What Students Learn and Don't Learn about Inferential Reasoning in their Introductory Statistics Courses

Sharon Lane-Getaz
St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057

Abstract
This observational study examines correct conceptions and misconceptions of inferential reasoning. Much has been said about the difficulties people have understanding inferential reasoning, in particular the difficulties interpreting *p*-values, statistical significance and confidence intervals. This study highlights some of the previous literature on this topic and illuminates the discussion by reporting on empirical research on introductory statistics students. Students' inferential reasoning was measured before and after their introductory level course using Reasoning about *P*-values and Statistical Significance (RPASS), a reliable and valid measure of inferential reasoning. Common confusions and difficulties are discussed as well as implications for future research, teaching and consulting.

Key Words: inference; misconceptions, statistics education, consulting

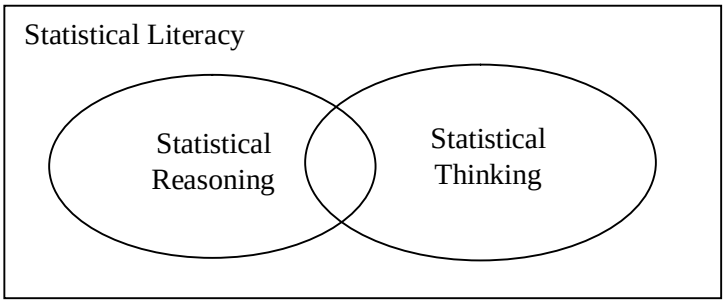
1. INTRODUCTION

People have many difficulties and misconceptions with understanding inference, statistical significance, hypothesis testing, *p*-values and confidence intervals (Garfield & Ahlgren, 1988; Lane-Getaz, 2007; Utts, 2003). It is difficult to get a grasp on the source of students' difficulties with inference since many of the concepts are interrelated. Individual items are not sufficient to assess student understanding of inferential reasoning. It is important to analyze patterns in multiple responses to better assess understanding. In addition, students, and statistical consulting clients as well, come to us with pre-existing, contradictory perspectives as has been documented in the literature (e.g., Konold, 1995; Tversky & Kahneman, 1992) A person may come to understand an isolated definition in an introductory statistics course and yet have difficulty differentiating that concept from another that sounds similar or is closely related that they come to learn during that same semester. Furthermore, time after instruction further complicates students' understanding and reasoning.

One way to define what students should understand about inference is to categorize inferential reasoning outcomes in a taxonomy. In statistics education a taxonomy that has emerged categorizes learning outcomes in terms of statistical literacy, statistical reasoning, and statistical thinking (Ben-Zvi & Garfield, 2004; Chance, 2002; delMas, 2002; Garfield, 2002; Rumsey, 2002). Statistical literacy is the ability to organize and work with data, tables, and representations of data. Statistical literacy includes "understanding basic concepts, vocabulary, and symbols, and understanding probability as a measure of uncertainty" (Ben-Zvi & Garfield, 2004, p.7). Statistical reasoning involves interpreting statistical information, connecting concepts and being able to explain statistical processes combining ideas of data and chance (Garfield, 2002). Statistical thinking requires understanding the "bigger picture" processes involved in conducting statistical investigations (Lane-Getaz, 2006; Pfannkuch & Wild, 2004). A statistical thinker moves beyond what is taught in the course to "spontaneously question and investigate the issues and data involved in a specific context" (Chance, p. 4). delMas (2002) depicts statistical reasoning and thinking as intersecting learning goals embedded

within statistical literacy. Using delMas' depiction of the taxonomy (as shown in Figure 1), once a student has attained sufficient statistical literacy about a particular concept--like the *p*-value--he or she can then develop the necessary reasoning and thinking to attain a deeper understanding of the concept. It is this perspective that informs how the results of this study will be interpreted.

