validity of inferences that may be drawn from the instrument.

2.3.1 Validity

Validity deals with the inferences drawn from responses to an item or instrument. To examine validity, evidence must be demonstrated that valid inferences can be drawn from item scores. The sources of validity evidence that are most relevant to RPASS test construction are content-related and construct-related validity evidence (Cronbach & Meehl, 1955; Messick, 1995).

Content-related evidence supports making inferences to the desired domain of knowledge. Evidence of content validity includes a test blueprint, the definition of the content domain, expert rater review, and a test of adequate length to sample across the content domain (Cronbach & Meehl, 1955).

Since there is no existing measure of understanding and misunderstanding of *P*-values and statistical significance as defined in Table 2.1, the desired construct, construct-related validity evidence must be gathered (Cronbach & Meehl, 1955). Evidence of construct-related validity includes correlations with related measures corrected for attenuation. For practical interpretations, correlations are corrected for measurement error in the comparison measures. For developmental purposes, it is informative to correct for attenuation in both the comparison measures and the test being evaluated to examine the theoretical correlation without measurement error (Muchinsky, 1996). As evidence of construct validity, convergent correlations should be moderate and positive with discriminant correlations near zero (Messick, 1989, 1995).

2.3.2 Reliability

Reliability of scores informs whether scores are likely to be repeatable. “Reliability describes the extent to which measurements can be depended on to provide consistent, unambiguous information” (Sax, 1997, p. 271). Reliability is interpreted as the proportion of score variation that can be accounted for by true score variance rather than error variance. One type of reliability is internal consistency. Internal consistency provides an estimate of reliability for a single test administration. Conceptually, the average of all possible split half correlations of the instrument with itself produces the commonly used reliability estimate known as Cronbach's coefficient alpha.

2.4 Existing Instruments that Assess Statistical Understanding

Before developing a new instrument for statistics education, it is important to determine whether research instruments exist that validly assess the desired content and produce reliable scores (Scheaffer & Smith, 2007). This section examines whether existing research instruments in statistical education with reported reliability and validity evidence may sufficiently assess this topic. The first subsection describes four assessments developed for assessing understanding of statistics and discusses why these instruments do not meet the current need. The second subsection reviews issues that arise in the interpretation of results from these kinds of statistical assessments. The final subsection summarizes the existing instruments, critiques the instruments relative to their use for this study, and discusses issues in assessing statistical understanding.

2.4.1 Existing Instruments

To answer the call for measures with reported reliability and validity to further research in statistical education (Garfield & Ben-Zvi, 2004; Shaughnessy, 1992), instruments with reported psychometric properties have been developed. These existing instruments include the Statistical Reasoning Assessment (SRA; Garfield, 2003); Assessment Resource Tools for Improving Statistical Thinking (ARTIST) topic scales (delMas, Ooms, Garfield, & Chance, 2006); Statistics Concepts Inventory (SCI; Allen, Stone, Rhoads, & Murphy, 2004); and the Comprehensive Assessment of Outcomes in a first Statistics course, the CAOS test (delMas, Garfield, Ooms, & Chance, in press).

2.4.1.1 Statistical Reasoning Assessment (SRA)

The Statistical Reasoning Assessment was one of the first instruments developed to measure statistical reasoning with reported validity evidence (see Garfield, 2003; Garfield & Chance, 2000). The SRA consisted of 20 multiple-choice items assessing correct conceptions and misconceptions of probability and statistics reasoning. The paper and pencil assessment included five topic areas: reasoning about data, statistical measures, uncertainty, samples, and association. Items were primarily related to probability, and no items were related to formal inference. The test was designed to measure eight correct reasoning skills and eight misconceptions. Expert ratings provided evidence of content validity. However, correlating results to classroom measures did not establish criterion-related validity. Internal consistency reliability was reportedly low. In a US and Taiwan gender comparison study, test-retest reliability was computed for the SRA as .70 and .75 for the misconceptions (Liu, 1998).

2.4.1.2 ARTIST, Assessment Resource Tools for Improving Statistical Thinking Topic Scales

The ARTIST topic scales (delMas et al., 2006) were developed to “assist faculty who teach statistics across various disciplines.” Instructors have used these scales to improve their courses, to assess student reasoning, to review before a test, or to assign extra credit. There are 11 online ARTIST topic scales, consisting of 7-15 multiple-choice items. The 11 topic scales include data collection, data representation, measures of center, measures of spread, normal distribution, probability, bivariate quantitative data, bivariate categorical data, sampling distributions, confidence intervals, and significance tests. The

ARTIST Test of Significance topic scale consists of 10 items. Only four of these items specifically address the *P*-value and statistical significance. The ARTIST Test of Significance topic scale has undergone expert review for content-related validity evidence. Reliability analyses are planned for the instrument but results have not been released (delMas et al., 2006).

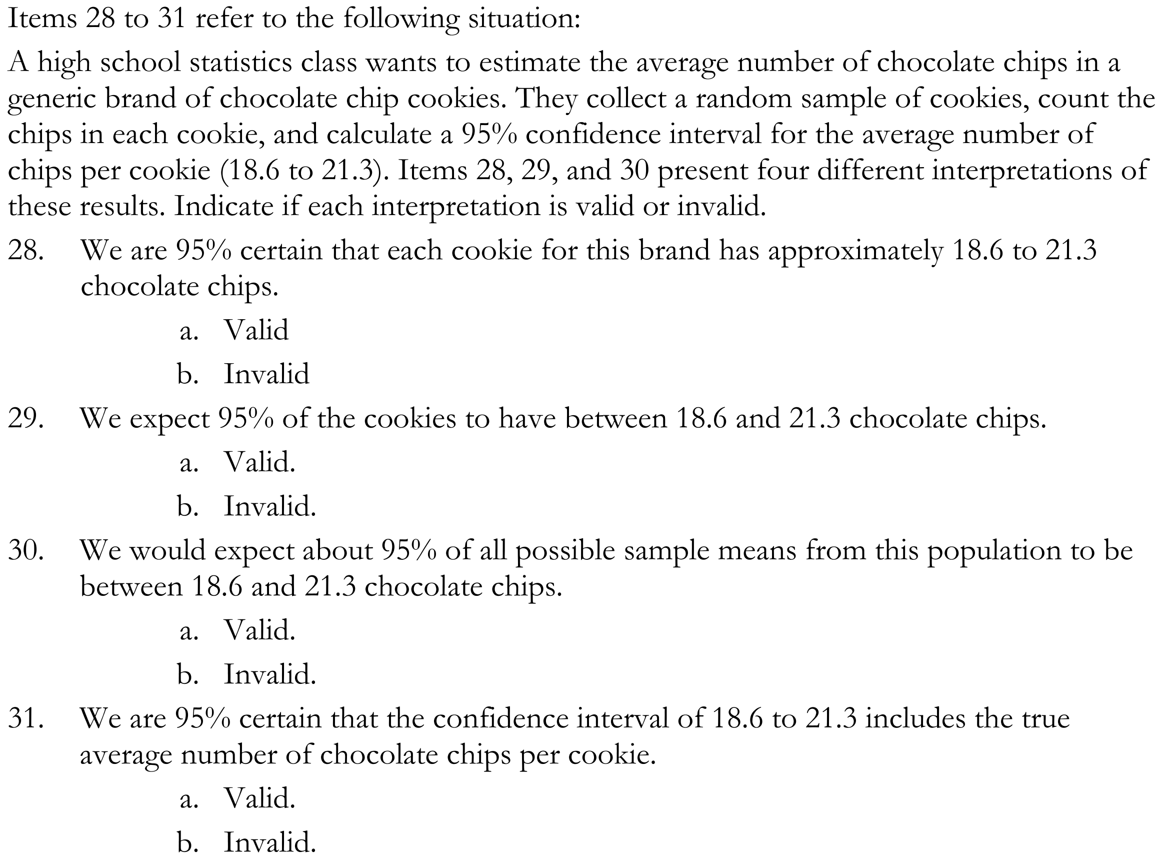
2.4.1.3 CAOS, the Comprehensive Assessment of Outcomes in a First Statistics Course

The ARTIST Comprehensive Assessment of Outcomes in a first Statistics course (CAOS) is a 40-item multiple-choice assessment designed to measure students' conceptual understanding at the end of an introductory course (delMas et al., in press). The validity of the test content was evaluated by statistics educators based on what the experts agreed students should know after completing any course in statistics. The items are designed to assess statistical literacy and reasoning. The reliability of the CAOS posttest was Cronbach's coefficient alpha of .82, with a national sample of 1470 students (delMas et al., in press). The national sample yielded a percent correct of 51.2% of the 40 items, on average, when given at the end of a first course in statistics.

Fourteen of the CAOS items were inference-related. Sixty percent or more of the respondents answered four of the 14 inferential items correctly (Items 19, 23, 31, and 34). For example, most respondents understood that small *P*-values are typically desirable (Item 19). Students also seemed to differentiate between statistical significance and practical significance.

However, interpreting respondents' correct responses to some CAOS items may be confounded by misconceptions. Many respondents correctly answered that a large, simple random sample typically resembles its parent population. However, there is a common misconception that all samples resemble the population from whence they came, regardless of the sample size (Kahneman & Tverksy, 1982; Tversky & Kahneman, 1982). Thus, results may reflect a true understanding of the law of large numbers, a misconception of representativeness, or some combination of the two.

Use of a multiple-true-false item type (Frisbie, 1992) allows students to select multiple and potentially conflicting responses for the same item prompt. DelMas and colleagues (in press) employ this item type in the CAOS test to probe students' interpretations of confidence levels (see Figure 2.1). The four TF items can be used to identify patterns which allows student understanding to be identified. For example, seventy-five percent of respondents identified the correct interpretation of a confidence interval, Item 31. Nevertheless, more than half of the respondents did not identify the two misinterpretations presented in Items 30 and 28 (delMas et al., in press).



Items 28 to 31 refer to the following situation:

A high school statistics class wants to estimate the average number of chocolate chips in a generic brand of chocolate chip cookies. They collect a random sample of cookies, count the chips in each cookie, and calculate a 95% confidence interval for the average number of chips per cookie (18.6 to 21.3). Items 28, 29, and 30 present four different interpretations of these results. Indicate if each interpretation is valid or invalid.

28. We are 95% certain that each cookie for this brand has approximately 18.6 to 21.3 chocolate chips.
    - a. Valid
    - b. Invalid
29. We expect 95% of the cookies to have between 18.6 and 21.3 chocolate chips.
    - a. Valid.
    - b. Invalid.
30. We would expect about 95% of all possible sample means from this population to be between 18.6 and 21.3 chocolate chips.
    - a. Valid.
    - b. Invalid.
31. We are 95% certain that the confidence interval of 18.6 to 21.3 includes the true average number of chocolate chips per cookie.
    - a. Valid.
    - b. Invalid.

Figure 2.1. A multiple-true-false CAOS item assessing confidence level interpretation

National CAOS results also indicated an increase in the percentage of students who held some misunderstandings after instruction. These misconceptions included confusing random assignment with random sampling, believing causation can be inferred from correlation, and believing rejection of a null hypothesis means the null is definitely false. Persistent misunderstandings from pretest to posttest included the need for randomization in an experiment, understanding factors that allow samples to be generalized to a population, and interpreting confidence levels.

2.4.1.4 Statistics Concepts Inventory (SCI)

The Statistics Concepts Inventory (SCI) consists of 32 multiple-choice questions assessing basic concepts and misconceptions in probability and statistics for engineering students (Allen, Stone, Rhoads, & Murphy, 2004). Evidence of content validity was based on results from an engineering faculty survey concerning the probability and statistics they wanted students to know. The SCI inference-related topics include hypothesis testing for means, *P*-values, Type I and Type II errors, and confidence intervals. The items were designed to focus on conceptual understanding rather than computational problem solving ability. Each item had one correct answer.

In engineering student-targeted statistics courses, the SCI scores correlated with final statistics course grades (summer 2003: $r = .60$, fall 2003: $r \leq .43$). However, when the SCI was administered in mathematics student-targeted courses, SCI scores did not correlate with final course grades (summer 2003: $r = -.023$, fall 2003: $r = -.054$). Reliability was reported using Cronbach's coefficient alpha based on scores from summer and fall of 2003. Pretest reliability ranged from .57 to .69, and posttest reliability ranged from .58 to .86, depending on groups assessed. Item discrimination was measured with Ferguson's delta, where 1 indicates that all scores are unique, suggesting there would be large variation in scores. All groups measured from fall 2002, summer 2003, and fall 2003 yielded Ferguson's delta above .92.

A possible strength of the SCI is that item content was based on an outline from the Advanced Placement (AP) Statistics curriculum. The test items were culled from unspecified textbooks and statistics journals. Item content was revised based on item analyses and feedback from engineering student focus groups. However, instrument content was focused on how statistics is taught to engineering students at this one university (Allen, Stone, Rhoads, & Murphy, 2004). In addition, the content–validity evidence was based on a College of Engineering survey of faculty from this one institution (*n* = 23). No discriminant validity evidence was reported. Without discriminant validity evidence, the instrument may measure some other ability trait (e.g., engineering ability).

2.4.1.5 Why these Assessments Do Not Meet the Current Need

None of these existing instruments address all the content identified in Table 2.1. The SRA does not include items on *P*-values and inference. Moreover, there is no reliability or validity evidence that the 14 inference-related CAOS items can be used as an isolated subscale. The current set of 10 ARTIST Test of Significance topic scale items has not been determined to be a reliable scale. SCI was designed to address the needs of engineering students and has not been validated with other disciplines or institutions. Thus, despite the widespread concern with how people understand and reason about statistical inference, there are no instruments with psychometric evidence of validity and reliability targeted to assess how students understand and misunderstand *P*-values and statistical significance to further research on this topic.

2.4.2 Issues in Assessing Statistical Understanding

In a 1995 article, Konold summarized 15 years of research in assessing misconceptions in probability and statistics. He cited three major findings. First, prior to instruction, students have theories or intuitions at odds with conventional thinking. Second, prior intuitions are difficult to alter. Third, altering prior conceptions is further complicated since students may reason from multiple and contradictory perspectives. Each of these issues is discussed in light of the challenges they present to assessing understanding of *P*-values and statistical significance.

2.4.2.1 Interfering, Persistent, and Contradictory Prior Intuitions

As was suggested by the national CAOS results, broader misconceptions about probability and statistics may interfere with assessing inferential understanding. As an example, Tversky and Kahneman (1982) enumerated four difficulties people may have with statistical significance if they reason using a “law of small numbers” heuristic. People may gamble the research hypothesis on small samples, become overly confident in early trends, have high expectations in the replicability of significant results, and find a causal “explanation” rather than statistical explanations for any inconsistency between sample results and expected population parameters.

Konold (1995) illustrated how a slight alteration in item wording for two similar items can invoke different reasoning approaches. In one case, students used an outcome-oriented approach to answer the item. That is, the students made a decision as if they were predicting the outcome of a single trial rather than an expected outcome over a series of trials. By a slight alteration in the item wording, the representativeness heuristic was employed (i.e., believing all samples reflect population characteristics). Research suggests that students simultaneously maintain different and, in some cases, contradictory reasoning schemata when encountering statistical test items (see delMas & Bart, 1989; Konold, 1995). Contradictory reasoning strategies appeared to be employed in reasoning about confidence levels based on the CAOS results (delMas et al., in press). An assessment of understanding of *P*-values and inference should allow students to give multiple answers to the same item prompt to potentially expose contradictory reasoning.

2.4.2.2 Patterns in Responses Across a Set of Items are Needed to Better Interpret Understanding

Individual item responses are not sufficient for interpreting students' understanding. It is important that researchers analyze patterns of responses rather than individual responses to better interpret the correct understandings, misconceptions, and heuristics students may use to answer statistical questions and problems (Garfield, 1998; Konold, 1995). Thus, it may be necessary to group RPASS item results in different ways to assess the patterns in student understanding and misunderstanding.

2.4.3 Summary and Critique of Existing Instruments and Issues

Appendix A, Table A3 presents a summary of the reliability and validity evidence for existing research instruments assessing statistical understanding. None of the instruments reviewed meet current needs. None of the instruments reported reliability or validity evidence specifically related to measuring students' understanding and misunderstanding of *P*-values and statistical significance. None of the instruments measured all 14 difficulties cited in Table 2.1. The SRA excluded *P*-values and inference from the content domain. The ARTIST Test of Significance topic scale had too few items ($n = 10$) to make reliability likely and only focused on a subset of the misconceptions cited. The SCI's content-related evidence was not gathered across disciplines. There is no content-related validity evidence to suggest the CAOS inferential items can be used as a subscale.

This leads to the need to develop and validate a new research instrument to assess all the cited difficulties and educators' desired understandings. In addition, the instrument must be designed to address known issues in assessing students' understanding of statistics. Thus, the RPASS must be designed to capture interfering, persistent, and contradictory prior intuitions and misconceptions about inference. Furthermore, the instrument analysis must look at patterns in responses across sets of items to facilitate better interpretation of results.

2.5 Formulation of the Problem Statement and Research Questions

The literature review suggests a need to develop and validate a new instrument that assesses all 14 difficulties and the desired correct understandings about *P*-values and statistical significance. Therefore, this dissertation addressed the following questions:

Research question 1: *Can a research instrument be developed that is a sufficiently reliable and valid measure of students' understanding of and difficulties with reasoning about P-values and statistical significance?*

Research question 2: *What does the proposed instrument indicate about students' understanding of and reasoning about P-values and statistical significance?*

The next chapter describes the qualitative and quantitative methods employed to develop and test the RPASS and to gather evidence of validity and reliability.

CHAPTER 3: METHODS

The review of the literature in the previous chapter suggested a need to develop a new assessment instrument, the Reasoning about *P*-values and Statistical Significance (RPASS) scale with reported psychometric properties. The purpose of the test is to facilitate statistics education research on students' conceptual understanding and misunderstanding of statistical inference and the effect of instructional approaches on this understanding. This chapter describes the methods used to develop the RPASS, to evaluate reliability and validity evidence, and to examine baseline data results.

3.1 Overview of the Study

3.1.1 Participants and Settings

The initial version of the RPASS was piloted in four University of Minnesota (U of M) statistics courses for non-majors at the end of fall semester of 2004 ($N = 333$). Five statistics education advisors from four universities were consulted to inform instrument development. Introductory statistics students from California Polytechnic State University (Cal Poly) participated in field tests and student interviews ($n = 61$). Ten expert raters from four colleges and universities evaluated the validity of the instrument content.

Baseline data were gathered at Cal Poly across five undergraduate, introductory statistics courses for non-majors at the end of the spring quarter of 2006 ($N = 224$). Participant descriptions are detailed within each of the five study phases presented in this chapter.

3.1.2 Study Phases and Timeline

Phases I through V of the study were designed to answer the first research question: *Can a research instrument be developed that is a sufficiently reliable and valid measure of students' understanding of and difficulties with reasoning about P-values and statistical significance?* Data collected in Phase IV were used to examine evidence of reliability and validity. These data were also analyzed to examine the second research question: *What does the proposed instrument indicate about students' understanding of and reasoning about P-values and statistical significance?* In Phase V the baseline data were re-analyzed to explore improving reliability and validity.

Phase I. An initial literature review was conducted in the spring of 2004. The preliminary test blueprint and RPASS-1 were developed based on this literature. RPASS-1 was piloted at the University of Minnesota in the fall of 2004.

Phase II. During 2005, RPASS-1 was modified to reflect the ongoing literature review, pilot results, and input from five statistics education advisors from four research universities. RPASS-2 was produced.

Phase III. Feedback on RPASS-2 was obtained from two field tests and 13 student interviews during winter quarter at California Polytechnic State University (Cal Poly). Instrument modifications produced RPASS-3A. During spring quarter, the RPASS-3A content was rated, modified, and re-rated by 10 subject matter experts from four colleges and universities to produce RPASS-3B and -3C. Deleting one item with redundant content produced the 27-item RPASS-4.

Phase IV. Baseline data were collected at Cal Poly across five introductory courses at the end of spring quarter 2006 to estimate RPASS-4 reliability. Two additional instruments were administered in two of the five courses. One instrument was used to evaluate convergent validity evidence, and the other was administered to evaluate discriminant validity evidence.

Phase V. An item analysis was conducted in Phase V using baseline data. An optimal subset of items (RPASS-5) was identified. Reliability and construct-related validity evidence were estimated for RPASS-5.

Table 3.1 summarizes the methods and RPASS versions by phase of the study.

Table 3.1

Overview of RPASS Development, Validation, and Baseline Data Collection by Phase

	Instrument development and content validation				Data collection and analysis	
	Phase I	Phase II	Phase III		Phase IV	Phase V
	2004	2005	winter 2006	spring 2006	spring 2006	fall 2006
Methods	Conduct literature review	Continue literature review	Administer field test	Review with expert raters	Collect baseline data	Conduct item analysis
	Develop blueprint	Consult advisors	Interview students	Modify instrument	Evaluate item results	
	Pilot instrument	Modify instrument	Modify and re-administer instrument	Validate content with expert raters	Evaluate reliability and validity	Estimate reliability and validity
Version	RPASS-1	RPASS-2	RPASS-3	RPASS-4	RPASS-4	RPASS-5

3.1.3 Instruments Used to Evaluate RPASS Construct Validity

In addition to administering the RPASS in Phase IV, two additional instruments were administered to gather data to examine construct validity. Fifty-six of the RPASS baseline participants completed the additional instruments. This section describes these instruments and the timeline for their administration.

Since no criterion measure existed that measured all the difficulties identified in Table 2.1, an instrument measuring similar content was constructed to evaluate evidence of convergent validity. A second instrument, measuring dissimilar content, was selected to evaluate evidence of discriminant validity. These measures were selected from the ARTIST (Assessment Resource Tools for Improving Statistical Thinking) website at https://app.gen.umn.edu/artist/ (delMas, et al., 2006).

To gather evidence of convergent validity, a five-part open-ended item was selected from the ARTIST website. This open-ended item, the rating scale used to grade the item responses, and the answer key used by the raters appear in Appendix B.1. The 14-item ARTIST Bivariate Quantitative Data topic scale selected to collect discriminant validity evidence appears in Appendix B.2.

The timeline for administering the instruments to examine construct-related validity is summarized in Table 3.2. The Bivariate Quantitative Data topic scale was administered online during week 9 of the 10-week term. The ARTIST open-ended item was administered as a paper and pencil test concurrent with the RPASS-4 online administration during finals week. The remainder of this chapter details participants and procedures by phase of the study.

Table 3.2

Timeline for Administering Instruments to Examine Construct Validity

	Week of the 10-week spring quarter[a]	
	Week 9	Finals week
Validity evidence	Discriminant	Convergent
Instruments administered	14-item ARTIST Bivariate Quantitative Data topic scale	5-part ARTIST open-ended item
	—	RPASS-4

Note. [a]Spring quarter 2006.

3.2 Instrument Development and Content Validation

3.2.1 Phase I. Test Blueprint Development and RPASS-1 Pilot

3.2.1.1 Development of the Preliminary Test Blueprint

Developing a test blueprint was the first step in defining the construct, conceptions, and misconceptions of *P*-values and statistical significance. *The Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999) number 3.2 and 3.3 specify that "the definition of the domain, and the test specification should be stated clearly" for a test under construction (p. 43). A test blueprint was developed to specify the content areas to be assessed. When mapped to test items, test blueprints provide content-related validity evidence (*The Standards*, 1.6). The preliminary test blueprint was based on the difficulties culled from the research literature that were categorized in Table 2.1.

3.2.1.2 RPASS-1A Instrument Development

The 16-item RPASS-1A instrument was initially developed per the preliminary item blueprint. RPASS-1A items originated from four multiple-choice (MC) items from the ARTIST Test of Significance topic scale (delMas et al., 2006). These four items were selected because they directly addressed *P*-values and statistical significance and could be linked to some of the difficulties identified in the literature review. The items appear in Appendix B.3. Each MC item was altered to create a set of multiple-true-false (MTF) item sets (Downing, 1992; Frisbie, 1992; Frisbie & Sweeney, 1982). True-false options were modified or added so each of the misconceptions identified in the literature review was assessed. The MC item stems were embellished or rewritten to create a more fully developed context or problem scenario. Established item writing guidelines were followed (Haladyna, Downing, & Rodriguez, 2002).

Measurement research related to item formats suggests MTF item sets may improve reliability and validity compared to equivalent MC test items (Frisbie, 1992; Frisbie & Sweeney, 1982). MTF item sets assess more content in the same period of time. RPASS-1A scores could vary from 0-16. The comparable multiple choice questions would have been scored from 0-4. When a sufficient number of items is used, TF test results are not adversely affected by guessing effects (Ebel, 1970). Furthermore, the MTF format allows students to concurrently select both correct and incorrect options for the same item prompt, potentially exposing contradictory conceptions.

Some research suggests confidence weightings can improve reliability (Ebel, 1965). After each of the RPASS-1 TF items, students were asked to rate their level of confidence in their answer on a scale from *0-100%* confident. Results were analyzed to assess if confidence weightings might explain variation in scores.

3.2.1.3 Feedback from Statistics Education Advisors

Participants

RPASS-1A was reviewed by five statistics education professionals: two from the University of Minnesota, one from the University of California at Los Angeles, one from Ohio State University, and one from California Polytechnic State University.

Procedures

To improve content validity and reduce measurement error, the 16-item RPASS-1A and a listing of the targeted concept or difficulty by item (the preliminary test blueprint) were emailed to the statistics education advisors for their review and feedback. Suggestions from these email exchanges were summarized. RPASS-1A was revised based on feedback from these advisors to produce the 17-item RPASS-1B.

3.2.1.4 Pilot of RPASS-1B

Participants

Eight instructors of four University of Minnesota (U of M) statistics courses were invited to have their students participate in the RPASS-1B instrument pilot. Both undergraduate and graduate students participated. The statistics courses were targeted for social science majors. Two undergraduate courses and the master's course were first courses in statistics. The doctoral level course was a second course in statistics. Of the 423 students invited, 333 completed the instrument (78.7%). Three hundred respondents provided demographic data as displayed in Table 3.3. The rows identify respondents by their class standing. Columns break out respondents by the type of statistics course taken.

Table 3.3

Number of RPASS-1 Pilot Respondents by Statistics Course

	Statistics course				
	Undergraduates		Graduate students		
Respondent class standing	Lower division	Upper division	Master's first	Doctoral second	Total
Freshmen or sophomore	82	8	2	0	92
Junior or senior	9	54	22	0	85
Master's	0	0	29	3	32
Doctoral	0	0	33	47	80
Non-degree	2	1	7	1	11
Not specified	10	2	15	6	33
Total	102	65	108	57	333

Procedures

Permission was obtained from the U of M Institutional Review Board (IRB) to pilot the RPASS-1B instrument in the fall of 2004. Following recommendations of the Tailored Design Method and social exchange theory (Dillman, 2000), a preliminary letter and link were sent to students via their course instructors. A completion reminder and follow-up thank you note were sent to students after the first week of a two-week administration period. A consent form was integrated into the online instrument to allow students to opt out of participation.

3.2.2 Phase II. Instrument Development from RPASS-1 to RPASS-2

Results from the RPASS-1 pilot informed further item modifications. The rationale for the addition, alteration, and removal of items was documented and reported. This new set of items was reviewed by the five statistics education advisors (described in Section 3.2.1.3) for editing and feedback. At the end of this process, the 25-item RPASS-2 was produced as it appears in Chapter 4.

In addition to understanding difficulties people have, it was important to include items that assess what students should know. The instructional outcomes from the Tools for Teaching and Assessing Statistical Inference website (Garfield et al., 2005) were mapped into outcomes of statistical literacy, reasoning, and thinking. The website grouping of instructional outcomes entitled the "*P*-value is a probability statement" was mapped onto statistical literacy. The "interpretation" and "relationships" groupings were mapped onto statistical reasoning, and the "cautions and limitations" grouping was mapped onto all three areas: statistical literacy, reasoning, and thinking. Based on this mapping, the learning goals for *P*-values and statistical significance were constructed for this project, and targeted learning goals were added to the test blueprint. The revised blueprint, reported in Chapter 4, guided further item selection and development.

3.2.3 Phase III. Instrument Development from RPASS-2 to RPASS-4

Permission was obtained from the U of M Institutional Review Board for Phases III through V. Phase III used instrument development procedures designed to reduce measurement error. RPASS-2 item modifications based on student feedback from field tests and interviews produced RPASS-3A. Two rounds of expert reviews of RPASS-3A produced RPASS-3B and RPASS-3C, respectively. Removal of one redundant item produced RPASS-4.

3.2.3.1 Administration of RPASS-2

Participants

The RPASS-2 was administered during the winter quarter of 2006 at Cal Poly. Students were recruited from three sections of an introductory statistics course for liberal arts (LibStat) students. Thirty-six students volunteered to take the RPASS-2. Twenty-two of the 36 participants were freshmen, with eight sophomores and four juniors. Of the two remaining students, one was a senior and the other was in a master's program. The 32 students who completed all the answers on the test comprised the sample.

Procedures

The RPASS-2 field test was administered in a computer lab on a drop-in basis during week 8 of a 10-week term, after the completion of a unit on inference. A link to the online test was sent to students via email from the instructor. As backup for those who hadn't received the link, the URL was posted in the lab. The test was accessible during lab times only. The online instrument was preceded by a consent form to ensure respondents were aware that results would be used for research purposes (see Appendix C.1). At the end of the consent form, students could opt to exit from the test. No students chose to do so.

The new scenario, items, and instructions in RPASS-2 were administered at Cal Poly. Comments were solicited from respondents to improve the clarity of the test instructions. Notes were made concerning any difficulties with the administration process and procedures.

Following the RPASS-2 field test, students were invited to participate in cognitive “think aloud” interviews or in-depth interviews. All test respondents and interviewees received two extra credit percentage points for participating. Feedback from the RPASS-2 field test and the subsequent interviews was used to produce RPASS-3A.

