IS THE *P*-VALUE REALLY DEAD? ASSESSING INFERENCE LEARNING OUTCOMES FOR SOCIAL SCIENCE STUDENTS IN AN INTRODUCTORY STATISTICS COURSE[1]

SHARON J. LANE-GETAZ
St. Olaf College
lanegeta@stolaf.edu

ABSTRACT

In reaction to misuses and misinterpretations of p-values and confidence intervals, a social science journal editor banned p-values from its pages. This study aimed to show that education could address misuse and abuse. This study examines inference-related learning outcomes for social science students in an introductory course supplemented with randomization and simulation content. Learning gains were measured across a suggested taxonomy of inference learning outcomes using the Reasoning about P-values and Statistical Significance (RPASS-10) scale. Three graphical comparisons of students' Pretest and Posttest proportions were encoded by learning gain or loss, an inference learning outcome taxonomy, or if a correct concept or misconception was assessed. What students learned and the difficulties that persisted shape recommendations for teaching and future research.

Keywords: *Statistics education research; p-values; inference; misconceptions; hypothesis tests; statistical significance*

1. INTRODUCTION

This classroom-based action research study was motivated, in part, to refute that inference procedures should be banned from applied social science as was suggested by Trafimow and Marks (2015). The myriad of misconceptions, misinterpretations and misuses that are associated with null hypothesis significance testing (NHST), *p*-values and statistical significance may be greatly reduced with a proper educational focus. Some of the criticisms are merely attempts to push a Bayesian agenda. Convoluted arguments against NHST have contributed to the misunderstanding of these procedures and clouded the real challenges that students have learning to use these procedures properly.

The review of the literature gave rise to a working taxonomy of difficulties that people have had with Null Hypothesis Significance Test (NHST) procedures, *p*-values and, to a lesser degree, confidence intervals. This taxonomy of inference learning outcomes provided a framework for assessing learning gains. The lower-order outcomes require that students: (1) recognize basic concepts and (2) differentiate similar concepts. The higher-order outcomes require that students: (3) interpret inferential results and (4) evaluate validity of the procedures and the inferences drawn from the procedures.

In this study student learning gains are measured across the taxonomy of inference learning outcomes. The study aims to improve inference learning outcomes in a statistical literacy-based social science course, merely by eliminating some probability content and adding randomization and simulation content. In addition, the study aims to refute the claim that inference procedures should be banned from applied social science journals

Statistics Education Research Journal, 16(1), 357-399, http://iase-web.org/Publications.php?p=SERJ

due to improper use. If introductory level social science students can achieve more than just statistical literacy by taking a one semester introductory course, this bodes well for proper use of inferential procedures among future researchers in the social sciences.

2. LITERATURE REVIEW

2.1. BACKGROUND

Researchers have discussed the many difficulties people have understanding inference, statistical significance, hypothesis testing, *p*-values and confidence intervals (Batanero, 2000; Garfield & Ahlgren, 1988; Lane-Getaz, 2007; Wainer & Robinson, 2003; Utts, 2003). Some researchers have documented that introductory statistics students come to class with pre-existing misconceptions about reasoning under uncertainty that may present an obstacle to teaching and learning inferential topics (e.g., Konold, 1995; Kahneman & Tversky, 1982). During an introductory statistics course, a student may come to understand an isolated definition well and yet have difficulty differentiating that concept from another, similar concept encountered later in the course (e.g., *p*-values and significance levels). These are among the reasons that inferential procedures may be difficult to learn, to teach and to assess. Isolated assessment questions are not sufficient to assess student understanding about inference but analyzing the patterns among multiple items may provide a clearer picture of inference learning outcomes.

There has been an ongoing controversy—particularly among some social scientists, psychologists, and education researchers—about the use of NHST procedures, ostensibly due to widespread misuses, misinterpretations, and misconceptions surrounding the procedures. Most recently, Trafimow and Marks (2015) called for a ban of NHST procedures from the *Journal of Basic and Applied Social Psychology*, joining a chorus of researchers in psychology, education and measurement (e.g., Carver, 1978, 1993; Cohen, 1994; Kline, 2004; Schneider, 2013; Wainer & Robinson, 2003; Wilkinson et al., 1999). Trafimow and Marks suggest that the *p*-value is dead—or at least call for its execution—among those who do research in applied social psychology. Twenty years ago educational researchers responded differently, suggesting that the problem might not be with the tests:

> The fact that many researchers 'are' now inappropriately using tests of statistical significance does not necessarily mean that researchers 'ought' to abandon statistical tests. (Thompson, 1996, p. 28)

2.2. USE OF NULL HYPOTHESIS SIGNIFICANCE TESTING

Sir Ronald Fisher (1929) who popularized the use of *p*-values and significance testing by applied researchers held that statistical tests are an essential part of an ongoing research investigation. Fisher continued:

> Their [the *p*-value's] function is to prevent us from being deceived by accidental occurrences…. It is common practice to judge a result significant, if it is of such a magnitude that it would have been produced by chance not more frequently than once in twenty trials. …The test of significance only tells him [the researcher] what to ignore, namely all experiments in which significant results are not obtained. He should only claim that a phenomenon is experimentally demonstrable when he knows how to design an experiment so that it will rarely fail to give a significant result. Consequently, isolated significant results which he does not know how to reproduce are left in suspense pending further investigation. (p. 189)

In response to the prevalent criticism of NHST procedures by psychologists, Krantz (1999) also describes the role of the significance test as a small part in a larger research investigation. Krantz argues that too often applied researchers believe they are using significance tests to validate a theory. Most NHST studies do not provide evidence that validate nor falsify a theory in the Popperian sense (see Popper, 1963). "What is missing from this viewpoint is that formal statistical methods are not the whole, but only a part of *inductive inference* (and in many areas of science only a minor part)" (Krantz, 1999, p. 1374).

Krantz attempts to clarify that NHST procedures have five distinct uses, to:

1. Check the adequacy of a provisional working model;
2. Evaluate important model parameters;
3. Show results that seem to confirm that a theory is not attributable to mere chance;
4. Test a serious (approximate) null hypothesis; and
5. Choose an appropriate action or policy.

It is the "confusion among these distinct uses [that] underlies both the abuses of hypothesis testing…and the overreaction of the critics" (p. 1375). However, when properly used and interpreted, *p*-values and NHST can move research in a verifiable, philosophically and scientifically sound direction. Statisticians develop statistical methods to be useful in the analysis of data; "if people misapply them, this is…a problem for education, not for statistical research. …The intrinsic value of statistical methods is judged by their costs and their benefits when properly used, not by the blunders of the poorly educated" (p. 1374).

2.3. THE TAXONOMY OF INFERENCE LEARNING OUTCOMES

Toward a proper understanding and use of inference procedures, psychology and statistics education researchers have documented important inferential concepts to learn and differentiate, as well as common misconceptions and misinterpretations to suppress. The review of the literature gave rise to a working taxonomy of difficulties that people have had with NHST procedures, *p*-values and, to a lesser degree, confidence intervals (see Table 1.) Students should be able to recognize six basic concepts related to the *p*-value, denoted as *Basic-1* through *Basic-6* for reference throughout this paper. Students should also be able to differentiate between seven interconnected concepts, denoted as *Connected-7* through *Connected-13*. They should also be able to interpret inferential results, denoted as *Logic-14* through *Logic-18*. Ultimately, students should be able to evaluate procedural and inferential validity, denoted as *Validity-19* and *Validity-20*. This taxonomy includes both correct concepts to be attained and the misconceptions that need to be suppressed.

These 20 inference learning outcomes are measured by the 36 items that comprise the *Reasoning about P-values and Statistical Significance* (*RPASS-10*) scale (Lane-Getaz, 2007; 2010; 2011; 2013). The *RPASS* scale was developed to assess the effects of different teaching methods on inference learning outcomes. The current version of this scale (version 10) appears in Appendix A and is further described in Section 3.3. Each of the 20 inference learning outcomes enumerated in Table 1 is tied to at least one *RPASS-10* item, either explicitly or implicitly. The mapping of the taxonomy of inference learning outcomes onto *RPASS-10* is documented in Appendix B.

Table 1. Taxonomy of inference learning outcomes with literature references

Inference learning outcome	Reference

Recognize basic concepts (Basic-1 through Basic-6)	
1. Recognize that a small *p*-value measures rareness or unusualness, when the null hypothesis is true.	Carver, 1978; Fisher, 1929; Saldanha & Thompson, 2006; Schneider, 2015
2. Recognize that a small *p*-value is indicative of statistical significance	Lane-Getaz, 2013
3. Recognize that the *p*-value is conditioned on the null hypothesis being true.	Ancker, 2006; Diaz & Batanero, 2009; Falk, 1986
4. Recognize that the *p*-value is indirectly related to sample size.	Mogie, 2004; Wilkerson & Olson, 1997
5. Recognize that the magnitude of the *p*-value depends on the direction of the alternative hypothesis.	Lane-Getaz, 2013
6. Recognize that the *p*-value may not be small; large *p*-values indicate the sample did not support the research hypothesis.	Lane-Getaz, 2013; Williams, 1999
Differentiate connected concepts (Connected-7 through Connected-13)	
7. Differentiate a *p*-value from the significance level (α).	Hubbard & Bayarri, 2003;
8. Differentiate between Type I and Type II (β)□error.	Schneider, 2015
9. Differentiate statistical significance from practical importance.	Tyler, 1931; Gliner, Leech, & Morgan, 2002
10. Differentiate strength of evidence (*p*-values) from the size of an effect.	Gliner, Leech, & Morgan, 2002
11. Differentiate sample statistics from population parameters.	Lane-Getaz, 2013; Mittag & Thompson, 2000
12. Differentiate reliability or repeatability from statistical significance; "1 - *p*-value" is not a measure of reliability. (Note: Not explicitly assessed in this study.)	Oakes, 1986; Haller & Kraus, 2002
13. Differentiate variation within (spreads) from variation between (effects).	Reading & Reed, 2010; Zieffler, Garfield, delMas, & Reading, 2008; Wild, Pfannkuch, Regan, & Horton, 2011
Interpret inferential results (Logic-14 through Logic-18)	
14. Interpret a confidence interval to assess statistical significance as a complement to—or in lieu of—NHST and *p*-values.	Lane-Getaz, 2013 Capraro, 2004; Cumming & Fidler, 2002
15. Suppress the misinterpretation of the *p*-value as the $P(H_o\|data)$; switching the null hypothesis with the data in the conditional probability; aka confusion of the converse. (Note: Not explicitly assessed in this study.)	Batanero, 2000; Cohen, 1994; Falk & Greenbaum, 1995; Lane-Getaz, 2007
16. Suppress the misinterpretation of the *p*-value as a deterministic proof by contradiction. Inferential logic introduces probabilistic thinking, Type I error and pre-conditions for inference to be satisfied; aka illusion of contrapositive proof by contradiction.	Batanero, 2000; Cohen, 1994; Falk, 2008; Falk & Greenbaum, 1995; Hagen, 1997; Kirk, 1996
17. Suppress the misinterpretation of the *p*-value as the probability that research results were "due to chance;" aka odds-against-chance fallacy.	Carver, 1978, p. 5
18. Suppress the misinterpretation of the *p*-value as the probability that one of the hypotheses (null or alternative) is true or false.	Oakes, 1986; Haller & Kraus, 2002
Evaluate procedural and inferential validity (Validity-19 through Validity-20)	
19. Evaluate validity of the procedure based on how well the necessary conditions for inference were met.	Hahn & Meeker, 1993

20. Evaluate the validity of inferences to be drawn based on how randomization was used in the study design, aka scope of inference.	Lane-Getaz, 2013; Ramsey & Schafer, 2002; Robinson, Levin, Thomas, Pituch, & Vaughn, 2007

2.4. RESEARCH QUESTIONS

To reiterate Krantz's (1999) admonition, "The intrinsic value of statistical methods is judged by their costs and their benefits when properly used, not by the blunders of the poorly educated" (p. 1374). Rather than concede to a broad-brushed dismissal of inference procedures, Krantz's statement prompts statistics education researchers to identify the obstacles to understanding, propose interventions and assess the impact on inference learning outcomes. A new generation of social scientists awaits a proper understanding of inference. To this end, this study will address these three research questions:

1. *Which inference learning outcomes did students learn during an introductory statistics course?*
2. *After instruction, which of the inference learning outcomes remained elusive?*
3. *What do posttest explanations reveal about persistently difficult inference learning outcomes?*

3. METHODS

3.1. SUBJECTS AND SETTING

The current study was conducted during spring semester of 2014 at a small liberal arts college of approximately 3000 students located in the US upper Midwest. Pretest and posttest data were collected in an introductory-level statistics literacy course aimed at students in the social sciences. Out of 79 students enrolled in the course, 69 completed both tests and consented to participate in this study, an 87% response rate. Respondents include: (53) females, (15) males and (1) did not provide a gender response. There were (28) first years, (29) sophomores, (9) juniors and (3) seniors. Nearly half of the students majored in or intended to major in psychology or sociology/anthropology (see Table 2). Most first and second year students have not yet declared majors.

Table 2. Respondents' major or, if not yet declared, intended major, N= 69

Major[a]	Count
Psychology	20
Sociology / Anthropology	12
Biology	6
Exercise Science, Political Science (4 each)	8
Social Work, English (3 each)	6
Economics, Music, Environmental Studies, Nursing, Spanish, Theater (2 each)	12
Dance, Philosophy, Art (1 each)	3
Undecided or n/a	2

Note. [a]First major mentioned in the class survey is reflected, if multiple majors were described.