Figure 1. Interrelationships in the taxonomy of Statistical Literacy, Reasoning and Thinking as described by delMas (2002)



Students with good statistical literacy should be able to recognize basic terms and representations. Furthermore, statistical reasoning moves beyond definitions, to employ a deeper understanding of the concepts. Students who apply appropriate inferential reasoning should be able to understand the interrelationships between the *p*-value, significance levels and confidence intervals. They should understand how sample size relates to the magnitude of the *p*-value, and they should understand that large sample sizes lead to statistical significance even when there are small, unimportant differences from a practical perspective (Garfield, delMas, & Chance, 2005). Students who learn to think statistically, will be ware of *p*-values and statistical significance in the broader context of a statistical investigation (Lane-Getaz, 2007). These students understand why *p*-values should be complemented with confidence intervals (see de Veaux, Velleman, & Bock, 2006; Cumming & Finch, 2005). Student who think statistically recognize that a *p*-value indicates that there may be an effect but that is does not indicate how large the effect is. These students should also understand that there is nothing sacred about setting the significance level at .05.

2. METHODS

2.1 Subjects and Setting

This study was conducted at a small liberal arts college of approximately 3000 students in the upper Midwest US, in a small town of "cows, colleges and contentment" during the spring semester of 2011.

The subjects in this study were enrolled in two introductory-level statistics courses aimed at students interested in the social sciences *($n_1 = 78$)* and those interested in the natural sciences *($n_2 = 60$, respectively)*. Out of 167 enrolled students138 completed the RPASS-9 Pretest and Posttest, and consented to participate in the study (an 83% response rate.) There were (94) females, (43) males and (1) student provided no response to the gender question. The breakdown of the students' year in school was (34) first years, (56) sophomores, (30) juniors and (18) seniors.

The approach to teaching introductory statistics differs in the two courses. The teaching in the course for students in the natural sciences (Stat-212) uses a randomization-based curriculum, *Investigating Statistical Concepts, Applications and Methods* (Chance & Rossman, 2006), a statistics education reform-based textbook. Stat-212 is taught in a computer-equipped classroom for 55 minutes 3 times per week using *R*

software to analyze data. Stat-212 course sections are limited to 26 students due to the size of the computer-equipped classroom. The second course (Stat-110) was designed to be a statistical literacy course and used the *Seeing through Statistics* textbook (Utts, 2005). Stat-110 is taught in a large lecture hall classroom twice per week for 80 minute classes of up to 80 students. There is also a hands-on SPSS computer lab once per week with up to 27 students per lab. Combining results from the two distinct courses provides a broad range of abilities and scores for analysis.

2.2 Measurement

The *Reasoning about P-values and Statistical Significance* (RPASS-9) scale measures introductory students' correct conceptions, misconceptions and difficulties with inferential concepts. RPASS (Lane-Getaz, 2007) was initially developed to assess fourteen of the documented difficulties people have reasoning about inference. The items were categorized into four groupings: Basic concepts (which addresses statistical literacy), relationships between concepts, the logic of inference (which both address statistical reasoning), and hypothesis testing (which requires some statistical reasoning and thinking). Both correct conceptions and misconceptions are assessed in each of these categories (see Lane-Getaz, 2007, 2013, 2014).

Taking a step back from formal inference, Zieffler and colleagues suggest a framework to support research on informal inferential reasoning. The two items that were added in RPASS-9 follow this suggestion by having students reason about the interplay between center and variation when comparing two boxplots. This approach is consistent with the research of Bakker & Gravemeijer (2004) who discuss how reasoning about distribution precedes inferential reasoning. In RPASS-9 adds two new items to assess students' informal inferential reasoning as described by Zieffler, Garfield, delMas and Reading (2008). The two new items assess how well students reason about variability when analyzing comparative boxplots. In addition to responding to the 37 multiple choice items, students were asked to explain their reasoning on 12 selected items for analysis in future research.