3.2.3.2 In-depth Interviews using RPASS-2

Participants

Five students who participated in the RPASS-2 field test volunteered to participate in follow-up in-depth interviews. The in-depth interviews were conducted during the week following the RPASS-2 field test, week 9 of the 10-week term. Interview participants were initially selected on a first come, first served basis. Three students volunteered for in-depth interviews. Two additional interviewees were purposefully selected based on the students’ perceived willingness to talk freely about their ideas and to obtain a cross section of ability levels in the interviews (Creswell, 1998).

Procedures

An interview protocol (Appendix C.2) was written to probe the strategies students employed when answering the RPASS-2 items. Students were provided a hardcopy version of the RPASS-2 section(s) being discussed. The objective was to assess student strategies for answering the items. Attempts were made to make the interview conversation as natural as possible. Each participant provided feedback on up to three of the five RPASS problem scenarios and related questions. Modifications were made to the RPASS based on feedback from the five interviews.

3.2.3.3 Cognitive Interviews Using RPASS-2 and RPASS-3A

Participants

Eight students who had not participated in the RPASS-2 field test were invited to participate in individual cognitive interviews. These students were from the same three sections of the LibStat course who participated in the RPASS-2 field test. Cognitive interviews were also held during week 9 of the 10-week term. Three interviews were conducted on March 6-7, 2006, using RPASS-2. Five interviews were conducted on March 8-9, 2006. Information from the cognitive interviews was used to modify online instructions or item wording and create RPASS-3A.

Procedures

The cognitive interviewees took the online RPASS test individually, reading the instructions, problem scenarios, and items aloud. The student-computer interaction was videotaped. These "think aloud" sessions were used to note students' difficulties with understanding online instructions or item wording. A digital video camera recorded the computer screen over the student's shoulder and recorded students' comments. The researcher encouraged each student to keep talking and probed students' understanding of instructions or item wording. Participants' difficulties were used to modify the online instructions. Cognitive interviews were conducted until no more substantive changes were needed based on subsequent participant comments. After the recommended instrument modifications had been made, a new group of students from the same LibStat course sections were solicited to participate in the field test of RPASS-3A.

3.2.3.4 Administration of RPASS-3A

Participants

Twenty students were recruited from the same three sections of the LibStat course to field test RPASS-3A. Only those students who did not participate in the RPASS-2 field test or in student interviews were eligible to participate. Twenty-five students initially volunteered. Twenty students completed all items and were included in the sample. Twelve of the 20 participants were freshmen, with two sophomores, five juniors, and one senior.

Procedures

The field test of RPASS-3A focused on whether students were able to independently read and follow the online procedures and instructions after the suggested modifications. Student volunteers were sent an email invitation and an internet link to access the RPASS outside of class or lab during finals week. Students were encouraged to contact the researcher if they had any difficulties understanding the items or the online procedures. This administration was not monitored.

3.2.3.5 Content Validation and Instrument Modifications: RPASS-3A to RPASS-4

Participants

All ten expert raters recruited to validate the RPASS content had obtained a doctorate in statistics or a closely related field. The raters were seven instructors of statistics from California Polytechnic State University, one from University of California at Los Angeles, one from Ohio State University, and one from Meredith College in North Carolina. The letter of invitation and instructions for raters appear in Appendix C.3 and C.4, respectively. Ten experts completed the validation process. As description of expert qualifications (*The Standards* 1.7), three experts were introductory statistics textbook authors. One of the raters was a prior chief reader for the Advanced Placement (AP) Statistics Examination, four were AP readers, and one was a test development committee member for the AP Statistics Exam. Experts also included a prior editor of the *Journal for Statistics Education,* an American Statistical Association (ASA) Waller Education Award winner, three ASA Founders Award winners, a Fellow of the ASA, and a president of the International Association for Statistical Education.

Procedures

Content validity assesses the extent to which a domain is described by the instrument content (Messick, 1989). Content-relevant evidence included restricting RPASS item selection to the test blueprint and obtaining content validity ratings from subject matter experts (*The Standards* 1.7). The feedback from the first round of ratings was prioritized. Experts' recommended changes were sorted from worst to best item. Items were added, modified, or deleted based on round one feedback to produce RPASS-3B. Individual meetings were held with each expert rater to review changes and receive a second round of ratings for problem items. After the second round of ratings and modifications, RPASS-3C was produced.

The correct conception and misconception items were rated as two separate subscales. Each rater was provided a three-column expert rater packet where each row consisted of an item, the learning objective or misconception measured by the item, and the rating scale. The 4-point rating scale ranged from 1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, to 4 = *strongly agree*. The aggregated expert ratings were judged to indicate a sufficient level of content validity if any item, or subset of items, had a mean rating of 3 or higher. When a content expert gave an item a rating less than 3, a discussion with the particular rater was scheduled to understand whether the item might need modification or possibly might need to be eliminated. The content experts used the same rating scale and procedures to judge the content validity of the two subscales of the RPASS: correct conceptions and misconceptions. Finally, the experts commented on whether the RPASS included extraneous or missing content. Experts were consulted to modify items or write new items to assess missing content.

Expert validity ratings for each of the RPASS subscales were used to assess structural validity (Messick, 1995). To obtain the final set of items, any redundant items were removed from the scale. The final set of 27 non-redundant items was denoted RPASS-4.

3.3 Baseline Data Collection and Analysis

3.3.1 Phase IV. Evaluation of RPASS-4 Baseline Scores, Reliability, and Validity

3.3.1.1 Administration of RPASS-4 Instrument

Participants

Two hundred ninety-six students were invited to participate in the RPASS administration. Two hundred twenty-four respondents completed all the items and were included in the sample. Students were tested either in week 10 of the quarter (*n* = 118) or during finals week (*n* = 106). The 118 students tested during week 10 were either from a course for science majors (SciStat) or a course for business majors (BusStat). The 106 students tested during finals week were from a course for mathematics majors (MathStat), agriculture majors (AgStat), or a general liberal arts course (LibStat).

Per *The Standards* 4.6, the RPASS respondents' demographics are reported in Tables 3.4 and 3.5. The rows of Table 3.4 list the college majors of the 224 respondents included in the sample. The rows of Table 3.5 list the class standing of the respondents included in the sample, with the introductory course taken indicated by the columns.

Table 3.4

Number of RPASS-4 Respondents by College Major and Course

	Baseline respondents by course					
	Week 10		Finals week			
College where respondent majors	BusStat	SciStat	LibStat	AgStat	MathStat	Total
Architecture and environmental design	10	0	0	3	0	87
Agriculture	3	36	6	42	0	13
Business	22	1	1	1	0	25
Engineering	0	1	0	0	0	1
Liberal arts	5	0	21	7	0	33
Science and math	4	35	7	2	13	61
Did not specify	1	0	0	2	1	4
Completed/Invited	45/67	73/108	35/43	57/64	14/14	224/296
Participation rate	67%	68%	81%	89%	100%	76%

Note. BusStat = Statistics for business students, SciStat = Statistics for science students, LibStat = Statistics for liberal arts students, AgStat = Probability and statistics for agriculture students, MathStat = Statistics for mathematics students.

Table 3.5

Number of RPASS-4 Respondents by Class Standing and Course

	Baseline respondents by course					
	Week 10		Finals week			
Respondent class standing	BusStat	SciStat	LibStat	AgStat	MathStat	Total
Freshman	24	21	19	13	2	79
Sophomore	5	27	6	15	2	55
Junior	12	19	6	19	5	61
Senior	3	5	3	8	4	23
Other	0	1	0	0	0	1
Not specified	1	0	1	2	1	5
Total	45	73	35	57	14	224

Procedures

The RPASS-4 instrument was administered in a computer lab with 24 computers. Three of the courses participating in this study were held in this lab. These students took the RPASS as part of their final exam. The two remaining courses were tested on a drop-in basis during specified lab hours over four days. Course instructors forwarded an email invitation to students in the SciStat and BusStat courses to participate in RPASS-4 data collection the last week of the quarter (week 10).

Participation incentives varied across the five courses. The SciStat instructor offered students six points for showing up, plus additional points for correct answers. The BusStat instructor awarded extra credit for participation. No incentives were offered in the other three courses since RPASS was administered as part of the final. Week 10 respondents were given written procedures for logging-in on the 24 lab computers (see Appendix C.5). All RPASS specific instructions were presented online.

Finals week respondents were administered the RPASS during their scheduled exam hour by their course instructor. All internet protocol (IP) addresses, completion dates, and times were consistent with administration times and locations. There was no apparent intrusion by uninvited respondents.

The mean proportion of correct responses was calculated for various subsets of items. Three different divisions of the items were used to create item subsets. The first divided items into those that measured correct conceptions and those that measured misconceptions. The second divided the items into four different content areas identified in the item blueprint (basic literacy, relationships between concepts, logic of inference, and belief in the truth or falsity of hypotheses). The third divided items according to the three learning outcomes (statistical literacy, statistical reasoning, and statistical thinking). For each subset of items, the mean proportion correct (mean $\hat{p}$) was computed by first computing for each student the proportion of correct responses for a specified subset of items, and then computing the average of these proportions across students. The mean proportion of correct responses across the three item groupings and the number of items per group were reported.

3.3.1.2 RPASS-4 Reliability

Participants

Two hundred twenty-four students who completed RPASS-4 comprised the sample.

Procedures

Score variation, item discrimination, item difficulty, and the number of items all impact score reliability for a particular group (see Cronbach, 1951; Ebel, 1967; Haertel, 2006). A reliability item analysis was conducted to report the reliability of the 27-item instrument (*N* = 224). Statistics reported include the proportion of correct responses (item difficulty), standard deviation of item difficulty, corrected item-total correlation, and the value for coefficient alpha if a particular item were deleted. The corrected item-total correlation is the point-biserial correlation (r_{pb}) with the total score corrected by subtracting the contribution for the item being assessed. Items that contribute to higher reliability of the total score should be positive and should correlate at .15 or higher for a dichotomous scale (i.e., corrected $r_{\text{pb}} \geq .15$).

3.3.1.3 RPASS-4 Validity Evidence

Participants

Convergent and discriminant validity evidence was collected in two of the five courses included in the RPASS-4 administration. Out of the 224 students who completed the RPASS-4 items, 56 students (37 AgStat and 19 LibStat) took both of the comparison instruments. These 56 students comprised this sample.

Procedures

The procedures used to collect, compute, and report convergent and discriminant validity evidence are described. Descriptive statistics were reported for each of the comparison measures (mean, standard deviation, median, and interquartile range).

Convergent validity evidence was collected using an open-ended item from the ARTIST website assessment builder tools database. The item assessed students' ability to interpret test of significance output and draw conclusions based on statistical results. The item was administered via paper and pencil. When this concurrent validity evidence was gathered by administering the open-ended item along with administration of the RPASS, approximately half of the respondents started with the paper and pencil open-ended item, and half started by taking the online RPASS. Students were assigned to a written or online start based on whether they sat to the right or left of the instructor. The grading rubric, the open-ended item, and the answer key appear in Appendix B.1 (per *The Standards* 3.22).

Two raters with previous experience as Advanced Placement Statistics Readers used a holistic, four-point rubric to independently score the item. The researcher was one of the two raters. The second rater was external to the project. Level of rater agreement was computed as the number of equal ratings divided by the total number of cases rated. The proportion of rater agreement reported was obtained before the raters discussed their discrepant item ratings. After this discussion and recalibration between raters, no ratings remained more than one point apart. For the ratings that remained one point apart, the mean of the two ratings was used to compute concurrent validity correlations. Ideally, the correlation between the open-item response ratings and students' RPASS scores would be at least moderate and positive.

Discriminant validity evidence was computed by correlating scores from the Bivariate Quantitative Data topic scale with RPASS-4 scores. Correlations with the Bivariate Quantitative Data topic scale were corrected for attenuation in the comparison measure. Both the bivariate scale and RPASS-4 were administered online to AgStat and LibStat students. A low to moderate correlation between RPASS-4 and the bivariate scale was expected because of the similarity of the testing methods and effects related to general intelligence that make the correlation appear higher.

After verifying necessary conditions, the RPASS-4 validity coefficients were presented in a correlation matrix. Correlations were structured similar to the multitrait-multimethod matrix (MTMM) (Campbell & Fiske, 1959). Validity correlation coefficients were presented as off-diagonal elements and reliabilities (or proportion of rater agreement for the open-ended item) as on-diagonal elements.

Validity coefficients were reported as uncorrected Pearson product moment correlations and as correlations corrected for attenuation related to the discriminant measure or RPASS-4 (per *The Standards* 1.18). Correlations were corrected for attenuation in the RPASS-4 for development purposes. Twice-corrected correlations estimate the theoretical correlations between the true scores for these instruments (i.e., scores without measurement error) and suggest whether improving reliability would also improve RPASS-4 correlations for construct validity evidence.

3.3.2 Phase V. Estimation of RPASS-5 Reliability and Validity Data

3.3.2.1 RPASS-5 Reliability

Participants

Responses from the 224 participants who took the RPASS-4 in Phase IV were re-analyzed in Phase V to identify an optimal subset of items, RPASS-5.

Procedures

An additional item analysis was conducted to identify a subset of RPASS-4 items that may have high internal consistency reliability when administered to a new sample.

The iterative process included removing the item with the lowest corrected item-total correlation, computing coefficient alpha, and then reassessing the correlations. At the end of this iterative process, all remaining items had corrected item-total correlations of .15 or higher. The resulting set of items was designated RPASS-5.

3.3.2.2 RPASS-5 Validity Evidence

Participants

The 56 participants who comprised this sample completed the RPASS-4 and the two additional instruments used to assess construct validity.

Procedures

Additional item analyses were conducted to estimate convergent and discriminant validity evidence using the RPASS-5 subset of items. Validity coefficients for RPASS-5 were summarized in a correlation matrix as in Phase IV. Validity coefficients were reported as uncorrected Pearson product moment correlations and as correlations corrected for attenuation related to the discriminant measure or RPASS-4.

3.4 Chapter Summary

The methods for the development and validation of the RPASS were chosen to facilitate using the instrument for research purposes. Methods were chosen to reduce measurement error and to examine content and construct validity. To ensure coverage of the content domain, a test blueprint was developed based on the literature to include content assessing both correct conceptions and misconceptions about this topic. To reduce measurement error, the instrument was modified based on student input from pilots and interviews. To improve item content, experts reviewed the instrument items and rated item content and the two subscales. To examine construct validity and reliability of scores, the instrument was administered to a large, interdisciplinary sample. Results from employing these methods are discussed in the next chapter.

CHAPTER 4: RESULTS AND ANALYSIS

This chapter reports the study results for each of the five phases as described in Chapter 3. First, results are examined from the instrument development and content validation phases (Phases I through III). Second, results are examined from the baseline data collection (Phases IV and V).

Results from Phases I through V provide evidence to examine the first research question: *Can a research instrument be developed that is a sufficiently reliable and valid measure of students' understanding of and difficulties with reasoning about P-values and statistical significance?* Baseline data results gathered in Phase IV are analyzed to address the second research question: *What does the proposed instrument indicate about students' understanding of and reasoning about P-values and statistical significance?*

4.1 Examining Results from Instrument Development and Content Validation

4.1.1 Phase I. Test Blueprint Development and RPASS-1 Pilot

4.1.1.1 Development of the Preliminary Test Blueprint

RPASS-1 test blueprint was developed based on difficulties identified in Table 2.1. The left columns in Table 4.1 list the 13 misconceptions or difficulties and five correct conceptions targeted. Like Table 2.1, item content was grouped by four content areas: basic literacy of statistical significance (B-1 to B-2), relationships between inferential concepts (R-1 to R-5), the logic of statistical inference (L-1 to L-3), and misinterpretations of *P*-value as the truth or falsity of hypotheses (H-1 to H-4).

Table 4.1

Preliminary Test Blueprint

Content areas	Blueprint category	Correct conceptions (C) or misconceptions (M)
Basic literacy		
Textbook definition	B-1	C
Simulation definition	B-1	C
Lay definition	B-1	C
P-value embedded in sampling variation	B-1	C
P-value as rareness measure	B-1	C
P-value as always low	B-2	M
Relationships between concepts		
Test statistic and *P*-value	R-1	M
Sample and population	R-2	M
Type I / α and *P*-value	R-3	M
Sample size and significance	R-4	M
Reliability and *P*-value	R-5	M
Logic of inference		
Inverse as true	L-1	M
Converse as true	L-2	M
Chance as cause of results observed	L-3	M
Belief in the truth or falsity of hypotheses		
Probability: alternative is true	H-1	M
Probability: alternative is false	H-2	M
Probability: null is true	H-3	M
Probability: null is false	H-4	M

4.1.1.2 RPASS-1A Instrument Development

A 16-item instrument (RPASS-1A) was developed based on the preliminary test blueprint. RPASS-1A included four problem scenarios. Four true-false or valid-invalid item sets were associated with each scenario. Eleven items assessed misconceptions or conceptual difficulties, and five items assessed correct conceptions. RPASS-1A also included a confidence rating scale for each item. Respondents were asked to rate their confidence in their answer by entering a number between 0 and 100, where 0 meant *no confidence* and 100 meant *completely confident*. RPASS-1A did not include items to assess confusion of samples with populations (R-2) and the misinterpretation of the *P*-value as the probability that the null is false (H-4).

4.1.1.3 Feedback from Statistics Education Advisors

Five statistics education advisors provided feedback on the RPASS-1A. Suggestions were made for altering each item. Advisors' recommendations were implemented to produce RPASS-1B as it appears in Appendix D.1. Changes implemented include the following:

1. Reworded items to rely less on terms such as "probability" and "frequency." For example, one might ask "how often" rather than "what is the frequency."

2. Reworded item assessing test statistic confusion (R-1) in the "interpreting results" section (Scenario 3). The term "*t*-value" reworded to read "test statistic."

3. Determined that the online presentation would mimic the format of a paper and pencil presentation. The online page would include one page per scenario with the associated item set. Answers could be influenced by answers on previous questions.

4. Clarified Scenario 2 for the "using tests of significance" answers. The answer would depend on whether or not the sample group was randomly selected.

In addition, two new item requirements were suggested. One suggestion was to assess confusing *P*-value with the population proportion. An item was added to assess this difficulty. It was noted that this confusion might depend on the book used and the order of topics in the course. The second suggestion was to assess whether students interpret a large *P*-value to mean that the null hypothesis is true. The item was not added at this juncture. This item suggestion seemed to overlap with several difficulties cited in the literature. It was similar to testing the difficulty *inverse as true* (L-1), in the sense that there was a deterministic connotation to the item suggestion. However, the suggested item also had the respondent interpret a large *P*-value. A correct response to RPASS-1B, Item 13 (*P-value as always small*) was judged as measuring the same or similar content.

4.1.1.4 Pilot of RPASS-1B

Figure 4.1 presents a histogram of the 17-item RPASS-1B total correct scores for the 333 students who completed the test (*M* = 8.8, *SD* = 2.6, *Mdn* = 9, *IQR* = 4). Scores ranged from 2 to 15 items answered correctly (Cronbach's coefficient $\alpha = .46$). Items were scored with a 1 if the respondent recognized a correct conception as correct or identified a misconception as incorrect. Otherwise, the item was scored 0. There was little variation in responses between the four courses (see Table 4.2 and Figure 4.2). Mean scores increased from lower division undergraduates to upper division undergraduates to master's level students. There was, however, no difference between means for master's students in their first statistics course and doctoral-level students in their second course.

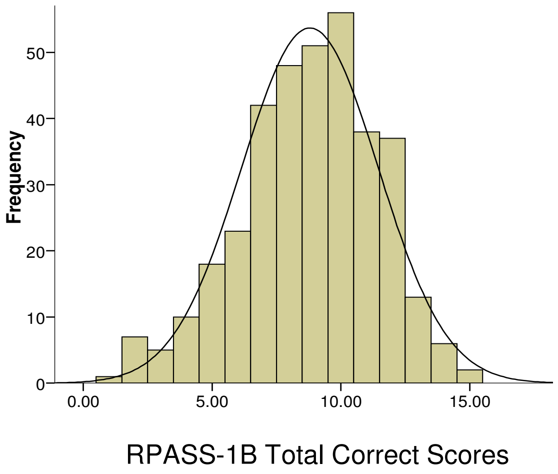


Figure 4.1. Histogram of 17-item RPASS-1B score distribution, *N* = 333

Table 4.2

Mean (SD), Median (IQR) for RPASS-1B Total Correct Scores by Course (N = 333)

	Lower Division *n* = 103	Upper Division *n* = 65	Masters First *n* = 108	Doctoral Second *n* = 57
Mean (*SD*)	7.7 (2.5)	8.8 (2.3)	9.5 (2.6)	9.5 (2.2)
Median (*IQR*)	8.0 (4)	9.0 (4)	10.0 (3.75)	9.0 (3.5)

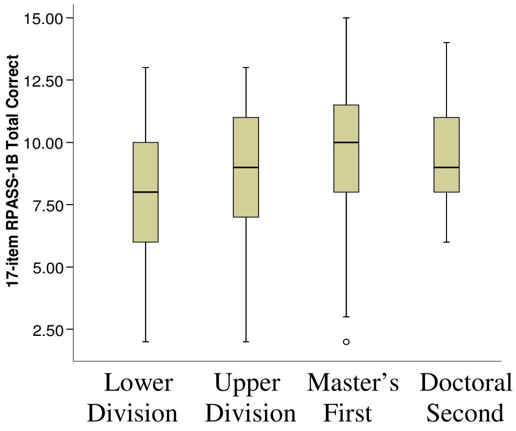


Figure 4.2. Box plots comparing 17-item RPASS-1B score distributions by course

4.1.2 Phase II. Instrument Development from RPASS-1 to RPASS-2

Based on the ongoing literature review, the item content blueprint was revised. Nine items were added to the RPASS to reflect the revised blueprint and potentially improve instrument reliability. Two items were added to assess the new specific content: dependence of *P*-values on the alternative hypothesis (B-1) (Lipson, 2002) and linking a two-tailed significance test to confidence interval construction (R-6) (Cumming & Finch, 2005). One item was added to assess whether *P*-values reflect the probability of results occurring in the sample versus the population (R-2). A fifth scenario with six associated true-false items was added to the RPASS based on Oakes' (1986) study. Four items assessed misinterpreting the *P*-value as the truth or falsity of hypotheses (H-1 to H-4). One item assessed the reliability fantasy (R-5), and one item assessed making a deterministic interpretation of statistically significance results (L-1).

One item was removed from the RPASS, the item that assessed confusing *P*-values with population proportions. This item was removed because no literature was found to substantiate its inclusion. The remaining content was cross-referenced to the literature and was also mapped to the three learning goals for a first course in statistics: statistical literacy, statistical reasoning, and statistical thinking. The revised test blueprint appears in Table 4.3.

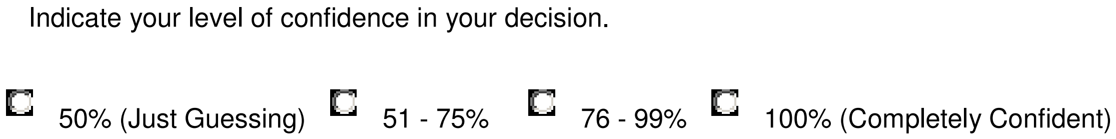
Table 4.3

Revised Test Blueprint with Content Areas, Learning Goals, Blueprint Category, and Correct Conception (C) or Misconception (M) Identified

		Learning goals for the items		
Content areas	Blueprint category	Statistical literacy	Statistical reasoning	Statistical thinking
Basic literacy				
Textbook, simulation or lay definition	B-1	C		
P-value embedded in sampling variation	B-1	C		
P-value as rareness measure	B-1	C		
P-value dependence on alternative	B-1		C	
P-value as always low	B-2		M	
P-value as strong or weak evidence	B-1[a]	C	C	
P-value and sampling error	B-1[a]		M	
P-value and differences or effects	B-1[a]	C		
P-value and practical significance	B-1[a]		C	
Relationships between concepts				
Test statistic and *P*-value	R-1	M		
Sample and population	R-2	M		
Type I / α and *P*-value	R-3		M	
Sample size and significance	R-4		M	
Reliability and *P*-value	R-5		M	
Confidence interval and significance	R-6		C	
Logic of inference				
Inverse as true	L-1			M
Converse as true	L-2			M
Chance as cause of results observed	L-3			M
Conclusions as independent of study design	L-4[a]			M
Belief in the truth or falsity of hypotheses				
Probability: alternative is true	H-1		M	
Probability: alternative is false	H-2		M	
Probability: null is true	H-3		M	
Probability: null is false	H-4		M	

Note. [a]Added during expert review as described in Section 4.1.3.5.

Based on feedback from a statistics education advisor, the confidence rating scale was altered. Any confidence rating of *50%* or less would constitute guessing. Therefore, the *0-100%* scale was modified to a *50-100%* scale. The new scale appears in Figure 4.3. The final version of the 25-item RPASS-2 appears in Appendix D.4, column 1.



Indicate your level of confidence in your decision.

□ 50% (Just Guessing) □ 51 - 75% □ 76 - 99% □ 100% (Completely Confident)

Figure 4.3. RPASS-2 confidence rating scale ranging from 50 to 100% confident

4.1.3 Phase III. Instrument Development from RPASS-2 to RPASS-4

4.1.3.1 Administration of RPASS-2

The 25-item RPASS-2 was administered to 36 students during the winter quarter of 2006. A histogram of the RPASS-2 total correct scores appears in Figure 4.4 for the 32 students who answered all 25 items (*M* = 13.3, *SD* = 3.2, *Mdn* = 13, *IQR* = 4.75). RPASS-2 correct responses ranged from 6 to 21 items.

Two wording changes were made based on comments from participants during the test administration. The fifth scenario was altered to spell out the phrase "degrees of freedom" rather than use the abbreviation "*df.*" RPASS-2 Item 14 was reworded. Two participants were unfamiliar with the word "awry," so alternative wording was explored in cognitive interviews. The phrase was reworded from "awry" to "gone bad" to suggesting that a "calculation error" was made by the researcher. (The final version of the item wording for RPASS-2 is detailed in Appendix D.4, column 1, item 14.)

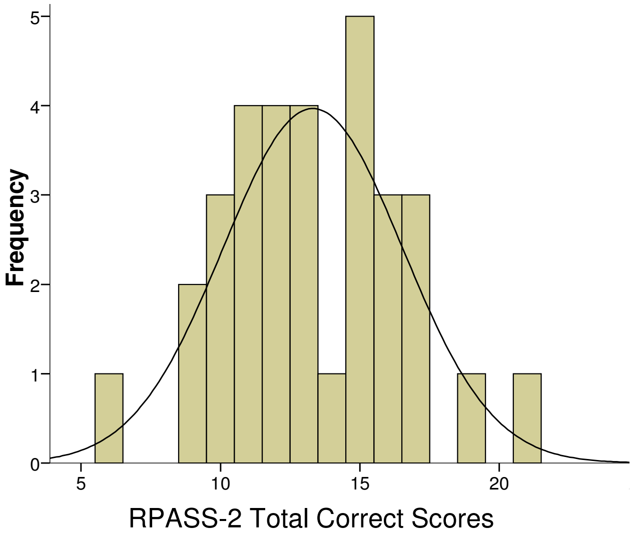


Figure 4.4. Histogram of 25-item RPASS-2 score distribution, *n* = 32

4.1.3.2 In-depth Interviews with RPASS-2

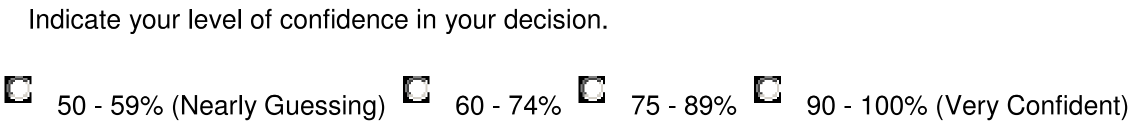
Results are reported from the in-depth student interviews. Four interviewees participated in the testing of RPASS-2. One student who volunteered for an in-depth interview had not previously taken the RPASS. Selected notes from all five interviews are reported in Appendix D.2. Some of the discussion during the in-depth interviews motivated RPASS-2 item modifications. These modifications are summarized in Table 4.3, along with the item changes motivated by cognitive interviews.

4.1.3.3 Cognitive Interviews using RPASS-2 and RPASS-3A

Relevant student comments from each of the cognitive interviews (CIs) are detailed in Appendix D.3. Eight student volunteers participated in "think aloud" cognitive interviews as described in Chapter 3. Five CIs were conducted with RPASS-2. Recommended changes were made to RPASS-2 Items 2-8, 10-14, 16, 17, and 23 to produce RPASS-3A as detailed in Appendix D.4. Three CIs were conducted with RPASS-3A.

The confidence rating scale was examined based on cognitive interview feedback. The *50-100%* confidence scale seemed to prohibit students from selecting the extremes of the scale. The scale was modified to broaden the range at the extremes of the scale. The low end of the scale was changed from *50%* to *50-59%* percent and the high end from *100%* to *90-100%* as depicted in Figure 4.5.

After subsequent interviews using this modified scale, the confidence scale was removed from RPASS-3A altogether. The deliberation of cognitive interview Student 329 suggested she was interpreting the confidence rating as the “probability of the item being valid or invalid.” She concluded “…So, I think I’ll go with *60-74%* that it’s Invalid.” This type of misinterpretation would confound results from the instrument. Moreover, the confidence ratings were not correlated with correct scores in previous field tests. So, confidence ratings were removed. For more details, see the student comments in the cognitive interview (CI) 329 (Appendix D.3, Table D6). Comments related to Item 12 illustrate how the confidence rating scale may have confounded RPASS results.



Indicate your level of confidence in your decision.

□ 50 - 59% (Nearly Guessing) □ 60 - 74% □ 75 - 89% □ 90 - 100% (Very Confident)

Figure 4.5. Confidence rating scale ranging from 50 to 100% confident altered to have broader ranges at the ends of the scale

RPASS-2 item options were modified based on feedback from cognitive interviews. Students 210 and 311 suggested that it was unclear what they were being asked to evaluate as being “Valid” or “Invalid.” Therefore, most options were augmented. For example, the response options read “Valid Action” or “Invalid Action” if the action taken was to be evaluated. After making these modifications, RPASS-3A items consisted of two item types: multiple TF for the definitional items of Scenario 1 and multiple-alternate choice for the remainder of the items (Downing, 1992).