3.2. TEACHING PHILOSOPHY AND COURSE CONTENT

This introductory statistics course was designed to improve statistical literacy for students who planned to major in the social sciences. There is an algebra prerequisite; no calculus is required. The 79-student class met twice per week (85 and 80 minute sessions). Lecture days were closely aligned with the text and included some in-class group activities, group quizzes, and investigations. The textbook, *Seeing through Statistics* (Utts, 2005) has been lauded for helping students develop statistical literacy, for taking a critical eye to statistics encountered in the real world, and for its emphasis on understanding statistics rather than computing them (Cavanaugh, 2007; Lamprecht, 1996). To meet the needs of the client disciplines, students participated in a 55-minute per week lab to learn the *Statistical Program for the Social Sciences (SPSS).* The class was split into three subsections for the labs. Students developed skills necessary for their culminating research projects with the help of two teaching assistants and the instructor. The final two weeks of lectures diverged from the textbook material to discuss creation and interpretation of *SPSS* output for *t*-tests, chi-square tests, ANOVA and regression, reinforcing the lessons from the lab sessions.

As a statistics education reformer, the instructor aimed to emphasize a stronger conceptual foundation for inference by adding randomization and simulation activities to the course. Simulation has been shown to improve conceptual understanding of inference (Lane-Getaz, 2013; Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011). Statistics education reformer Cobb (2007) recommended stressing the randomization and simulation process in the introductory course. He also coined a handy phrase, to stress “The Three R's of inference,” namely to: Randomize the data, Repeat the process, and Reject models that put your data in the tails of the null distribution. After multiple semesters teaching the course without the randomization and simulation content, a two-week module was piloted during spring 2014 in lieu of the probability content (i.e., omitting Utts (2005), Chapters 16-18). “Randomize, Repeat, and Reject” was the drumbeat for the added lectures and labs. Refer to Appendix C for the schedule of textbook chapters, lecture topics and lab topics by day.

Two lab sessions were devoted to the randomization test: one for a categorical response variable and the other for a quantitative response variable. For the categorical lab students shuffled playing cards to gain a concrete experience with the randomization process. After creating a null distribution with the class data, students used an online applet to repeat the random assignment many more times and ultimately to make a rejection decision (see Dolphin Study, Rossman (2008), at http://www.rossmanchance.com/applets/). During the quantitative lab students randomized quantitative response data. They shuffled pieces of paper with the quantitative values written on them to conduct the randomization by hand, computed and plotted the means, and produced a null distribution from the class data. Again, students used an online applet to repeat many times, then made a rejection decision (http://www.rossmanchance.com/applets/randomization20/Randomization.html).

3.3. MEASUREMENT

RPASS-10 Pretest and Posttest Pretest and posttest data were collected using the *Reasoning about P-values and Statistical Significance* (*RPASS-10*) scale. The *RPASS* was designed as a research tool to assess the effects of different teaching methods on inference learning outcomes (see Lane-Getaz, 2013). The *RPASS-10* Pretest was administered during the first weekly lab session of the semester. Similarly, the *RPASS-10* Posttest was administered during the last weekly lab session. *RPASS-10* scenarios and items appear in Appendix A.

***RPASS-10* Item content, formats and scoring** *RPASS-10* items are a sampling of conceptual knowledge from across the domain of inference learning outcomes. The 36 *RPASS-10* items explicitly assess eighteen of the concepts listed in the taxonomy of inference learning outcomes (Table 1). Eighteen items assess whether the respondents *recognize basic concepts*. Twelve items assess if they *differentiate connected concepts*. Five items assess how the respondents *interpret inferential results*. Two items assess whether they can *evaluate procedural and inferential validity*.

The *RPASS-10* instrument presents contextual scenarios followed by multiple-choice items. There are 24 two-option items, 9 three-option items and 3 four-option items. All 36 items are dichotomously scored, right (1) or wrong (0). In addition, students are asked to explain their reasoning for 19 of the 36 items that were chosen due to previous challenges students had with the particular item or concept.

Refer to Appendix B for a listing of the *RPASS-10* scenario numbers and item numbers grouped within the taxonomy of inference learning outcomes, ordered by the Posttest proportion correct. Also noted is the Pretest proportion correct, the learning gain (or loss) for the item, and a categorization as a correct concept (C) or misconception (M).

3.4. PROCEDURES

Graphical and numerical summaries and reliability estimates are reported for the *RPASS-10* Pretest and Posttest total score distributions. The internal consistency reliability estimates are reported using Cronbach's coefficient α and Guttman's Lambda-6 ($\lambda6$). Guttman's $\lambda6$ estimate tends to be higher than Cronbach's α for dichotomously scored items, as is the case with *RPASS-10.* Beyond this initial analysis of the total score distributions and the reliability of the total scores, an item level analysis was conducted to report what the students learned in light of each of the three research questions.

Which of the inference learning outcomes did students learn during an introductory statistics course? Two plots of item responses help to examine this first research question. The items were plotted as a coordinate pair (p_1, p_2), where p_1 is the proportion of respondents answering correctly on the Pretest and p_2 is the Posttest proportion answering correctly. The $y = x$ line superimposed on the plot delineates items with no change (no learning gain nor loss). A "canoe-shaped" 95% confidence band along $y = x$ demarcates the area of plausible variation, if there were no change in Pretest to Posttest proportions (Lane-Getaz, 2014). Items appearing outside the confidence bands indicate statistically significant differences in the proportions answering correctly. Items within the confidence bands suggest insignificant differences. The margin of error includes the Wilson adjustment to maintain the 95% nominal rate (Agresti & Caffo, 2000). Because no inferential conclusions were drawn, no family-wise corrections were made. The two canoe plots provide two lenses for examining item results: one plot was encoded by statistically significant learning gains (or losses) and the other plot was encoded by the inference learning outcomes taxonomy. A cross-tabular summary provides the count and proportion of learning gains or losses broken out by the inference learning outcome taxonomy.

After instruction, which of the inference learning outcomes remained elusive? Misconceptions may or may not be elusive concepts for these respondents. The teaching method used may have suppressed some common difficulties on the one hand and on the other hand may have inadvertently introduced other challenges. A third canoe plot was produced to examine potential patterns among correct concepts and misconceptions related to learning gains and losses. A cross-tabular summary reports the count and proportion of learning gains or losses by a categorization of the items as correct concept or misconception.

Elusive concepts were defined as any items with Posttest proportions with less than 70% of respondents answering correctly. Four quadrants were delineated on the canoe plots at the lines for a 70% correct response on Pretest and Posttest. Generally, items with low Posttest proportions, that appear in the lower two quadrants of the canoe plot ($p_2 <$ .70), indicate persistent difficulties. The most surprising results are items with a high Pretest ($p_1 \geq .70$) and a low Posttest proportion ($p_2 < .70$) that appear in the lower right quadrant of the canoe plot. These item results suggest that something in the course may have introduced a misconception or difficulty. The upper right quadrant items indicate concepts the respondents knew on both the Pretest and Posttest. These items are indicative of prior knowledge, ostensibly from statistics exposure within the K-12 curriculum. Finally, items in the upper left quadrant with low Pretest and high Posttest are most desirable, indicating learning during the course.

What do Posttest explanations reveal about persistently difficult inference learning outcomes? The analysis of respondents to answer this question was limited to the intersection of: (1) 19 items where an explanation was requested; (2) seven of these items that surfaced as elusive; and (3) items with the lowest Posttest proportion correct within an inference learning outcome category (see Appendix A). The outcomes discussed include: *recognizing a basic concept* (item 3b-3), *differentiating a connected concept* (item 3c-1), and *interpreting an inferential result* (item 2-5). There was no explanation request associated with the difficult item in the *evaluating validity* category. Examining incorrect themes in respondents' explanations may help identify the crux of the respondents' confusion. Once the difficulty is documented, a teaching intervention can be designed and tested to attempt to address the misconception or difficulty.

4. RESULTS

The 69 respondents answered 72.7% of the 36 RPASS-10 items correctly for the Posttest, on average (*M* = 26.16, *SD* = 4.32, *Med* = 26, *IQR* = 5) compared to 55% correct on the Pretest (*M* = 19.8, *SD* = 3.72, *Med* = 26, *IQR* = 5). The average gain was 6.36 items. Pretest and Posttest distributions for the total correct scores appear in Figure 1 and 2 as comparative boxplots and density plots, respectively. Internal consistency reliability, the proportion of variation in *RPASS-10* scores that are attributed to true score variation (rather than error), are Posttest: $\lambda 6$ = .88, α = .69, and Pretest: $\lambda 6$ = .73, α = .42. Further results of this study are organized by the three research questions.

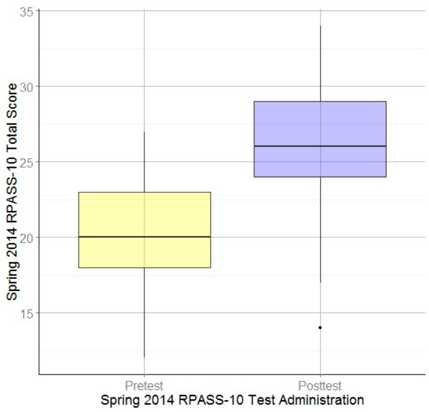


Figure 1. Boxplot of RPASS-10 Total Scores: Pretest and Posttest, N = 69

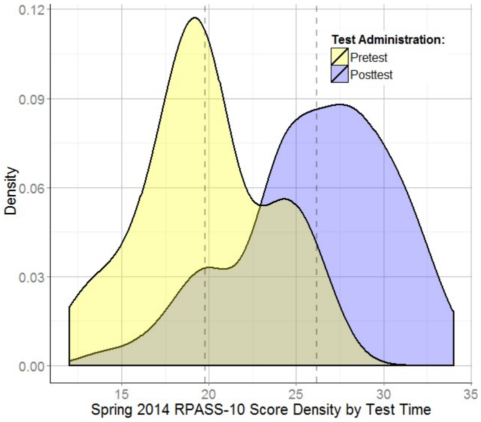


Figure 2. Density plot of RPASS-10 Total Scores: Pretest and Posttest, N = 69

4.1. RESULT 1: INFERENCE LEARNING OUTCOMES THAT STUDENTS LEARNED DURING THE INTRODUCTORY STATISITICS COURSE

Delving deeper than aggregated scores, an item-level analysis reveals more about what respondents learned. Item results are reported by statistically significant learning gains or losses and by the taxonomy of inference learning outcomes. The "canoe plot" in Figure 3 plots the proportion of students answering each item correctly on the Pretest (p_1) and Posttest (p_2) as an ordered pair (p_1, p_2), encoded by statistically significant learning gains or losses.

There were 20 items above the 95% confidence band that indicated statistically significant learning gains. Fourteen items within the confidence bands showed no significant change. Two items below the bands showed a statistically significant learning loss. It is important to note that there were items with a low Posttest ($p_2 < 70\%$) and a statistically significant learning gain. Even though these concepts are apparently difficult, considerable progress was made in a one-semester course. See Appendix B for a listing of the *RPASS-10* items with the proportion correct for the Posttest and Pretest, and the Learning Gains by item listed within the taxonomy of inference learning outcomes.

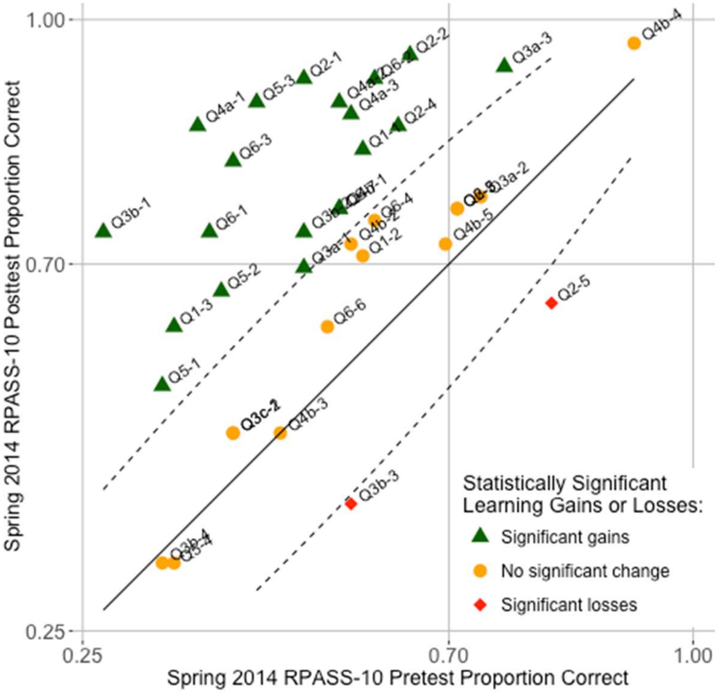


Figure 3. 36 RPASS-10 items plotted as ordered pairs (p_1, p_2) with Pretest Proportion (p_1) and Posttest Proportion (p_2), encoded by Learning Gains or Losses, N = 69

In Figure 4 the item results are encoded by the taxonomy of inference learning outcomes. Furthermore, Table 3 provides a cross-tabulation of the count and proportion of *RPASS-10* items by significant learning gains or losses broken out by the inference learning outcome taxonomy.