2.3 Analysis procedure

To illustrate the broad range of students' abilities that were included in this study the RPASS-9 Pretest and Posttest scores are reported by course and in the aggregate using descriptive and graphical boxplot summaries. The proportion of students answering each item correctly on the RPASS Pretest and Posttest are compared graphically in "item plots" as described in Lane-Getaz 2014. Each RPASS item is plotted on the item plot based on the Pretest proportion on the x axis and the Posttest proportion on the y axis. The item plot also includes a 95% confidence band representing the area of plausible variation, if there were no difference in proportions (e.g., $\pi_{Posttest}$-$\pi_{Pretes1}$ = 0). Items outside the band indicate the proportion answering correctly differs significantly from Pretest to Posttest. The margin of error for the confidence band is computed with a Wilson adjustment ($\tilde{p}_i = (X_i + 1)/(n_i + 2)$) to maintain the 95% nominal rate (Agresti & Caffo, 2000). No family-wise correction is made for multiple comparisons since the graphical presentation is intended to be used descriptively.

3. RESULTS

3.1 Descriptive comparisons

The 138 respondents answered 70% of the 37 RPASS-9 Posttest items correctly, on average (M = 26.1, SD = 5.1. The Posttest respondents answer five more items correctly on average compared to the Pretest (M = 21.0, SD = 4.2). The Pretest and Posttest

distributions appear in Figure 2 for the two courses combined. By combining results for the two courses, a broad range of responses can be analyzed. Broken out by course, one can see that while both groups improve from Pretest to Posttest, there are greater gains in the Stat-212 course (Figure 3).

Figure 2. RPASS-9 Posttest and Pretest scores for the courses combined, $N = 138$.

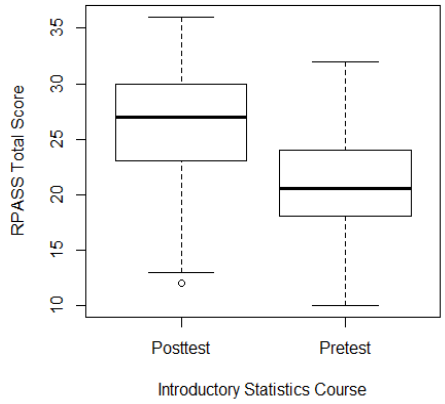
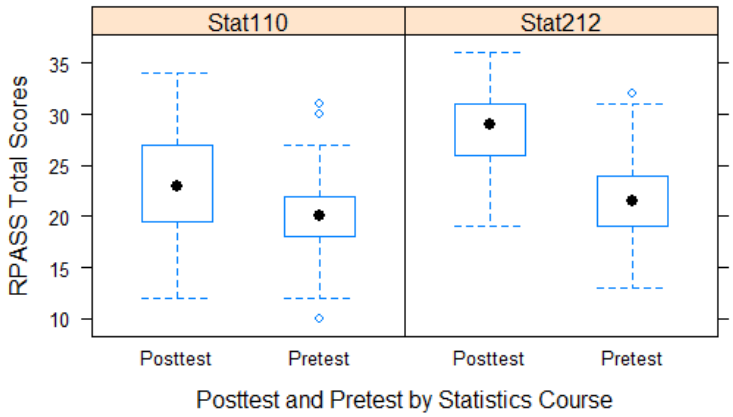


Figure 3. RPASS-9 Posttest and Pretest score distributions broken out by course: Stat-110: $N_1 = 78$, Stat-212 $N_2 = 60$.



3.2 Item proportion correct for Pretest and Posttest

The item plot in Figure 4 displays the proportion of students answering each item correctly on the Pretest and Post-test for the combined courses. For reference, a 95% confidence band is superimposed on the item plot to represent the expected area of variation if there were no difference in proportions ($\pi_{Posttest}$ - $\pi_{Pretest}$ = *0*). The 95% confidence band delineates items with significant differences in the observed proportions (outside the band) from those with insignificant differences (within the band).

There are 23 of 37 RPASS-9 items that appear above the 95% confidence band; 13 items appear within the band and one item appears below the band. The 23 items above the band signify that students' inferential reasoning improved on the Posttest compared to their reasoning on the Pretest, more than one would expect by a "lucky guess." The item number, the reasoning assessed by the item and the difference between the proportion of students who answered the item correctly (Posttest – Pretest) are listed in Table 1. The one item below the band, Item 3b-4 will be discussed in detail.

Figure 4. RPASS-9 Posttest Items plotted by the Proportion of Correct Responses for the Pretest on the *x* axis and Posttest on the *y* axis (37 items, $N = 138$), with 95% confidence band for $\pi_{Posttest} = \pi_{Pretest}$.