Table 4.4 highlights seven item changes and two scenario changes that converted RPASS-2 to RPASS-3A. In the first column is the item wording before the changes were made (RPASS-2). The second column presents the rationale for making the change and a reference to the interview that motivated the change. Column three presents the item wording after the changes were made (RPASS-3A). The complete list of item modifications from the 25-item RPASS-2 to the 25-item RPASS-3A appears in Appendix D.4.

Table 4.4

Item Modifications from RPASS-2 to RPASS-3A with Rationale for Change

RPASS-2	Rationale for change	RPASS-3A
3. If a simulation of the experiment were conducted, the *P*-value of .001 is the long-run frequency of obtaining the experimental results or something more extreme due to chance. □ True □ False	Clarified what the simulation entailed; Added “probability” to clarify “long-run frequency;” “Due to chance” wording confounds, was deleted. (CI: 230, 107)	3. Simulating the experiment with a random model (to model no difference), p = .001 is the long-run frequency (i.e., the probability) of obtaining the experimental results or results even more extreme than those observed. □ True □ False
4. This *P*-value tells me the chances are 1 in 1000 of observing data this surprising (or more surprising) than what I observed, if the null hypothesis is true. □ True □ False	Student was not sure if “surprising” is equivalent to “extreme;” Changed to “rare.” (ID: 118)	4. This *P*-value tells me the chances are 1 in 1000 of observing data this rare (or more rare) than what I observed, if the null hypothesis is true. □ True □ False
5. This *P*-value may reflect a statistically significant difference between two samples but says nothing about the populations from which they come. □ True □ False	Wording was changed; Students had difficulty understanding what was being asked / stated. (CI: 107, 210, 329, 311)	5. The *P*-value may reflect a difference between the two samples observed but has no bearing on whether there is a statistically significant difference in the population. □ True □ False

Note. Numbers in parentheses identify the In-Depth (ID) or Cognitive Interview (CI) that motivated the change.

Table 4.4 (continued)

RPASS-2	Rationale for change	RPASS-3A
6. Action: The district researchers pre-determined the number of students to sample to ensure their *P*-value could detect whether the improvement could be attributed to the new Head Start-like program. □ Valid □ Invalid	Deleted “Head Start” reference, not necessarily familiar to Ss. (ID: 322) Omitted “predetermined,” since Ss misinterpreted it to mean researcher selected the sample and did not have an SRS. (CI: 230, 107, 303) Altered valid-invalid to an alternate choice (AC) format for clarity. (CI: 210, 311)	6. Action: The district researchers carefully planned how many students should be included in the study, since they were concerned about how the size of their random sample would impact *P*-value. □ Valid Action □ Invalid Action
7. Action: The district researchers determined how often they would obtain a score of 102 or higher just by chance to "definitively prove" whether the program had a positive impact. □ Valid □ Invalid	The word “determined” was unclear; Added: “conducted a statistical test to determine” to clarify. (CI: 230, 210) Altered valid-invalid to AC format. (CI: 210, 311)	7. Action: The district researchers conducted a statistical test to determine how often they would obtain a score of 102 or higher just by chance to "definitively prove" whether the program had a positive impact. □ Valid Action □ Invalid Action
8. Action: After checking the necessary conditions, the district researchers proceeded to determine if random chance was the “cause of the results observed.” □ Valid □ Invalid	Student said, “what are you asking here?” (CI: 210) Added, “conducted a test of significance.” Altered valid-invalid to AC format. (CI: 210, 311)	8. Action: After checking the necessary conditions, the district researchers conducted a test of significance to determine if random chance was the “cause of the results observed.” □ Valid Action □ Invalid Action

Note. Numbers in parentheses identify the In-Depth (ID) or Cognitive Interview (CI) that motivated the change.

Table 4.4 (continued)

RPASS-2	Rationale for change	RPASS-3A
14. Interpretation: The researcher interprets a large *P*-value for his hair growth treatment to mean that the experiment has gone bad. □ Valid □ Invalid	Reworded from "gone bad" to "there was a calculation error." Item was still problematic after change. (ID: 304, 322; CI: 230, 210) Altered valid-invalid to AC format. (CI: 210, 311)	10. Interpretation: The researcher assumes that getting a large *P*-value for his hair growth treatment clearly means that there was a calculation error. □ Valid Interpretation □ Invalid Interpretation
Scenario 4: A researcher believes that an SAT preparation course will improve SAT scores. The researcher invites a random sample of students to take the online prep course, free of charge. ….	Added, "All of students agree to participate" to suggest researchers have an SRS. (CI: 311)	Scenario 4: A researcher believes that an SAT preparation course will improve SAT scores. The researcher invites a random sample of students to take the online prep course, free of charge. All of these students agree to participate. ….
Scenario 5: Suppose you have a treatment which 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, degrees of freedom df = 18, p = 0.01). Please mark each of the statements below as "true" or false."	Removed any reference to "degrees of freedom" and added more contextual information to the scenario. (CI: 107, 329, 311) Driving school context added.	Scenario 5: Suppose you have a driving school curriculum which you suspect may alter performance on passing the written exam portion of the driver's test. 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, degrees of freedom df = 18, p = 0.01). Please mark each of the statements below as "true statement" or "false statement."

Note. Numbers in parentheses identify the In-Depth (ID) or Cognitive Interview (CI) that motivated the change. ID and CI codes refer to interviews detailed in Appendixes D.2 (ID) and D.3 (CI).

4.1.3.4 Administration of RPASS-3A

Total correct scores are reported for 20 new students who volunteered to take the 25-item RPASS-3A (*M* = 15.6, *SD* = 3.0, *Mdn* = 16, *IQR* = 4.5 items). RPASS-3A correct responses ranged from 10 to 21 items. The RPASS-3A total score distribution is presented in a histogram in Figure 4.6. RPASS-3A had the same number of items as the RPASS-2 but with the items reworded based on the student interviews. After the item modifications, total scores were slightly higher for RPASS-3A (*M = 15.6*) compared to RPASS-2 (*M* = 13.3), and the spread statistics were comparable (*SD* = 3.0 and *SD* = 3.2, respectively). Both versions assessed 7 correct conceptions and 18 misconceptions. No changes were made to the RPASS-3A after this field test.

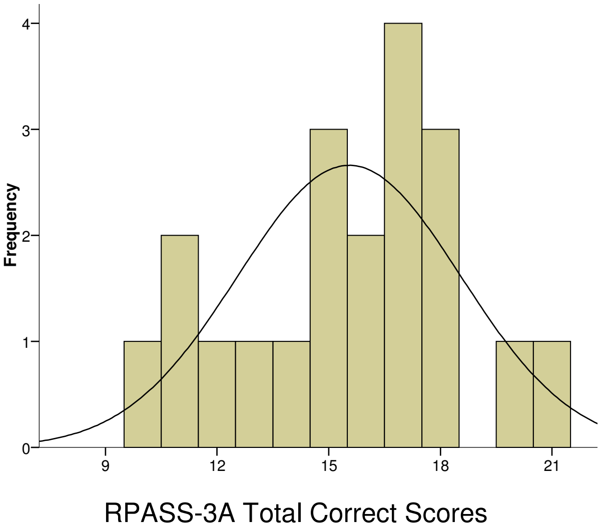


Figure 4.6. Histogram of 25-item RPASS-3A score distribution, *n* = 20

4.1.3.5 Content Validation and Instrument Modifications: RPASS-3A to RPASS-4

As described in Section 3.2.3.5, ten content experts completed two rounds of content validation. RPASS-3A was submitted to the content experts for the first round of review and validity ratings. A four-point rating scale was used, where 3 meant the raters *agreed* to the items' validity, and 4 was *strongly agreed*. The aggregate mean was sufficient after the first round of ratings (*M* = 3.2). However, six items received mean ratings under 3. Appendix D.5 Part A lists the experts' suggestions for the seven most problematic items (RPASS 3A Items 3, 7, 8, 13, 18, 21, and 22).

Appendix D.6 lists experts' specific suggestions related to missing or extraneous content. Before adding any new content, item suggestions were cross-referenced to the research literature. Eight new items were developed with expert input and feedback (RPASS-3B Items 5, 6, 7, 13, 18, 21, 23, and 32). Four items were added to the basic literacy blueprint category (B-1). Three items assessed understanding *P*-values and statistical significance as the strength or weakness of evidence against the null hypothesis. One item dealt with practical significance. One new item linking sample size and significance was added to the relationships between concepts blueprint category (R-4). One item was added to the logic of inference category (L-4) linking study design to the conclusions drawn that can be drawn from statistically significant results (see Garfield et al., 2005). In addition to these item changes, Scenarios 3 and 5 were modified. The scenarios were amended to clarify whether the conditions for inference were satisfied (e.g., randomization of treatments or random sampling of subjects; see Appendix D.5, Part B).

One item that assessed the reliability fantasy (RPASS-3A, Item 15) was removed from RPASS-3A. The item was judged to be redundant with RPASS-3A, Item 24 (see rater Comment 1 in Appendix D.6 Expert Rater Assessment of Missing or Extraneous Content). Appendix D.7 details the item wording changes from the 25-item RPASS-3A (in the first column) to the 32-item RPASS-3B (in the second column).

During the second round of expert validation, the 32-item RPASS-3B was reviewed with raters in one-on-one meetings. During this second review, two raters reiterated their concern about the length of the RPASS and item redundancy (see Appendix D.6, Comments 1 and 5). Therefore, four redundant items were deleted from the scale, namely RPASS-3B Items 8, 19, 20, and 28 (equivalent to RPASS-3A Items 6, 16, 17, and 23).

A suggestion was made to alter some items to have three-option choices (rather than two) to encourage statistical reasoning, where appropriate. RPASS-3C Items 3, 6, 16, 19, and 28 were altered to have three-options. For example, RPASS-3C, Item 3 assessed understanding that the magnitude of the *P*-value is dependent on the alternative hypothesis. This item required a “true” or “false” response in RPASS-3B. The new wording in RPASS-3C required students to decide if the *P*-value would be larger, smaller or the same when conducting a two-tailed test compared to a one-tailed test. The change to this item wording to accommodate the three options appears in Figure 4.7.

RPASS-3B, Item 3. Statement: In general, a two-tailed alternative hypothesis would yield a larger *P*-value than .001 for the same value of the test statistic.

○ True ○ False

RPASS-3C, Item 3. Statement: If the students had conducted a two-tailed test instead of a one-tailed test on the same data, how would the *P*-value have changed?

○ *P*-value would be larger.
○ *P*-value would be smaller.
○ *P*-value would not change.

Figure 4.7. Example of item modification from two options (RPASS-3B) to three options (RPASS-3C)

Table 4.5 presents the expert item ratings for the initial review of the 25-item RPASS-3A (*M* = 3.2, *Mdn* = 3.1) and for the final review of the 28-item RPASS-3C (*M* = 3.6, *Mdn* = 3.7). Only two items were rated with a mean score of less than 3 in the final review: chance as cause of results (Item 12, *M* = 2.9) and converse as true (Item 15, *M* = 2.9). The two remaining items with mean ratings of 2.9 were judged sufficient for retention. All ten experts *agreed* or *strongly agreed* that the collection of items assessed the stated learning objectives or misconceptions (correct conceptions: *M* = 3.3, *Mdn* = 3; misconceptions: *M* = 3.2, *Mdn* = 3). After content validation, one additional redundant item was removed from the scale (RPASS-3C, Item 1) due to its similarity with RPASS-3C, Item 22. The resultant 27-item scale (RPASS-4) assessed 13 correct conceptions and 14 misconceptions.

Table 4.5

Mean Validity Ratings by Item from Ten Expert Raters for RPASS-3A and RPASS-3C

RPASS-3A initial review		RPASS-3C final review	
Correct conception or misconception	Rating	Correct conception or misconception	Rating
1. Probability: null is true	3.9	1. Probability: null is true	4.0[a]
2. Textbook definition	3.9	2. Textbook definition	4.0
3. Simulation definition	2.8	3. *P*-value dependence on alternative	4.0[c]
4. Lay definition	3.4	4. Lay definition	3.7
5. Sample and population	3.1	5. Conclusions and study design	3.7[b]
6. Sample size and significance	3.3[a]	6. Smaller the *P*-value	4.0[b,c]
7. Inverse as true	2.6	7. *P*-value and standard error	3.8[b]
8. Chance as cause of results	2.8	8. *P*-value in sampling variation	3.3
9. *P*-value in sampling variation	3.0	9. Confidence interval and significance	3.7
10. Confidence interval and significance	3.0	10. Inverse is true	3.1
11. *P*-value as rareness measure	3.4	11. Strong statistical evidence	4.0[c]
12. Test statistic and *P*-value	3.4	12. Chance as cause of results	2.9
13. Converse as true	2.9	13. *P*-value as rareness measure	3.5
14. *P*-value as always low	3.1	14. Test statistic and *P*-value	3.5
15. Reliability and *P*-value	3.6[a]	15. Converse as true	2.9
16. Probability: alternative is true	3.9[a]	16. *P*-value as always low	3.6[c]
17. Probability: alternative is false	3.7[a]	17. Weak statistical evidence	4.0[b]
18. Type I / α and *P*-value	2.4	18. Practical significance	4.0[b]
19. Probability: null is false	3.1	19. Type I / α and *P*-value	3.4[c]
20. Probability: null is true	3.8	20. Large difference or effect	4.0[b]
21. Probability: alternative is false	2.6	21. Probability: null is false	3.3
22. Probability: alternative is true	2.9	22. Probability: null is true	4.0
23. Probability: alternative is false	3.0[a]	23. Probability: alternative is false	3.0
24. Reliability and *P*-value	3.6	24. Probability: alternative is true	3.0
25. *P*-value dependence on alternative	3.0	25. Reliability and *P*-value	3.0
		26. Simulation definition	3.7
		27. Sample and population	3.7
		28. Sample size and significance	3.5[b,c]

Note. [a]Item deleted during expert validation. [b]Item added during expert validation. [c]Item altered to three-option item type.

4.2 Examining Results from the Baseline Data Collection

4.2.1 Phase IV. Evaluation of RPASS-4 Baseline Scores, Reliability, and Validity

4.2.1.1 Results from the RPASS-4 Administration

Item responses are summarized for the baseline data to examine the second research question: *What does the proposed instrument indicate about students' understanding of and reasoning about P-values and statistical significance?* Two hundred twenty-four respondents completed all 27-items. A histogram of the RPASS-4 total correct scores appears in Figure 4.8 (*M* = 15.9, *SD* = 3.1, *Mdn* = 16, *IQR* = 2). Items were scored with a 1 if the respondent recognized a correct conception item as correct or identified a misconception item as incorrect, 0 otherwise. The items assessed 13 correct conceptions and 14 misconceptions. Scores ranged from 7 to 25 items answered correctly.

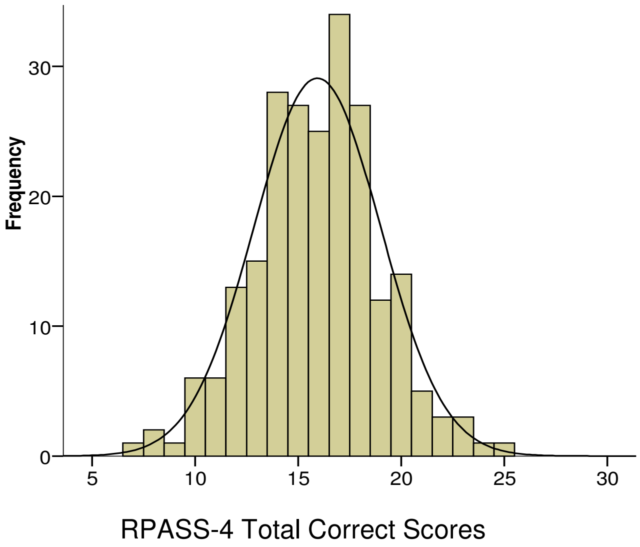


Figure 4.8. Histogram of 27-item RPASS-4 score distribution, *N* = 224

Table 4.6 summarizes the mean proportion of correct responses across three item groupings and the number of items per group. The first subset groups items by whether a correct conception or misconception was assessed. The second subset groups items based on the four content areas defined by the test blueprint. The third subset is based on the three learning goals for a first statistics course: statistical literacy, statistical reasoning, and statistical thinking.

Table 4.6

Mean Proportion of Correct Responses for Three Item Groupings: Correct Conceptions and Misconceptions, Content Areas, and Learning Goals (N = 224)

Three item groupings	Mean proportion of correct responses ($\mu_{\hat{p}}$)	Number of items
Correct conception and misconception items		
Correct conceptions	.66	13
Misconceptions	.55	14
Content areas defined by the test blueprint		
Basic literacy	.68	13
Relationships between concepts	.55	6
Logic of inference	.48	4
Belief in the truth or falsity of hypotheses	.55	4
Learning goals for statistics instruction		
Statistical literacy	.71	9
Statistical reasoning	.57	14
Statistical thinking	.48	4

Table 4.7 reports the proportion of correct responses (RPASS-4 item difficulty) and corrected item-total correlation by item. Learning goals and correct conception or misconception are also identified. Items are sorted by proportion of correct responses within blueprint category.

Table 4.7

RPASS-4 Proportion of Correct Responses, Corrected Item-total Correlation, and Learning Goals, sorted by Proportion Correct within Blueprint Category (N = 224)

		Proportion	Corrected	Learning goal for the item[a]		
RPASS-4 correct conception or misconception	Blueprint category	of correct responses	item-total correlation	Statistical literacy	Statistical reasoning	Statistical thinking
5. Smaller the *P*-value	B-1[b]	.78	.26		C	
19. Large difference or effect	B-1	.76	.21	C		
15. *P*-value as always low	B-2[b]	.76	.32		M	
25. Simulation definition	B-1	.75	.09	C		
1. Textbook definition	B-1	.74	.23	C		
10. Strong statistical evidence	B-1	.74	.24	C		
12. *P*-value as rareness measure	B-1	.74	.24	C		
7. Sampling variation	B-1	.72	.06	C		
3. Lay definition	B-1	.69	.11	C		
17. Practical significance	B-1	.67	-.06		C	
2. Dependence on alternative	B-1[b]	.54	.10		C	
16. Weak statistical evidence	B-1	.53	.06		C	
6. *P*-value and standard error	B-1	.46	.02		M	
18. Type I / α and *P*-value	R-3[b]	.67	.42		M	
13. Test statistics and *P*-value	R-1	.65	.08	M		
26. Sample and population	R-2	.63	.14	M		
8. CI and significance	R-6	.58	-.16		C	
24. Reliability and *P*-value	R-5	.40	.01		M	
27. Sample size and significance	R-4[b]	.37	.11		C	
11. Chance as cause of results	L-3	.69	.32			M
4. Conclusions and study design	L-4	.51	.18			M
14. Converse as true	L-2	.37	.18			M
9. Inverse as true	L-1	.35	-.17			M
23. Probability: alternative is true	H-1	.61	.07		M	
22. Probability: alternative is false	H-2	.60	-.08		M	
20. Probability: null is false	H-4	.55	.15		M	
21. Probability: null is true	H-3	.44	-.15		M	

Note. [a]Correct conception (C) or misconception (M). [b]Three-option item.

4.2.1.2 RPASS-4 Reliability

The RPASS-4 reliability across the five introductory courses was low (Cronbach's coefficient $\alpha = .42$, $N = 224$). Thus, 42% of the variation in RPASS scores can be attributed to true score variation. The remainder of the variation is attributable to measurement error. The maximum validity coefficient of RPASS-4 is .65 (the square root of reliability). In Appendix E.2, item statistics are reported from the reliability analysis, including proportion of correct responses, corrected item-total correlation, and an estimate for coefficient alpha if the item were deleted from the scale. Items are sorted by proportion of correct responses within blueprint category.

4.2.1.3 RPASS-4 Validity Evidence

Descriptive statistics for the instruments used to assess construct validity are reported in Table 4.8. Assumptions to make inferential decisions using Pearson product-moment correlations (*r*) were not violated. Assuming the distributions are normally distributed in the population, inferences can be drawn from the sample item scores to these respondents' true scores.

Table 4.8

Mean (SD), Median (IQR) for Instruments Used to Assess Construct Validity (n = 56)

	Convergent	Discriminant	
	Concurrent: Five-part open-ended item	Bivariate Quantitative Data topic scale	27-item RPASS-4
Mean (*SD*)	2.43 (.90)	60.20 (12.98)	16.64 (3.15)
Median (*IQR*)	2.50 (1.0)	57.10 (19.60)	17.00 (4.75)

Table 4.9 presents the convergent and discriminant validity evidence in a correlation matrix for the 56 respondents who were administered all three instruments. Rows and columns indicate the instruments used. Off-diagonal elements are the validity coefficients (correlations). The on-diagonal elements are the instrument reliabilities for the Bivariate Quantitative Data topic scale and RPASS-4. The proportion of rater agreement is reported for the open-ended item.

The convergent validity coefficient was positive, weak, and statistically significantly different from zero. The discriminant validity correlation coefficient was positive and very weak and was not statistically significantly different from zero, as desired. Both the RPASS and the Bivariate Quantitative Data topic scale were administered online. The lack of correlation with the discriminant measure suggests plausible rival interpretations – such as general statistics knowledge or general intelligence – do not explain the relationships found (Messick, 1989). In addition, similarities and dissimilarities in the method of testing (online for the Bivariate Quantitative Data topic scale and RPASS-4 versus paper and pencil for the open-ended item) do not explain correlations or lack of correlations found (Campbell & Fiske, 1959).

For development purposes, validity coefficients were corrected for attenuation due to measurement error. The corrected coefficient for convergent validity was moderate, and the discriminant coefficient remained weak. This pattern of correlation with the convergent measure and lack of correlation with the discriminant measure provides some evidence of construct-related validity.

Table 4.9

RPASS-4 Reliability and Validity Coefficients for AgStat and LibStat Respondents[a]

	Convergent	Discriminant	
Instrument	Concurrent five-part open-ended item	Bivariate Quantitative topic scale	27-item RPASS-4
Open-ended item			
Proportion of rater agreement	.82 (88)[b]		
Bivariate Quantitative Data topic scale			
Pearson's *r*	.20	.25[d] (57)[b]	
Corrected for comparison attenuation	—		
Corrected for RPASS-4 attenuation	.29		
RPASS-4			
Pearson's *r*	.38**	.09	.46[c]
Corrected for comparison attenuation	—	.18	
Corrected for RPASS-4 attenuation	.56	.27	

Note. [a]Off-diagonal elements are validity, *n* = 56 listwise unless otherwise noted. [b]Sample size noted in parentheses. [c]Internal consistency reliability estimated using Cronbach's coefficient alpha. [d]Internal consistency reliability estimated using K-R 20. ***p* < .01, 2-tailed

4.2.2 Phase V. Estimation of RPASS-5 Reliability and Validity Data

4.2.2.1 RPASS-5 Reliability

Using the baseline data collection in Phase IV, an item analysis was conducted, removing items with corrected item-total correlations less than .15 as described in Chapter 3. In Table 4.10 the proportion of correct responses (item difficulty), *SD* of item difficulty, corrected item-total correlation, and alpha-if-item-deleted are reported for the resultant 15-item scale (RPASS-5). RPASS-5 has an expected Cronbach's coefficient alpha of .66 (*M* = 9.6, *SD* = 2.9, *Mdn* = 10, *IQR* = 4). The maximum validity coefficient of RPASS-5 is .81 (the square root of reliability).

Table 4.10

RPASS-5 Proportion of Correct Responses, Corrected Item-total Correlation, and α-if-item-deleted, Sorted by Proportion Correct within Blueprint Category (α = .66, N = 224)

RPASS-5[a] correct conception (C) or misconception (M)		Blueprint category	Proportion of correct responses	*SD*	Corrected item-total correlation	α-if-item-deleted
5. Smaller the *P*-value	C	B-1[b]	.78	.41	.29	.647
19. Large difference or effect	C	B-1	.76	.43	.29	.646
15. *P*-value as always low	M	B-2[b]	.76	.43	.40	.633
10. Strength of evidence	C	B-1	.74	.44	.35	.639
1. Textbook definition	C	B-1	.74	.44	.28	.648
12. *P*-value as rareness measure	C	B-1	.74	.44	.25	.652
3. Lay definition	C	B-1	.69	.46	.16	.664
18. Type I / α and *P*-value	M	R-3	.67	.47	.52	.612
26. Sample and population	M	R-2[b]	.63	.48	.19	.661
27. Sample size and significance	C	R-4	.37	.48	.17	.663
11. Chance as cause of results	M	L-3	.69	.46	.38	.634
4. Conclusions and study design	M	L-4	.51	.50	.24	.654
14. Converse as true	M	L-2	.37	.48	.24	.653
23. Probability: alternative is true	M	H-1	.61	.49	.21	.658
20. Probability: null is false	M	H-4	.55	.50	.23	.655

Note. [a]RPASS-4 item numbers are reported. [b]Three-option item.

4.2.2.2 RPASS-5 Validity Evidence

Validity coefficients were re-estimated and reported for RPASS-5 using Pearson's *r* (see Table 4.11). When corrected for measurement error in RPASS-5, the expected convergent correlation coefficient was positive and moderate. For development purposes, discriminant correlations were corrected for attenuation in the discriminant measure and for RPASS-5. The corrected correlation coefficient for the discriminant measure remained weak, as desired.

Table 4.11

15-item RPASS-5 Estimated Reliability and Validity Coefficients[a]

	Convergent	Discriminant	
Instrument	Concurrent five-part open-ended item	Bivariate Quantitative topic scale	15-item RPASS-5
Open-ended item			
Proportion of rater agreement	.82 (88)[b]		
Bivariate Quantitative Data topic scale			
Pearson's *r*	.20	.25[d] (57)[b]	
Corrected for comparison attenuation	—		
Corrected for RPASS-5 attenuation	.25		
RPASS-5			
Pearson's *r*	.40**	.06	.66[c]
Corrected for comparison attenuation	—	.12	
Corrected for RPASS-5 attenuation	.49	.15	

Note. [a]Off-diagonal elements are validity, n = 56 listwise unless otherwise noted. [b]Sample size noted in parentheses. [c]Internal consistency reliability estimated using Cronbach's coefficient alpha. [d]Internal consistency reliability estimated using K-R 20. $**p < .01$, 2-tailed

4.3 Chapter Summary

The development, testing, and validation of the RPASS produced a 27-item scale. All experts *agreed* or *strongly agreed* that the subscales measured the stated learning objectives or misconceptions, providing content-related validity evidence. Baseline scores for the 27-item RPASS-4 were examined for evidence of reliability, convergent and discriminant validity evidence. Total score reliability was low. Only 42% of the variation in RPASS-4 scores can be attributed to true score variation. Even though the uncorrected convergent validity coefficient was positive and statistically significant, it was weak. The low internal consistency reliability seems to have attenuated the true convergent correlation because the corrected convergent correlation was moderate.

Furthermore, the discriminant correlation coefficient remained weak when corrected for attenuation, as desired. The pattern of correlations provides some construct-related validity evidence. The subsequent item analysis of the Phase IV data identified a subset of 15 RPASS-4 items (designated RPASS-5) with a higher estimated internal consistency reliability ($\alpha = .66$). RPASS-5 maintained similar evidence of construct validity as obtained for RPASS-4.

CHAPTER 5: DISCUSSION

This chapter makes a final assessment of the quality of RPASS-4 based on the evidence collected and reported in Chapter 4. Furthermore, the chapter summarizes what was learned about the respondents' reasoning about *P*-values and statistical significance based on the fifteen most reliable items. Each of the two research questions is addressed, and conclusions are drawn with caution due to the limitations of the study. Finally, directions for future RPASS research are suggested.

This study developed the Reasoning about *P*-values and Statistical Significance (RPASS) scale. The results provided strong evidence of content validity, but weak evidence of construct-related validity. The RPASS was designed to assess conceptual understanding and misunderstanding of *P*-values and statistical significance and to facilitate future research about the effects of instructional approaches on this understanding. Four versions of the RPASS were administered in this study to develop, test, and validate the instrument. The fourth version, RPASS-4, was administered to 224 students across five introductory statistics courses to gather baseline data. Thirteen correct conceptions and 14 misconceptions were assessed. Convergent and discriminant validity evidence were gathered in two of the five courses ($n = 56$). These data facilitate examining each of the two research questions:

1. *Can a research instrument be developed that is a sufficiently reliable and valid measure of students' understanding of and difficulties with reasoning about P-values and statistical significance?*

2. *What does the proposed instrument indicate about students' understanding of and reasoning about P-values and statistical significance?*

5.1 Research Question 1

The RPASS-4 content-related validity evidence was judged to be sufficient. All ten experts *agreed* or *strongly agreed* that the two subscales (correct conceptions and misconceptions) measured the stated learning objectives or misconceptions. All but two RPASS-4 items were rated *agreed* or *strongly agreed* (3 or above) by every rater. The two remaining items with mean ratings of 2.9 were retained pending further data gathering (Items 12 and 15).

The pattern of corrected validity correlation coefficients provided some evidence of construct-related validity. RPASS-4 had a weak but statistically significant positive correlation with the convergent measure. Correcting for RPASS-4 measurement error yielded evidence of a moderate correlation, as desired. RPASS-4 had a weak positive correlation with the discriminant measure (whether corrected or uncorrected). This pattern of correlations provides some evidence that RPASS-4 measures the desired construct: students' understanding and misunderstanding of *P*-values and statistical significance.

Nevertheless, the total score reliability was low for RPASS-4. While 42% of the variation in RPASS scores could be explained by true score variation, the remaining 58% of variation can be attributed to measurement error. Measurement error needs to be reduced to use the RPASS total score for research purposes. To further explore reliability, an item analysis was conducted to identify RPASS-4 items with weak or negative corrected item-total correlations. The twelve items identified with low correlations (corrected $r_{\text{pb}} < .15$) may need further development to improve reliability. Omitting these twelve items with low correlations seemed to produce a more reliable total score, the 15-

item RPASS-5 (estimated $\alpha = .66$). Furthermore, RPASS-5 retained a similar pattern of convergent and discriminant correlations as was obtained for RPASS-4.