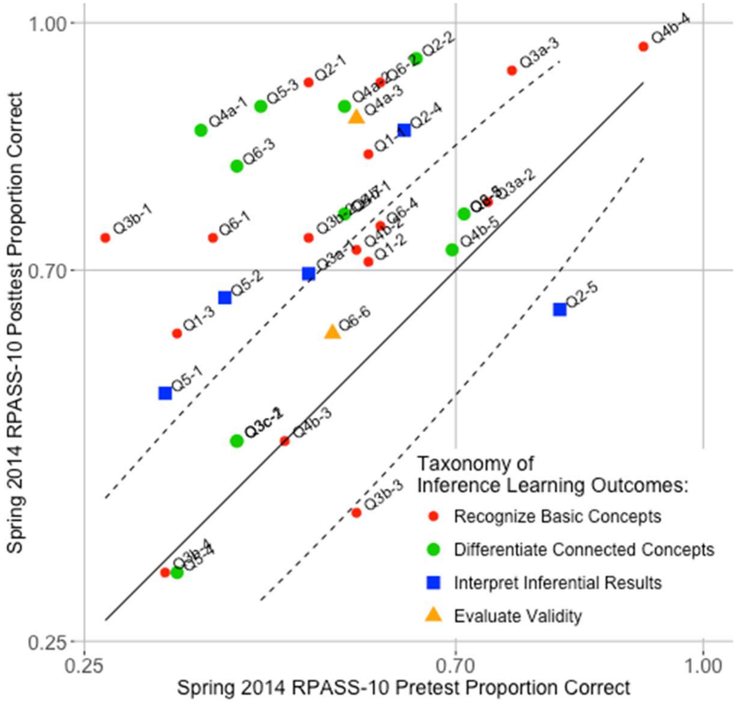


Figure 4. 36 RPASS-10 items plotted as ordered pairs (p_1, p_2) with Pretest Proportion (p_1) and Posttest Proportion (p_2), encoded by Inference Learning Outcomes, N = 69

Table 3. Cross-tabulation of the Count and Proportion of 36 RPASS-10 items by Significant Learning Gain or Loss broken out by Inference Learning Outcome

	Inference Learning Outcome								
Significant Learning Gain or Loss	Recognize Basic Concepts		Differentiate Connected Concepts		Interpret Inferential Results		Evaluate Validity		Total
Learning Gain	9	.50	6	.55	4	.80	1	.50	20
No change	8	.44	5	.45	0	.00	1	.50	14
Learning Loss	1	.06	0	.00	1	.20	0	.00	2
Total	18		11		5		2		36

There were significant gains in the proportion of respondents answering items correctly across all four categories of the inference learning outcomes. Respondents *recognized basic concepts*, *differentiated connected concepts, interpreted inferential results* and *evaluated validity*. Items contributing to these statistically significant learning gains are described.

Gains recognizing "basic concepts" Nine of the 18 *basic concept* items appear above the confidence bands, indicating statistically significant learning gains in the proportion of respondents answering correctly. Respondents learned to recognize four concepts:

- That a small *p*-value measures rareness or unusualness, when the null hypothesis is true (*Basic-1:* 1-1, 2-1, 6-1).
- That a small *p*-value is indicative of statistical significance (*Basic-2:* 6-2).

- That the magnitude of the *p*-value depends on the direction of the alternative hypothesis (*Basic-5:* 1-3, 4b-1, 3b-2) (Note: For Item 1-3, despite statistically significant gains the Posttest proportion was still lower than the threshold, $p_2 < 70\%$.)
- That the *p*-value may not be small; large *p*-values indicate the sample did not support the research hypothesis (*Basic-6:* 3a-3, 3b-1).

Gains differentiating "connected concepts" Six of the 11 items associated with *connected concepts* showed significant learning gains during the course. The items are associated with three connected concepts. Respondents were able to differentiate:

- *P*-values from significance level (α) (*Connected-7:* 4a-1).
- Concepts of Type I (α) and Type II (β) error (*Connected-8:* 6-7).
- Strength of evidence (*p*-values) from size of an effect (*Connected-10:* 2-2, 4a-2, 5-3, 6-3).

Gains interpreting "inferential results" Four of the five *interpret inferential results* concepts are measured by *RPASS-10.* Two misinterpretations were successfully suppressed with statistically significant gains. Respondents were able to suppress:

- The misinterpretation of the *p*-value as the deterministic *contrapositive proof by contradiction* (*Logic-16:* 3a-1).
- The misinterpretation of the *p*-value as the probability that research results were "due to chance;" aka *odds-against-chance fantasy* (*Logic-17:* 2-4).

Gains evaluating "validity" There was a statistically significant gain in the proportion of respondents who evaluated the validity of inferences to be drawn. Even with a small *p*-value, respondents understood that inferential validity is limited by how randomization was used in the study design (*Validity-20*: 4a-3).

4.2. RESULT 2: INFERENCE LEARNING OUTCOMES THAT REMAINED ELUSIVE AFTER INSTRUCTION

A third canoe plot facilitates identifying patterns among correct concepts and misconceptions that may be related to learning gains and losses (Figure 5). *RPASS-10* items measure 11 correct concepts, 12 known misconceptions and 13 items measure a combination of both. Table 4 reports the count and proportion of *RPASS-10* items with statistically significant learning gains or losses broken out by a categorization of the item as a correct concept, misconception or if the items measures a combination of both.

The cross-tabular results suggest that the number of items with significant learning gains primarily stems from suppressing or overturning known misconceptions. Seventy-five percent of respondents made statistically significant gains relative to known misconceptions. Items that remained elusive—in the lower two quadrants of the canoe plots—include correct concepts, misconceptions as well as combinations of the two.

Table 5 provides a cross-tabulation of the count and proportion of *RPASS-10* items with Posttest proportions above or below the 70% criterion and broken out by the four inference learning outcomes categories. Seven concepts emerge that seem to remain elusive, at least one from each of the inference learning outcome categories. Three of these items (3b-3, 3c-2, and 2-5) are explored further in Section 4.3 by examining respondent Posttest explanations.

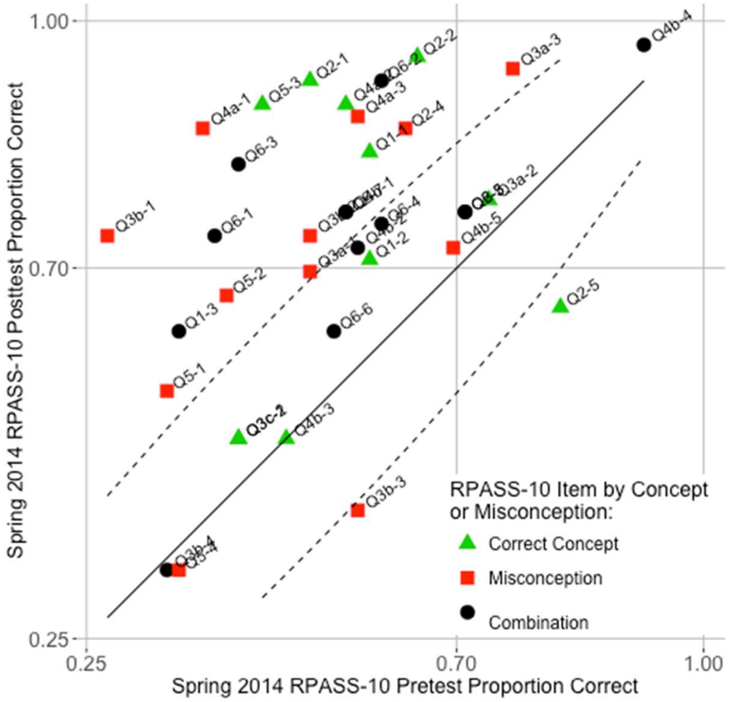


Figure 5. 36 RPASS-10 items plotted as ordered pairs (p_1, p_2) with Pretest Proportion (p_1) and Posttest Proportion (p_2), encoded by Concept or Misconception, N = 69

Table 4. Cross-tabulation of the Count and Proportion of 36 RPASS-10 items by Significant Learning Gain or Loss broken out by Correct Concept or Misconception

Significant Learning	Correct Concept or Misconception Assessed						
Gain or Loss	Correct Concept		Misconception		Combination		Total
Learning Gain	5	.45	9	.75	6	.46	20
No change	5	.45	2	.17	7	.54	14
Learning Loss	1	.10	1	.08	0	.00	2
Total	11		12		13		36

Table 5. Cross-tabulation of the Count and Proportion of 36 RPASS-10 items by High or Low Posttest proportion broken out by Inference Learning Outcome

	Inference Learning Outcome								
Posttest proportion correct	Recognize Basic Concepts		Differentiate Connected Concepts		Interpret Inferential Results		Evaluate Validity		Total
High Posttest ($p_2 \geq .70$)	14	.78	8	.73	2	.40	1	.50	25
Low Posttest ($p_2 < .70$)	4	.22	3	.27	3	.60	1	.50	11
Total	18		11		5		2		36

Two "basic concepts" seemed to remained elusive The first elusive *basic concept* was recognition that the magnitude of the *p*-value depends on the direction of the alternative hypothesis (*Basic-5*). Even though item 1-3 seems to have been difficult, there were statistically significant gains for this item. Furthermore, the remaining three *Basic-5* items did not corroborate this difficulty (see Appendix B). The second elusive concept was recognition that the p-value may not be small; that large p-values indicate the sample did not support the research hypothesis (*Basic-6*). Three of the five *Basic-6* items (3b-3, 3b-4, 4b-3) had low Posttest proportions, which is indicative of a systematic underlying misconception. Furthermore, there was a significant learning loss for item 3b-3. Posttest explanations for item 3b-3 are further examined to shed light on the source of this apparent difficulty in Section 4.3.

Two "connected concepts" seemed to remain elusive An isolated item suggests that respondents struggle to differentiate random samples from population parameters (*Connected-11:* 5-4). Regretfully, no Posttest explanations were requested for this item to examine the source of the confusion. However, there was additional evidence of this confusion in response to using confidence intervals to assess significance as reported in Section 4.3 and discussed in Section 5.3. It was difficult to tell if the students failed to make the correct inference from the sample statistics to the population parameter because they thought it was redundant or if the decision they made had another source of confusion. A new item needs to be added to the RPASS scale to better tease out this student difficulty in addition to asking students to explain their reasoning for this item.

The other elusive connected concept outcome was corroborated by two items on the test to assess respondents' consideration of within variation and between variation (*Connected-13:* 3c-1, 3c-2). Respondent Posttest explanations for item 3c-1 are also examined in Section 4.3 and discussed in Section 5.3 to identify patterns in student thinking.

Two "inferential results" concepts seemed to remain elusive Interpreting a confidence interval to assess statistical significance as a complement to—or in lieu of—NHST was an elusive concept for many respondents (*Logic*-14). It is of concern that there was a significant learning loss for the item assessing this outcome (2-5). Explanations for this item were examined to provide some insight for a teaching intervention. A second interpret inferential results outcome presented challenges. Two items measure these well-documented misinterpretations of the p-value. One is the probability that null hypothesis is false (*Logic*-18: 5-1). The other is the probability that the alternative hypothesis is true (*Logic*-18: 5-2). While these concepts remained elusive for some, these two items had statistically significant gains, which provide evidence that these misconceptions can be overturned or at least suppressed, given proper focus.

One "validity" concept seemed to remain elusive Respondents simply failed to evaluate whether the necessary conditions for inference were met for item 6-6 (*Validity*-19). This is not a surprising outcome. Checking conditions for inferential procedures is discussed and practiced in this course. However, greater emphasis and focus is placed on checking conditions for inferential procedures in the second course.

4.3. RESULT 3: WHAT POSTTEST EXPLANATIONS REVEALED ABOUT PERSISTENTLY DIFFICULT LEARNING OUTCOMES

Seven elusive concepts emerged from the results of 11 items with low Posttest proportions ($p_2 < .70$). Two of these concepts are of less concern because there were statistically significant learning gains for the associated items (*Basic-5:* 1-3, *Logic-18:* 5-1 and 5-2). Four items did not include a prompt to "Explain your reasoning" (*Basic-6:* 3b-4, 4b-3, *Connected-11:* 5-4 and *Validity-19:* 6-6). The results of the remaining four items illuminate three elusive inference learning outcomes: one *basic concept* (*Basic-6:* 3b-3), one *connected concept* (*Connected-13:* 3c-1 and 3c-2), and one *interpretation of inferential results* concept (*Logic-14: 2-5*). The following results examine each of the three elusive concepts. Of particular interest are the two items on the *RPASS-10* Posttest with statistically significant learning losses (*Basic-6:* 3b-3 and *Logic-14:* 2-5).

Posttest explanations for recognizing that the p-value may not be small (Basic-6) There was a statistically significant learning loss for item 3b-3, one of the five items associated with "Recognize the *p*-value may not be small." In total, three items assessing this concept had low Posttest proportions, suggesting there is a real obstacle for these respondents. For item 3b-3 (see Appendix A, Scenario 3b, item 3), respondents read that student researchers expected to show that Radium-226 levels in the soil were less than the EPA maximum safe criterion of 4 pCi/L. The mean Radium-226 from the sample of 32 soil specimens was actually higher than expected, namely 4.1 pCi/L. The center of the null distribution was 4.0 pCi/L with the observed mean identified at 4.1 pCi/L. With conditions checked, respondents were asked whether the *p*-value could be illustrated by shading the area to the left of the sample mean (which is correct). Among those who answered correctly, the most prevalent theme provided by 10 respondents, is reflected by this statement:

> Yes because they were conducting a one-tailed hypothesis test stating that the levels of pCi/L would be lower than 4 pCi/L, but their results showed the levels being greater than 4 pCi/L. Therefore, the *p*-value is the area to the left of the their observed sample mean.

Among those who answered the item incorrectly, 23 respondents wrote an explanation much like this one, "The *p*-value could be estimated by shading the area to the RIGHT of the observed mean." [Emphasis in original.] Explanations such as these, reveal a clear obstacle shading the area for a *p*-value for a one-sided hypothesis test when the observed statistic is in the opposite direction from what was hypothesized by the researcher.