Assessing inferential topics that are more commonly taught across courses should reduce guessing and improve reliability of scores. Even though omitting the twelve low or negatively correlating items does not sample all of the specific content, each of the four major content areas are represented and the elimination of noisy items seems to produce a more reliable scale. The RPASS-5 items may cover inference topics more common across the different courses. Lengthening RPASS-5 with additional items that cover the same content should increase score variation and improve reliability (Cronbach, 1951). Converting some of the RPASS-5 items from two-option to three-option items, where appropriate, may also reduce guessing and improve reliability (Rodriguez, 2005).

5.2 Research Question 2

The low internal consistency of the RPASS-4 score suggests that item results based on the total score are not reliable. Therefore in this discussion, inferences about the respondents' true understanding are limited to the fifteen most reliable items (RPASS-5) and conclusions drawn are limited to the aggregate group. Scores are not sufficiently reliable to make inferences about an individual respondent's true understanding.

To infer what respondents did or did not understand, item responses are discussed across four item groupings: (1) based on whether a correct conception or misconception was assessed; (2) based on the four content areas defined by the blueprint; (3) based on the three learning goals for a first statistics course (statistical literacy, statistical reasoning, and statistical thinking); and (4) based on the proportion of correct responses by item (item difficulty). Item difficulty is categorized as least difficult ($\hat{p} \geq .70$), as moderately difficult ($.55 < \hat{p} < .70$) or as most difficult ($\hat{p} \leq .55$).

5.2.1 Item Results Grouped by Correct Conceptions and Misconceptions

This study may contribute to research investigating whether instructional interventions can overturn misconceptions (e.g., delMas & Bart, 1989; Garfield et al., 2005; Kahneman & Tversky, 1982; Konold, 1995; Nisbett et al., 1993; Tversky & Kahneman, 1982). After instruction, the respondents in this study seemed to harbor contradictory inferential conceptions, as theorized by Konold (1995). Previous research results suggest that targeted training and assessment can bring about the level of conceptual change required to develop correct conceptions (e.g., delMas & Bart, 1989; Garfield et al., 2004). Targeted instruction may be needed to supplant what appear to be commonly held and persistent misconceptions about *P*-values and statistical significance. If RPASS measurement error can be reduced, the instrument may help identify whether targeted instruction is effective.

5.2.2 Item Results Grouped by Four Content Areas Defined by the Blueprint

Three of the four content areas were moderately to very difficult for these respondents: understanding relationships between inferential concepts, applying the logic of inference, and belief in the truth or falsity of hypotheses. Items related to basic literacy were understood by the largest proportion of respondents.

5.2.3 Item Results Grouped by Three Learning Goals

Analyzing items based on the three targeted learning goals, most respondents seemed to exhibit a good understanding in terms of statistical literacy. Statistical reasoning and thinking were more difficult. RPASS results seem to support the relationships between statistical literacy, statistical reasoning, and statistical thinking as described by delMas (2002). However, the categorization of items by learning goals needs to be externally reviewed by statistical education experts to evaluate if these categorizations are consistent across raters.

5.2.4 Item Results Grouped by Three Levels of Difficulty

Grouping the 15 most reliable items by the proportion of correct responses (difficulty) provides a reliable measure of respondents' understanding and misunderstanding. The least difficult items identify content for which most respondents had obtained a good understanding. The moderately difficult items identify content that only a little more than half of the respondents understood. The most difficult items identify content that many of the respondents (45% or more) did not understand.

Least difficult items

Among the six least difficult items ($\hat{p} \geq .70$), respondents linked concepts of variation to the magnitude of the *P*-value. A large proportion of respondents consistently linked statistical significance to small *P*-values, strong evidence, rare observed events, or large differences or effects (Items 5, 10, 12, and 19). Item 5 had three-options, making correct guessing even less likely. Contrary to Williams' (1999) suggestion that introductory students seemed to believe *P*-values are always low, 76% of baseline respondents recognized the correct interpretation of a large *P*-value (Item 15). Respondents also recognized a textbook definition of the *P*-value (Item 1). These items correlated well with the corrected total score (corrected r_{pb} = .25 to .40).

Moderately difficult items

The proportion of correct responses was between .55 and .70 for five items. Three of these moderately difficult items assessed misconceptions. The misconception items included Item 11, incorrectly concluding that significant results imply that chance *caused* the observed results rather than concluding that results *could be* explained by chance (see Daniel, 1998). Item 11 results correlated well with the corrected total score (corrected r_{pb} = .38). The RPASS item with the highest corrected item-total correlation also assessed a misconception, Item 18 (corrected r_{pb} = .52). A small majority of baseline respondents were able to differentiate *P*-values from the Type-I error rate (Garfield & Ahlgren, 1988).

Scores for Item 26 indicate some respondents correctly associated *P*-values with a statistically significant difference in the population, rather than the observed difference between samples. Item 26 had one of the weakest corrected item total correlations (corrected r_{pb} = .19) of the 15 items being discussed. Results for Item 26 were similar to the "neutral results" observed by Mittag and Thompson (2000). Item 23 was fashioned after the 1986 Oakes study. Most respondents were able to avoid this misinterpretation of the *P*-value as the probability that the alternative hypothesis is true (corrected r_{pb} = .21).

The only correct conception item in this moderately difficult group required that respondents recognize an informal or lay definition of the *P*-value (69% answered the item correctly). This item had the weakest correlation with the corrected total score of the 15 items under discussion (corrected r_{pb} = .16).

Most difficult items

The four items classified as most difficult had guessing-like responses ($\hat{p} \leq .55$). Three of these items were misconceptions, two with corrected r_{pb} = .24. Item 14 suggests respondents indiscriminately attribute a finding from a sample to the entire population (see Batanero, 2000; Nisbett et al., 1993). Of course, the objective of significance testing is to make inferences about population parameters from sample statistics. However, the data should be sampled randomly to be able to make an unbiased generalization. Results suggest that the importance of random sampling for inference was not well understood. Results from Item 4 also suggest respondents did not understand the importance of randomization of treatments in order to draw causal conclusions. These results are consistent with similar observations made by delMas et al. (in press) based on the national CAOS results. The third misconception item in this grouping was Item 20 fashioned after the 1986 Oakes study (corrected $r_{\text{pb}} = .23$). Item 20 assessed the misinterpretation of a small *P*-value as the probability that the null hypothesis is false.

There was one correct conception item among these most difficult items: the three-option Item 27. Respondents did not show evidence of understanding the influence of sample size on statistically significant results, if all else remained the same. Unlike the AERA member survey (Mittag & Thompson, 2000) and the study of intermediate level graduate students (Wilkerson & Olson, 1997), these introductory students did not link larger sample sizes with smaller *P*-values. The corrected item-total correlation for this item was relatively low (corrected $r_{\text{pb}} = .17$).

One might speculate that these prevalent misconceptions or misunderstanding may not be emphasized by instruction. Possible interventions or item improvement are discussed in the limitations and conclusions sections that follow.

5.3 Limitations

5.3.1 Construct Validity

The evidence gathered to examine convergent and discriminant validity were consistent with concluding that RPASS-4 measures the desired construct. However, when the coefficients were uncorrected for measurement error, the convergent validity coefficient was weak. Four factors may have influenced this low validity coefficient. First, no criterion measure existed to provide an adequate comparison for RPASS-4 scores. Second, there was no available content validity evidence for the five-part open-ended item. Third, the RPASS and the five-part open-ended item do not seem to measure exactly the same content domain. Only parts 4 and 5 of the open-ended item directly overlap with RPASS blueprint content. The first three parts measure slightly different content than the RPASS. Fourth, the low reliability of the RPASS limited the convergent validity correlations. This is evidenced by the fact that correcting the convergent coefficient for RPASS-4 or RPASS-5 attenuation yielded more moderate correlations. The maximum validity coefficient of RPASS-4 was .65 and the maximum validity coefficient for the RPASS-5 was .81 (the square root of reliability).

5.3.2 Reliability

A limitation of this study was the low reliability. The low reliability attenuated correlations of RPASS-4 and RPASS-5 scores with the convergent comparison measure. Pedhazur and Schmelkin (1991) suggest the following:

> [Lower] reliability coefficients are tolerable in early stages of research, that higher reliability is required when the measure is used to determine differences among groups, and that very high reliability is essential when the scores are used for

> making important decisions about individuals (e.g., selection and placement decisions). (p. 109)

However, as Cronbach and Shavelson (2004) warn, coefficient alpha can underestimate “the expected relationship between observed scores and true scores” for some specific tests (p. 403). Coefficient alpha may not be “strictly appropriate for many tests constructed according to a plan that allocates some fraction of the items to particular topics or processes” (Cronbach & Shavelson, 2004, p. 403). The RPASS was constructed in a stratified manner (i.e., correct conceptions and misconceptions) and this may have constrained internal consistency reliability.

5.4 Implications for Future Research

Future research should build on results for each of the two research questions. In terms of RPASS development, reliability improvement should be pursued. Improving RPASS reliability may also strengthen the correlations for construct-related validity evidence. For this study, reliability was estimated with Cronbach’s coefficient alpha to facilitate comparison of reliability across studies. To improve reliability, items that were judged to be redundant and removed from the scale might be reintroduced to lengthen the scale with items that correlate with the RPASS-5 content. Improving discrimination should also be pursued. Item discrimination can be influenced by items that are too difficult or too easy and, therefore, do not discriminate between those who generally understand the topic and those who do not. High item discrimination tends to yield higher reliability (Ebel, 1967).

Since the RPASS was constructed in a stratified manner, future research might investigate the construction of a stratified alpha formula. Alpha might be stratified by content area, learning goal, or the correct conception - misconception grouping. Internal consistency reliability may also be constrained by students' inconsistent reasoning on these kinds of items as discussed by Konold (1995). If inconsistent student reasoning across content areas limits the internal consistency reliability of scores, a test-retest (stability) correlation may be a better measure of score reliability than internal consistency.

Reliability may be improved by reducing measurement error related to item quality, administration, or motivation. First, to reduce measurement error related to item quality, items with weak (or negative) corrected item-total correlations should be reviewed to improve readability. Interviews with students who generally scored well on the RPASS but did not answer the low correlated items correctly might provide insight to improve item wording. For example, for Item 27 there could have been some confusion about wording of the item. This item was among the most difficult items and was only a marginal contributor to the RPASS-5 total score (corrected r_{pb} = .17). The item was intended to measure understanding of the impact of a larger sample size on the *P*-value if "all else remained the same." Most students responded that the *P*-value would "remain the same." This item might be reworded to read, "If the difference between the sample statistic and the hypothesized population parameter remained the same but the sample size was increased, what impact would there be on the *P*-value?" This rewording might be clearer. Second, there is unlikely to be measurement error reduction associated with RPASS administration since test instructions are all conveyed online rather than through a test administrator. There may, however, be some measurement error reduction related to student motivation. Efforts were made to motivate students by employing social exchange theory recommendations from The Tailored Design Method for mail and internet surveys (see Dillman, 2000). All respondents were offered extrinsic motivation to participate: extra credit, homework points, or points toward their final exam grade. However, students taking the test as part of a final exam grade may have been more motivated to do well.

Subsequent reliability studies should explore converting some of the two-option items to three-option items, where appropriate. While 73% of the two-option items had high corrected item-total correlations, four of the five three-option items (80%) were among the 15 most reliable items. Some of the research questions that might be pursued in future research about the development of the RPASS include exploring whether RPASS-5 measures a unitary construct; investigating if item content with lower item-total correlations differs from content with higher item-total correlations; having external experts rank item content in terms of importance; and having external subject matter experts map items to the learning goals: statistical literacy, statistical reasoning, and statistical thinking.

Another line of future research should use the RPASS to explore respondents' inferential reasoning. In future studies, Item Response Theory (IRT) may provide more information about respondents' abilities along with information about items. If a factor analysis of scores from the next version of the RPASS indicates it measures a unitary construct, then IRT may be a useful tool for analyzing results. Large samples of respondents are typically recommended for IRT estimation procedures to converge.

Administration of RPASS with and without instructional interventions may facilitate evaluating the effectiveness of new teaching approaches on inferential understanding. Some reform-based courses have integrated the *P*-value and statistical inference topics throughout the introductory course in order to improve students' inferential reasoning (e.g., Chance & Rossman, 2006; Lane-Getaz & Zieffler, 2006). Since random assignment of teaching methods is rarely feasible, concurrent administration of the CAOS test could provide a statistical control for a comparative study.

Future research questions about inferential reasoning might include exploring how repeated administration of the RPASS impacts student learning; investigating how inferential reasoning may be reflected in RPASS scores obtained before, during, and at the end of an introductory course; examining how students' and instructors' correct conceptions and misconceptions compare; or exploring what connections exist, if any, between students' understanding of random sampling and random assignment and their RPASS responses.

5.5 Conclusions

After multiple iterations of development and testing, strong content- and weak construct-related validity evidence was demonstrated for RPASS-4 and RPASS-5. However, reliability of scores remained too low to evaluate individual students' understanding (i.e., $\alpha < .70$). Even if higher reliability is achieved, the RPASS should not be used as a classroom assessment for assigning individual grades or diagnosing individual students' understanding or misunderstanding. The RPASS is intended for use as a research instrument to evaluate teaching methods with aggregate scores as the unit of measure, not individual students.

Relative to the first research question, the conclusions about instrument development seem clear. The content of the 15-item RPASS-5 appears to be sufficient for assessing introductory students' understanding. The higher internal consistency reliability of RPASS-5 ($\alpha = .66$) suggests RPASS-5 may behave like a unitary construct. A confirmatory factor analysis could explore this assertion. Most respondents had adequate exposure to these inferential concepts across disciplines, teachers, books, and teaching methods. Adding items that cover the same content as RPASS-5 should improve internal consistency reliability enough to use the instrument for research purposes ($\alpha \geq .70$). Even though the RPASS-4 content that was eliminated may be useful when assessing the understanding of respondents with stronger statistical backgrounds, the RPASS-5 item set appears to be more useful for assessing introductory students' understanding.

Relative to the second research question concerning student's inferential reasoning, the conclusions are more speculative. The Guidelines for Assessment and Instruction in Statistics Education (GAISE) urge educators to teach concepts and develop statistical literacy (ASA, 2005). RPASS results suggest that statistical literacy was stressed in these introductory courses and that literacy may be necessary to demonstrate statistical reasoning or thinking about inference. In addition to stressing statistical literacy, statistics educators might target instruction and assessment to address prevalent inferential misconceptions for introductory students ($\hat{p} \leq .55$), including misapplications of inferential logic, missing the link between study design and conclusions, and misinterpreting the *P*-value as the probability the null hypotheses is false. Statistics educators might engage students in discussions about the logic of inference. Students might be asked to routinely link study design (e.g., random sampling, random allocation, or sample size) to the interpretation of significant results. Differentiating frequentist interpretations of *P*-values and statistical significance from the subjective (or Bayesian) perspective, may help students discriminate *P*-values from the probability of hypotheses.

This research makes two contributions to statistics education research. First, evidence of introductory students' correct conceptions and misconceptions about *P*-values and statistical significance was documented using an instrument with reported psychometric properties. Second, with improved reliability RPASS could be used in a research program to link results across studies that examine students' inferential understanding as called for by Garfield (2006), Scheaffer and Smith (2007), and Shaughnessy (1992).

REFERENCES

Allen, K., Stone, A., Rhoads, T. R., & Murphy, T. J. (2004). The statistics concepts inventory: Developing a valid and reliable instrument. *Proceedings of the 2004 American Society for Engineering Education Annual Conference and Exposition* (pp. 1-15). Salt Lake City, UT. Retrieved May 14, 2007 from http://www.asee.org/acPapers/2004-301_Final.pdf

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

American Statistical Association. (2005). *GAISE college report.* Retrieved March 27, 2007, from http://www.amstat.org/education/gaise/GAISECollege.pdf

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423-437.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning*, *2*(1&2), 75-97.

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

Bennett, D. (1998). *Randomness*. Cambridge, MA: Harvard University Press.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistician, 37*(219), 325-335.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105.

Capraro, R. M. (2004). Statistical significance, effect size reporting, and confidence intervals: Best reporting strategies. A forum for researchers. *Journal for Research in Mathematics Education,* 57-62.

Chance, B. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education, 10*(3). Retrieved March 20, 2007, from http://www.amstat.org/publications/jse/v10n3/chance.html

Chance, B. L., delMas, R. C., & Garfield, J. B. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 295-323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Chance, B. L., & Rossman, A. J. (2006). *Investigating statistical concepts, applications, and methods*. Belmont, CA: Brooks/Cole – Thomson Learning.

Cobb, G. (1992). Teaching statistics. In L. A. Steen (Ed.), *Heeding the call for change: Suggestions for curricular action* (pp. 3-43). Washington, DC: The Mathematical Association of America.

Cobb, G. (2000). Teaching statistics: More data, less lecturing. In T. L. Moore (Ed.), *Resources for undergraduate instructors: Teaching statistics, MAA Notes* (Vol. 52; pp. 3-8). Washington, DC: Mathematical Association of America.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003.

Collins, L. B., & Mittag, K. C. (2005). Effect of calculator technology on student achievement in an introductory statistics course. *Statistics Education Research Journal, 4*(1), 7-15. Retrieved March 20, 2007, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ4(1)_Collins_Mittag.pdf

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions.* Thousand Oaks, CA: Sage Publications.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 391-418.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391-418.

Cumming, G., & Finch, S. (2005). Confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170-180.

Daniel, L. G. (1998). Statistical significant testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5*(2), 23-32.

delMas, R. (2002). Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education, 10*(3). Retrieved March 20, 2007, from http://www.amstat.org/publications/jse/v10n3/delmas_discussion.html

delMas, R., & Bart, W. (1989). The role of an evaluation exercise in the resolution of misconceptions of probability. *Focus on Learning Problems in Mathematics, 11*(3), 39-54.

delMas, R. C., Garfield, J. B., Ooms, A., & Chance, B. (in press). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*.

delMas, R. C., Ooms, A., Garfield, J. B., & Chance, B. (2006). Assessing students' statistical reasoning. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*. Voorburg, The Netherlands: International Statistical Institute. Retrieved April 27, 2007, from http://www.stat.auckland.ac.nz/~iase/publications/17/6D3_DELM.pdf

De Veaux, R. D., Velleman, P. F., & Bock, D. E. (2006). *Intro stats* (2nd ed.). Belmont, CA: Brooks/Cole – Thomson Learning.

Devore, J., & Peck, R. (2006). *Statistics: The exploration and analysis of data* (5th ed.). Belmont, CA: Brooks/Cole – Thomson Learning.

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method* (2nd ed.). New York: John Wiley & Sons, Inc.

Downing, S. M. (1992). True-false, alternate-choice, and multiple-choice items. *Educational Measurement: Issues and Practice, 11*(3), 27-30.

Ebel, R. L. (1965). Confidence weighting and test reliability. *Journal of Educational Measurement, 2*(1), 49-57.

Ebel, R. L. (1967). The relation of item discrimination to test reliability. *Journal of Educational Measurement, 4*(3), 125-128.

Ebel, R. L. (1970). The case of true-false test items. *The School Review, 78*(3), 373-389.

Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist, 53*(7), 798-799.

Falk, R., & Greenbaum, C. (1995). Significance tests die hard. *Theory & Psychology, 5*(1), 75-98.

Fischbein, E., & Gazit, A. (1983). Does the teaching of probability improve probabilistic intuitions? In D. R. Grey, P. Holmes, V. Barnett & G. M. Constable (Eds.), *Proceedings of the First International Conference on Teaching Statistics* (pp. 738-751). University of Sheffield: Organizing Committee of First International Conference on Teaching Statistics.

Fisher, R. A. (1951). *Design of experiments* (6th ed.). Edinburgh, England: Oliver and Boyd.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice, 11*(4), 21-26.

Frisbie, D. A., & Sweeney, D. C. (1982). The relative merits of multiple true-false achievement tests. *Journal of Educational Measurement, 19*(1), 29-35.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education, 10*(3). Retrieved March 20, 2007, from http://www.amstat.org/publications/jse/v10n3/garfield.html

Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1). Retrieved March 27, 2007, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ2(1).pdf

Garfield, J. (2006). Collaboration in statistics education research. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics.* Retrieved March 20, 2007, from http://www.stat.auckland.ac.nz/~iase/publications/17/PL2_GARF.pdf

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*(1), 44-63.

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematics Thinking and Learning, 2*(1&2)*,* 99-125.

Garfield, J. B., delMas, R. C., & Chance, B. (2005). *Tools for teaching and assessing statistical inference.* Retrieved October, 12, 2005, from http://www.tc.umn.edu/~delma001/stat_tools/

Good, P. I., & Hardin, J. W. (2003). *Common errors in statistics (and how to avoid them).* Hoboken, NJ: John Wiley & Sons.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.; pp. 65-110). Westport, CT: American Council on Education and Praeger Publishers.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*(1), 15-24.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1-20.

Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (*p*'s) versus errors (α) in classical statistical testing. *The American Statistician, 57*(3), 171-178.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson view in textbooks. *Journal of Experimental Education, 61*(4), 317-333.

Kahneman, D., & Tversky, A. (1982). Subjective probability: A judgment of representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 32-47). Cambridge: Cambridge University Press.

Kirk, R. E. (1996). Practical significance: A concept whose time as come. *Educational and Psychological Measurement, 56*(5), 746-759.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education, 3*(1). Retrieved March 20, 2007, from http://www.amstat.org/publications/jse/v3n1/konold.html

Lane-Getaz, S. J. (2006). What is statistical thinking and how is it developed? In G. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth yearbook* (pp. 272-289). Reston, VA: National Council of Teachers of Mathematics.

Lane-Getaz, S. J., & Zieffler, A. S. (2006). Using simulation to introduce inference: An active-learning approach. *2006 Proceedings of the American Statistical Association*, Statistical Education Section [CD-ROM]. Alexandria, VA: American Statistical Association.

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In B. Philips (Ed.), *Developing a statistically literate society: Proceedings of the Sixth International Conference on Teaching Statistics*. Voorburg, The Netherlands: International Statistical Institute. Retrieved March 27, 2007, from http://www.stat.auckland.ac.nz/~iase/publications/1/6c1_lips.pdf

Lipson, K., Kokonis, S., & Francis, G. (2003). Investigation of students' experiences with a web-based computer simulation. *Proceedings of the 2003 IASE Satellite Conference on Statistics Education and the Internet.* Berlin. Retrieved March 20, 2007 from http://www.stat.auckland.ac.nz/~iase/publications/6/Lipson.pdf

Liu, H. J. (1998). *A cross-cultural study of sex differences in statistical reasoning for college students in Taiwan and the United States*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Menon, R. (1993). Statistical significance testing (SST) should be discontinued in mathematics education research. *Mathematics Education Research Journal, 5*(1), 4-18.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

Messick, S. (1995). Validity of inferences from persons' responses and performances as scientific inquiry into score meaning**.** *American Psychologist*, *50*(9), 741-749.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(4), 14-20.

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

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

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*(1), 63-75.

National Research Council. (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1993). The use of statistical heuristics in everyday inductive reasoning. In R. Nisbett (Ed.), *Rules for reasoning*. Mahwah, NJ: Lawrence Erlbaum.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* Chichester, England: Wiley.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 17-46). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24(*2), 3-13.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57*(5), 416-428.

Rumsey, D. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education, 10*(3). Retrieved March 20, 2007, from http://www.amstat.org/publications/jse/v10n3/rumsey2.html

Saldanha, L. (2004). *Is this sample unusual?: An investigation of students exploring connections between sampling distributions and statistical inference.* Unpublished doctoral dissertation, Vanderbilt University, Nashville, TN. Retrieved March 27, 2007, from http://www.stat.auckland.ac.nz/~iase/publications/dissertations/04.Saldanha.Dissertation.pdf

Saldanha, L. A., & Thompson, P. W. (2002). Students' scheme-based understanding of sampling distributions and its relationship to statistical inference. In D. Mewborn (Ed.), *Proceedings of the Twenty-fourth Annual Meeting of the International Group for the Psychology of Mathematics Education* (pp. 1-7). Athens, GA.

Saldanha, L. A., & Thompson, P. W. (2006). Investigating statistical unusualness in the context of a resampling activity: Students exploring connections between sampling distributions and statistical inference. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*. Retrieved March 20, 2007, from http://www.stat.auckland.ac.nz/~iase/publications/17/6A3_SALD.pdf

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed.). Belmont, CA: Wadsworth Publishing Company.

Scheaffer, R., & Smith, W. B. (2007). *Using statistics effectively in mathematics education research: A report from a series of workshops organized by the American Statistical Association with funding from the National Science Foundation*. Alexandria, VA: American Statistical Association.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465-494). Reston, VA: National Council of Teachers of Mathematics.

Tempelaar, D. T., Gijselaers, W. H., & Schim van der Loeff, S. (2006). Puzzles in statistical reasoning. *Journal of Statistics Education*, *(14)*1. Retrieved March 20, 2007, from http://www.amstat.org/publications/jse/v14n1/tempelaar.html

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61*(4), 361-377.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools, 5*(1), 33-38.

Tversky, A., & Kahneman, D. (1982). Belief in the law of small numbers. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 23-31). Cambridge: Cambridge University Press.

Utts, J. M., & Heckard, R. F. (2004). *Mind on statistics* (2nd ed.). Belmont, CA: Brooks/Cole–Thomson Learning.

Vallecillos-Jimenez, A., & Holmes, P. (1994). Students' understanding of the logic of hypothesis testing. In J. Garfield (Ed.), *Research Papers from the Fourth International Conference on Teaching Statistics* (pp. 1-12). Minneapolis: University of Minnesota.

Wainer, H., & Robinson, D. H. (2003). Shaping up the practice of null hypothesis significance testing. *Educational Researcher, 32*(7), 22-30.

Watkins, A. E., Scheaffer, R. L., & Cobb, G. W. (2004). *Statistics in action: Understanding a world of data*. Emeryville, CA: Key Curriculum Press.

Wilkerson, M., & Olson, J. R. (1997). Misconceptions about sample size, statistical significance, and treatment effect. *The Journal of Psychology, 131*(6), 627-631.

Wilkinson, L., & APA Task Force on Statistical Significance. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604.

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

APPENDIX A

This appendix provides tabular summaries of key studies discussed in the literature review. Table A1 summarizes seven observational studies investigating understanding of this topic. Table A2 summarizes three empirical studies investigating instructional interventions to improve understanding. Table A3 summarizes the reliability and validity evidence for four statistical education research instruments.

Table A1

Observational Studies Investigating Understanding of P-values and Statistical Significance

Authors	Study focus	Sample size, *N*	Subjects or participants	Number of items or tasks	Contribution
Falk & Greenbaum (1995)	Misinterpretations of statistical significance and significance logic	53	University students with two previous statistics courses	5 T-F items	Demonstrated that merely citing or warning students of common misconceptions is insufficient for helping overturn them.
Haller & Krauss (2002)	Misinterpretations of statistical significance and significance logic	113	Methods professors, scientists, and students of psychology	6 T-F items	Extended Oakes (1986) study; showed both teachers and students share misinterpretations.
Mittag & Thompson (2000)	Perceptions of statistical significance	225	AERA members from 12 divisions	29 items rated on a 5-point scale	Provided contemporary snapshot of education researchers' perceptions of statistical significance.
Oakes (1986)	Misinterpretations of statistical significance and significance logic	70	Lecturers, psychologists, fellows, post-grads; 2+ years research experience	6 T-F items plus researchers' interpretation	Provided evidence of academic psychologists' misinterpretations; only three psychologists correctly marked all false.

Note. Six of the seven observational studies reviewed were quantitative, using surveys or questionnaires to collect data. Williams (1999) used mixed methods.

Table A1 continued

Authors	Study focus	Sample size, *N*	Subjects or participants	Number of items or tasks	Contribution
Vallecillos-Jimenez & Holmes (1994)	Belief that significant results *prove* the truth or falsity of one of the hypotheses	436	Cross-disciplinary students taking a theoretical or practical statistics course, per major	20 T-F items	Approximately a third of respondents reflected a belief that significance tests *prove* the truth or falsity of hypotheses; however, many misunderstood the item wording.
Wilkerson & Olson (1997)	Whether interpretations of significance tests reflect understanding of relationships between treatment effects, sample size, and Type-I and Type-II error	52	Graduate students pursing a masters, Ph.D., or Ed.D.	6 items	Demonstrated graduate-level researchers studied did not understand the influence of sample size on treatment effects but did link sample size influence on statistical significance.
Williams (1999)	Conceptual and procedural understanding of inference	18	Undergraduates in a university introductory statistics course	3 tasks	Concluded that introductory students studied had major problems expressing statistical ideas with accuracy that may mask conceptual knowledge.

Note. Six of the seven observational studies reviewed were quantitative, using surveys or questionnaires to collect data. Williams (1999) used mixed methods.

Table A2

Empirical Studies Investigating Interventions to Improve Inferential Understanding

Authors	Methodology	Sample size, *N*	Subjects or participants	Number of items or tasks	Contribution
Collins & Mittag (2005)	Quantitative, comparative study	Treatment (*n* = 22): Inference-capable graphing calculator No treatment (*n* = 47): Graphing calculator without inference capability	Undergraduates in a university-level introductory course in statistics	6 items on final exam	Concluded that the use of inference capable calculators did not appear to be related to student performance on inference-related exams.
Lipson, Kokonis, & Francis (2003)	Qualitative interview research	6	Undergraduates; after descriptive statistics and before learning inference	n/a	Suggested that students progress through four developmental stages when using computer simulations to make inferences.
Saldanha & Thompson (2002, 2006)	Teaching experiment methodology	27	11th and 12th graders in a non-Advanced Placement statistics course	n/a	Suggested that students who perceive a multiplicative conception of sampling develop a better understanding of inference.