Posttest explanations for differentiating variation within versus between *(Connected-13)* There was no significant improvement in the proportion of respondents who could differentiate within variation from between variation. Respondents were asked to choose the *most* convincing evidence among four sets of boxplots that would distinguish between two group's running times (see Appendix A, Scenario 3c, item 1). Note that the second item (3c-2) had equivalent explanations to those for item (3c-1). One group of runners was randomized to use a standard training program and the other group added weight training to the standard training program. The Posttest explanations provided by 30 respondents who answered item 3c-1 correctly were similar to this:

> Although the mean scores are farther apart on boxplots c, boxplots a have the least amount of variablility *[sic]*. This shows that there is no huge number skewing the data, the fact that the whole group got basically the same time shows that the methods differ greatly because less variance tells us that the relationship was stronger and there were little to no big outliers.

Among those who answered item 3c-1 incorrectly, some 25 respondents wrote an explanation similar to this one:

> The medians on Boxplot C have the largest difference in distance and since the median is the average, the average difference is quite large between the weight and standard training compared to the other boxplots.

These explanations revealed that many students only attended to the centers when comparing the two distributions and failed to account for the spreads of the distributions.

Posttest explanations for using confidence intervals to assess significance (Logic-14) There was a statistically significant learning loss for the item associated with this learning outcome. Posttest respondents were asked whether a confidence interval could be used to assess statistical significance as a complement to—or in lieu of—NHST (see Appendix A, Scenario 2, item 5). Of the forty-five Posttest respondents who answered this item correctly, 39 wrote explanations and 27 were much like this one:

> If the researchers found that the 95% confidence interval did not include 100, then they would know that even taking into account plausible variation in population means, the result of 102 would not have been observed if the null hypothesis were true, therefore their results are statistically significant. They would also know the direction in which the observed results were different from the null hypothesis (whether the interval was above or below 100).

Twelve of the remaining explanations were patently wrong or vague, much like this one “Confidence intervals should include the hypothesized or observed mean in the range to show statistical significance.” Note the conflation of the sample and population implied in this quote, which may be the source of students’ difficulty with this item.

Of those who answered item 2-5 incorrectly, their explanations reflected their confusion. One of the two most prevalent explanations was written by nine respondents who wrote something much like this: “The confidence interval isn't testing statistical significance, but rather seeing if their observed mean falls into the range of likely true means.” Again, note the confusion between the sample and population in this statement. The next most common explanation was written by seven respondents who essentially wrote: “Because they should have done a 95% confidence interval test, centered around 100 and then looked at how many standard deviations away the mean of 102 was from the center to assess the statistical significance.” These students apparently wanted to conduct a hypothesis test.

5. DISCUSSION

The results of this study show that during a one-semester introductory statistics course most social science students achieved statistically significant gains for 20 of the 36 *RPASS-10* items. There were learning gains across the taxonomy of inference learning outcomes. Respondents improved in their ability to *recognize basic concepts* (*Basic-1, Basic-2, Basic-5, Basic-6*), to *differentiate connected inferential concepts* (*Connected-7, Connected-8, Connected-10*), to *interpret results* (*Logic-16 and Logic-17*) and to *evaluate inferential validity (Validity-20)*. The respondents’ inference learning outcomes are discussed in light of each of the three research questions.

5.1. DISCUSSION 1: INFERENCE LEARNING OUTCOMES THAT STUDENTS LEARNED DURING THE INTRODUCTORY COURSE

Respondents achieved significant gains recognizing “basic concepts” They recognized *p*-value definitions, statistical significance, that the *p*-value is dependent on the direction of the test, and the *p*-value may not be small (*Basic-1, Basic-2, Basic-5, and Basic-6*). There were nine basic concept items that showed significant learning gains from Pretest to Posttest. There were no learning gains associated with two concepts that were answered correctly on the Pretests and Posttests recognizing how *p*-values are indirectly related to sample size and the p-values are conditioned on the null being true (*Basic-3, Basic-4*)

Respondents achieved significant gains differentiating "connected concepts" More respondents were able to differentiate connected concepts. Inconsistent with some of the existing research on common misconceptions (Garfield & Ahlgren, 1988), respondents did differentiate p-values from the significance level α (see *Connected-7*).In some research circles p-values have errantly been dubbed "observed significance levels", which is believed to have contributed to this confusion in the past (Hubbard & Bayarri, 2003; Schneider, 2015). Careful use of language seems to obviate the confusion. Respondents also differentiated concepts of Type I error from Type II error (*Connected-8*).

The remaining connected concept is well supported by multiple item responses. Four items—all with significant gains—provide evidence that respondents were able to differentiate between the strength of evidence (p-values) and the size of an effect by the end of the course (*Connected-10*). These results are encouraging in light of the research of Gliner, Leech, and Morgan (2002) who found that few of the textbooks that they reviewed covered well the relationships between p-values, effects, and evidence. These results also run counter to the results reported by Wilkerson and Olson (1997) who found that graduate student researchers tend to confuse relationships between samples sizes, significance, and effects. For example, if two studies produce the same p-value, the graduate students failed to understand that the study with the *smaller* sample size has a *larger* treatment effect (also see Gregiore, 2001). These topics are handled well in Utts (2005), Chapter 13.4.

Respondents achieved significant gains interpreting "inferential results" Respondents came to a more subtle understanding of NHST than merely to reject the null hypothesis when $p < \alpha$. Most respondents improved in their ability to suppress known misinterpretations that can plague how they *interpret inferential results.* They suppressed the misinterpretation of the p-value as a deterministic proof by contradiction (*Logic-16*). This misinterpretation is sometimes referred to as the *illusion of probabilistic proof by contradiction*, the misapplication of definitive contrapositive logic to a probabilistic event (see Cohen, 1994; Falk, 2008; Falk & Greenbaum, 1995; Hagen, 1997; Kirk, 1996). The Boolean logic of contrapositive proof hinges on the idea that if all squares are rectangles, and if we don't have a rectangle; then we don't have a square. Probabilistic logic follows this general pattern as well. However, there remains uncertainty in statistics. One cannot reject the null hypothesis with 100% certainty. Inferential logic is fuzzy. There is the potential for Type I error. More importantly, a small p-value should always be suspect. One must evaluate whether the data collection methods warrant an inferential procedure. One must evaluate if conditions for the procedure are sufficiently met. If conditions are not checked, then the illusion of probabilistic proof by contradiction is likely. Although Boolean logic is a good foundation to interpret results, the logic of inference requires looking more broadly at the statistical process as a whole (see Falk, 2008; Hagen, 1997; O'Brien, 1973; Rubel, 2007; Tversky & Kahneman, 1974). These respondents seemed to understand this fairly well.

Similarly, respondents suppressed the misinterpretation of the p-value as the probability that research results were "due to chance." This phrase is nearly correct; if the probabilities are computed assuming the null is true. Carver (1978), a critic of significance test procedures, wrote that a small p-value (e.g., less than .05 or 1 in 20) has been misinterpreted as the "odds-against-chance fantasy" (not to be confused with the "odds against chance fallacy" aka the "gambler's fallacy," that a win will soon come after a series of losses). Regretfully, Carver's discussion of his "odds-against-chance fantasy" conflates probability with odds and *p*-values with significance levels but is worth the thought experiment. Carver cited a psychology textbook (Hebb, 1966), in which the "statistically significant results" were interpreted to mean the odds are 19 to 1 that chance caused the results observed. If the *p*-value is .05 or 1/20, are the "odds against" 19 to 1? The odds against what? This statement is mute about the null hypothesis being true. Carver explains this by using quasi-Bayesian language assigning 100% probability to the claim that the null hypothesis is true. He states:

> It is therefore impossible for the *p* value to be the probability that chance caused the mean difference between two research groups since (a) the *p* value was calculated by assuming that the probability was 1.00 that chance did cause the mean difference, and (b) the *p* value is used to decide whether to accept [*sic*] or reject the idea that probability is 1.00 that chance caused the mean difference. (p. 5)

Carver was correct to critique the book author. However, his critique obfuscates the underlying issue. It seems likely that Carver's motivation was to promote Bayesian probability in lieu of Frequentist probability, not to educate the reader on the proper interpretation of a *p*-value. Nevertheless, Carver did make a salient and important point. When one fails to reject the null hypothesis, he or she should not conclude that the results were therefore "caused by chance." What one can say, when the null hypothesis is rejected, is that the results were *not* likely to be "caused by chance." This is a subtle but important distinction. These respondents achieved significant gains recognizing that "the cause of the results obtained was clearly due to chance" was an invalid interpretation for item 2-4.

Respondents achieved significant gains evaluating "validity" Respondents did learn one aspect of evaluating validity, that inferences must be based on how the data were gathered. The presence or absence of randomization in the study design determines the scope of inferences that can be drawn (*Validity-20*). Specifically, the respondents recognized that no casual conclusion could be drawn—even if a small *p*-value were obtained—unless the researcher had randomized the treatments in the design of the study. This is an important, higher-order learning outcome, often referred to as the *scope of inference*. Some researchers claim that failing to attend to the scope of inference may underlie many of the misuses of NHST procedures (e.g., Hahn & Meeker, 1993).

5.2. DISCUSSION 2: INFERENCE LEARNING OUTCOMES THAT REMAINED ELUSIVE AFTER INSTRUCTION

As reported in the results, respondents wavered in their understanding of seven concepts across the taxonomy of inference learning outcomes: two *basic concepts,* two *connected concepts,* two *interpretation* concepts, and one *validity* concept. Each of these elusive concepts is further discussed based on the results from the associated problematic item and in light of the results from closely related items.

Discussion of two "basic concepts" that remained elusive For item 1-3, respondents were confused about whether the *p*-value is doubled or cut in half when one computes a two-tailed *p*-value given a one-tailed *p*-value. The three other items in the *Basic-5* inference learning outcome category were answered sufficiently well (4b-1, 3b-2, 4b-2). Item 1-3 is more computational than the other three items, which are presented in light of a graphical representation of a null distribution (see Appendix A). Because conceptual understanding is the goal of this course, this is not a surprising result. This computational difficulty is also reported by Aquilonious and Brenner (2015), “Mary and Nan divided all their *p*-values by two, independent of whether they were working a two-tailed test or not” (p. 23). This apparent confusion may be related to a general difficulty understanding one-sided versus two-sided tests in general. This difficulty may also be related to confusion that the significance level is split on either side of the distribution for a two-sided hypothesis test. Item 1-3 gives numeric results for a one-sided *p*-value, which requires doubling the *p*-value to produce the results for a two-sided test. Although there remains room for improvement for this item, there were statistically significant gains, which is encouraging.

A more disturbing confusion surrounds the basic interpretation of large *p*-values. Respondents must recognize that the *p*-value may not be small; large *p*-values indicate the sample did not support the research hypothesis (*Basic-6*). Respondents had considerable difficulties interpreting large *p*-values, particularly when the sample obtained was in the opposite direction than hypothesized, as is the case for item 3b-3. Williams (1999) also noted that her introductory students always expected *p*-values to be relatively small. Item 3b-3 is one of two *RPASS-10* items with a significant learning loss from the Pretest. In addition, there are low Posttest results for two additional items among the five that assess the *Basic-6* concept (3b-4 and 4b-4), suggesting a real obstacle exists. Respondent explanations for 3b-3 may shed light on the problem.

Discussion of two “connected concepts” that remained elusive Only one item is related to this confusion between samples and populations *(Connected-11*). This item (5-4) is among the lowest Posttest proportions on the test (p_2 = .33). For item 5-4 respondents had to decide if a statistically significant difference between two random samples suggests a true population difference. There was considerable confusion differentiating sample and population effects. Mittag and Thompson (2000) reported a similar finding in their survey of education researchers. The American Educational Research Association respondents believed *p*-values test the probability of results occurring in the sample, rather than the probability of results occurring in the hypothesized population. The low Posttest proportion correct in this study is consistent with results from previous introductory statistics classes (Lane-Getaz, 2013). This confusion also tends to emerge from at least one other item (i.e., using confidence intervals to assess significance, *Logic-14:* 2-5) as will be discussed.

There are two complementary items that assess an apparent confusion between variation within and variation between (*Connected-13:* 3c-1 and 3c-2). As a result these items have identically low Pretest and Posttest proportions (i.e., p_1 = .43, p_2 = .49). Looking at the explanations for one of these items provided some insight into student thinking. Respondent explanations are explored for item 3c-1 in Section 4.3 and are further discussed in Section 5.3.

Discussion of two “inferential results” concepts that remained elusive One elusive interpretation involved assessing statistical significance with a confidence interval (*Logic-14*). The one item assessing use of confidence intervals to assess significance (2-5) was one of the two items with a statistically significant learning loss. To better understand these results, respondent explanations for item 2-5 were discussed in Section 4.3. This item may be difficult for these students because the sample mean in this scenario is close to the hypothesized mean, which seems to further complicate the interpretation for these respondents. The explanations that they provided suggest that a less complex confidence interval item is needed to better corroborate and contrast results with this item. The confusion may be an in ability to differentiate population parameters from sample statistics when constructing a confidence interval.

The second elusive interpretation dealt with assigning probabilities to the null or alternative hypotheses (*Logic-18*). Some 67% of respondents correctly suppressed the misinterpretation that the *p*-value is the probability that the alternative hypothesis is true (5-2). Of course the *p*-value is a probability concerning data, not a probability related to either of the hypotheses. Even if one were to set aside this point for a moment, the fact that a third of the respondents believed that the small *p*-value of .01 for this item—which suggests rejecting the null hypothesis—means that there is also a low probability of the alternative being true reveals a deeper misunderstanding of the inferential logic.