Table A3

Existing Instruments in Statistical Education: Reliability and Validity

Authors	Instrument	Number of items	Number of inference items	Subjects or participants	Reliability		Validity	
					Procedure	Evidence	Procedure	Evidence
Garfield (2003); Liu (1998)	Statistical Reasoning Assessment (SRA)	20	0	Undergraduates in a first statistics course	Test-retest	.70	Content	Expert ratings
delMas, Ooms, Garfield, & Chance (2006)	ARTIST Test of Significance topic scale (TOS)	10	10	Undergraduates in a first statistics course	n/a [a]	n/a	Content	Expert ratings
delMas, Garfield, Ooms, & Chance (in press)	Comprehensive Assessment of Outcomes in a first Statistics course (CAOS)	40	14	Undergraduates in a first statistics course	Internal consistency	.82[b]	Content	Unanimous external expert validity ratings
Allen, Stone, Rhoads, & Murphy (2004)	Statistics Concepts Inventory (SCI)	32	n/a	Undergraduates in an engineering-targeted statistics course, and a math major-targeted statistics course	Internal consistency	.58 - .86[b]	Content	Engineering faculty survey; student focus groups
							Concurrent	Correlation with engineering course grades, no correlation with math course grades

Note. [a]n/a = not available , [b]Cronbach's coefficient α.

APPENDIX B

B.1 ARTIST Five-part Open-ended Item (Key) and Holistic Scoring Rubric

Parts 1 to 5 refer to the following situation:

The Northwestern University Placement Center in Evanston, Illinois, conducts a survey on starting salaries for college graduates and publishes its observations in The Northwestern Lindquist-Endicott Report. The two variables, "Liberal_Arts" and "Accounting," give the starting annual salaries obtained from *independent random samples* of Liberal-Arts graduates and Accounting graduates. Data are in thousands of dollars. This is a summary of the statistical test conducted for these data:

$$H_0 : \mu_{Liberal\_arts} - \mu_{Accounting} = 0$$

$$H_a : \mu_{Liberal\_arts} - \mu_{Accounting} \neq 0$$

$$t - score = -2.01$$

$$df* = 9$$

$$P - value = 0.0741$$

**Degrees of freedom*

1. What conditions should have been checked before conducting this test *and why*?

   <u>Conditions:</u>
   - *Was the sample an SRS?*
   - *Was there a sufficiently large sample size?*
   - *Did the distribution of these samples look normal (which implies the population may be normal)? Are there other reasons to suggest the population distribution may be normal (e.g., measurement error or natural variation)?*

   <u>Why do we check these conditions?</u> We check these conditions to see if we can apply the Central Limit Theorem which states that the sampling distribution of the means for independent random samples will behave normally, if either the population is normal or if the sample size is sufficiently large (e.g. n $\geq$ 30)

2. Informally state in words what these hypotheses mean in terms of this problem description.

   <u>The null hypothesis</u> states that the population mean of starting salaries for liberal arts graduates and the population mean of starting salaries for accounting graduates are equal; there is no difference between these two population means.

   <u>The alterative hypothesis</u> states that the population mean of starting salaries for liberal arts graduates and the population mean of starting salaries for accounting graduates are not equal; there is a difference between these two population means.

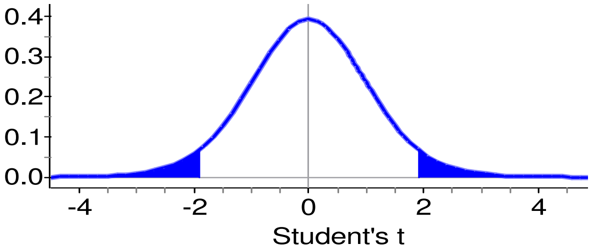
3. Given these hypotheses what type of test does it appear was performed?

   2-sample t-test, comparing means test or difference of means test is acceptable.

Appendix B.1 (continued)

Subject Number

4. a. Draw a sketch of the sampling distribution of the test statistic. Clearly identify the appropriate center of the sampling distribution. Using the information provided about the test, indicate the location of the observed test statistic, and shade the region that represents the *P*-value.



b. Interpret the *P*-value obtained.

Assuming that the mull hypothesis is true, the P-value of 7.4% is the probability of obtaining a |t| ≥ 2.01 or a t-test statistic this extreme or more extreme than that observed.

c. If the evidence against the null hypothesis were stronger, would you expect the *P*-value to be larger or smaller?

Smaller.

5. a. Given the results of the test that was performed what is your conclusion about these hypotheses? *Please state your conclusion in the context of the problem.*

Given the variability in the two samples, the difference between sample means is not large enough to conclude that starting salary is related to college major for graduates of Northwestern University.

It is customary in social science research to use .05 as a significance level to reject the null hypothesis. The P-value of .074 does not meet this criteria but remains of some interest, particularly with such a small sample size. Even though this evidence is not strong enough to reject the null, there may be a difference between starting salaries in the population. We would need a new sample with a larger sample size to make a more informed conclusion.

b. The article concludes that choosing an accounting major clearly causes one to earn a higher salary. Please critique the conclusion from the article.

*First of all, no *causal* conclusions can be drawn from an observational study. No variables were manipulated or controlled in this study. Secondly, if the P-value were small enough to reject the null, then we could at best say there is a link (association or relationship) between major and starting salary.*

Appendix B.1 (continued)

Holistic Scoring Rubric for the Five-part Open-ended Item

Tables B1 through B5 are the rating criteria used independently by each of the two raters.

Table B1

Criteria for Rating Open-ended Item Part 1

Criteria	Rating
Provides 3 conditions (SRS, sample size, population shape) and why	E
Provides 2 conditions (SRS, sample size) and why	P
Provides 1 or no condition (SRS, sample size or population shape)	I

Table B2

Criteria for Rating Open-ended Item Part 2

Criteria	Rating
Describes null hypothesis as no difference; alternative as difference	E
Describes null hypothesis as no difference	P
Provides any other answer	I

Table B3

Criteria for Rating Open-ended Item Part 3

Criteria	Rating
Lists 2-sample *t*-test, comparing means test or difference of means test	E
Lists *t*-test, significance test or *P*-value test	P
Lists any other answer	I

Appendix B.1 (continued)

Table B4

Criteria for Rating Open-ended Item Part 4

Criteria	Rating
Draws a distribution centered at 0, shows t = 2.01 and -2.01, Shades areas, interprets *P*-value, Smaller	E
Draws a distribution centered at 0, shows t = 2.01 and -2.01, Shades both areas, interprets *P*-value	E
Draws a distribution centered at 0, shows t = -2.01 and -2.01, Shades both areas, Smaller	P
Draws a distribution centered at 0, shows t = -2.01, Shades left area, interprets *P*-value	P
Draws a distribution centered at 0, shows t = -2.01, Shades left area	P
Provides any other answer	I

Table B5

Criteria for Rating Open-ended Item Part 5

Criteria	Rating
Concludes evidence is not strong enough to reject the null, no causal conclusions (not an experiment)	E
Concludes *P*-value is not small enough to reject null hypothesis (not < .05), no causal conclusions	E
Concludes *P*-value < .10 is of interest but not small enough to reject the null, no causal conclusions	E
Concludes evidence is not strong enough to reject the null, incorrect causality statement	P
Concludes *P*-value is not small enough to reject null hypothesis (not < .05), incorrect causality statement	P
Concludes *P*-value < .10 is of interest but not small enough to reject the null, incorrect causality statement	P
Concludes that the results depend on the significance level, which was not stated.	P

In general, an E is earned for an essentially correct answer, a P is earned for a partially correct answer, and an I is earned for an incorrect answer or no answer. Put simply, an E is one point and 2 Ps are required for one point as well.

B.2 ARTIST Bivariate Quantitative Data Topic Scale

1. Sam is interested in bird nest construction, and finds a correlation of .82 between the depth of a bird nest (in inches) and the width of the bird nest (in inches) at its widest point. Sue, a classmate of Sam, is also interested in looking at bird nest construction, and measures the same variables on the same bird nests that Sam does, except she does her measurements in centimeters, instead of inches. What should her correlation be?

   o Sue's correlation should be 1, because it will match Sam's exactly.

   o Sue's correlation would be 1.64(.82) = 1.3448, because you need to change the units from inches to centimeters and 1 inch = 1.64 centimeters.

   o Sue's correlation would be 82, the same as Sam's.

2. The correlation between height and weight for a certain breed of plant is found to be .75. What percentage of the variability in plant weight is NOT explained by height?

   o 1-.75 = 25 or 25%

   o $(.75)^2$ = .5625 or 56.25%

   o $1-(.75)^2$ = .4375 or 43.75%

   o $(1-.75)^2$ = .0625 or 6.25%

3. A student was studying the relationship between how much money students spend on food and on entertainment per week. Based on a sample size of 270, he calculated a correlation coefficient (r) of .013 for these two variables. Which of the following is an appropriate interpretation?

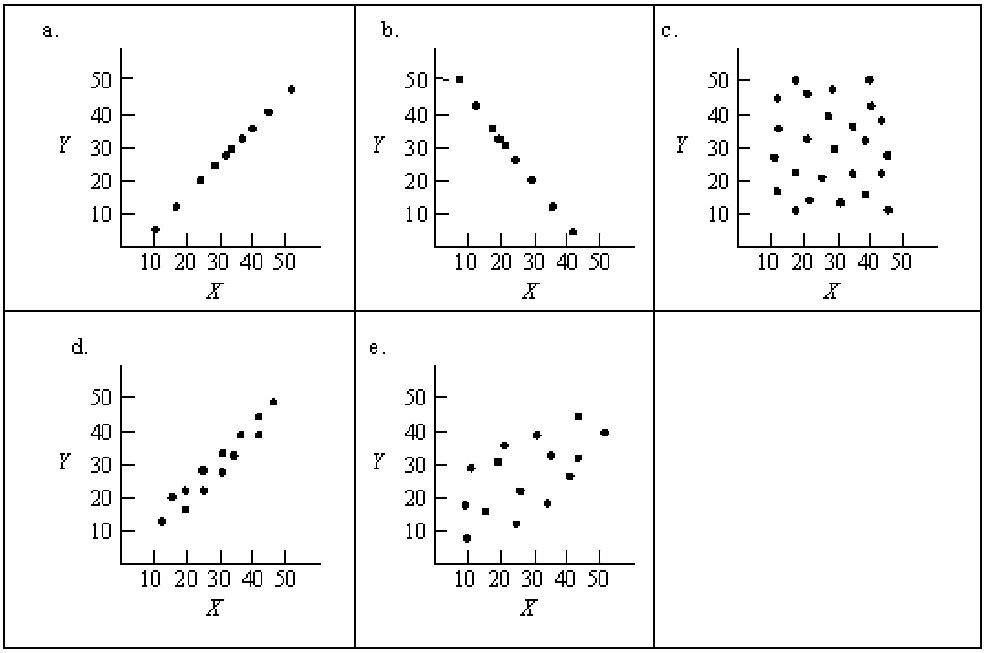
   o This low correlation of .013 indicates there is no relationship.

   o There is no linear relationship but there may be a nonlinear relationship.

   o This correlation indicates there is some type of linear relationship.

Appendix B.2 (continued)

Items 4 to 6 refer to the following situation:
Consider the five scatterplots that are shown below:



4. Select the scatterplot that shows a correlation of zero?

o a

o b

o c

o d

o e

5. Select the scatterplot that shows a correlation of about .60?

o a

o b

o c

o d

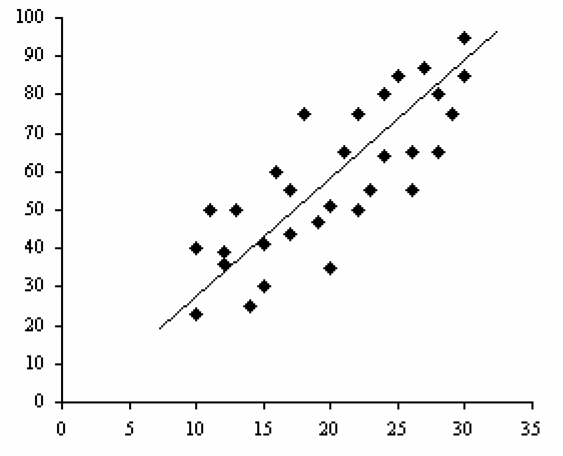
o e

Appendix B.2 (continued)

6. Select the scatterplot that shows the strongest relationship between the X and Y variables?

  - o a
  - o b
  - o a and b
  - o a and d
  - o a, b, and d

Items 7 and 8 refer to the following situation:
A statistics instructor produces the following scatterplot and regression line to see if her students' exam scores can be predicted from their scores on a standard test of mathematical ability.



7. What do the numbers on the horizontal axis represent?

  - o Statistics exam scores
  - o The number of people earning each exam score
  - o The response variable
  - o Mathematics ability scores

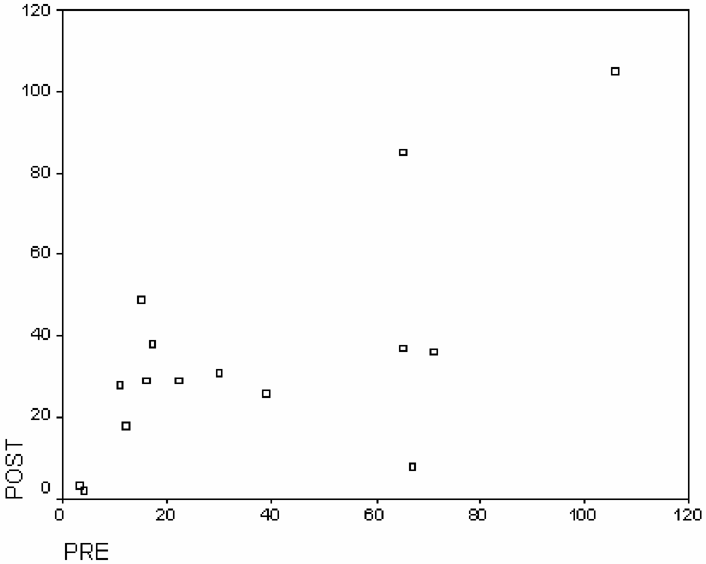
8. What do the numbers on the vertical axis represent?

  - o Statistics exam scores
  - o The number of people earning each exam score
  - o Mathematics ability scores

9. A random sample of 25 Real Estate listings for houses in the Northeast section of a large city was selected from the city newspaper. A correlation coefficient of -.80 was found between the age of a house and its list price. Which of the following statements is the best interpretation of this correlation?

   - o Older houses tend to cost more money than newer houses.
   - o Newer houses tend to cost more money than older houses.
   - o Older houses are worth more because they were built with higher quality materials and labor.
   - o New houses cost more because supplies and labor are more expensive today.

Items 10 to 12 refer to the following situation:
Dr. Jones gave students in her class a pretest about statistical concepts. After teaching about hypotheses tests, she then gave them a posttest about statistical concepts. Dr. Jones is interested in determining if there is a relationship between pretest and posttest scores, so she constructed the following scatterplot and calculated the correlation coefficient.



10. Which of the following is the best interpretation of the scatterplot?

   - o There is a moderate positive correlation between pretest and posttest scores.
   - o There is no correlation between pretest and posttest scores.
   - o All of the students' scores increased from pretest to posttest.

Appendix B.2 (continued)

11. Locate the point that shows a pretest score of 107. This point, which represents John's scores, is actually incorrect. If John's scores are removed from the data set, how would the correlation coefficient be affected?

    ○ The value of the correlation would decrease.

    ○ The value of the correlation would increase.

    ○ The value of the correlation would stay the same.

12. It turns out that John's pretest score was actually 5, and his posttest score was 100. If this correction is made to the data file and a new correlation coefficient is calculated, how would you expect this correlation to compare to the original correlation?

    ○ The absolute value of the new correlation would be smaller than the absolute value of the original correlation.

    ○ The absolute value of the new correlation would be larger than the absolute value of the original correlation.

    ○ The absolute value of the new correlation would be the same as the absolute value of the original correlation.

    ○ It is impossible to predict how the correlation would change.

13. A statistics student gathered data on a large numbers of cars of a particular model, from new cars to those that were up to 10 years old. Using the data on car ages (in years) and car prices (in US dollars) he found a linear relationship and produced the following regression model:<p>Predicted Price = 5620 - 440 * Age<p>A friend asked him to predict the price of a 5 year old model of this car, using his equation. Which of the following is the most correct response to provide?

    ○ Plot a regression line, find 5 on the horizontal axis, and read off the corresponding value on the y axis.

    ○ Substitute 5 in the equation and solve for "price".

    ○ Both of these methods are correct.

    ○ Neither of these methods is correct.

Appendix B.2 (continued)

14. A statistics instructor wants to use the number of hours studied to predict exam scores in his class. He wants to use a linear regression model. Data from previous years shows that the average number of hours studying for a final exam in statistics is 8.5, with a standard deviation of 1.5, and the average exam score is 75, with a standard deviation of 15. The correlation is .76. Should the instructor use linear regression to predict exam scores for a student who studied 10 hours for the final?

    ○ Yes, there is a high correlation, so it is alright to use linear regression.

    ○ Yes, because linear regression is the statistical method used to make predictions when you have bivariate quantitative data.

    ○ Linear regression could be appropriate if the scatterplot shows a clear linear relationship.

    ○ No, because there is no way to prove that more hours of study causes higher exam scores.

B.3 Items Used in this Study from the ARTIST Test of Significance Topic Scale

Four ARTIST Test of Significance (TOS) topic scale items were modified to develop RPASS-1A and RPASS-1B multiple-true-false item sets. The resultant RPASS-1B item set appears in Appendix D.1. The four items in this appendix were selected from the TOS topic scale because they were directly related to understanding *P*-values and statistical significance.

Choose the best answer to each of the following items:

1. A research article gives a p-value of .001 in the analysis section. Which definition of a p-value is the most accurate?
   a. the probability that the observed outcome will occur again.
   b. the probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true.
   c. the value that an observed outcome must reach in order to be considered significant under the null hypothesis.
   d. the probability that the null hypothesis is true.

2. If a researcher was hoping to show that the results of an experiment were statistically significant they would prefer:
   a. a large p-value
   b. a small p-value
   c. p-values are not related to statistical significance

3. It is reported that scores on a particular test of historical trivia given to high school students are approximately normally distributed with a mean of 85. Mrs. Rose believes that her 5 classes of high school seniors will score significantly better than the national average on this test. At the end of the semester, Mrs. Rose administers the historical trivia test to her students. The students score an average of 89 on this test. After conducting the appropriate statistical test, Mrs. Rose finds that the p-value is .0025. Which of the following is the best interpretation of the *p*-value?
   a. A *p*-value of .0025 provides strong evidence that Mrs. Rose's class outperformed high school students across the nation.
   b. A *p*-value of .0025 indicates that there is a very small chance that Mrs. Rose's class outperformed high school students across the nation.
   c. A *p*-value of .0025 provides evidence that Mrs. Rose is an exceptional teacher who was able to prepare her students well for this national test.
   d. None of the above.

4. A researcher conducts an experiment on human memory and recruits 15 people to participate in her study. She performs the experiment and analyzes the results. She obtains a *p*-value of .17. Which of the following is a reasonable interpretation of her results?

   a. This proves that her experimental treatment has no effect on memory.

   b. There could be a treatment effect, but the sample size was too small to detect it.

   c. She should reject the null hypothesis.

   d. There is evidence of a small effect on memory by her experimental treatment.

APPENDIX C

C.1 Online Consent Form for RPASS

You are invited to complete a test on "Reasoning about *P*-values and Statistical Significance." You were selected as a possible participant because you are currently taking or have taken post-secondary statistics courses.

Background Information:
The purpose of the RPASS instrument is to inform statistics education research about students' understanding and interpretation of *P*-values and statistical significance.

Procedures:
If you choose to participate, you will be asked to take a 15-20 minute test which includes five main sections: 1) Defining *P*-values, 2) Using Tests of Statistical Significance, 3) Interpreting Results, 4) Drawing Conclusions about Statistical Significance, and 5) Tying *P*-values back to Hypotheses.

Risks and Benefits of Being in the Study:
Although there are no known risks to participating, the potential benefit will be to contribute to improved statistics instruction in college classes. You may be able to earn extra credit in your statistics course as a participant in this study.

Confidentiality:
The test results will be used in the aggregate, so that you will remain completely anonymous and your results are confidential. Once your responses are entered into an electronic file, your original test results will be destroyed.

Voluntary Nature of the Study:
Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota, Cal Poly or other cooperating institutions. If you decide to participate, you are free to not answer any questions or withdraw at any time without affecting those relationships.

Contacts and Questions:
The researcher conducting this study is Sharon Lane-Getaz under the advisement of Professors Joan Garfield, Ph.D. (Educational Psychology--Statistics Education) and Michael Rodriguez, Ph.D. (Educational Psychology--Measurement & Evaluation). If you are willing to participate or have any questions you are encouraged to contact me, Sharon Lane-Getaz via my University of Minnesota, email: lane0139@umn.edu. You may also contact my measurement advisor, Michael Rodriguez, at 612-624-4324.

If you have any questions or concerns regarding the study and would like to talk to someone other than the researchers, contact Steve Davis, Chair of the Cal Poly Human Subjects Committee, at (805) 756-2754, sdavis@calpoly.edu, or Susan Opava, Dean of Research and Graduate Programs, at (805) 756-1508, sopava@calpoly.edu. You may also contact the Research Subjects' Advocate line at the University of Minnesota, D528 Mayo, 420 Delaware Street S.E., Minneapolis, Minnesota 55455; telephone (612) 625-1650.

C.2 In-depth Interview Protocol

1. Thank you for agreeing to participate in this interview today. I would like to tell you about this *Reasoning about P-values and Statistical Significance (RPASS)* project and what we will do.
2. This project is intended to gather information on how students understand and misunderstand concepts of *P*-value and statistical significance.
3. As you remember we took this RPASS assessment as a field test in our class.
4. Now to help me further understand how well the RPASS captures what you know about this topic, I want to have you read through one of the five scenarios you read about to respond to the RPASS questions (e.g., (1) Defining *P*-values, (2) Using Tests of Significance, (3) Interpreting Results, (4) Drawing Conclusions about Statistical Significance, and (5) Tying *P*-values back to Hypotheses.
5. With your written consent, I would like to audiotape our discussion from this point, so that I can review our discussion in greater detail later without having to stop and take very detailed notes now. You will not be identified personally in any of the work that I produce from this interview.

If student does NOT agree to sign they will be thanked for their time thus far and will be invited to go, no questions asked. If student DOES sign the consent form, then I will inform them that I will start the audiotape as soon as you are ready. I expect this interview will last no more than 1 hour. Do you have any questions or concerns at this time?

6. After you read through the scenario, I am interested in how you responded to each of the questions associated with the scenario and why you chose your response.
7. In order to better understand your reasoning I may ask a few probing or clarifying questions to make your perspective more clear for me.
8. Does this process sound acceptable to you?
9. Do you have any questions? If not, we can begin.

C.3 Expert Invitation Letter

April 1, 2006

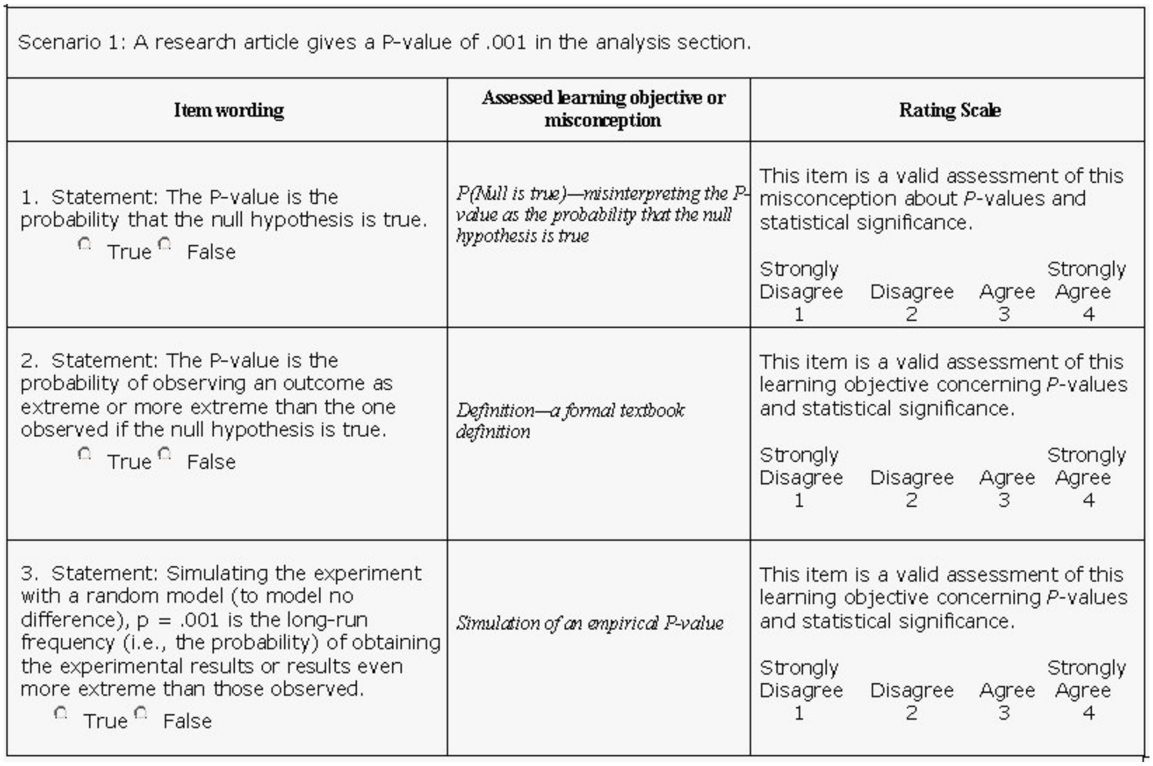
Professor, Ph. D.
Department of Statistics
University

Dear Professor,

I am continuing my dissertation research on how students understand *P*-values and statistical significance. With this letter I am formally soliciting your expert opinion on the current version of my research instrument now titled, *Reasoning about P-values and Statistical Significance* (RPASS). As a statistics educator your expert opinion on how these items measure students' understanding or misunderstanding of the construct is invaluable. Fourteen difficulties that students have understanding and interpreting *P*-values and statistical significance were culled from the research literature as the basis for this study. These difficulties and some proper conceptions were used to identify and/or develop the RPASS item set. As an expert rater you are being asked to assess the validity of the RPASS instrument in relation to these specific learning objectives and misconceptions.

If you are willing to participate, please email me to confirm your interest at slanegetaz@msn.com. Later this spring—after readability modifications are completed from student interview data—you will receive the expert rater packet. The rater packet is organized much like the RPASS instrument with 5 main sections: 1) Defining *P*-values, 2) Using Tests of Statistical Significance, 3) Interpreting Results, 4) Drawing Conclusions about Statistical Significance, and 5) Tying *P*-values back to Hypotheses. Each of the 5 sections starts with a context or scenario that students read and (4-7) questions to which they respond. As you can see in the sample page below, column 1 contains the RPASS "item wording." Column 3 contains the "rating scale" you complete to rate the degree to which the item is a valid measure of the stated learning objective or misconception that appears in column 2.

Appendix C.3 (continued)



Scenario 1: A research article gives a P-value of .001 in the analysis section.

Item wording	Assessed learning objective or misconception	Rating Scale
1. Statement: The P-value is the probability that the null hypothesis is true. ○ True ○ False	*P(Null is true)—misinterpreting the P-value as the probability that the null hypothesis is true*	This item is a valid assessment of this misconception about *P*-values and statistical significance. Strongly Disagree 1 · Disagree 2 · Agree 3 · Strongly Agree 4
2. Statement: The P-value is the probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true. ○ True ○ False	*Definition—a formal textbook definition*	This item is a valid assessment of this learning objective concerning *P*-values and statistical significance. Strongly Disagree 1 · Disagree 2 · Agree 3 · Strongly Agree 4
3. Statement: Simulating the experiment with a random model (to model no difference), p = .001 is the long-run frequency (i.e., the probability) of obtaining the experimental results or results even more extreme than those observed. ○ True ○ False	*Simulation of an empirical P-value*	This item is a valid assessment of this learning objective concerning *P*-values and statistical significance. Strongly Disagree 1 · Disagree 2 · Agree 3 · Strongly Agree 4

After you rate each item individually, you will be asked to rate the RPASS document as a whole, and to suggest how to improve any items for which you "strongly disagreed" or "disagreed" that the item validly assessed the objective or misconception stated. There is a space provided for comments on what concepts may be missing or what can be removed, and any other suggestions you may have regarding the instrument.

Should you agree to participate, you will need to reach your conclusions independently so that your ratings can be included in the research project data. The turnaround for the document will be 2 weeks. Feel free to ask me questions as they arise. I sincerely hope that you will be able to contribute to my research.

Your time and input is greatly appreciated!

Sincerely,

Sharon Lane-Getaz
Ph. D. Candidate, University of Minnesota

C.4 Expert Rater Packet (Instructions for First Round of Ratings)

Reasoning about *P*-values and Statistical Significance: Expert Rater Packet

Welcome to the content validity rating of the "Reasoning about *P*-values and Statistical Significance" (RPASS) assessment. The purpose of the RPASS instrument is to inform statistics education research about students' understanding and interpretation of *P*-values and statistical significance. The RPASS has 5 main sections:

1) Defining *P*-values,
2) Using Tests of Statistical Significance,
3) Interpreting Results,
4) Drawing Conclusions about Statistical Significance, and
5) Tying *P*-values back to Hypotheses.

Each of the 5 sections has a context or scenario to be read and a series of items that students classify as true or false (valid or invalid).

As an expert rater, you are contributing to the validation of this instrument. Please rate the content validity of each item individually and finally, the RPASS instrument as a whole. At the beginning of each section is a scenario. Following each scenario are 4-7 items to be rated.

For each RPASS item there are three columns on this form:

Column a. contains the RPASS "Item wording,"

Column b. is the "Assessed learning objective or misconception" related to the item, and

Column c. is the "Validity rating" scale you complete by circling the rating that corresponds to the degree to which you believe the item is a valid assessment of the stated learning objective or misconception.

If there are any items for which you "Strongly disagree" or "Disagree" that the item is a valid assessment; i.e., you rate the item with a 1 or 2, please explain in the space provided at the end of this form why the item does not assess the stated learning objective or misconception and suggest how the item might be improved.

Your time and input is greatly appreciated! Feel free to ask me questions as they arise. If you prefer, the rating packet can be completed online. Email me for a link and access code for the online Expert Rater form.

Please return the completed document to me within two weeks.

C.5 Student E-mail Invitation

Dear student of statistics,

I am both on faculty here at California Polytechnic State University and am a Ph.D. candidate at the University of Minnesota. I am inviting you to participate in my dissertation research, "Reasoning about P-values and Statistical Significance (RPASS)." I am collecting data to inform statistics education research about students' understanding and interpretation of this concept. I would appreciate it if you would take 20-30 minutes to complete the online assessment—completely anonymously. Whether or not you participate is optional from my perspective; however, your instructors may choose to tie this activity to a course requirement or award extra credit for your participation.

At the online site, you will see the consent form, which I encourage you to review before completing the assessment. You can access the RPASS assessment by clicking on the URL below:

http://s.education.umn.edu/COST/TakeSurvey.asp?SurveyID=5238mlL18om5G

Thanks for considering this request -- I hope you can help me in my research effort! Once you come to the lab I will provide a code for you to use during the session.

Sincerely,

Login Sheet

Sequence Number ____

Welcome to the studio lab!

Press CTRL-Alt-Delete to login to the local network.
The student ID is typically displayed and you need only enter the password:

&*)(*#&*)

Your instructor has sent you an email with a URL to the assessment site. Please access the URL and navigate to the site. The test monitor will give you a code to use to access the assessment that will be used for passing a grade back to your instructor.

Access Code ________

Thank you for participating!

APPENDIX D

D.1 17-item RPASS-1B Instrument

The RPASS-1B instructions, four problem scenarios, associated multiple-true-false item sets, and confidence rating scales:

Defining *P*-values

Based on your opinion, please click the circle next to "True" or "False" to indicate whether you think the following definitions are true or false. Then, enter a value from 0 (not confident) to 100 (completely confident) indicating how confident you feel about your decision.

Scenario 1:
A research article gives a *P*-value of .001 in the analysis section. Do you think the following definition is true or false?

1. The *P*-value is the probability that the null hypothesis is true.
   □ True □ False
   Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
   The value must be between 0 and 100, inclusive.
   ______

2. The *P*-value is the probability of observing an outcome as extreme as or more extreme than the one observed if the null hypothesis is true.
   □ True □ False
   Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
   The value must be between 0 and 100, inclusive.
   ______

3. If a simulation of the experiment were conducted, the *P*-value of .001 is the long-run frequency of obtaining the experimental results or something more extreme due to chance.

   □ True □ False

   Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
   The value must be between 0 and 100, inclusive.

   ______

4. A *P*-value tells me the chances are 1 in 1000 of observing data this surprising (or more surprising) than what I observed, if the null hypothesis is true.

   □ True □ False

   Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
   The value must be between 0 and 100, inclusive.

   ______

5. The *P*-value is the proportion of a population (.1%) that has a particular characteristic of interest.

   □ True □ False

   Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
   The value must be between 0 and 100, inclusive.

   ______

Appendix D.1 (continued)

Using *P*-values

Based on your opinion, please click the circle next to "Valid" or "Invalid" to indicate whether you think the following conclusions are valid (true or correct) or invalid (false or incorrect). Then, enter a value from 0 (not confident) to 100 (completely confident) indicating how confident you feel about your decision.

Scenario 2:

The district administrators of an experimental program similar to Head Start are interested in knowing if the program has had an impact on the reading readiness of first graders. Assume that the historical, pre-implementation mean Reading Readiness score for all first graders is 100 and the population standard deviation is 15. A random sample of current first graders who have been through the preschool program scored a mean Reading Readiness of 102.

6. The district researchers would have pre-determined the number of students to sample to ensure their *P*-value could detect if the improvement is attributable to the Head Start-like program.

   ○ Valid ○ Invalid

   Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
   The value must be between 0 and 100, inclusive.

   ______

7. The district researchers should determine how often they would obtain a score of 102 or higher just by chance to "definitively prove" whether the program had a positive impact.

   ○ Valid ○ Invalid

   Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
   The value must be between 0 and 100, inclusive.

   ______

8. After checking the necessary conditions, the district researchers should proceed to determine if random chance “caused the results observed.”
   □ Valid □ Invalid
   Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
   The value must be between 0 and 100, inclusive.
   ______

9. The district researchers should compare the sample group’s mean to its sampling distribution based upon assuming the population mean is 100.
   □ Valid □ Invalid
   Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
   The value must be between 0 and 100, inclusive.
   ______

Appendix D.1 (continued)

Interpreting *P*-values

Based on your opinion, please click the circle next to "Valid" or "Invalid" to indicate whether you think the following interpretations are valid (true or correct) or invalid (false or incorrect). Then, enter a value from 0 (not confident) to 100 (completely confident) indicating how confident you feel about your decision.

Scenario 3:

An ethical researcher is hoping to show that his new hair growth treatment had statistically significant results. How should this researcher interpret results from the research study?

10. Assuming the hair treatment had no effect, the researcher should interpret the rareness of obtaining his research results due to chance.

    ○ Valid ○ Invalid

    Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
    The value must be between 0 and 100, inclusive.

    ______

11. The researcher should interpret the results as statistically significant as long as the test statistic for his hair growth treatment is less than the .05 significance level (alpha).

    ○ Valid ○ Invalid

    Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
    The value must be between 0 and 100, inclusive.

    ______

12. If the results from the hair growth treatment are statistically significant, the researcher should interpret his small *P*-value to mean the hair growth treatment "caused" the hair growth observed in the study.

    ○ Valid ○ Invalid

    Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
    The value must be between 0 and 100, inclusive.

    ______

13. The researcher should interpret a large *P*-value for his hair growth treatment to mean that the treatment effects cannot be attributed to chance.

    ○ Valid ○ Invalid

    Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
    The value must be between 0 and 100, inclusive.

    ______

Appendix D.1 (continued)

Drawing Conclusions from *P*-values

Based on your opinion, please click the circle next to "Valid" or "Invalid" to indicate whether you think the following conclusions are valid (true or correct) or invalid (false or incorrect). Then, enter a value from 50 (50% confident, just guessing) to 100 (100%, completely confident) indicating how confident you feel about your decision.

Scenario 4:

A researcher conducts an appropriate hypothesis test where she compares the scores of a random sample of students' SAT scores to a national average (500). She hopes to show that the students' mean score will be higher than average. The researcher finds a *P*-value for her sample of .03.

14. The researcher concludes that there is a 97% chance that repeating her research will yield the same or similar results.

    ○ Valid ○ Invalid

    Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
    The value must be between 0 and 100, inclusive.

    ______

15. The researcher concludes that there is a .03 probability that her research hypothesis (that the students have higher than average scores) is true.

    ○ Valid ○ Invalid

    Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
    The value must be between 0 and 100, inclusive.

    ______

Appendix D.1 (continued)

16. The researcher concludes that there is only a 3% probability that her research hypothesis (that there is no difference between population means) is wrong.

    □ Valid □ Invalid

    Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
    The value must be between 0 and 100, inclusive.

    ______

17. The researcher concludes that there remain 3 chances in 100 that the observed results would have occurred even if the SAT preparation program had no effect.

    □ Valid □ Invalid

    Indicate your level of confidence in your decision. (Enter 0 = "not confident" to 100 = "completely confident.")
    The value must be between 0 and 100, inclusive.

    ______

D.2 In-depth Interview Results: Selected Notes from Five Interviews

Tables D1 through D4 contain selected notes from the five in-depth interviews conducted during March 6 – 9, 2006 (ID: 304, 322, 124, 118, and 123). For each table the first column is the student's identification code. The second column is the RPASS-2 item number and the short name for the conception or misconception being assessed. The last two columns report comments or notes from the student or interviewer, respectively.

Table D1

In-depth Interview Notes 1 of 5, with No RPASS-2 Score Available, ID: 304

In-depth Interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
ID: 304	1. Probability: null is true	"I marked it TRUE. Well, the way I looked at it is…that's completely wrong. …The smaller the *P*-value the more evidence that the null is false. I think I got it mixed up. There's nothing wrong with that problem, I just read it wrong."	
	3. Simulation definition	"Pretty wordy. … You gotta take a long time to read and dissect this one. Do you want me to figure out this one now? I might have guessed on that one."	Item seems to be wordy when read aloud by students.
	5. Sample and population	Student selected TRUE. "Easy... Yeah, you have to look at the practical."	Answered incorrectly.
	7. Inverse is true 8. Chance as cause of results observed	Student underlined "definitely prove" and "cause of the results observed." "You can't prove anything, it's just evidence."	Student liked the quotes around "definitely prove" and "cause of the results observed" for emphasis.

Note. No RPASS-2 test score available.

Table D1 (continued)

In-depth Interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
ID: 304	14. *P*-value as always low	Student underlined “gone bad.” Discussed whether “gone awry” or “gone bad” was better phrasing. “Does this mean ‘error in computations or something?’” “It’s not necessarily ‘gone bad’ it could be there’s no difference between his hair growth treatment and the placebo.”	“Large *P* means it’s going to support the null hypothesis.” The wording on this item remains problematic.
	19. Probability: null is false; Inverse as true	“This is a wordy question.” Student underlined “absolutely disproved the null hypothesis.” “You can stop right there.” Student struck out “which assumed there is no difference between the control and experimental group means in the population.”	Need to simplify the question wording. Item also appears to be measuring two different difficulties. *P*(Null is false) and Inverse as true (proof).
	23. Probability: alternative is false	“It seems too personal for a question on a test.” Student suggested moving “you know” at the end of the first phrase.	As written this item was misinterpreted. “You know” was perceived as a casual use rather than the intended “you understand.”
	25. *P*-value dependence on alternative	Student drew two normal-shaped distributions shading both tails on one distribution and one tail of the other. Student selected TRUE. “Cause you’ve got 2 areas here and here as opposed to one area.”	

Note. No RPASS-2 test score available.

Table D2

In-depth Interview Notes 2 of 5, RPASS-2 Score 14 / 25, ID: 322

In-depth Interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
ID: 322	n/a. Confidence scale	After being prompted about the confidence rating the scale the student responded, "I thought the confidence stuff was OK. It was just there. …I don't think I ever chose 50%, and 100% only once."	It may be better to have a more evenly distributed scale with more of a range on the top and bottom of the scale. Few students have chosen the extremes of the scale.
	n/a	"Honestly I've never heard of Head Start before. I kind of figured it was a way to improve reading is that what it is? … It's worded pretty well."	Remove reference to "Head Start."
	7. Inverse as true	"VALID because …" "Quotes kind of imply that it's kind of bogus or something? I didn't think anything of the quotes until you pointed them out."	"I am not sure if quotes are the right thing to use. I want to draw people's attention to these words."
	8. Chance as cause of results observed	"I don't really know about that one. What do you mean by checking the necessary conditions?" …. "I guess it's kind of implied because in class we always did the checking conditions. …" "The quotes don't really change my answer."	"Conditions to be able to conduct a test of significance." "Did the quotes around 'cause of the results observed' jump out at you?"
	10. Confidence interval and significance	Student stopped suddenly at the end phrase "in relation to the population mean." "…OK that's VALID, I guess."	Edit the wording for this question.

Note. 56% correct on RPASS-2.

Table D2 (continued)

In-depth Interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
ID: 322	11. *P*-value as rareness measure	Student read the question twice…pregnant pause: “What gets me on this one is ‘assuming the hair growth treatment had no effect.’ …You’re using it as the null hypothesis?”	Avoiding the hypothesis testing language may add confusion.
	13. Converse is true	“INVALID because there could be other things going on there. Actually no, if that’s what he’s studying and it’s the only variable.”	“… If I were to say the researcher controlled for the other variables then this would be VALID. If I say that this is an observational study…then INVALID.”
	14. *P*-value as always low	Student tripped on the word “awry” when reading the item aloud. “That was one I pretty much just guessed on. I wasn’t sure what you meant by…”had gone awry.” “Calculation error might make more sense.”	“The researcher expected a small *P*-value and therefore he thought he made a ‘calculation error.’ However, a large *P*-value could be the right answer. This evidence supports the null hypothesis.”
	16. Probability: alternative is true	“I think it’s VALID. I think it’s a different way to think about it.”	“Would it make a difference if it said null and alternative versus ‘research hypothesis?’”
	18. Type I / alpha and *P*-value	“I guess I’d say that’s VALID. I guess it’s just another one that’s kind of wordy but I don’t know what I could do to change it because it gets to the point it just has a long….”	Seems wordy.

Note. 56% correct on RPASS-2.

Table D3

In-depth Interview Notes 3 of 5, RPASS-2 Score 12 / 25, ID: 124

In-depth Interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
ID: 124	11. *P*-value as rareness measure	"I was trying to think of what the null hypothesis would be in this scenario. *Ho:* That it worked; *Ha:* It didn't work. Would the *P*-value represent the percent of people that it affects?" "A small *P*-value would support the null hypothesis and a large one wouldn't." "It helps me to see the language in problems."	This student repeatedly confused the null and alternative hypotheses. The student thought it would help to have the null hypotheses spelled out in these kinds of questions. The student had mis-remembered the decision criteria for rejection of the null hypothesis.
	12. Test statistic and *P*-value	"I remembered from class that it was less than the significance level, then we could consider it significant.. .05 or less was supporting the null hypothesis, anything greater than that I could refute it because it's above that level." "The change would be helpful if you didn't know what test statistics were." "The *z* has to do with standard deviation, what the value's distance from the mean is."	Student didn't notice test statistic was used to evaluate against significance level rather than *P*-value. Suggested adding "*t* or *z*" after the word test statistic to see if she would change her answer. Adding this fact did not seem to help the student see the problem as intended.

Note. 48% correct on RPASS-2.

Table D3 (continued)

In-depth Interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
ID: 124	13. Converse is true	Student said if it was statistically significant and less than .05, then it "would have caused the hair growth." … "I was looking at it in terms of these two statements being true. If they are statistically significant then I'd say he could interpret the *P*-value." Student drew a normal-shaped distribution to represent the sampling distribution representing the null hypothesis and shaded the right tail region. The student said this drawing indicated the data would "support the null hypothesis that the treatment made a difference."	Student was using information from previous Items 11 and 12 (to answer this question. This could have a domino effect on wrong answers. Perhaps the test directions need to explicitly say that each of the items are independent of the others. The students' confusion between null and alternative hypotheses seems to confound her misunderstandings.
	15. Reliability and *P*-value	"The *P*-value is .03, isn't her *P*-value…? She was saying that this program was going to improve SAT scores. … The alternate said it will." "…The null would be that mean for those who complete the program would be about the same. That the program wouldn't dramatically help students do any better on the SAT. You've got .03 …I'm going on the idea that .05 or less supports the null hypothesis." " …For some reason it makes me think of confidence intervals…I think that it's the whole 97% issue that's doing it. …This one I'm not really sure I'm understanding correctly on the question."	I prompted the student to clarify what she thought the null and alterative hypotheses were before answering the question. Again the student has mis-remembered the decision criteria. "We actually don't talk about this concept in class."

Note. 48% correct on RPASS-2.

Table D3 (continued)

In-depth Interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
ID: 124	16. Probability: alternative is true	"So, is the research hypothesis supposed to be the alternate or the null?" "I think that this would be INVALID…The null would be that it would be that it's not goin' to have an impact and since the *P*-value is so small, that it would support the null."	Perhaps the wording should be changed to simply say that "the alternative hypothesis?' Student continues to confuse the null and alternative hypothesis in her interpretation and decision criteria.
	17. Probability: alternative is false	"I put VALID because I was thinking of it as….Oh, but the research hypothesis is the alternative, so that would change my answer. Doesn't the *P*-value represent that there's a 3% chance the null hypothesis is true?" "If there's a 3% chance that the alternative is wrong. There's a 97% chance that the alternative is accurate."	Student indicates multiple misconceptions and language confusion that confound her results.

Note. 48% correct on RPASS-2.

Table D4

In-depth Interview Notes 4 of 5, RPASS-2 Score 17 / 25, ID: 118

In-depth Interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
ID: 118	3. Simulation definition	"Simulation just means like doing like a model, redrawing a sample or modeling doing something like that? … What do you mean by 'long run frequency?'" "I'd say it's TRUE."	Perhaps the word 'probability" should be used in lieu of 'long run frequency."
	4. Lay definition	"Are we thinking 'surprising' is the same as 'extreme?'" "Rare is not the same as extreme but if the *P*-value 1 in 100 it is very rare and we have to reject the null? I like rare or more rare I think it's TRUE. Because if you think about it extreme would be rare.	"What if I use 'rare' instead of 'surprising?'"
	6. Sample size and significance	"Does 'pre-determined' mean that it wasn't a random sample? I think that's true, that's VALID. They are just picking a sample size."	"I'm trying to say that they planned ahead and figured out what would be a good sample size."
	7. Inverse is true	"INVALID they can't 'definitely prove' something. "The quotes help me. [As an alternative] I'd go with underline or bold."	Unclear whether to keep quotes around phrases that help the student to understand the purpose of the item.

Note. 68% correct on RPASS-2.

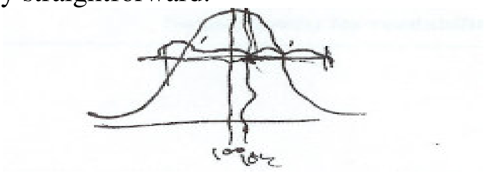
Table D5

In-depth Interview Notes 5 of 5, RPASS-2 Score 16 / 25, ID: 123

In-depth Interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
ID: 123	6. Sample size and significance	"This is probably gonna be VALID." "When you had really large samples it kind of messed up. It made it so that you could kind of get a certain result, if you used, like a thousand for your 'n.' You wouldn't randomly decide how many to have in your sample. You randomly choose them."	
	7. Inverse is true	"…I don't know exactly what 'definitively prove' would mean." "… We aren't proving anything with these. I think I wanted say INVALID." "It is clear that there are 2 parts and the second part is about definitively prove. Perhaps if you say 'fully prove.'" "Even just say 'prove.'"	"If I put quotes around it some people may interpret it to mean that I don't really mean definitively prove." Remove quotes and "definitively."
	8. Chance as cause of results observed	"We don't determine cause and effect but that is not what this is asking. This is random chance was the 'cause of the results observed.'" "…I don't think they say 'yes, random chance caused these results' or 'no, random chance caused the results.' They are gonna say 'this is the probability that random chance would cause these results' and that probability is strong enough for us to make this relationship." "It's probably INVALID."	

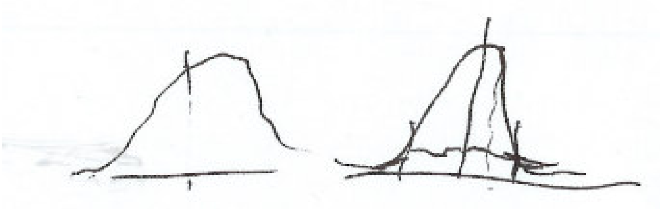
Note. 64% correct on RPASS-2.

Table D5 (continued)

In-depth Interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
ID: 123	9. *P*-value embedded in sampling variation	That makes sense and I guess I could draw it to explain what I'm thinking. …There's the population of all your first graders and … it's a given in this question that the mean is 100. …And so if I'm the researcher one of the things, once I take my sample the first thing I'm gonna do is find the sample's mean and plot it in this distribution. It was 102, so … and that's pretty straightforward."  	
	10. Confidence interval and significance	"..okay, so the 95% confidence interval for the sample mean … is gonna be… you have from 102, you'll make an interval around that. It's gonna look, probably something like that [see above]. … and if the interval captures the population mean (which in my rough sketch it probably would), …then this is equivalent to a two-tailed test to see if the sample mean is statistically significant at the .05 level. Um, I think it's equivalent in con—in what it says about the 95% confidence interval and statistically significant at the .05 level for the two-tailed test…. But it's not completely equivalent, they're two different tests because they say it in a slightly different way but the conclusions of them are equivalent. …The difference is that with the confidence interval you can tell how close you are, where here you can just tell where you're at. So, I'm gonna say that this is VALID."	

Note. 64% correct on RPASS-2.

Table D5 (continued)

In-depth Interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
ID: 123			"Are you thinking of standard errors?"
		"…The distance between the mean of the sample and the mean of the population are at least 2 standard errors for both of them [if the sample is in the tail]."	
	12. Test statistic and *P*-value	"The researcher interprets the results as statistically significant and the *t*- or *z*-test statistic <= .05." "…Yes, this makes sense so far…if the *t* is less than .05. If your mean is exactly on that spot, does that mean you are in it or not?" "…Oh, you mean "how many standard errors"? I'd say it would have to be 2 or greater." "…I remember it should be .01 if you are doing medical treatment but for a hair growth treatment…."	"If you wanted to give me a *t*-test statistic that deviated from the sampling distribution mean…"
	14. *P*-value as always low	"Even if he thought there was something wrong with the research. I learned this with the project. This is the data you have and you can't get rid of it. You still have to use it and take note of it."	

Note. 64% correct on RPASS-2.

D.3 Cognitive Interview Results: Selected Notes from Eight Interviews

Table D6 contains selected notes from the first three cognitive interviews that utilized RPASS-2 (March 6-7, 2006). Table D7 contains notes from the next five cognitive interviews testing RPASS-3A (March 8-9, 2006). For both tables the first column is the student's identification code. The second column is the RPASS item number and the conception or misconception being assessed. The last two columns are the student and interviewer comments or notes, respectively.

Table D6

Cognitive Interview Notes Using RPASS-2, CI: 230, 107, 210

Cognitive interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
CI: 230	3. Simulation definition	"I have no idea what that means." "'Long run frequency and 'simulation.' What are you simulating?"	
	6. Sample size and significance	"How can you predetermine a simple random sample? The number of people in the sample you can predetermine that? They don't say which ones."	The word "predetermined" seems to be distracting student from the intended content.
	14. *P*-value as always low	"A small *P*-value means rare. A large *P*-value means this is a natural occurrence; it happens often. He's testing to see if … if the hair treatment had no effect on his hair."	
		"…I don't understand what you meant by the experiment has 'gone bad.' Meaning that … the hair stuff didn't work or it did work."	So you think I should say that the hair growth treatment "did work or didn't work," instead of saying the "experiment has gone bad"?
CI: 107	2. Textbook definition	"'As extreme …as' or more extreme …"	The phrase "as extreme as" seems awkward. Student paused twice on the second "as."

Table D6 (continued)

Cognitive interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
CI: 107	3. Simulation definition	“The ‘long-run frequency’ is not really understandable.”	Change the wording to “probability” since long-run frequency term may confound the intended content to be measured.
	5. Sample and population	“I don’t really understand the question.” “Oh… Okay. Does it represent the population or the samples? I think there needs to be a little more in depth about the question.”	“What is it about this question that is difficult to understand?” Question seems to get at the intended content.
	6. Sample size and significance	“So it’s, the question is saying that kids scored, the mean scored above 100? …Predetermined? So, they picked out students? It sounds like they picked out students that had good scores.”	“Did they pick out the students or determined the *number* of students?” “They predetermined ‘*how many* students to sample.’ Is that better stated?”
	7. Inverse as true	“Valid or Invalid? … This is weird because it just says the district researchers determined how’ but it doesn’t say HOW. They just show one random sample of 102 and it doesn’t show like. That’s only one test, you can’t really prove just by that. It doesn’t say how many times or how often they would attain that score.” “Well, I guess they could determine if they could get one sample, I guess they could get more.” “[The quotation marks] mean that’s it’s a sure thing. …That’s it’s a good program for first graders.”	“See where I have the quotation marks there? …What does that mean to you?” [The quotation marks were around the word “prove.”]

Table D6 (continued)

Cognitive interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
CI: 107	9. *P*-value embedded in sampling variation	“ ..That sounds valid. I have no idea what it’s trying to…. Okay, yeah. It doesn’t seem like it’s asking anything. It’s just saying that’s what their going to do.” “Is this action normally something they would do to test this?”	“I want to know if this is a valid action or not.”
	11. *P*-value as rareness measure	“Valid. I have to have a picture and it’s, there’s not enough. How could I say this? …My confidence level for these questions. Some of these questions are a little confusing. I don’t think I could be completely confident.” “Um, probably 91 to 100% But most of my confidence is pretty much in the 85 to 90 range. But with 76 in the group, it would seem that I am not that confident.”	“This confidence rating, would you ever choose 100%?” “What if the scale read 76-90% and 91- 100% what would you pick?”
	17. Probability: alternative is false	“It’s kind of hard to understand the hypothesis if it’s worded this way. … ‘cause, you know in class it has the hypothesis and it had the alternative hypothesis and it’s clearly....”	
	20. Probability: null is true	“So the null hypothesis is that the treatment doesn’t have any effect on anyone. And the alternative is that there is an effect. FALSE. 51%-75% …Honestly I don’t remember what ‘degrees of freedom’ is. I don’t know what that means for this answer.”	May want to delete degrees of freedom from the item, since it may confound results.

Table D6 (continued)

Cognitive interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
CI: 210	21. Probability: alternative is false	“…The null hypothesis is that there’s no effect.” “…Maybe what’s tripping me up is that the whole experiment seems unreal.” “So then the null is the driving classes didn’t make a difference and the alternative is they helped.” “So, ‘absolutely disproved your alternative hypothesis’….” “I don’t like this question.”	“If the null hypothesis is that there’s no difference between the control and experimental groups, would that change your answer?” “Let’s say my treatment is sending you to driving school and I measure the task…parallel parking. Do I see a difference for those who go to driving school versus those who don’t?” This item apparently needs more context.
	5. Sample and population	“I’m not exactly sure if it’s asking …when it’s referring to populations. If when they are saying that ….if the population was half men and half women. I’m a little confused. It’s just my lack of knowledge….” “I understand the questions and I’m not sure of the answer. It’s speaking of the sample. I will go with FALSE but I’m just guessing.”	“I’m asking, ‘Does the *P*-value tell about the sample difference or about the population difference?’”
	7. Inverse is true	“…immediately when I read this question, it’s kind of like ‘how did they determine how often they would obtain the score’ …I don’t know, that’s what, just kind of what, like immediately makes me think, whether it’s a reasonable question or not.” “…That would be better. That would be VALID.”	“So, if I said the district researchers ‘conducted a test to determine…’ would that be clearer?

Table D6 (continued)

Cognitive interview	RPASS-2 conception or misconception	Student comments or notes	Interviewer comments or notes
CI: 210	8. Chance as cause of results observed	“That seems OK to me. But when you’re saying VALID or INVALID, is it like…is that an OK thing to do?” “Usually when I see VALID or INVALID it’s the answer and here it’s their techniques.” “For example, VALID ACTION or INVALID ACTION as the option? That would definitely make it stick.”	“Is there something I should say there to say ‘is this an OK thing to do?’” “Would YES or NO be better than VALID or INVALID?” Options need to be embellished to clarify what is being asked.
	9. *P*-value embedded in sampling variation	“I would think that’s a VALID ACTION. I’m just iffy with the whole subject of *P*-values at the moment. I don’t understand the question but not that it’s worded improperly.”	
	14. *P*-value as always low	“…‘gone bad’…I would think that…is VALID. I think it was low *P*-value would reject the null hypothesis, so a high *P*-value would accept the null hypothesis and I’m assuming that his, the null hypothesis, is that it doesn’t do anything. So, I don’t know if it’s ‘gone bad’ but I don’t know that it’s ‘gone good’ for him.” “Maybe that it ‘hasn’t gone well.’ …if they heard the word ‘awry’ they might have understood it but seeing it on paper may not be as familiar.” “Then, I’d question the whole thing.”	“Can you suggest an alternate wording other than ‘gone bad’? The ‘gone bad’ wording was suggested by a student during field testing as an alternative to ‘gone awry.’” “I wanted to suggest that the researcher ‘thought there was a miscalculation.’”

Table D7

Cognitive Interview Notes Using RPASS-3A, CI: 303, 199, 310, 329, 311

Cognitive interview	RPASS-3A conception or misconception	Student comments or notes	Interviewer comments or notes
CI: 303	6. Sample size and significance	"They predetermined … simple random sample? If it is a simple random sample, then it would be VALID."	Changed "pre-determined" to "planned how many" and added "randomly" to sample.
CI: 199	24. Reliability and *P*-value	"Not sure what this question means. I don't see the cor…. I don't know if it has to do with confidence levels. I'll say FALSE but I'm merely guessing."	"The topic of reliability is beyond scope of our class."
CI: 310	16. Probability: alternative is true	"OK but the way she took the test wasn't right because she invited people and if people were going to take the test free of charge, then people would probably already have some information. I'm gonna say it was invalid because of the way the test was conducted."	Student seemed to appreciate the importance of considering how the data was collected.
	Scenario 4	"There was one question that was wrong. The SAT scores one, where she invited people free of charge to take this online course. What are we supposed to do? Should we assume…?"	[That the sample is random.]
		"The people who would be involved would be the students who care. Students who would care would be better students."	"Your concern is that the researcher has 'stacked the deck?'"

Table D7 (continued)

Cognitive interview	RPASS-3A conception or misconception	Student comments or notes	Interviewer comments or notes
CI: 329	5. Sample and population	[Deleted the old wording “from which they come.”] “Um, I’m sure the beginning of this is true. …So, I guess I’m trying to figure out if the *P*-value is related to the population as well. Testing the difference between two samples. …has nothing to do with… it does have to do with the population. So, I’m going to say TRUE.”	
	8. Chance as cause of results observed	“…I assume that that is what they want to do. …I agree to that because that’s why they’re running this test. …Yeah, I believe that, not very confident.”	
	12. Test statistics and *P*-value	“…interprets the *t*-test statistic of 4% …as significant’ 4% would say that it’s not very significant at all. So, I’d say that’s INVALID. …If it’s only 4% then it would be more rare than this one up here. Then that would mean that the results up here are TRUE. …There’s a 4% chance that these results are true, that they’re significant, which makes me think the hair growth does have an effect, they’re statistically significant. I would say that would make the results not significant because 4% is so low. …So, I think I’ll go with 60-74% that it’s Invalid.”	Student never seemed to notice that a percentage was being used for the *t*-test statistic. This item may not measure the intended content. Student’s reasoning suggests they are interpreting the *t*-value (confusing it with *P*-value) and interpreting it as the probability the “results are true.” This student is also confused by the confidence ratings as if they represent the probability that the answer chosen is true or false. Confidence ratings may need to be removed since they may confound results interpretation rather then contribute to it.