On a related item, only 55% of the respondents suppressed the misinterpretation of the small *p*-value as the probability that the null hypothesis is false. If one ignores the issue of assigning a probability to the null hypothesis, this misinterpretation suggests that 45% of the respondents believed that the small *p*-value meant that the null hypothesis is false. Because small *p*-values are indeed inconsistent with the null, this misinterpretation suggests an emergent understanding of inferential logic. Recognizing that most students are attaining at least a partially correct understanding is encouraging and in particular, because both of these items had statistically significant gains.

These quasi-Bayesian misinterpretations of the *p*-value are commonly cited as reasons to do away with NHST and the *p*-value altogether (see Cohen, 1994; Oakes, 1986). Cohen attributes this common misinterpretation to the idea that most researchers actually want to know the probability associated with the hypotheses. However, the *p*-value does not provide that measure. This point must be explicitly emphasized, because students tend to seek a short cut interpretation of the *p*-value, which often leads to their demise.

Discussion of one "validity" concept that remained elusive Many respondents did not attend to the need to check conditions (*Validity-19*). Respondents should respond that no statistical test should be performed because conditions had not been met (p_2 = .62). The fact that 38% of the respondents got this wrong is not particularly surprising because condition checking is emphasized to a lesser degree in this introductory course compared to the course for students in the natural sciences. There may be an opportunity to place a greater emphasis on why condition checking is needed within the randomization and simulation modules or in transition from using randomization and simulation distributions to using theoretical null distributions. Failing to check the necessary conditions for inference is arguably at the crux of many difficulties people have with the proper use and interpretation of statistical results (Hahn & Meeker, 1993).

5.3. DISCUSSION 3: WHAT POSTTEST EXPLANATIONS REVEALED ABOUT PERSISTENTLY DIFFICULT INFERENCE LEARNING OUTCOMES

Respondent explanations revealed a difficulty with p-values that are not small, particularly if in the opposite direction from hypothesized (*Basic-6:* 3b-3). Depending on the specific scenario, respondents struggled with *p*-values that were large. For item 3b-3 respondents failed to select the option to shade the *p*-value in the direction hypothesized by the researcher. The alternative hypothesis was left-sided; therefore, shading left would produce a *p*-value larger than .50. The *p*-value was, in fact, greater than .50 (which is explicitly asked in item 3b-4). A one-sided alternative hypothesis seemed to present a considerable hurdle when the results were in the opposite direction than the researcher expected. Respondents' explanations showed a clear tendency to "want to shade RIGHT, to the closest tail," regardless. This difficulty appears to be related to recognizing that the magnitude of the *p*-value depends on the direction of the alternative hypothesis (*Basic-5*). Emphasizing two-sided tests accompanied by a confidence interval rather than one-sided tests may circumvent some of this confusion.

This item's results were inconsistent with two *Basic-6* items with high Posttest proportions (3a-3, 3b-1). In item 3a-3 respondents were given a large *p*-value of .72 and correctly responded that the sample data did not support the research hypothesis (p_2 = .94). For item 3b-1 respondents saw a null distribution with a sample mean near the hypothesized value and most (p_2 = .74) correctly responded that the results were not statistically significant. The problem seems to clearly point to the directional challenge with the problem scenario provided for 3b-3.

Respondent explanations revealed a difficulty wrestling with variation within and between (*Connected-13:* 3c-1). Respondents showed no improvement in their ability to differentiate within variation from between variation. There are two related items in Appendix A (3c-1 and 3c-2 that assess students' ability to employ informal inferential reasoning as described by Zieffler and colleagues (2008) and Wild et al. (2011). Out of the 34 respondents who correctly answered item 3c-1 on the Posttest, there were 29 explanations. Twenty-seven described both the variation within and variation between the groups when comparing the boxplot distributions. There were 33 respondents who answered this item (3c-1) incorrectly and 32 explanations. Among these, 25 were primarily focused on the largest difference in centers (effects) with no regard or some confusion about the impact of the magnitude of the spreads. It seems clear that understanding variation within and variation between is not a precursor to a basic understanding of inference, rather, this understanding appears to be an indicator of a deeper (higher-order) understanding of inference. What is particularly interesting about this confusion is that there was no mention of NHST or *p*-values for this item. Maligning of NHST procedures seems to be a scapegoat for a conceptual confusion that may lie much deeper. This conceptual confusion is explored in a Ben-Zvi (2004) qualitative study that identifies seven stages of development toward noticing and differentiating within and between variation in a graph.

Respondent explanations revealed a difficulty using confidence intervals to assess statistical significance (*Logic-14:* 2-5). Respondents must recognize that confidence intervals can be used to determine statistical significance by assessing whether a particular null hypothesis is included in the range of plausible parameter values (Cumming & Fidler, 2002). After the course some Posttest respondents (35%) remained confused about this. Among those who answered correctly, the written explanations reflect fairly clear reasoning. Among those who answered the item incorrectly, the two most prevalent explanations were either that confidence intervals simply do not assess statistical significance; that "we need to do a test for that" or that the center of the confidence interval should be centered at the null hypothesis, rather than the sample mean. The lack of clarity that the confidence interval is constructed around the sample mean to predict plausible values for the population mean may be exacerbated by a fundamental lack of differentiation of the sample mean statistic from the population mean parameter.

Respondents' Posttest explanations suggest that once they learn that *p*-values are used to determine statistical significance, they suppress their instinctive Pretest notions that confidence intervals can signify statistical significance as well. Students may have correctly learned that there is more information provided by the confidence interval than one gets from the *p*-value alone. Rewording the item to reflect this fact may change the results. However, only one of the explanations seems to reflect that there was additional information provided by the confidence interval. What is more likely is that they have not connected the inter-relationship between confidence intervals and *p*-values well. The first definition for statistical significance in the textbook is in the context of a *p*-value being used to determine whether an association is statistically significant using a chi square test (Utts (2005), Chapters 10, 12 and 13). This first encounter seems to trump the later encounters with statistical significance and confidence intervals in Chapters 20-21. Chapter 20 "Estimating Proportions with Confidence" focuses on constructing a confidence interval for a proportion and provides examples and a case study that puts these confidence intervals in context. Chapter 21, "The Role of Confidence Intervals in Research", focuses on constructing a confidence interval for a mean and for a difference between means. This chapter also highlights case studies that use these procedures. The order in which students learn inferential topics may be influencing the study results.

5.4. LIMITATIONS

Although each *RPASS* item was intended to measure just one inference learning outcome, student explanations reveal that multiple misconceptions, misinterpretations or difficulties crop up in thinking through a specific problem. Some items are classified as a combination of correct concepts and misconceptions to reflect this overlap. In addition, the concepts listed in the taxonomy of inference learning outcomes are clearly not all-inclusive. There are likely to be concepts and misconceptions related to inference that have not yet been documented in the existing literature. Thus, the *RPASS-10* items are just a sample of items from the inference content domain.

Respondent explanations illustrated how student thinking can diverge from what is expected. Even though an item was written to assess one particular concept on face value, student thinking may reveal how other difficulties or misconceptions come into play. Explanations also highlight that their reasoning contains both correct concepts and misconceptions at the same time. The analysis of student explanations reveals surprising sources of confusion that were not anticipated. For example, what appears to be a difficulty with estimating and shading the *p*-value was apparently due to an underlying difficulty with a one-sided hypothesis when given evidence in the opposite direction than hypothesized.

Further, it should not be inferred that the misconceptions and difficulties that these particular respondents have suppressed would be the same for introductory students in other disciplines. Historically, students in the introductory course for the natural sciences scored statistically higher on the *RPASS* Pretests and Posttests compared to the social science course students, on average (e.g., Lane-Getaz, 2011; Lane-Getaz, 2013). Thus, one might expect different results from students with a stronger quantitative preparation.

6. SUMMARY AND CONCLUSION

The results of this study showed that during a one semester course most students achieved statistically significant gains for 20 of the 36 *RPASS-10* items. Students were able to *recognize basic concepts*, *differentiate between similar inferential concepts, interpret results*, and *evaluate validity*—there were learning gains across the inference learning outcome taxonomy (Table 1). Few of the difficulties cited in the literature presented persistent challenges for these introductory students. In fact most of the gains can be attributed to suppressing known misconceptions. It is not reasonable to expect that students will learn all the subtleties of NHST in a one-semester course; however, these students moved well beyond merely recognizing basic concepts and connecting related concepts. Given a strong foundation, respondents are better prepared to develop higher-order thinking about inference. There is clear evidence of improved interpretations and evaluation of study design. In a subsequent course, should they choose to take one, respondents are better prepared to further improve their inferential thinking. As Krantz (1999) suggested, the problem with the use of inference procedures may not be with the tests, but with how they are taught. There is no need to toss out the proverbial baby with the bathwater. The teaching of inference has been improved and can continue to improve.

6.1. RECOMMENDATIONS FOR TEACHING

Emphasize that the p-value is an integral part of a larger statistical process. Students were able to recognize basic concepts and differentiate between similar inferential concepts after taking this course. Greater exposure in subsequent courses would likely improve interpretations. However, for many of these social science students this course is their one and only statistics course. It seems important to intentionally engage these students in some high-order thinking about the entire statistical process: more interpretation, less computation, and more stress on evaluation of the validity of the study design. Utts (2005) does this fairly well in her textbook. To interpret a *p*-value properly, it is imperative to attend to the scope of inference (Ramsey & Shafer, 2002) and use of the proper inferential logic when interpreting results. Is the study design sufficient to justify using these inferential procedures and to draw the desired inferences? Results suggest that a better emphasis is needed on checking if necessary conditions are sufficiently met. Although this is emphasized in our second courses in statistics, at this level students still need to recognize that condition checking is necessary before any conclusions can be drawn from these procedures.

Emphasize the concept of the distribution under the null hypothesis. When researchers conduct a study, they typically only have results from a single sample and "you need to have something to compare to." The null distribution provides that point of comparison, granted the null hypothesis is the worse case scenario; i.e., there is no effect. However, the null model does provide a point of comparison. In this course a week and a half of classroom lectures and two hands-on computer labs were devoted to randomization and simulation content, to reinforce the concept of the null distribution as a classroom experiment. Adding simulation may have contributed to the inference learning outcomes observed. The aggregate *RPASS-10* results from this study compare favorably to *RPASS* Posttest Means (and *SDs*) observed in previous studies:

RPASS-7: 23.2 (4.5) out of 34 items, $n = 55 = 68\%$ (Lane-Getaz, 2013),
RPASS-8: 22.0 (5.3) out of 35 items, $n = 19 = 66\%$ (Lane-Getaz, 2011),
RPASS-9: 23.0 (5.0) out of 37 items, $n = 60 = 61\%$ (Lane-Getaz, 2014),
RPASS-10: 26.0 (4.3) out of 36 items, $n = 69 = 72\%$ (current study).

Of course, no causal conclusion can be drawn. Nevertheless, the favorable, correlational link between the higher average score and the introduction of the randomization and simulation content to the course is encouraging.

Emphasize variation within and between. Wild et al. (2011) describe a staged path to develop concepts of statistical inference in a first course. They describe methods of visualizing variation within and between groups. To help students visualize sampling variation over the course of repeated sampling, Wild and colleagues depict comparative boxplots along with shadowy memories of the previous boxplots distribution from all the samples. Another interesting example offered by Wild and colleagues considers comparative boxplots all with the same spread but with different effects. The student must answer: "When can I make the call that [boxplot] B tends to give larger values than [boxplot] A?" Wild et al. suggest that the ability to make this type of comparison is one of the "big ideas of statistical inference" that provides a foundation for more formal statistical methods. They argue that this "informal" foundation to inference should (ideally) be developed in the K-12 curriculum. Students can deal with the concepts of variation within and between early and often by visually comparing side-by-side boxplots in grade school—with no mention of NHST, confidence intervals or *p*-values. This may be a developmental milestone toward preparing students to compare distributions using confidence intervals as suggested by Cumming and Finch (2005).

Emphasize confidence intervals (CI) and two-sided tests to assess statistical significance The CI estimates population parameters or true effects, given the sample data observed. Because the CI provides a range of plausible values for the population parameter, it can also identify the exclusion of a particular null hypothesis value, signifying statistical significance. Beyond this, the CI provides complementary information that *p*-values cannot provide alone, namely, the bounds for the estimated population value or effect. NHST critics often recommend that NHST be replaced by the use of confidence intervals (e.g., Hubbard & Armstrong, 2006; Robinson & Wainer, 2002). In actuality, the CI is an equivalent inferential procedure. "A point estimate, together with a *t* statistic for a particular hypothesized parameter value, can be readily converted to an (approximate) confidence interval, and vice-versa" (Krantz, 1999, p. 1372). Even though understanding CIs is challenging for introductory students (see Chance & McGaughey, 2014), one simple curricular change might be routinely to have students note whether a given null hypothesis value is contained in the confidence interval or not. Ask students "what does the CI provide that the *p*-value does not provide alone?" Unlike the *p*-value, the confidence interval provides a measure of the effect in the units of the original problem.

Furthermore, the problem that some respondents had sorting out one-sided versus two-sided tests, as was noted for item 1-3, may suggest that this detail requires greater emphasis. An existing class activity could be revised to focus on the relationship between one-sided and two-sided *p*-values. Another option might be to eliminate one-sided tests in the introductory course altogether by recommending two-sided tests in all cases with a confidence interval to determine the magnitude and direction of the effect. In the context of ongoing research, it seems reasonable to look not only for what the researcher might expect but also for something unusual that was not expected. Using confidence intervals appears to be a crucial step in the direction of emphasizing the importance of understanding effects (and in a second course, effect size.) Schneider (2013) noted that:

> Some researchers have called for a ban on NHST (e.g., Hunter, 1997). Censoring is not the way forward, but neither is status quo. What we need is statistical reforms as suggested for example by Wilkinson et al. (1999), Kline (2004) and Cumming (2012). Here emphasis is on parameter estimation, i.e., effect size estimation with confidence intervals. Important publication guidelines such as APA (2010) still sanction the use of NHST, albeit with strong recommendations to report measures of effect size and confidence intervals around them (e.g., APA, 2010, p. 34). (p. 60)

If we teach it, that is what they will use.

Emphasize the differentiation of sample statistics from population parameters. Students were unclear whether *p*-values reflect a difference in the sample statistics or a true difference in the population parameters (also see Wild et al., 2010). Students did not understand that the *p*-value indicates whether the sample effect is big enough to claim a population effect. This may be a semantic problem more than a conceptual one. However, this confusion may be related to a difficulty differentiating samples from populations in general. Although students may be able to think of a sample as a subset of the population, they seem to have difficulty when these concepts have to be translated to specific means in a problem context. In the midst of teaching randomization and simulation methods, there is an opportunity to emphasize the distinction between the many samples, and the population value that is estimated at the center of the null distribution (or some hypothesized population parameter value). The sample mean, the evidence for the sample, is just one mean in the null distribution that can be concretely differentiated from the parameter value at the center of the null distribution.

6.2. DIRECTIONS FOR FUTURE RESEARCH

The randomization and simulation module was inserted into this introductory course in lieu of the probability chapters with the goal of improving conceptual understanding of inference. Regretfully, the introduction of the randomization and simulation content occurred in the second half of the course (see Appendix C), after the students had already been exposed to the concept of the *p*-value to determine statistical significance using Chi-square tests in Chapter 13 of Utts (2005). For future research the order of the topics will be altered to introduce randomization and simulation content at the beginning of the course. Early and frequent exposure to inference-related outcomes may further deepen students' inferential understanding.

With the earlier introduction of the randomization and simulation content, there will be an opportunity to place a greater emphasis on differentiating the concepts of samples and their statistics from the populations and their parameter values. The current study suggests that these concepts are elusive for students. Further research is needed to identify the crux of the confusion that students exhibited when differentiating sample statistics from population parameters. Routinely using confidence intervals to estimate the range of the population parameter or population effect may help shed light on this distinction as well.

An intervention will be piloted to improve understanding of within variation and between variation. An existing randomization lab activity that compares two groups using side-by-side boxplots will be updated. Students will be asked to "predict statistical significance" before obtaining a *p*-value and confidence interval. Students will be prompted by questions to consider how much the boxes in the boxplots overlap and to what degree. Using boxplots to compare samples provides a developmental stepping-stone to use confidence intervals to population effects (Cumming & Finch, 2005).

There is also an existing activity that compares confidence intervals to signify statistical significance. This activity will be revisited to emphasize that a NHST can be used to determine significance. Furthermore, it will be emphasized that the confidence interval method provides this distinction of significance as well as an estimate for the effect. A series of questions will be written to prompt student thinking. Examining results and explanations from related *RPASS* items may provide insight into student thinking related to these inference learning outcomes.

To assess curricular interventions the *RPASS* will need two modifications. An item needs to be added specifically to assess how a confidence interval is constructed, namely, centered on the sample mean with some margin of error added and subtracted from either side. An additional item needs to assess whether there is general confusion of a sample mean from a population mean. Furthermore, explanations need to be requested for all items that were identified as difficult for these students as well as the new or modified items.

Researchers interested in improving inference learning outcomes might consider:

- Further research on teaching interventions to improve use of confidence intervals,
- Development of assessments to assess the use and understanding of confidence intervals,
- Research in the use and understanding of exclusively employing two-tailed tests in lieu of one-tailed tests.

6.3. CONCLUSION

The decision by Trafimow and Marks (2015) to ban NHST procedures, *p*-values and confidence from the *Journal of Basic and Applied Social Psychology* sends a disturbing message to students that these procedures are of no use in applied social science research. Many of the criticisms of NHST procedures have been addressed in the literature (Cox et al., 1977; Hagen, 1997; Krantz, 1999; Mogie, 2004; among others). Furthermore, results of this study and others (Chance & McGaughey, 2014; Kalinowski, Fidler, & Cumming, 2008; Lane-Getaz, 2013; Reaburn, 2014) suggest that with proper instruction introductory students can overturn many of the documented misconceptions and misinterpretations of NHST, *p*-values, and statistical significance. It is counterproductive to bemoan that *p*-values and statistical significance are a challenge for people to understand and use. With careful teaching, evaluation, and modifications to teaching, future social science researchers may be better prepared to evaluate if the procedures are being used properly and to interpret properly inferential results in the context of a broader investigative research agenda. Rumors of the *p*-value's death are greatly exaggerated.

REFERENCES

Agresti, A., & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, *54*, 280–88.

American Psychological Association. (2010). *Publication manual of the APA* (6th ed.). Washington, DC: APA.

Ancker, J. S. (2006). The language of conditional probability. *Journal of Statistics Education,* *14*(2). Retrieved from http://www.amstat.org/publications/jse/v14n2/ancker.html

Aquilonious, B. C., & Brenner, M. E. (2015). Students' reasoning about *p*-values. *Statistics Education Research Journal, 14*(2), 7-27. Retrieved from http://iase-web.org/Publications.php?p=SERJ

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. (2004). Reasoning about variability in comparing distributions. *Statistics Education Research Journal*, *3*(2), 42-63.

Capraro, R. M. (2004). Statistical significance, effect size reporting, and confidence intervals: Best reporting strategies. *Journal for Research in Mathematics Education, 35*(1), 57-62.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*(3), 378–399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*(4)*,* 287–292.

Cavanaugh, J. (2007, February). Reviews of books and teaching materials [Review of the books *Seeing through statistics*, *and Mind on statistics*]. *The American Statistician, 61*(1)*,* 95. Retrieved from http://www.tandfonline.com/doi/abs/10.1198/000313007X169532

Chance, B., & McGaughey, K. (2014). Impact of a simulation/randomization-based curriculum on students understanding of *p*-values and confidence. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education.* (Proceedings of the Ninth International Conference on Teaching Statistics, Flagstaff, USA). Voorburg: The Netherlands: International Association for Statistical Education and International Statistical Institute. Retrieved from http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_6B1_CHANCE.pdf

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum?, *Technology Innovations in Statistics Education, 1*(1). Retrieved from http://repositories.cdlib.org/uclastat/cts/tise/

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

Cox, D. R., Spjøtvoll, E., Johansen, S., van Zwet, W. R., Bithell, J. F., Barndorff-Nielsen, O., & Keuls, M. (1977). The role of significance tests. *Scandinavian Journal of Statistics, 4*(2)*,* 49–7.

Cumming, G. (2012). *Understanding the new statistics. Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Cumming, G., & Fidler, F. (2002). The statistical re-education of psychology. In B. Phillips (Ed.), *Developing a statistically literate society. (*Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.) Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots6/6c3_cumm.pdf

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*(2), 160–18.

Daniel, L. G. (1998). Statistical significance 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.

Diaz, C., & Batanero, C. (2009). University students' knowledge and biases in conditional probability reasoning. *International Electronic Journal of Mathematics Education, 4*(3), 131-162.

Falk, R. (1986). Conditional probabilities: Insights and difficulties. In R. Davidson, & J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics,* Victoria, Canada, (pp. 292–297). Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots2/Falk.pdf

Falk, R. (2008). Probabilistic reasoning is not logical. *Mathematics Magazine, 81*(4), 268–275.

Falk, R., & Greenbaum, C. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5*(1), 75–98.

Fisher, R. A. (1929). The statistical method in psychical research. *Proceedings of the Society for Psychical Research, 39,* 189-192.

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.

Gliner, J., Leech, N., & Morgan, G. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education, 71*(1), 83–92.

Gregiore, T. G. (2001). Biometry in the 21st Century: Whither statistical inference? *Proceedings of the Conference on Forest Biometry and Information Science* (pp. 1-14). London, U.K: The University of Greenwich. Retrieved from http://cms1.gre.ac.uk/conferences/iufro/proceedings/gregoire.pdf

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

Hahn, G. J., & Meeker, W. Q. (1993). Assumptions for statistical inference. *The American Statistician, 47*(1), 1–11.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1–2.

Hebb, D. O. (1966). *A textbook of psychology.* Philadelphia: W. B. Saunders.

Hubbard, R., & Armstrong, J. S. (2006). Why we don’t really know what “statistical significance” means: Implications for educators. *Journal of Marketing Education, 28*(2), 114-120. Retrieved from http://journals.sagepub.com/doi/pdf/10.1177/0273475306288399

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

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8,* 3–7.

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 Univ. Press.

Kalinowski, P., Fidler, F., & Cumming, G. (2008). Overcoming the probability of the inverse fallacy: A comparison of two teaching interventions. *Experimental Psychology, 4*(4), 152-158.

Kirk, R. E. (1996). Practical significance: A concept whose time has 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 from http://www.amstat.org/publications/jse/v3n1/konold.html

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 94*(448), 1372-1381.

Lamprecht, B. (1996). Book review: Seeing through statistics. *The Statistics Teacher Network, 43,* 1-3. Retrieved from http://www.amstat.org/education/stn/pdfs/STN43.pdf

Lane-Getaz, S. J. (2007). Toward the development and validation of the reasoning about *p*-values and statistical significance scale, In B. Phillips & L. Weldon (Eds.), *Proceedings of the ISI / IASE Satellite Conference on Assessing Student Learning in Statistics*, Voorburg, The Netherlands: ISI. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/sat07/Lane-Getaz.pdf

Lane-Getaz, S. J. (2010). Linking the randomization test to reasoning about *p*-values and statistical significance. In C. Reading (Ed.), *Data and context in statistics education:*

Towards an evidence-based society. (Proceedings of the 8th International Conference on the Teaching of Statistics, Ljubljana, Slovenia, July). Voorburg, The Netherlands: International Statistical Institute. Retrieved from https://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_C210_LANEGETAZ.pdf

Lane-Getaz, S. J. (2011). A comparison of students' inferential reasoning in three college courses. In *Joint Statistical Meetings (JSM) Proceedings,* Statistical Education Section. Alexandria, VA: American Statistical Association. Retrieved from http://www.meetingproceedings.us/2011/asa-jsm/contents/papers/303371_70177.pdf

Lane-Getaz, S. J. (2013). Development of a reliable measure of students' inferential reasoning ability. *Statistics Education Research Journal*, *12*(1), 20-47. Retrieved from http://iase-web.org/documents/SERJ/SERJ12(1)_LaneGetaz.pdf

Lane-Getaz, S. J. (2014). What students learn and don't learn about inferential reasoning in their introductory statistics courses. In *Joint Statistical Meetings (JSM) Proceedings*, Statistical Consulting Section. Alexandria, VA: American Statistical Association. Retrieved from https://www.amstat.org/membersonly/proceedings/2014/data/assets/pdf/311030_86608.pdf

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

Mogie, M. (2004). In support of null hypothesis significance testing. *Proceedings of the Biological Sciences (Supplement)*, *271*(3), S82-S84.

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

O'Brien, T. C. (1973). Logical thinking in college students. *Educational Studies in Mathematics, 5*(1), 71–79.

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge and Kegan Paul.

Ramsey, F., & Schafer, D. (2002). *The statistical sleuth: A course in methods of data analysis* (2nd Ed.). Pacific Grove, CA: Duxbury Press.

Reaburn, R. (2014). Introductory statistics course tertiary students' understanding of *p*-values. *Statistics Education Research Journal, 13*(1), 53-65. Retrieved from http://iase-web.org/documents/SERJ/SERJ13%281%29_Reaburn.pdf

Reading, C., & Reed, J. (2010). Reasoning about variation. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society.* (Proceedings of the 8th International Conference on the Teaching of Statistics, Ljubljana, Slovenia, July). Voorburg, The Netherlands: International Statistical Institute. Retrieved from www.stat.auckland.ac.nz/~iase/publications.php

Robinson, D. H., & Wainer, H. (2002). On the past and future of null hypothesis significance testing. *The Journal of Wildlife Management*, *66*(2), 263-271. Retrieved from http://www.jstor.org/stable/3803158

Robinson, D. H., Levin, J. R., Thomas, G. D., Pituch, K. A., & Vaughn, S. (2007). The incidence of "causal" statements in teaching-and-learning research journals. *American Educational Research Journal, 44*(2), 400-413. Retrieved from http://www.jstor.org/stable/30069442

Rossman, A. J. (2008), Reasoning about informal statistical inference: One statistician's view, *Statistics Education Research Journal, 7*(2), 5-19. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Rossman.pdf

Rubel, L. H. (2007). Middle school and high school students' probabilistic reasoning on coin tasks. *Journal for Research in Mathematics Education, 38*(5), 531-556.

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, Salvador, Brazil). Voorburg, The Netherlands: International Association for Statistical Education and the International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/17/6A3_SALD.pdf

Schneider, J. (2013). Caveats for using statistical significance tests in research assessments. *Journal of Infometrics, 7*(1), 50-62.

Schneider, J. (2015). Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. *Scientometrics, 102*, 411–432.

Thompson, B. (2006). Critique of *p*-values. *International Statistical Review, 74*(1), 1-14.

Tintle, N., VanderStoep, J., Holmes, V., Quisenberry, B., & Swanson, T. (2011), Development and assessment of a preliminary randomization-based introductory statistics curriculum, *Journal of Statistics Education, 19*(1). Retrieved from http://www.amstat.org/publications/jse/v19n1/tintle.pdf

Trafimow, D., & Marks, M. (2015). Publishing models and article dates explained. *Basic and Applied Social Psychology, 37*(1), 1-2.