Table D7 (continued)

Cognitive interview	RPASS-3A conception or misconception	Student comments or notes	Interviewer comments or notes
CI: 311	6. Sample size and significance	[The language had been changed to '…planned how many students should be sampled…']	
		"I'd say it's like a VALID statement." I don't understand exactly what this question…."	"OK, so tell me what is confusing about it."
		"Well, I understand that depending on the sample size their *P*-value could change. So, I'd say that that's VALID, that they would see that's that. I don't understand how they'd do it."	'…Would it help if I had 'Valid Action - Invalid Action?'"
		"I guess I don't understand what the question is saying."	
		"I would think so…"	"How about 'the district researchers were *concerned* about how many students should be sampled?'"
		"When it says they 'planned' how many students, I think maybe they inflamed their sample size to get statistical significance."	
		"I'm confident that the *P*-value is influenced by sample size and the researchers need to take that under consideration."	
	n/a	"I have a concern that the researcher invites a random sample to this. …A random sample is important. That also gives them a choice, too. In collecting data that's going to be there, so it's always volunteering." "I don't know how many high school students would turn down a free course."	"The 'free of charge' phrase is there to suggest subjects would be willing participants. But you're the second person who brought up the issue of having volunteers from an SRS, is no longer an SRS." Decided to add that "all invited accepted the invitation."

D.4 RPASS-2 and RPASS-3A Items as Modified from Student Interviews

Table D8

Item Modifications from RPASS-2 to RPASS-3A based on Interview Input with Rationale for Change

RPASS-2	Rationale for change	RPASS-3A
Scenario 1: A research article gives a *P*-value of .001 in the analysis section.	No change	Scenario 1: A research article gives a *P*-value of .001 in the analysis section.
1. The *P*-value is the probability that the null hypothesis is true. □ True □ False	Deleted the confidence ratings from for ALL items. (CI: 329)	1. Statement: The *P*-value is the probability that the null hypothesis is true. □ True □ False
2. The *P*-value is the probability of observing an outcome as extreme as or more extreme than the one observed if the null hypothesis is true. □ True □ False	Awkwardly worded, deleted second "as." (CI: 107, 210)	2. Statement: The *P*-value is the probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true. □ True □ False
3. If a simulation of the experiment were conducted, the *P*-value of .001 is the long-run frequency of obtaining the experimental results or something more extreme due to chance. □ True □ False	Added "probability" after long run frequency; Clarified the simulation; "Due to chance" deleted. (CI: 230, 107)	3. Statement: Simulating the experiment with a random model (to model no difference), $p = .001$ is the long-run frequency (i.e., the probability) of obtaining the experimental results or results even more extreme than those observed. □ True □ False
4. This *P*-value tells me the chances are 1 in 1000 of observing data this surprising (or more surprising) than what I observed, if the null hypothesis is true. □ True □ False	Is "surprising" same as extreme?; Changed to "rare." (ID: 118)	4. Statement: This *P*-value tells me the chances are 1 in 1000 of observing data this rare (or more rare) than what I observed, if the null hypothesis is true. □ True □ False

Note. Numbers in parentheses identify the In-Depth (ID) or Cognitive Interview (CI) that motivated the change.

Table D8 (continued)

RPASS-2	Rationale for change	RPASS-3A
5. This *P*-value may reflect a statistically significant difference between two samples but says nothing about the populations from which they come. □ True □ False	Difficulty understanding what was asked / stated; Wording was changed. (CI: 107, 210, 329, 311)	5. Statement: The *P*-value may reflect a difference between the two samples observed but has no bearing on whether there is a statistically significant difference in the population. □ True □ False
Scenario 2: The district administrators of an experimental program similar to Head Start are interested in knowing …. A random sample of current first graders who have been through the preschool program scored a mean Reading Readiness of 102.	Added LARGE before random sample to better describe conditions for Item 9. (ID: 18)	Scenario 2: The district administrators of an experimental program similar to Head Start are interested in knowing …. A large random sample of current first graders who have been through the preschool program scored a mean Reading Readiness of 102.
6. Action: The district researchers pre-determined the number of students to sample to ensure their *P*-value could detect whether the improvement could be attributed to the new Head Start-like program. □ Valid □ Invalid	Deleted reference to "Head Start" (ID: 322) Omitted "predetermined" (CI: 230, 107, 303) Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	6. Action: The district researchers carefully planned how many students should be included in the study, since they were concerned about how the size of their random sample would impact *P*-value. □ Valid Action □ Invalid Action

Note. Numbers in parentheses identify the In-Depth (ID) or Cognitive Interview (CI) that motivated the change.

Table D8 (continued)

RPASS-2	Rationale for change	RPASS-3A
7. Action: The district researchers determined how often they would obtain a score of 102 or higher just by chance to "definitively prove" whether the program had a positive impact. □ Valid □ Invalid	How this was “determined” was unclear; Added “conducted a statistical test to determine” (CI: 230, 210) Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	7. Action: The district researchers conducted a statistical test to determine how often they would obtain a score of 102 or higher just by chance to "definitively prove" whether the program had a positive impact. □ Valid Action □ Invalid Action
8. Action: After checking the necessary conditions, the district researchers proceeded to determine if random chance was the “cause of the results observed.” □ Valid □ Invalid	Student said, “what are you asking here?” Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	8. Action: After checking the necessary conditions, the district researchers conducted a test of significance to determine if random chance was the “cause of the results observed.” □ Valid Action □ Invalid Action
9. Action: The district researchers should compare the sample group’s mean to its sampling distribution based upon assuming the population mean is 100. □ Valid □ Invalid	Need to indicate sample size in Scenario 2 to verify if conditions for inference have been satisfied (ID: 118) Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	9. Action: Assuming the population mean is 100, the district researchers assessed where the sample group’s mean would appear in its sampling distribution. □ Valid Action □ Invalid Action

Note. Numbers in parentheses identify the In-Depth (ID) or Cognitive Interview (CI) that motivated the change.

Table D8 (continued)

RPASS-2	Rationale for change	RPASS-3A
10. Action: The researcher builds a 95% confidence interval for this sample mean to assess if the interval captures the population mean; this is equivalent to testing if the sample mean is statistically significant at the .05 level in relation to the population mean. □ Valid □ Invalid	Wordy. Broke into two sentences; deleted “in relation to population mean.” (ID: 322) Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	10. Action: The researcher builds a 95% confidence interval for the sample mean. If the interval captures the population mean, this is equivalent to a 2-tailed test to see if the sample mean is statistically significant at the .05 level. □ Valid Action □ Invalid Action
Scenario 3: An ethical researcher is hoping to show that his new hair growth treatment had statistically significant results. How should this researcher interpret results from this one-tailed test?	No change	Scenario 3: An ethical researcher is hoping to show that his new hair growth treatment had statistically significant results. How should this researcher interpret results from this one-tailed test?
11. Interpretation: Assuming the hair treatment had no effect, the researcher interprets the *P*-value as an indicator of how rare it would be to obtain the observed results if generated by a random model. □ Valid □ Invalid	Corrected wording by adding “or more extreme” after observed. Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	11. Interpretation: Assuming the hair treatment had no effect, the researcher interprets the *P*-value as an indicator of how rare it would be to obtain the observed results if generated by a random model. □ Valid Interpretation □ Invalid Interpretation

Note. Numbers in parentheses identify the In-Depth (ID) or Cognitive Interview (CI) that motivated the change.

Table D8 (continued)

RPASS-2	Rationale for change	RPASS-3A
12. Interpretation: The researcher interprets the results as statistically significant as long as the test statistic for his hair growth treatment is less than the .05 significance level (alpha). □ Valid □ Invalid	Reworded, no one seemed to notice issue being assessed. (CI: 329) Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	12. Interpretation: The researcher interprets the t-test statistic of .04, as a 4% probability of obtaining the results observed or those more extreme, if the null is true. □ Valid Interpretation □ Invalid Interpretation
13. Interpretation: If the results from the hair growth treatment are statistically significant, the researcher interprets the *P*-value to mean the hair growth treatment "caused" the hair growth observed in the study. □ Valid □ Invalid	NOTE: 68% of winter test respondents answered this item incorrectly; Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	13. Interpretation: If the results from the hair growth treatment are statistically significant, the researcher interprets the *P*-value to mean the hair growth treatment "caused" the hair growth observed in the study. □ Valid Interpretation □ Invalid Interpretation
14. Interpretation: The researcher interprets a large *P*-value for his hair growth treatment to mean that the experiment has gone bad. □ Valid □ Invalid	Rewordings from “awry” to “gone bad” to “calculation error” were made. (ID: 304, 322, CI: 230, 210) Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	14. Interpretation: The researcher assumes that getting a large *P*-value for his hair growth treatment clearly means that there was a calculation error. □ Valid Interpretation □ Invalid Interpretation

Note. Numbers in parentheses identify the In-Depth (ID) or Cognitive Interview (CI) that motivated the change.

Table D8 (continued)

RPASS-2	Rationale for change	RPASS-3A
Scenario 4: A researcher believes that an SAT preparation course will improve SAT scores. The researcher invites a random sample of students to take the online prep course, free of charge. ….	Added “All of these students agree to participate” to suggest an SRS (CI: 311)	Scenario 4: A researcher believes that an SAT preparation course will improve SAT scores. The researcher invites a random sample of students to take the online prep course, free of charge. All of these students agree to participate. ….
15. Action: The researcher concludes that there is a 97% chance that repeating her research will yield the same or similar results. ○ Valid ○ Invalid	Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	15. Conclusion: The researcher concludes that there is a 97% chance that repeating her research will yield the same or similar results. ○ Valid Conclusion ○ Invalid Conclusion
16. Action: The researcher concludes that there is a .03 probability that her research hypothesis (that the students have higher than average scores) is true. ○ Valid ○ Invalid	Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	16. Conclusion: The researcher concludes that there is a .03 probability that her research hypothesis (that the students have higher than average scores) is true. ○ Valid Conclusion ○ Invalid Conclusion
17. Action: The researcher concludes that there is only a 3% probability that her research hypothesis (that there is a difference between population means) is wrong. ○ Valid ○ Invalid	Parenthetically added “the alternative” to clarify research hypothesis (CI: 107, 210) Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	17. Conclusion: The researcher concludes that there is only a 3% probability that her research hypothesis (the alternative) is wrong. ○ Valid Conclusion ○ Invalid Conclusion

Note. Numbers in parentheses identify the In-Depth (ID) or Cognitive Interview (CI) that motivated the change.

Table D8 (continued)

RPASS-2	Rationale for change	RPASS-3A
18. Action: Since alpha is .05, the researcher concludes that there remain 3 chances in 100 that the observed results would have occurred even if the SAT preparation program had no effect. □ Valid □ Invalid	Perhaps significance level should be used in lieu of “alpha” Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	18. Conclusion: Since significance level is .05, the researcher concludes that there remain 3 chances in 100 that the observed results would have occurred even if the SAT preparation program had no effect. □ Valid Conclusion □ Invalid Conclusion
Scenario 5: Suppose you have a treatment which 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, degrees of freedom *df* = 18, *p* = 0.01). Please mark each of the statements below as "true" or false."	Removed reference to degrees of freedom and added the driving school context. (CI: 107, 329, 311)	Scenario 5: Suppose you have a driving school curriculum which you suspect may alter performance on passing the written exam portion of the driver's test. 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, degrees of freedom *df* = 18, *p* = 0.01). Please mark each of the statements below as "true statement" or "false statement."
19. Conclusion: You have absolutely disproved the null hypothesis which assumed there is no difference between the control and experimental group means in the population. □ True □ False	Clarified the wording. Item should measure only one objective. Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	19. Statement: You have absolutely disproved the null hypothesis which assumed there is no difference between the control and experimental group means in the population. □ True Statement □ False Statement
20. Conclusion: You have found the probability of the null hypothesis being true. □ True □ False	Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	20. Statement: You have found the probability of the null hypothesis being true. □ True Statement □ False Statement

Note. Numbers in parentheses identify the In-Depth (ID) or Cognitive Interview (CI) that motivated the change.

Table D8 (continued)

RPASS-2	Rationale for change	RPASS-3A
21. Conclusion: You have absolutely disproved your alternative hypothesis (that there is a difference between population means). □ True □ False	Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	21. Statement: You have absolutely disproved your alternative hypothesis (that there is a difference between population means). □ True Statement □ False Statement
22. Conclusion: Reasoning logically, you can determine the probability of the experimental (i.e., the alternative) hypothesis being true. □ True □ False	Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	22. Statement: Reasoning logically, you can determine the probability of the experimental (i.e., the alternative) hypothesis being true. □ True Statement □ False Statement
23. Conclusion: You know, if you decided to reject the null hypothesis, the probability that you are making the wrong decision. □ True □ False	Suggested moving “you know” to avoid misinterpretation. (ID: 304) Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	23. Statement: If you decided to reject the null hypothesis, you know the probability that you are making the wrong decision. □ True Statement □ False Statement
24. Conclusion: You can conclude that if the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions. □ True □ False	Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	24. Statement: You can conclude that if the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions. □ True Statement □ False Statement

Note. Numbers in parentheses identify the In-Depth (ID) or Cognitive Interview (CI) that motivated the change.

Table D8 (continued)

RPASS-2	Rationale for change	RPASS-3A
25. Conclusion: Assuming the sampling distribution is symmetrical, a two-tailed alternative hypothesis would yield a larger *P*-value than a 1-tailed alternative for the same value of the test statistic. □ True □ False	Altered valid-invalid to alternate-choice format for clarity (CI: 210, 311)	25. Statement: Assuming the sampling distribution is symmetrical, a two-tailed alternative hypothesis would yield a larger *P*-value than a 1-tailed alternative for the same value of the test statistic. □ True Statement □ False Statement

Note. In-depth Interview (ID) and Cognitive Interview (CI) codes refer to specific interview comments. See Appendixes D.2 In-depth Interviews and D.3 Cognitive Interviews for the relevant interview notes grouped by interviewee student numbers.

D.5 Summary of Expert Rater Item Suggestions for RPASS-3A

This appendix summarized expert rater comments and feedback on the 25-item RPASS-3A after the first round of expert review. Part A lists the suggestions for the seven most problematic items (Items 3, 7, 8, 13, 18, 21, and 22). Part B lists the changes made to four of the five problem scenarios (Scenarios 1, 2, 3, and 5). Item changes based on these suggestions are included in the 32-item RPASS-3B as it appears in Appendix D.7.

Part A. Seven items are discussed where the item was rated under 3, on the 4 point rating scale; i.e., experts "disagreed" or "strongly disagreed" with the item's validity. Some raters did not rate all the items in the first round of review but nevertheless made suggestions for item improvement. Below are the consolidated suggestions for the seven most problematic items.

Item 3. *Simulation definition:* Simulating the experiment with a random model (to model no difference), p = .001 is the long-run frequency (i.e., the probability) of obtaining the experimental results or results even more extreme than those observed.

1: Depends more on understanding simulations and "random model" than measures intended topic.

2: Wording doesn't seem correct. The simulation should not be done with any random model but rather with the model specified by the null hypothesis (see Q2).

3: I don't think that all courses cover simulation (e.g., my own). This would make it difficult for my students … to understand this item.

4: Felt "*p*" was undefined. Recommendation: "*p*-value = .001"

5: Wording is problematic; given scenario not enough information. Also, lots of intro students won't understand "random model." Scenario for this question needs more info here – context for term "difference" in questions.

Item 7. *Inverse as true.* The district researchers determined how often they would obtain a score of 102 or higher just by chance to "definitively prove" whether the program had a positive impact.

1: (The term "definitely prove" is circled.) Since this is never true, don't think it is really a Boolean contra-positive.

2: I think … can answer just with the rule "statistics doesn't prove." Recommendation: Change misconception.

Item 8. *Chance as cause of results observed.* After checking the necessary conditions, the district researchers conducted a significance test to determine if random chance was the “cause of the results observed.”

1: Again not sure I understand the misconception. *P*-value does give info about plausible role of chance…?

2: More general. Bigger misconception on “causality”

3: Is action performing the test or reaching the stated conclusion? One (conducting the test) is valid, and one (attributing chance as cause) is invalid.

4: Insert “for inference” before “the necessary conditions” in the stem.

Item 13. *Converse is true.* If the results from the hair growth treatment are statistically significant, the researcher interprets the *P*-value to mean the hair growth treatment "caused" the hair growth observed in the study.

1: This is a bigger, more general misconception on causality.

2: Depends on experimental design. If design is sound …would accept this interpretation.

3: Don't know what type of study conducted.

Item 18. *Type I / α and P-value.* Since alpha is .05, the researcher concludes that there remain 3 chances in 100 that the observed results would have occurred even if the SAT preparation program had no effect.

1: Without “since statement” seems OK and students could easily ignore beginning phrase because remainder is good -- more a question in logic than stats.

2: Wording of the problem is not clear. If your intention is to mix *P*(Type I error) and *P*-value, you might instead write "...there is a 3% chance that we incorrectly reject *Ho* when *Ho* is in fact true."

3: (Did not rate the item.) With regard to "alpha," you haven't used or defined that term/symbol in the stem, so I suggest not testing whether students know that term/symbol. I suggest saying "significance level" as you said in the stem.

Appendix D.5 (continued)

Item 21. *Alternative is false.* You have absolutely disproved your alternative hypothesis (that there is a difference between population means).

1: Too wordy remove the term "absolutely."

2: Wording is too strong -- student might choose false only because of word "absolutely." Remove "absolutely." The goal is to see if students' logic is right, not if they recognized that statistical evidence isn't proof.

3: Circled “population means” and suggested to change to “treatment means.”

Item 22. *Probability: alternative is true.* Reasoning logically, you can determine the probability of the experimental (i.e., the alternative) hypothesis being true.

1: I don’t think this gets at the intent. Why “reasoning logically”? Why “you can determine”? Recommendation: “The probability that the experimental (i.e., the alternative) hypothesis is true is .01.”

2: “Reasoning logically” is vague and might suggest to a student that somehow we could find this probability, even if the student might not know it can’t come from the p-value. Recommendation: Replace” reasoning logically” with “based on the p-value.”

3: The question doesn’t test the misconception P-value = $P(H_a$ is true) but the misconception that we can even calculate $P(H_a$ is true). Which one is the goal?

Appendix D.5 (continued)

Part B. Problem scenario changes motivated by expert comments and suggestions.

Scenario 1: Needed more context for RPASS-3A, Item 25 to assess whether respondents understand that the *P*-value is tied to the alternative hypothesis.

Before: A research article gives a *P*-value of .001 in the analysis section.

After: In the analysis section of a research article, the *P*-value for a 1-tailed statistical significance test is .001.

Scenario 2: Delete reference to "Head Start." Delete second reference to "score." Grammatical error: "have" changed to "has."

Before: The district administrators of an experimental program similar to Head Start are interested in knowing if the program had an impact on the reading readiness of first graders. Historically, before implementing the new program, the mean score for Reading Readiness score for all first graders was 100 and the population standard deviation is 15. A random sample of current first graders who attended the new preschool program have a mean Reading Readiness score of 102.

After: The district administrators of an experimental program are interested in knowing if the program had an impact on the reading readiness of first graders. Historically, before implementing the new program, the mean score for Reading Readiness for all first graders was 100 and the population standard deviation is 15. A random sample of current first graders who attended the new preschool program has a mean Reading Readiness score of 102.

Appendix D.5 (continued)

Scenario 3: This scenario was altered to add reference to a 2-sample test. Wording was altered to imply that there was no random assignment of subjects to treatments, therefore causality should not be inferred.

Before: An ethical researcher is hoping to show that his new hair growth treatment had statistically significant results. How should this researcher interpret results from this one-tailed test?

After: An ethical researcher conducts a two sample test. He compares the hair growth results for a group of volunteers who try his treatment to a second group who does not use the treatment. He hopes to show his new hair growth treatment had statistically significant results. How should this researcher interpret results from this one-tailed 2-sample test?

Scenario 4: No changes.

Scenario 5: Random assignment and random selection were added to this scenario to minimize confusion. A suggestion was also made to remove the word “simple” because it may imply a “1-sample *t*-test.”

Before: Suppose you have a driving school curriculum which you suspect may alter performance on passing the written exam portion of the driver's test. 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, degrees of freedom df = 18, p = 0.01). Please mark each of the statements below as "true" or "false."

After: Suppose you have a driving school curriculum which you suspect may alter performance on passing the written exam portion of the driver's test. You compare the means of randomly selected and randomly assigned control and experimental groups (20 subjects in each group). You use a 2-sample test of significance and obtain a test statistic which is 2.7 standard errors from the mean of the sampling distribution (p = 0.01). Please mark each of the statements below as "true" or "false."

D.6 Expert Rater Assessment of Missing or Extraneous Content

This appendix summarizes comments made by experts after the first round of instrument review concerning whether there was missing or extraneous RPASS content. Part I asked the expert raters to assess missing content. Part II asked the expert raters to assess extraneous content. Missing or extraneous content are potential sources of invalidity. Researcher notes are inserted in [comment boxes]. Comment numbers do not correspond to numbers used in Appendix D.5.

Part I. What do you think may be missing from the content of the RPASS assessment related to *P*-values and statistical significance? Please describe.

1. Prefer focus on contextual applications to recognizing definitions. Not sure anything is missing, instead worry will be too long to use with my students.
2. Test for the misconception that a large *P*-value implies the null hypothesis is true.
3. You may want to add another question or two on Type I/II errors. Would a question on power be appropriate? Here's one question I've used you might find interesting:

   Scenario: (After experiment is defined) … and *P*-value = .72, no value of alpha is specified.

   Expert statement: When asked to determine the strength of evidence against *Ho*, the researcher states this is not possible since alpha was not given. (FALSE!) The point of the question is to illustrate that a big *P*-value provides weak evidence against *Ho* whether alpha is supplied or not. [An item was altered to assess a large *P*-value. The concept of Type I error is included but not referenced by name.]
4. Good idea, but some faulty or confusing questions. [Subsequent one-on-one meetings were held to address these issues.]
5. More concrete context in wording of items. Fewer "definition" wordings. Most of the questions I marked "3" instead of "4" were because the question was a little too abstract. It would be better to include the relevant numbers in the item wording. [Specific *P*-values were added to item wordings.]

6. I think the RPASS has great potential, but there are some difficulties in wording of scenarios and questions that may limit its use and validity. [Subsequent one-on-one meetings were held to address these issues.]

7. I think your items do a good job of assessing the objectives that you've laid out. …(M)any fundamental concepts associated with truly understanding the concepts of significance and *p*-value are not assessed here. [Eight specific suggestions were made for additional item content. New items were added and problem scenarios altered to address these suggestions.] The specific learning objectives included:

    1. The smaller the *P*-value, the stronger the evidence of a difference or effect.
    2. The smaller the *P*-value, the stronger the statistical significance of a difference or effect.
    3. The bigger the difference in observed results between two groups, the smaller the *P*-value, and so the more significant the results, if all else remained the same.
    4. The larger the sample size, the smaller the *P*-value, and so the more significant the observed results, if all else remains the same.
    5. The more variability in the sample results, the larger the *P*-value, if all else remains the same.
    6. The method for determining the *P*-value depends on the way randomness was used in the collection of the data. (Altered problem Scenarios 3 and 5.)
    7. One can draw a causal conclusion from a small *P*-value with a randomized experiment but not with an observational study.
    8. A small *P*-value does not mean that there is necessarily a large difference or effect.

Appendix D.6 (continued)

Part II. What part of the RPASS content may be extraneous in terms of understanding *P*-values and statistical significance? Please describe.

1. Reliability and "proof" statements…more than [you] need?
2. Only the simulation question.
3. I made notes on the questions. Sometimes the context affects the interpretation of the questions.
4. The CI/HT (Confidence Interval/Hypothesis Testing) duality isn't directly related to *P*-values and significance.
5. There appears to be some repetition of questions (e.g., RPASS-3A Item 1 and Item 20).
6. Why does scenario 3 need to be one-sided? It doesn't seem to be used in any question. [Researcher note: Scenario 3 was re-written and excluded the unneeded one-tailed test detail.]
7. I suggest you eliminate all mentioning of specific tests. Your goal of assessing understanding of concepts of significance and p-value is very commendable, and that understanding should be independent of a specific test procedure. [Researcher note: References to specific tests were eliminated.]
8. The statement in question 28 now begins with a sentence fragment. The two sentences could be combined into one. [Researcher note: This comment was offered after the final round of reviews and was corrected before administration of RPASS-4.]

D.7 RPASS-3A, RPASS-3B, and RPASS-3C Item Modifications Based on Expert Rater Input

Table D9

Item Modifications from RPASS-3A to RPASS-3B toRPASS-3C based on Expert Rater Input

RPASS-3A item number and wording	RPASS-3B item number and wording	RPASS-3C item number and wording
1. Statement: The *P*-value is the probability that the null hypothesis is true. □ True □ False ORIGIN RPASS-1B PILOT	1. Statement: The *P*-value is the probability that the null hypothesis is true. □ True □ False NO CHANGE	1. Statement: The *P*-value is the probability that the null hypothesis is true. □ True □ False ITEM DELETED, REDUNDANT (*Null is true*)
2. Statement: The *P*-value is the probability of observing an outcome as extreme as or more extreme than the one observed if the null hypothesis is true. □ True □ False ORIGIN RPASS-1B PILOT	2. Statement: The *P*-value is the probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true. □ True □ False NO CHANGE	2. Statement: The *P*-value (.001) is the probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true. □ True □ False
25. Statement: Assuming the sampling distribution is symmetric, a two-tailed alternative hypothesis would yield a larger *P*-value than a 1-tailed alternative for the same value of the test statistic. □ True Statement □ False Statement POST RPASS-1B PILOT ADDITION	3. Statement: In general, a two-tailed alternative hypothesis would yield a larger *P*-value than .001 for the same value of the test statistic. □ True □ False MOVED and ALTERED ITEM, OPTIONS	3. Statement: If the students had conducted a 2-tailed test instead of a 1-tailed test on the same data, how would the *P*-value have changed? □ *P*-value would be larger. □ *P*-value would be smaller. □ *P*-value would not change.

Table D9 (continued)

RPASS-3A item number and wording	RPASS-3B item number and wording	RPASS-3C item number and wording
4. Statement: This *P*-value tells me the chances are 1 in 1000 of observing data this rare (or more rare) than what I observed, if the null hypothesis is true. □ True □ False ORIGIN RPASS-1B PILOT	4. Statement: This *P*-value tells me the chances are 1 in 1000 of observing data as rare (or more rare) than what the researchers observed, if the null hypothesis is true. □ True □ False ALTERED ITEM	4. Statement: This *P*-value tells me the chances are 1 in 1000 of observing data this rare (or more rare) than what I observed, if the null hypothesis is true. □ True □ False
n/a. EXPERT RATER ADDITION	5. Statement: A causal conclusion can be drawn from a *P*-value this small, regardless of whether this was a randomized comparative experiment or an observational study. □ True □ False NEW ITEM and LEARNING OBJECTIVE	5. Statement: A causal conclusion can be drawn from a *P*-value this small, regardless of whether this was a randomized comparative experiment or an observational study. □ True □ False
n/a. EXPERT RATER ADDITION	6. Statement: The smaller the *P*-value, the stronger the evidence of a difference or effect. □ True □ False NEW ITEM and LEARNING OBJECTIVE	6. Statement: Assume that the students obtained an even smaller *P*-value. □ This is stronger evidence of a difference or effect. □ This is weaker evidence of a difference or effect. □ There is no change in the amount of evidence of a difference or effect.