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

Tyler, R. W. (1931). What is statistical significance? *Educational Research Bulletin, 10*(5), 115-118+142.

Utts, J. (2003). What educated citizens should know about statistics and probability. *The American Statistician, 57*(2), 74-79.

Utts, J. (2005). *Seeing through Statistics*. Belmont, CA: Brooks/Cole.

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

Wild, C. J., Pfannkuch, M., Regan, M., & Horton, N. J. (2010). Inferential reasoning: learning to "make a call" in theory. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society*. (Proceedings of the 8th International Conference on the Teaching of Statistics, Ljubljana, Slovenia, July). Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots8/ICOTS8_8B1_WILD.pdf

Wild, C. J., Pfannkuch, M., Regan, M., & Horton, N. J. (2011) Towards more accessible conceptions of statistical inferences. *Journal of the Royal Statistical Society*, *174*(2), 247-295. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2010.00678.x/epdf

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

Wilkinson, L. & The Task Force on Statistical Inference APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and expectations. *American Psychologist, 14*(8), 594-604.

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

Zieffler, A., Garfield, J., delMas R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40-58. Retrieved from http://iase-web.org/documents/SERJ/SERJ7%282%29_Zieffler.pdf

SHARON J. LANE-GETAZ
Department of Mathematics, Statistics and Computer Science
and the Department of Education
St. Olaf College
1520 St. Olaf Avenue
Northfield, MN 55057
USA

APPENDIX A. REASONING ABOUT P-VALUES & STATISTICAL SIGNIFICANCE SCALE (36 ITEM RPASS-10)

Section 1: Defining P-values (3 items)

Scenario 1: 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 article reported. To test their claim, data from a random sample of 81 students is analyzed using a one-tailed test. The analysis produces a P-value of .048.

1. Statement: The P-value of .048 is the probability that the students' random sample would have a mean as extreme or more extreme as what they had observed, if the results based on the research article (the null hypothesis) were indeed true.

○ True
○ False

2. Statement: This P-value tells the students that the long run probability is 48 in 1000 of observing data at least as unusual as what was observed, if the null hypothesis were true.

○ True
○ False
Please explain your reasoning in the space below:

3. Statement: Assume a student had conducted a two-tailed test instead of a one-tailed test on the same data, how would the P-value (.048) have changed?

○ The two-tailed P-value would be smaller (i.e., the P-value would be .024).
○ The two-tailed P-value be the same as the one-tailed (i.e., the P-value would be .048).
○ The two-tailed P-value would be larger than the one-tailed (i.e., the P-value would be .096).
Please explain your reasoning in the space below:

2014
: Reasoning about P-values & Statistical Significance (36 item RPASS-10)

Section 2: Using Tests of Statistical Significance (5 items)

Scenario 2: The district administrators of an experimental program are interested in knowing if the program had improved 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 actions and interpretations of district researchers.

1. 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
Please explain your reasoning on this item:

2. Interpretation: In their presentation to the district administration, the researchers explained that when comparing the observed results to the general population, the stronger the evidence that the reading readiness program had an effect, the smaller the P-value that would be obtained.

○ Valid Interpretation
○ Invalid Interpretation

APPENDIX A. CONTINUED… REASONING ABOUT P-VALUES & STATISTICAL SIGNIFICANCE SCALE (36 ITEM RPASS-10)

Section 2 continued…

3. Interpretation: The P-value should be interpreted as the conditional probability of having obtained a mean Reading Readiness score of 102 or higher, conditioned on the population mean being 100.

- o Valid Interpretation
- o Invalid Interpretation

Please explain your reasoning in the space below:

4. Interpretation: After checking the conditions necessary for inference, the district researchers found they had statistically significant results. They interpreted the small P-value to mean that the cause of the results obtained was clearly due to chance.

- o Valid Interpretation
- o Invalid Interpretation

5. Action: Conditions for inference were acceptable, so the district researcher constructed a 95% confidence interval (centered at 102) to estimate the range of plausible population means that could have produced the observed results. The researcher evaluated whether the interval captured the hypothesized mean of 100 (or if the entire range of values was greater or less than 100). This approach assesses statistical significance (much like a two-tailed test at the .05 level or a one-tailed test at the .025 level).

- o Valid Action
- o Invalid Action

Please explain your reasoning in the space below:

(36 itemRPASS-10)

Section 3: Interpreting Results (9 items)

Scenario 3a: A researcher conducts a two-sample test. He compares the mean hair growth results for one class section of students who agreed to try his treatment to a second class section's mean who do not use the treatment. He hopes to show that there is a statistically significant difference between the two group means. How should this researcher interpret results from this two-sample test?

1. Interpretation: If the class section that had the treatment has more hair growth compared to the no treatment group and the P-value is small, the researcher interprets the P-value to mean that all people that use the product will have more hair growth than people who do not.

- o Valid Interpretation
- o Invalid Interpretation

Please explain your reasoning in the space below:

2. Interpretation: Assume the conditions for inference were met. The researcher interprets the P-value as an indicator of how rare (or unusual) it would be to obtain the observed results or something more extreme, if the hair treatment had no effect.

- o Valid Interpretation
- o Invalid Interpretation

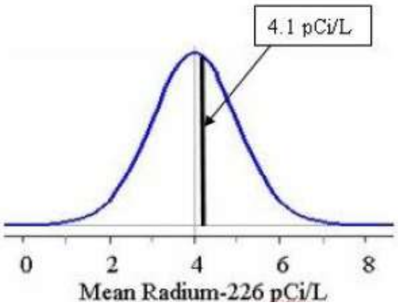
3. Interpretation: Assume the conditions for inference were met and the researcher obtains a large P-value of .72. How should this be interpreted?

- o There is a calculation error because
- o P-values are not supposed to be this large.
- o The sample data did not support the research hypothesis.

APPENDIX A. CONTINUED… REASONING ABOUT P-VALUES & STATISTICAL SIGNIFICANCE SCALE (36 ITEM RPASS-10)

Section 3 continued...

Scenario 3b. Radium-226 is a naturally occurring radioactive gas. For public safety, the Environmental Protection Agency (EPA) has set the maximum exposure level of Radium-226 at a mean of 4 pCi/L (picocuries per liter). Student researchers at a southern Florida university expected to show that Radium- 226 levels were less than 4 pCi/L. However, these student researchers collected 32 soil specimens with a mean Radium-226 measured at 4.1 pCi/L. Students checked the necessary conditions and conducted a hypothesis test at the .05 level. Estimate the P-value given the sketch below of the distribution of means and the observed mean of 4.1 pCi/L.



1. Interpretation: Based on the estimated P-value, the students' sample mean was statistically significant.
   - Valid Interpretation
   - Invalid Interpretation
   - Other (please specify)

2. Interpretation: The estimated P-value for the students' sample mean is greater than .05.
   - Valid Interpretation
   - Invalid Interpretation␟

   Please explain your reasoning in the space below:

3. Interpretation: The estimated P-value for the students' sample can be illustrated by shading the area to the left of the observed sample mean of 4.1 pCi/L in the sampling distribution of means represented above.
   - Valid Statement
   - Invalid Statement␟

   Please explain your reasoning on this item:

4. Interpretation: The estimated P-value for the students' sample is actually greater than .5.
   - Valid Interpretation
   - Invalid Interpretation

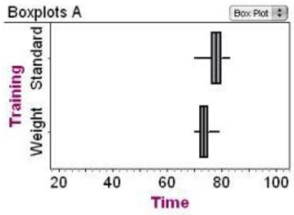
APPENDIX A. CONTINUED… REASONING ABOUT P-VALUES & STATISTICAL SIGNIFICANCE SCALE (36 ITEM RPASS-10)

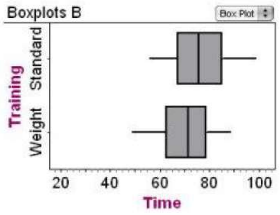
Section 3 continued...

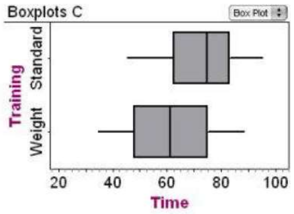
Scenario 3c: A group of 100 athletes are preparing to run a race. They are all pretty similar in their height, weight, and strength. They are randomly assigned to one of two groups. One group gets an additional weight-training program. The other group gets the regular training program without weights. All the students from both groups run the race and their times are recorded. The data are used to compare the effectiveness of the two training programs.

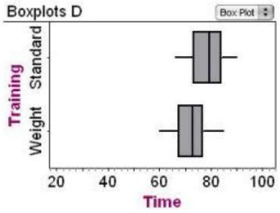
Presented below are some possible graphs that show boxplots for different scenarios, where the running times are compared for the students in the two different training programs (one with weight training and one with standard training).

Examine each pair of graphs and think about whether or not the sample data would lead you to believe that the difference in running times is caused by these two different training programs. (Assume that everything else was the same for the students and this was a true, well-designed experiment.)


Boxplots A


Boxplots B


Boxplots C


Boxplots D

1. Which set of boxplots provides the MOST convincing evidence that the difference between the two groups of athletes is due to the training program.

o Boxplots A
o Boxplots B
o Boxplots C
o Boxplots D

Please explain your reasoning on this item:

2. Which set of boxplots provides the LEAST convincing evidence that the difference between the two groups of athletes is due to the training program.

o Boxplots A
o Boxplots B
o Boxplots C
o Boxplots D

Please explain your reasoning on this item:

APPENDIX A. CONTINUED… REASONING ABOUT P-VALUES & STATISTICAL SIGNIFICANCE SCALE (36 ITEM RPASS-10)

Section 4: Drawing Conclusions about Statistical Significance (9 items)

Scenario 4a: 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.

1. Conclusion: Recall that the significance level is .05 and the P-value is .03.

o The .05 suggests the mean prep course score is higher than 500.
o The .03 suggests the mean prep course score is higher than 500.
o Since .03 is smaller than .05, the evidence suggests the mean prep course score is not statistically significantly higher than 500.

2. Conclusion: If there were an even greater difference between the mean scores of students who took the SAT preparation course and the historical average, we would obtain an even smaller P-value than .03.

o Valid Conclusion
o Invalid Conclusion

3. Conclusion: A causal conclusion can be drawn about the effectiveness of the review course based on a P-value this small, regardless of whether this was a randomized comparative experiment or an observational study.

o True
o False

Please explain your reasoning in the space below:

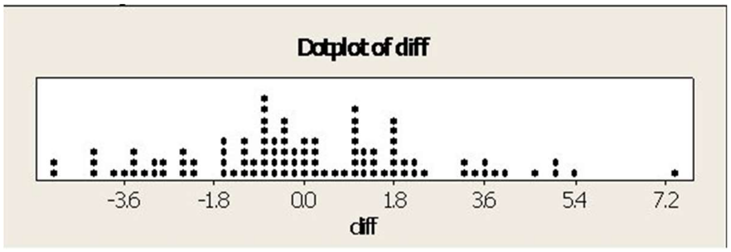
2014: Reasoning about P-values & Statistical Significance (36 item RPASS-10)

Scenario 4b: Researchers hypothesized that female students suffering from bulimia would have a greater fear of a negative evaluation by others than female students who had more normal eating habits. To investigate this theory, two samples of female subjects were recruited to participate in a psychological study. One sample consisted of 11 "bulimic" females; the other sample of 14 female subjects had "normal" eating habits. The response variable in this study was based on a questionnaire taken by each subject that measured her “fear of negative evaluation” (FNE). The mean difference in FNE scores between the "bulimic" group and the "normal" group was 3.68 points.

A statistics class was asked to assess if this difference of 3.68 was statistically significant (at the .05 level)? The statistics students decided to randomly reassign the observed FNE scores to two groups (Bulimic and Normal) 100 times, as if there were no difference in the two groups. For every random re-assignment, the statistics students computed differences between the mean FNE scores (mean Bulimic FNE score – mean Normal FNE score). They plotted these 100 mean differences in a dot plot to assess how much the mean difference would vary, just by chance. The distribution of mean differences appears below. Using this distribution, students estimated a P-value and assessed the statistical significance of the observed mean difference of 3.68.

APPENDIX A. CONTINUED… REASONING ABOUT P-VALUES & STATISTICAL SIGNIFICANCE SCALE (36 ITEM RPASS-10)

Section 4 continued…



1. Conclusion: One of the statistics students explained to the group that the appropriate P-value should be 7/100 or .07 for a one-tailed hypothesis. He concluded that this P-value is sufficient evidence to reject at the .10 significance level but insufficient to reject at the .05 level set for this study.

○ Valid Conclusion
○ Invalid Conclusion
Please explain your reasoning in the space below:

2. Action: Another student in the group counted the mean differences in the above distribution as large or larger than +3.68 and estimated the P-value to be 13/100 or .13.

○ Valid Action
○ Invalid Action
Please explain your reasoning on this item:

3. Conclusion: Given the observed difference in mean FNE scores of 3.68 in this study, the statistics students rejected the hypothesis that, on average, there was no difference in FNE scores between the two groups.

○ Valid Conclusion
○ Invalid Conclusion

4. Conclusion: Assuming the statistics students failed to reject the null hypothesis, the study may not have had a large enough sample size to detect a statistically significant difference.

○ Valid Conclusion
○ Invalid Conclusion

5. Conclusion: Larger sample sizes (e.g., 25 per group) would probably produce results that were statistically significant, regardless of whether they were practically significant.

○ Valid Conclusion
○ Invalid Conclusion
Please explain your reasoning in the space below:

APPENDIX A. CONTINUED… REASONING ABOUT P-VALUES & STATISTICAL SIGNIFICANCE SCALE (36 ITEM RPASS-10)

Section 5: Tying P-values back to Hypotheses (4 items)

Scenario 5: 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 mean scores of subjects who were randomly assigned to control or treatment groups (20 subjects in each group). The treatment group used the new curriculum. The control group did not. You use a 2-sample test of significance and obtain a P-value of 0.01.