Table D9 (continued)

RPASS-3A item number and wording	RPASS-3B item number and wording	RPASS-3C item number and wording
n/a. EXPERT RATER ADDITION	7. Statement: If there were more variability in the study results, we would expect to obtain a larger *P*-value than .001, if all else remained the same. ○ True ○ False NEW ITEM and LEARNING OBJECTIVE	7. Statement: If there were more standard errors between the observed sample mean and the hypothesized mean of the population, we would expect to obtain a larger *P*-value than .001, if all else remained the same. ○ True ○ False
9. Action: Assuming the population mean is 100, the district researchers assessed where the sample group's mean would appear in its sampling distribution. ○ Valid Action ○ Invalid Action ORIGIN RPASS-1B PILOT	11. Action: The district researchers found how likely the sample group's mean of 102 would be in the sampling distribution of mean scores, assuming that the population mean really is 100. ○ Valid Action ○ Invalid Action ALTERED ITEM	8. Action: The district researchers found how likely a sample mean of 102 or higher would be in the sampling distribution of mean scores, assuming that the population mean really is 100. ○ Valid Action ○ Invalid Action
10. Action: The researcher builds a 95% confidence interval for the sample mean. If the interval captures the population mean, this is equivalent to a 2-tailed test to see if the sample mean is statistically significant at the .05 level. ○ Valid Action ○ Invalid Action POST RPASS-1B PILOT ADDITION	12. Action: Since conditions for a confidence interval were reasonable, the researcher constructs a 95% confidence interval around the sample mean of 102. If the interval captures the hypothesized population mean, this is equivalent to a 2-tailed test to see if the sample mean is statistically different from 100 (at the .05 level). ○ Valid Action ○ Invalid Action ALTERED ITEM	9. Action: Since conditions for inference were acceptable, the researcher constructed a 95% confidence interval around the sample mean of 102. If the interval captures the hypothesized population mean, this is equivalent to a 2-tailed test to see if the sample mean is statistically different from 100 (at the .05 level). ○ Valid Action ○ Invalid Action

Table D9 (continued)

RPASS-3A	RPASS-3B	RPASS-3C
Item number and wording	item number and wording	item number and wording
6. Action: The district researchers carefully planned how many students should be included in the study, since they were concerned about how the size of their random sample would impact *P*-value. ○ Valid Action ○ Invalid Action ORIGIN RPASS-1B PILOT	8. Procedure: Before sampling the students, the district researchers calculated a minimum sample size that should be included in the study, since they were concerned about how the size of the random sample would impact the *P*-value. ○ Valid Procedure ○ Invalid Procedure ALTERED ITEM and OPTIONS	n/a: ITEM DELETED, REDUNDANT *(Sample size)*
7. Action: The district researchers conducted a statistical test to determine how often they would obtain a score of 102 or higher just by chance to "definitively prove" whether the program had a positive impact. ○ Valid Action ○ Invalid Action RPASS-1B PILOT	9. Action: The district researchers conducted a significance test to determine how often they would obtain a score of 102 or higher just by chance to prove whether the program had a positive impact. ○ Valid Action ○ Invalid Action ALTERED ITEM and MISCONCEPTION	10. Action: The district researchers used a significance test to determine how often they would obtain a sample mean score of 102 or higher just by chance in order to prove whether the program had a positive impact. ○ Valid Action ○ Invalid Action
n/a: EXPERT RATER ADDITION	13. Interpretation: The smaller the *P*-value for the reading readiness group results, the stronger the statistical significance of the reading readiness program results. ○ Valid Interpretation ○ Invalid Interpretation NEW ITEM and LEARNING OBJECTIVE	11. Interpretation: The stronger the evidence that the reading readiness program had an effect, the smaller the *P*-value that would be obtained when comparing the group results to the general population. ○ Valid Interpretation ○ Invalid Interpretation

Table D9 (continued)

RPASS-3A item number and wording	RPASS-3B item number and wording	RPASS-3C item number and wording
8. Action: After checking the necessary conditions, the district researchers conducted a test of significance to determine if random chance was the “cause of the results observed.” ○ Valid Action ○ Invalid Action ORIGIN RPASS-1B PILOT	10. Interpretation: After checking the conditions necessary for inference, the district researchers conducted a significance test to conclude whether random chance caused the results observed. ○ Valid Interpretation ○ Invalid Interpretation ALTERED ITEM and OPTIONS	12. Interpretation: After checking the conditions necessary for inference, the district researchers found they had statistically significant results. They interpreted the small *P*-value they obtained to mean that random chance caused the results observed. ○ Valid Interpretation ○ Invalid Interpretation
11. Interpretation: Assuming the hair treatment had no effect, the researcher interprets the *P*-value as an indicator of how rare it would be to obtain the observed results if generated by a random model. ○ Valid Interpretation ○ Invalid Interpretation ORIGIN RPASS-1B PILOT	14. Interpretation: Assuming the hair treatment had no effect, the researcher interprets the *P*-value as an indicator of how rare it would be to obtain the observed results. ○ Valid Interpretation ○ Invalid Interpretation ALTERED ITEM	13. Interpretation: The researcher interprets the *P*-value as an indicator of how rare it would be to obtain the observed results or something more extreme, assuming the hair treatment had no effect. ○ Valid Interpretation ○ Invalid Interpretation

Table D9 (continued)

RPASS-3A item number and wording	RPASS-3B item number and wording	RPASS-3C item number and wording
12. Interpretation: The researcher interprets the t-test statistic of .04, as a 4% probability of obtaining the results observed or those more extreme, if the null is true. ○ Valid Interpretation ○ Invalid Interpretation ORIGIN RPASS-1B PILOT	15. Interpretation: The researcher interprets a test statistic of .04, as a .04 probability of obtaining the results observed or those more extreme, if the null hypothesis is true. ○ Valid Interpretation ○ Invalid Interpretation ALTERED ITEM	14. Interpretation: Suppose that the researcher calculates a test statistic of .04. He interprets this as a .04 probability of obtaining the results observed or one more extreme, if the null hypothesis is true. ○ Valid Interpretation ○ Invalid Interpretation
13. Interpretation: If the results from the hair growth treatment are statistically significant, the researcher interprets the *P*-value to mean the hair growth treatment "caused" the hair growth observed in the study. ○ Valid Interpretation ○ Invalid Interpretation ORIGIN RPASS-1B PILOT	16. Interpretation: If the volunteers' have statistically significant hair growth compared to the no treatment group, the researcher interprets the *P*-value to mean the hair treatment caused the hair growth observed. ○ Valid Interpretation ○ Invalid Interpretation ALTERED ITEM	15. Interpretation: If the volunteers have longer hair growth compared to the no treatment group, the researcher interprets the results *P*-value to mean there is more hair growth in a population who uses his treatment. ○ Valid Interpretation ○ Invalid Interpretation
14. Interpretation: The researcher assumes that getting a large *P*-value for his hair growth treatment clearly means that there was a calculation error. ○ Valid Interpretation ○ Invalid Interpretation ORIGIN RPASS-1B PILOT	17. Interpretation: The researcher obtained a large *P*-value of .72 and assumes that the large *P*-value means that there was a calculation error. ○ Valid Interpretation ○ Invalid Interpretation ALTERED ITEM	16. Interpretation: Suppose the researcher obtains a large *P*-value of .72. What should he conclude? ○ There is a calculation error. ○ The sample data does not support the research hypothesis. ○ There is a 72% chance that the treatment is effective.

Table D9 (continued)

RPASS-3A item number and wording	RPASS-3B item number and wording	RPASS-3C item number and wording
15. Conclusion: The researcher concludes that there is a 97% chance that repeating her research will yield the same or similar results. □ Valid Conclusion □ Invalid Conclusion ORIGIN RPASS-1B PILOT	n/a. ITEM DELETED, REDUNDANT *(Reliability)*	n/a. ITEM DELETED, REDUNDANT *(Reliability)*
16. Conclusion: The researcher concludes that there is a .03 probability that her research hypothesis (that the students have higher than average scores) is true. □ Valid Conclusion □ Invalid Conclusion ORIGIN RPASS-1B PILOT	19. Conclusion: The researcher concludes that there is a .03 probability that her research hypothesis (that the students have higher than average scores) is true. □ Valid Conclusion □ Invalid Conclusion NO CHANGE	n/a. ITEM DELETED, REDUNDANT *(Alternative is true)*
n/a. EXPERT RATER ADDITION	18. Interpretation: Assume there was random assignment of subjects to groups and the *P*-value was found to be large (*P*-value = .72). Without a specified significance level, the researcher cannot determine the strength of evidence against the null hypothesis. □ Valid Statement □ Invalid Statement NEW ITEM and LEARNING OBJECTIVE	17. Action: Assume the *P*-value was found to be large (*P*-value = .72). Without a specified significance level, the researcher can still state that the results are compatible with the null hypothesis. □ Valid Statement □ Invalid Statement

Table D9 (continued)

RPASS-3A item number and wording	RPASS-3B item number and wording	RPASS-3C item number and wording
17. Conclusion: The researcher concludes that there is only a 3% probability that her research hypothesis (the alternative) is wrong. ○ Valid Conclusion ○ Invalid Conclusion ORIGIN RPASS-1B PILOT	20. Conclusion: The researcher concludes that there is only a .03 probability (3%) that her research hypothesis (the alternative) is wrong. ○ Valid Conclusion ○ Invalid Conclusion ALTERED ITEM	n/a. ITEM DELETED, REDUNDANT *(Probability: alternative is false)*
n/a. EXPERT RATER ADDITION	21. Conclusion: The small *P*-value does not necessarily mean that there is a large practical improvement in scores. ○ Valid Conclusion ○ Invalid Conclusion NEW ITEM and LEARNING OBJECTIVE	18. Conclusion: The small *P*-value does not necessarily mean that there is a large improvement in scores. ○ Valid Conclusion ○ Invalid Conclusion
18. Conclusion: Since alpha is .05, the researcher concludes that there remain 3 chances in 100 that the observed results would have occurred even if the SAT preparation program had no effect. ○ Valid Conclusion ○ Invalid Conclusion ORIGIN RPASS-1B PILOT	22. Conclusion: Since significance level is .05, the researcher concludes that there is a 3% chance that he has incorrectly rejected the null hypothesis, when the null hypothesis is in fact true. ○ Valid Conclusion ○ Invalid Conclusion ALTERED ITEM	19. Conclusion: Recall that the significance level is .05 and the *P*-value is .03. ○ The .05 suggests the prep course mean scores are higher than 500. ○ The .03 suggests the prep course mean scores are higher than 500. ○ Since .03 is smaller than .05, the evidence suggests the prep course is not helpful.

Table D9 (continued)

RPASS-3A Item number and wording	RPASS-3B item number and wording	RPASS-3C item number and wording
n/a. EXPERT RATER ADDITION	23. Conclusion: If there were an even greater difference between the scores of the students who took the SAT preparation course and the historical average, we would expect to obtain an even smaller *P*-value than .03. ○ Valid Conclusion ○ Invalid Conclusion NEW ITEM and LEARNING OBJECTIVE	20. Conclusion: If there were an even greater difference between the scores of these students who took the SAT preparation course and the historical average, we would expect to obtain an even smaller *P*-value than .03. ○ Valid Conclusion○ Invalid Conclusion
19. Statement: You have absolutely disproved the null hypothesis which assumed there is no difference between the control and experimental group means in the population. ○ True Statement ○ False Statement POST RPASS-1B PILOT ADDITION	24. Statement: The small *P*-value implies the null hypothesis is false. Thus, there is no difference between the control and experimental group means in the population. ○ True Statement ○ False Statement ALTERED ITEM	21. Statement: The small *P*-value of .01 is the probability that the null hypothesis (that there is no difference between the two population means) is false. ○ True Statement ○ False Statement

Table D9 (continued)

RPASS-3A Item number and wording	RPASS-3B item number and wording	RPASS-3C item number and wording
20. Statement: You have found the probability of the null hypothesis being true. □ True Statement □ False Statement POST RPASS-1B PILOT ADDITION	25. Statement: You have found the probability of the null hypothesis being true as p = .01. □ True Statement □ False Statement ALTERED ITEM	22. Statement: The *P*-value of .01 is the probability of the null hypothesis (that the new driving school curriculum had no effect) being true. □ True Statement □ False Statement
21. Statement: You have absolutely disproved your alternative hypothesis (that there is a difference between population means). □ True Statement □ False Statement POST RPASS-1B PILOT ADDITION	26. Statement: You have shown the experimental hypothesis (that there is a difference between population means) is false. □ True Statement □ False Statement ALTERED ITEM	23. Statement: You have shown the experimental hypothesis (that there is a difference between population means) is false. □ True Statement □ False Statement ALTERED ITEM
22. Statement: Reasoning logically, you can determine the probability of the experimental (i.e., the alternative) hypothesis being true. □ True Statement □ False Statement POST RPASS-1B PILOT ADDITION	27. Statement: The probability that the experimental (i.e., the alternative) hypothesis is true is .01. □ True Statement □ False Statement ALTERED ITEM	24. Statement: The probability that the experimental (i.e., the alternative) hypothesis is true is .01. □ True Statement □ False Statement

Table D9 (continued)

RPASS-3A Item number and wording	RPASS-3B item number and wording	RPASS-3C item number and wording
23. Statement: If you decided to reject the null hypothesis, you know the probability that you are making the wrong decision. □ True Statement □ False Statement POST RPASS-1B PILOT ADDITION	28. Statement: If you decided to reject the null hypothesis, you know the probability that you are making the wrong decision. □ True Statement □ False Statement NO CHANGE	n/a. ITEM DELETED, REDUNDANT *(Alternative is false)*
24. Statement: You can conclude that if the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions. □ True Statement □ False Statement POST RPASS-1B PILOT ADDITION	29. Statement: You can conclude that if the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions. □ True Statement □ False Statement NO CHANGE	25. Statement: You can conclude that if the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions. □ True Statement □ False Statement
3. Statement: Simulating the experiment with a random model (to model no difference), p = .001 is the long-run frequency (i.e., the probability) of obtaining the experimental results or something more extreme than those observed. □ True □ False ORIGIN RPASS-1B PILOT	30. Statement: Simulating the experiment with a random process (to model no difference between means), the *P*-value of .001 is the long-run frequency (i.e., the probability) of obtaining the experimental results or results even more extreme than those observed. □ True Statement □ False Statement MOVED and ALTERED ITEM	26. Statement: Repeating the randomization process many times to model no difference between means, the *P*-value of .01 is the long-run relative frequency (i.e., the probability) of obtaining results at least as extreme as those observed. □ True Statement □ False Statement

Table D9 (continued)

RPASS-3A item number and wording	RPASS-3B item number and wording	RPASS-3C item number and wording
5. Statement: The *P*-value may reflect a difference between the two samples observed but has no bearing on whether there is a statistically significant difference in the population. ○ True ○ False POST RPASS-1B PILOT ADDITION	31. Statement: The *P*-value may reflect a difference between the two samples observed but has no bearing on whether there is a statistically significant difference in the population. ○ True Statement ○ False Statement MOVED and ALTERED ITEM	27. Statement: The *P*-value may reflect a difference between the two samples observed but has no bearing on whether there is a statistically significant difference in the population. ○ True Statement ○ False Statement
n/a. EXPERT RATER ADDITION	32. Statement: If there were an even larger sample size, the researchers expect that they would obtain a smaller *P*-value than .001. ○ True Statement ○ False Statement NEW ITEM	28. Statement: If there were a larger sample size and the sample results turned out the same, how would the *P*-value change? ○ There would be a larger *P*-value. ○ There would be a smaller *P*-value. ○ The *P*-value would remain the same.

APPENDIX E

E.1 27-item RPASS-4 Instrument: Instructions, Scenarios, Items, and Scores

Tables E1 through E5 report the number and proportion of correct responses by item for the 224 respondents who answered all the items. Prior to each table are the online instructions and the problem scenario as seen by the respondents. Each table presents item results for one of the five RPASS-4 scenarios: Defining *P*-values, Using Tests of Significance, Interpreting Significant Results, Drawing Conclusions about Statistical Significance, and Tying *P*-values back to Hypotheses. The table rows list item numbers and complete item wordings by scenario. The last two columns describe the correct conception or misconception being assessed and the proportion who answered the item correctly. There were six items related to Scenario 1 (Defining *P*-values) as detailed in Table E1.

Defining P-values

In this section the questions are related to scenario #1. Read scenario #1. Following the scenario are statements that may be true or false. For each of the statements following the scenario, determine if you think the statement is 'True' or 'False,' then click the appropriate button.

A research article reports that the mean number of minutes students at a particular university study each week is approximately 1000 minutes. The student council claims that students are spending much more time studying than this magazine reported. To test their claim the students check a random sample of 81 students and they find a *P*-value of .001.

Table E1

Defining P-values Items: Number and Proportion of Correct Responses

	RPASS-4item wording as seen by respondents	Correct conception or misconception	Correct responses	
			Number	Proportion
1.	The *P*-value (.001) is the probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true. □ True □ False	*Textbook definition* (B-1)—Recognizing a formal textbook definition of the *P*-value without a context.	165	.74
2.	Statement: If the students had conducted a 2-tailed test instead of a 1-tailed test on the same data, how would the *P*-value have changed? □ *P*-value would be larger □ *P*-value would be smaller □ *P*-value would not change	*P-value dependence on alternative* (B-1)—Understanding the *P*-value depends on whether one has a one-tailed or two-tailed alternative hypothesis.	120	.54
3.	This *P*-value tells me the chances are 1 in 1000 of observing data as rare (or more rare) than what the researchers observed, if the null hypothesis is true. □ True □ False	*Lay definition* (B-1)—Recognizing an informal description of the *P*-value embedded in a context.	155	.69

Table E1 (continued)

	RPASS-4item wording as seen by respondents	Correct conception or misconception	Correct responses	
			Number	Proportion
4.	Statement: A causal conclusion can be drawn from a *P*-value this small, regardless of whether this was a randomized comparative experiment or an observational study. □ True □ False	*Conclusions as independent of study design* (L-4)—Understanding one can draw causal conclusions from a small *P*-value in randomized experiments but not with observational studies.	114	.51
5.	Statement: Assume that the students obtained an even smaller *P*-value. □ This is stronger evidence of a difference or effect. □ This is weaker evidence of a difference or effect. □ There is no change in the amount of evidence of a difference or effect.	*Smaller the P-value* (B-1)—Understanding the smaller the *P*-value, the stronger the evidence of a difference or effect.	175	.78
6.	Statement: If there were more standard errors between the observed sample mean and the hypothesized mean of the population, we would expect to obtain a larger *P*-value than .001, if all else remained the same. □ True □ False	*P-value and standard error* (B-1) —Understanding the more variation between the sample and the hypothesized population, the smaller the *P*-value, if all else remains the same.	104	.46

Note. N = 224. Mean proportion of correct responses for defining *P*-values items, $\mu_{\hat{p}}$ = .62.

Appendix E.1 (continued)

There were five items related to Scenario 2 (Using Tests of Significance) as detailed in Table E2. The instructions and scenario read:

Using Tests of Significance

> In this section the questions are based on scenario #2. Read scenario #2. Following the scenario are some possible actions based on the statistical results. For each of the statements following the scenario, determine if you think the action, procedure or interpretation described is valid or invalid, then click the appropriate button.
>
> > The district administrators of an experimental program are interested in knowing if the program has had an impact on the reading readiness of first graders. Historically, before implementing the new program, the mean score for Reading Readiness for all first graders was 100. A large random sample of current first graders who attended the new preschool program had a mean Reading Readiness score of 102. Assess the following things that the district researchers might have done.

Table E2

Using Tests of Significance Items: Number and Proportion of Correct Responses

	RPASS-4 item wording as seen by respondents	Correct conception or misconception	Correct responses	
			Number	Proportion
7.	Action: The district researchers found how likely a sample mean of 102 or higher would be in the sampling distribution of mean scores, assuming that the population mean really is 100. □ Valid Action □ Invalid Action	*P-value embedded in sampling variation* (B-1)—Embedding the *P*-value in a multiplicative conception of sampling variation.	162	.72
8.	Action: Since conditions for inference were acceptable, the researcher constructed a 95% confidence interval around the sample mean of 102. If the interval captures the hypothesized population mean of 100, this is equivalent to a 2-tailed test to see if the sample mean is statistically different from 100 (at the .05 level). □ Valid Action □ Invalid Action	*Confidence interval and significance* (R-6)—Understanding the equivalence of a 95% confidence interval and a two-tailed test of significance conducted at the .05 significance level.	131	.58
9.	Interpretation: The district researchers used the significance test to determine how often they would obtain a sample mean score of 102 or higher just by chance in order to prove whether the program had a positive impact. □ Valid Interpretation □ Invalid Interpretation	*Inverse as true* (L-1)—Believing statistics provide definitive proof.	78	.35

Table E2 (continued)

	RPASS-4 item wording as seen by respondents	Correct conception or misconception	Correct responses	
			Number	Proportion
10.	Interpretation: The stronger the evidence that the reading readiness program had an effect, the smaller the *P*-value that would be obtained when comparing the group results to the general population. □ Valid Interpretation □ Invalid Interpretation	*Strong statistical evidence* (B-1)—Understanding the stronger the statistical evidence of a difference or effect, the smaller the *P*-value.	166	.74
11.	Interpretation: After checking the conditions necessary for inference, the district researchers found they had statistically significant results. They interpreted the *P*-value they obtained to mean that random chance caused the results observed. □ Valid Interpretation □ Invalid Interpretation	*Chance as cause of results observed* (L-3)—Interpreting the *P*-value as the probability that observed results are due to chance or caused by chance.	154	.69

Note. N = 224. Mean proportion of correct responses for using tests of significance items, $\mu_{\hat{p}}$ = .62.

Appendix E.1 (continued)

There were five items related to Scenario 3 (Interpreting Results Items) as detailed in Table E3. The instructions and scenario read:

Interpreting Results

> In this section the questions are based on scenario #3. Read scenario #3. Following the scenario are some possible interpretations based on the statistical results. For each of the statements following the scenario, determine if you think the interpretation or statement is valid or invalid, then click the appropriate button." Scenario 3 had five items associated with it. The scenario read:
>
> > A researcher conducts a two sample test. He compares the hair growth results for a group of volunteers who try his treatment to a second group who does not use the treatment. He hopes to show his new hair growth treatment had statistically significant results. How should this researcher interpret results from this 2-sample test?

Table E3

Interpreting Results Items: Number and Proportion of Correct Responses

	RPASS-4 item wording as seen by respondents	Correct conception or misconception	Correct responses	
			Number	Proportion
12.	Interpretation: The researcher interprets the *P*-value as an indicator of how rare it would be to obtain the observed results or something more extreme, assuming the hair treatment had no effect. □ Valid Interpretation □ Invalid Interpretation	*P-value as rareness measure* (B-1)—Understanding the *P*-value can be considered a rareness measure	165	.74
13.	Interpretation: Suppose that the researcher calculates a test statistic of .04. He interprets this as a .04 probability of obtaining the results observed or one more extreme, if the null hypothesis is true. □ Valid Interpretation □ Invalid Interpretation	*Test statistics and P-value* (R-1)—Confusing the test statistic and its associated probability value	145	.65
14.	Interpretation: If the volunteers have longer hair growth compared to the no treatment group, the researcher interprets the *P*-value to mean there would be more hair growth in a population who uses his treatment. □ Valid Interpretation □ Invalid Interpretation	*Converse as true* (L-2)—Misusing the Boolean logic of the converse (*a->b* replaced with *b->a*)	83	.37

Table E3 (continued)

	RPASS-4 item wording as seen by respondents	Correct conception or misconception	Correct responses	
			Number	Proportion
15.	Interpretation: Suppose the researcher obtains a large *P*-value of .72. What should he conclude? □ There is a calculation error. □ The sample data does not support the research hypothesis. □ There is a 72% chance that the treatment is effective.	*P-value as always low* (B-2)—Believing the *P*-value is always a low number (or is always desired to be low a number)	171	.76
16.	Action: Assume the *P*-value was found to be large (*P*-value = .72). Without a specified significance level, the researcher can still state that the results are compatible with the null hypothesis. □ Valid Statement □ Invalid Statement	*Weak statistical evidence* (B-1)—Understanding large *P*-values provide weak evidence against the null hypothesis.	118	.53

Note. N = 224. Mean proportion of correct responses for interpreting results items, $\mu_{\hat{p}}$ = .61.

Appendix E.1 (continued)

There were seven items related to Scenario 4 (Drawing Conclusions about Statistical Significance) as detailed in Table E4. The instructions and scenario read:

Drawing Conclusions

In this section the questions are based on scenario #4. Read scenario #4. Following the scenario are some possible actions based on the statistical results. For each of the statements following the scenario, determine if you think the conclusion drawn is valid or invalid, then click the appropriate button.

A researcher believes that an SAT preparation course will improve SAT scores. The researcher invites a random sample of students to take the online prep course, free of charge. All of these students agree to participate. The researcher then conducts a statistical significance test (.05 significance level) to compare the mean SAT score of this random sample of students who took the review course to a historical average (500). She hopes that the students have a higher mean score than the historical average. The researcher finds a *P*-value for her sample of .03.

Table E4

Drawing Conclusions about Statistical Significance Items: Number and Proportion of Correct Responses

	RPASS-4 item wording as seen by respondents	Correct conception or misconception	Correct responses	
			Number	Proportion
17.	Conclusion: The small *P*-value does not necessarily mean that there is a large improvement in scores. □ Valid Conclusion □ Invalid Conclusion	*Practical significance* (B-1) – Understanding a small *P*-value does not necessarily mean that there is a large or practical difference or effect.	150	.67
18.	Conclusion: Recall that the significance level is .05 and the *P*-value is .03. □ The .05 suggests the prep course mean scores are higher than 500. □ The .03 suggests the prep course mean scores are higher than 500. □ Since .03 is smaller than .05, the evidence suggests the prep course is not helpful.	*Type I / α and P-value* (R-3)—Confusing significance level alpha or Type I error rate with the *P*-value.	149	.67
19.	Conclusion: If there were an even greater difference between the scores of these students who took the SAT preparation course and the historical average, we would expect to obtain an even smaller *P*-value than .03. □ Valid Conclusion □ Invalid Conclusion	*Large difference or effect and P-value* (B-1)—Understanding the bigger the difference in observed results between two groups, the smaller the *P*-value and more significant the results, if all else remained same.	170	.76

Table E4 (continued)

	RPASS-4 item wording as seen by respondents	Correct conception or misconception	Correct responses	
			Number	Proportion
20.	Statement: The small *P*-value of .01 is the probability that the null hypothesis (that there is no difference between the two population means) is false. □ True Statement □ False Statement	*Probability: null is false* (H-4)—Misinterpreting *P*-value as the probability the null hypothesis is false.	124	.55
21.	Statement: The *P*-value of .01 is the probability of the null hypothesis (that the new driving school curriculum had no effect) being true. □ True Statement □ False Statement	*Probability: null is true* (H-3)—Misinterpreting the *P*-value as the probability that the null hypothesis is true.	98	.44
22.	Statement: The probability is .01 that the experimental hypothesis (that the new driving school curriculum has an effect) is false. □ True Statement □ False Statement	*Probability: alternative is false* (H-2)—Misinterpreting the *P*-value as the probability the alternative hypothesis is false	135	.60
23.	Statement: The probability that the experimental (i.e., the alternative) hypothesis is true is .01. □ True Statement □ False Statement	*Probability: alternative is true* (H-1)—Misinterpreting *P*-value as the probability the alternative hypothesis is true.	136	.61

Note. *N* = 224. Mean proportion of correct responses for drawing conclusions about statistical significance items, $\mu_{\hat{p}}$ = .61.

Appendix E.1 (continued)

There were four items related to Scenario 5 (Tying *P*-values back to Hypotheses) as detailed in Table E5. The instructions and scenario read:

Tying P-values back to Hypotheses

The student online instructions for tying *P*-values back to hypotheses read: “In this section the questions are based on scenario #5. Read scenario #5. Following the scenario are some possible conclusions to be made about the null and alternative hypotheses. For each of the statements following the scenario, determine if the conclusion described is valid or invalid, then click the appropriate button.”

Suppose you have a new driving school curriculum which you suspect may alter performance on passing the written exam portion of the driver's test. You compare the means of randomly selected and randomly assigned control and experimental groups (20 subjects in each group). You use a 2-sample test of significance and obtain a *P*-value of 0.01. Please mark each of the following statements as ‘true’ or ‘false.’

Table E5

Tying P-values Back to Hypotheses Items: *Number and Proportion of Correct Responses*

	RPASS-4 item wording as seen by respondents	Correct conception or misconception	Correct responses	
			Number	Proportion
24.	Statement: You can conclude that if the experiment were repeated a great number of times, you would obtain a statistically significant result on 99% of occasions. □ True statement □ False statement	*Reliability and P-value* (R-5)—Believing the *P*-value is related to repeatability of the results; believing 1 - *P*-value is the reliability of the results.	89	.40
25.	Statement: Repeating the randomization process many times to model no difference between means, the *P*-value of .01 is the long-run relative frequency (i.e., the probability) of obtaining experimental results at least as extreme as those observed. □ True Statement □ False Statement	*Simulation definition* (B-1)—Understanding that an empirical *P*-value can be obtained using a simulation.	169	.75
26.	Statement: The *P*-value may reflect a difference between the two samples observed but has no bearing on whether there is a statistically significant difference in the population. □ True Statement □ False Statement	*Sample and population* (R-2) Confusing whether statistically significant results refer to a sample or a population.	141	.63

Table E5 (continued)

	RPASS-4 item wording as seen by respondents	Correct conception or misconception	Correct responses	
			Number	Proportion
27.	Statement: If there were a larger sample size and the sample results turned out the same, how would the *P*-value change? □ There would be a larger *P*-value. □ There would be a smaller *P*-value. □ The *P*-value would remain the same.	*Sample size and significance* (R-4)—Understanding larger sample sizes yield smaller *P*-values, and more statistically significant observed results, if all else remains the same.	82	.37

Note. N = 224. Mean proportion of correct responses for tying *P*-values back to hypotheses, $\mu_{\hat{p}}$ = .54.

E.2 RPASS-4 Reliability Analysis

Table E6

RPASS-4 Proportion of Correct Responses, Corrected Item-total Correlation, and α-if-item-deleted, Sorted by Proportion Correct within Blueprint Category (α = .42, N = 224)

RPASS-4 correct conception (C) or misconception (M)		Blueprint category	Proportion of correct responses	*SD*	Corrected item-total correlation	α-if-item-deleted
5. Smaller the *P*-value	C	B-1[a]	.78	.41	.26	.380
19. Large difference or effect	C	B-1	.76	.43	.21	.387
15. *P*-value as always low	M	B-2[a]	.76	.43	.32	.368
25. Simulation definition	C	B-1	.75	.43	.09	.408
10. Strong statistical evidence	C	B-1	.74	.44	.24	.381
12. *P*-value as rareness measure	C	B-1	.74	.44	.24	.381
1. Textbook definition	C	B-1	.74	.44	.23	.383
7. *P*-value in sampling variation	C	B-1	.72	.45	.06	.414
3. Lay definition	C	B-1	.69	.46	.11	.404
17. Practical significance	C	B-1	.67	.47	-. 06	.435
2. Dependence on alternative	C	B-1[a]	.54	.50	.10	.406
16. Weak statistical evidence	C	B-1	.53	.50	.06	.414
6. *P*-value and standard error	M	B-1	.46	.50	.02	.424
18. Type I / α and *P*-value	M	R-3[a]	.67	.47	.42	.342
13. Test statistics and *P*-value	M	R-1	.65	.48	.08	.411
26. Sample and population	M	R-2	.63	.48	.14	.399
8. Confidence interval & significance	C	R-6	.58	.49	-.16	.457
24. Reliability and *P*-value	M	R-5	.40	.49	.01	.425
27. Sample size and significance	C	R-4[a]	.37	.48	.11	.404
11. Chance as cause of results	M	L-3	.69	.46	.32	.364
4. Conclusions and study design	M	L-4	.51	.50	.18	.390
14. Converse as true	M	L-2	.37	.48	.18	.391
9. Inverse as true	M	L-1	.35	.48	-.17	.457
23. Probability: alternative is true	M	H-1	.61	.49	.07	.412
22. Probability: alternative is false	M	H-2	.60	.49	-.08	.442
20. Probability: null is false	M	H-4	.55	.50	.15	.396
21. Probability: null is true	M	H-3	.44	.50	-.15	.456

Note. RPASS-4 average scale difficulty 16 correct / 27 items = .60, *SD* = 3 items; assessed 13 correct conceptions and 14 misconceptions. [a]Three-option item.