1. 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

Please explain your reasoning in the space below:

2. Statement: The probability that the experimental (i.e., the alternative) hypothesis is true is .01.

- True Statement
- False Statement

3. Statement: Assume you had obtained an even smaller P-value (than .01). A smaller P-value...

- is stronger evidence of a difference or effect of the new driving school curriculum.
- is weaker evidence of a difference or effect of the new driving school curriculum.
- suggests no change in the difference or effect of the new driving school curriculum.

4. Statement: The P-value of .01 reflects a mean difference in scores between the treatment and control groups being studied but does not suggest that there is a true mean difference in the broader populations.

- True Statement
- False Statement

APPENDIX A. CONTINUED… REASONING ABOUT P-VALUES & STATISTICAL SIGNIFICANCE SCALE (36 ITEM RPASS-10)

Section 6. The remaining questions are multiple-choice. Please select the best option among the choices provided.

1. A research article gives a P-value of .001 in the analysis section. Which definition of a P-value is the most appropriate? The P-value is...

○ the probability that the observed outcome will occur again.
○ the probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true.
○ the value that an observed outcome must reach in order to be considered significant under the null hypothesis.
○ the probability that the null hypothesis is true.

Please explain your reasoning in the space below:

2. If a researcher was hoping to show that the results of an experiment were statistically significant they would prefer:

○ a large P-value
○ a small P-value
○ 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. The appropriate statistical test was conducted and Mrs. Rose finds the P-value is .0025. Assuming this were a random sample, which of the following is the best interpretation of the P-value?

○ A P-value of .0025 provides strong evidence that, on average, Mrs. Rose's class outperformed high school students across the nation.
○ A P-value of .0025 indicates that there is a very small chance that, on average, Mrs. Rose's class outperformed high school students across the nation.
○ A P-value of .0025 provides evidence that Mrs. Rose is an exceptional teacher who was able to prepare her students well, on average, for this national test.
○ 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?

○ This proves that her experimental treatment has no effect on memory.
○ There could be a treatment effect, but the sample size was too small to detect it.
○ She should reject the null hypothesis.
○ There is evidence of a small effect on memory by her experimental treatment.

Please explain your reasoning in the space below:

APPENDIX A. CONTINUED… REASONING ABOUT P-VALUES & STATISTICAL SIGNIFICANCE SCALE (36 ITEM RPASS-10)

Section 6 continued…

5. A newspaper article claims that the average age for people who receive food stamps is 40 years. You believe that the average age is less than that. You take a random sample of 100 people who receive food stamps, and find their average age to be 39.2 years. You find that this is significantly lower than the age of 40 stated in the article ($p < .05$). What would be an appropriate interpretation of this result?

○ The statistically significant result indicates that the majority of people who receive food stamps is younger than 40.

○ Although the result is statistically significant, the difference in age is not of practical importance.

○ An error must have been made.

○ This difference is too small to be statistically significant.

6. A newspaper article stated that the US Supreme Court received 812 letters from around the country on the subject of whether to ban cameras from the courtroom. Of these 812 letters, 800 expressed the opinion that cameras should be banned. A statistics student was going to use this sample information to conduct a test of significance of whether more than 95% of all American adults feel that cameras should be banned from the courtroom. What would you tell this student?

○ This is a large enough sample to provide an accurate estimate of the American public's opinion on the issue.

○ The necessary conditions for a test of significance are not satisfied, so no statistical test should be performed.

○ With such a large number of people favoring the notion that cameras be banned, there is no need for a statistical test.

7. Food inspectors inspect samples of food products to see if they are safe. This can be thought of as a hypothesis test, where: Ho: the food is safe (in the population), and Ha: the food is not safe (in the population). Identify whether the following statement is a Type I (Alpha), a Type II (Beta) error, or neither.

Statement: "The inspector says the food is not safe but it actually is safe."

○ The inspector rejects the null hypothesis when he shouldn't have (i.e., a Type I or alpha error)

○ The inspector fails to reject the null hypothesis when he should have (i.e., a Type II or beta error)

○ Not an error

Please explain your reasoning in the space below:

APPENDIX B. RPASS-10 ITEMS LISTED BY POSTTEST PROPORTION CORRECT WITHIN TAXONOMY OF INFERENCE LEARNING OUTCOME

Table B. 36 RPASS-10 scenario and item numbers ordered by Posttest proportion correct within inference learning outcome; if correct concept (C) or misconception (M) was assessed, Pretest proportion correct and Learning Gain by item, N = 69

RPASS	Inference learning outcome category,			Proportion correct		
Item	description of learning outcome, and references		C/M	Posttest	Pretest	Gain
2-1	Basic-1	Recognize a small p-value measures rareness or unusualness, when H_o is true. (Carver, 1978; Fisher, 1929; Saldanha & Thompson, 2006; Schneider, 2015)	C	.93	.52	.41[a]
1-1	Basic-1	As above	C	.84	.59	.25[a]
3a-2	Basic-1	As above	C	.78	.74[b]	.04[c]
6-1	Basic-1	As above	C/M	.74	.41	.33[a]
1-2	Basic-1	As above	C	.71	.59	.12[c]
6-2	Basic-2	Recognize a small p-value is indicative of statistical significance. (Lane-Getaz, 2013)	C/M	.93	.61	.32[a]
2-3	Basic-3	Recognize that a p-value is conditioned on the null hypothesis being true. (Falk, 1986; Ancker, 2006)	C/M	.77	.71[b]	.06[c]
4b-4	Basic-4	Recognize that the p-value is indirectly related to sample size. (Wilkerson & Olson, 1997)	C/M	.97	.93[b]	.04[c]
6-4	Basic-4	As above	C/M	.75	.61	.14[c]
4b-1	Basic-5	Recognize that the magnitude of the p-value depends on the direction of the alternative hypothesis. (Lane-Getaz, 2013)	C/M	.77	.57	.20[a]
3b-2	Basic-5	As above	M	.74	.52	.22[a]
4b-2	Basic-5	As above	C/M	.72	.58	.14[c]
1-3	Basic-5	As above	C/M	.62[d]	.36	.26[a]
3a-3	Basic-6	Recognize that the p-value may not be small; large p-values indicate the sample obtained did not support the research hypothesis. (Lane-Getaz, 2013; Williams, 1999)	M	.94	.77[b]	.17[a]
3b-1	Basic-6	As above	M	.74	.28	.46[a]
4b-3	Basic-6	As above	C	.49[d]	.49	.00[c]
3b-3	Basic-6	As above	M	.41[d]	.58	-.17[e]
3b-4	Basic-6	As above	C/M	.33[d]	.35	-.02[c]
4a-1	Connected-7	Differentiate p-values from significance level (α). (Hubbard & Bayarri, 2003)	M	.87	.39	.48[a]
6-7	Connected-8	Differentiate between concepts of Type I (α) error & Type II (β) error. (Schneider, 2013)	C/M	.77	.57	.20[a]
6-5	Connected-9	Differentiate statistical significance from practical importance. (Tyler, 1931; Gliner, et al., 2002)	C/M	.77	.71[b]	.06[c]
4b-5	Connected-9	As above	M	.72	.70[b]	.02[c]

Note. [a]One of 20 items with a significant learning gain. [b]One of seven items with pretest proportion $p_1 \geq 70\%$. [c]One of 14 items with no significant learning gain or loss. [d]One of 11 items with posttest proportion $p_2 < 70\%$. [e]One of two items with a significant learning loss.

APPENDIX B. CONTINUED… RPASS-10 ITEMS WITHIN THE TAXONOMY

Table B continued… 36 RPASS-10 scenario and item numbers ordered by Posttest proportion within inference learning outcome; if correct concept (C) or misconception (M) was assessed, Pretest proportion correct and Learning Gain by item, N = 69

RPASS	Inference learning outcome category,			Proportion correct		
Item	description of learning outcome, and references		C/M	Posttest	Pretest	Gain
2-2	Connected-10	Differentiate strength of evidence (*p*-values) from the size of an effect. (Gliner, Leech, & Morgan, 2002)	C	.96	.65	.31[a]
5-3	Connected-10	As above	C	.90	.46	.44[a]
4a-2	Connected-10	As above	C	.90	.57	.33[a]
6-3	Connected-10	As above	C/M	.83	.43	.40[a]
5-4	Connected-11	Differentiate sample statistics from population parameters.(Lane-Getaz, 2013; Mittag & Thompson, 2000)	M	.33[d]	.36	-.03[c]
3c-1	Connected-13	Differentiate variation within (spreads) from variation between (effects). (Reading & Reed, 2010; Zieffler, et al., 2008; Wild et al., 2011)	C	.49[d]	.43	.06[c]
3c-2	Connected-13	As above	C	.49[d]	.43	.06[c]
2-5	Logic-14	Interpret confidence intervals to signify statistical significance as a complement to—or in lieu of—NHST and *p*-values. (Lane-Getaz, 2013; Capraro, 2004; Cumming & Fidler, 2002)	C	.65[d]	.83[b]	-.17[e]
3a-1	Logic-16	Suppress the misinterpretation that the *p*-value provides a deterministic proof, *the illusion of contrapositive proof by contradiction*. (Batanero, 2000; Falk & Greenbaum, 1995; Oakes, 1986)	M	.70	.52	.17[a]
2-4	Logic-17	Suppress the misinterpretation that the *p*-value is the probability that the research results were “due to chance;” aka *odds against chance fantasy*. (Carver, 1978; Daniel, 1998)	M	.87	.64	.23[a]
5-1	Logic-18	Suppress the misinterpretation of the *p*-value as the probability that the null hypothesis is false. (Oakes, 1986)	M	.55[d]	.35	.20[a]
5-2	Logic-18	Suppress the misinterpretation of the *p*-value as the probability that the alternative hypothesis is true. (Oakes, 1986)	M	.67[d]	.42	.25[a]
6-6	Validity-19	Evaluate how well the necessary conditions for inference were met. (Hahn & Meeker, 1993)	C/M	.62[d]	.55	.07[c]
4a-3	Validity-20	Evaluate the validity of inferences to be drawn based on how randomization was used in the study design aka s*cope of inference*. (Lane-Getaz, 2013; Ramsey & Shafer, 2002; Robinson et al., 2007)	M	.88	.58	.30[a]

Note. [a]One of 20 items with a significant learning gain. [b]One of seven items with pretest proportion $p_1 \geq 70\%$. [c]One of 14 items with no significant learning gain or loss. [d]One of 11 items with posttest proportion $p_2 < 70\%$. [e]One of two items with a significant learning loss.

APPENDIX C. SPRING 2014 INTRODUCTORY STATISTICS COURSE FOR STUDENTS IN THE SOCIAL SCIENCES: LECTURE AND LAB TOPICS

Table C. Spring 2014 Introductory statistics course topics for social science students

Chapter[a]	Date	Lecture topics	Lab topics by week
1	2/11	Course Introduction and Overview	
1-2	2/13	Benefits & Risks of Using Statistics Reading the News	1. *RPASS-10* Pretest
3	2/18	Measurements, Mistakes & Misunderstandings	
4	2/20	How to Get a Good Sample	2. Canceled (Snow)
5-6	2/25	Experiments and Observational Studies Getting the Big Picture	
7-8	2/27	Summarize and Display Measurement Data; Bell-Shaped Curves; Moodle Quiz #1	3. *SPSS:* Creating tables and graphs for categorical data[c]
9	3/4	Plots, Graphs, and Pictures	
10-11	3/6	Relationships Between Measurement Variables; Relationships can be Deceiving	4. *SPSS:* Creating summaries & graphs for quantitative data[c]
12	3/11	Relationships between Categorical Variables	
13	3/13	Statistical Significance for 2 X 2 Tables; Chi Square	5. Kids' Feet case study: categorical & quantitative data[c]
n/a	3/18	Moodle Quiz #2; Wrap Up and Exam Review	
n/a	3/20	Midterm Exam, Thursday 3/20	6. Randomization test: Dolphin Therapy
n/a		*Spring Break: March 22-30*	
RM-1[b]	4/1	Introduction to Simulation & Randomization	
RM-2[b]	4/3	Simulation & Randomization continued	7. "Second Chance" Midterm (online)
RM-3[b] 18.4	4/8	Wrap up Randomization Positive Predictive Value	
19	4/10	The Diversity of Samples from the Same Population	8. Randomization test: Fish Oil Diet
19-20	4/15	Estimating Proportions with Confidence	
22	4/17	Rejecting Chance—Testing Hypotheses in Research	9. Project Overview
21	4/22	Role of Confidence Intervals in Research Hypothesis Testing—Examples and Case Studies	
	4/24	Role of CIs and HTs in Research continued… Moodle Quiz #3	10. Project: Exploratory Data Analysis
23	4/29	*t*-tests and ANOVA in *SPSS*	
24	5/1	Paired *t*-test and inference for Regression	11. Project Meetings
n/a	5/6	Project Presentations: Group I	
n/a	5/8	Project Presentations: Group II Moodle Quiz #4	12. *RPASS-10* Posttest
n/a	5/13	Project Presentations: Group III	

Note. [a]*Seeing through Statistics*, (Utts, 2005). Lecture podcasts were posted on Moodle. [b]One of three in-class Randomization Module lectures (RM-1: Categorical Response Randomization test, RM-2: Quantitative Response Randomization test and RM-3: Randomization Summary). [c]Five minute screen videos were posted to model "how-to" solutions for each of the *SPSS*-based labs.