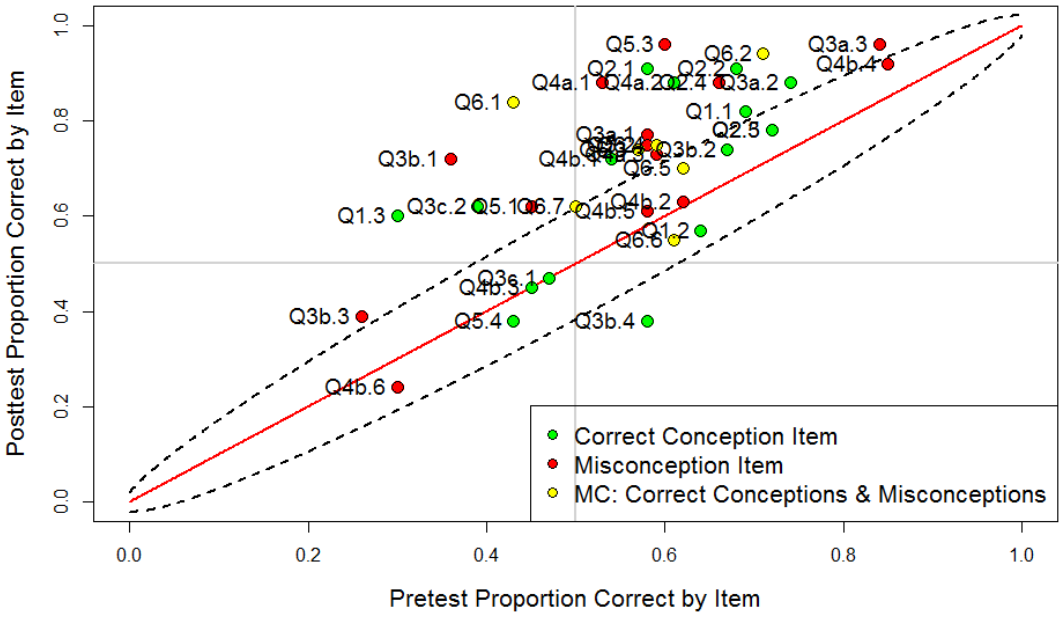


Table 1: Description of Inferential Reasoning Assessed on the 23 RPASS-9 Posttest Items, where the Posttest proportion (p_2) correct significantly exceeded the Pretest proportion (p_1)

RPASS-9 item	*Description of inferential reasoning or difficulty assessed*	p_2-p_1
6-1	Selects a textbook definition of a *p*-value given multiple choices.	.41[a]
3b-1	Uses a density curve and an observed value to estimate if the observed value (or more extreme) is statistically significant.	.36[a]
5-3	Reasons smaller *p*-value, stronger the evidence of a difference or effect.	.36
4a-1	Confuses *p*-value with significance level *a*.	.35
2-1	Recognizes *p*-value in terms of variation in a sampling distribution.	.33
1-3	Understands magnitude of *p*-value depends if test is one- or two-sided.	.30[a]
4a-2	Reasons greater evidence of a difference or effect, smaller the *p*-value.	.27
2-2	Understands stronger evidence of difference or effect, smaller *p*-value.	.23
3c-2	Employs graphical reasoning about variation	.23
6-2	Understands a small *p*-value suggests results are statistically significant.	.23[a]
2-4	Believes the *p*-value is the probability observed results are due to chance or caused by chance, if the null is true.	.22
3a-1	Believes statistics provide definitive proof; misuses the deterministic	.19

	Boolean logic of contrapositive proof.	
4b-1	Interprets a *p*-value for a one-tailed hypothesis.	.18
5-1	Misinterprets a *p*-value as the probability the null hypothesis is false.	.17[a]
5-2	Believes *p*-value is the probability that the alternative hypothesis is true.	.17
6-3	Understands stronger evidence of difference or effect, smaller *p*-value.	.17
6-4	Reasons about impact of a small sample size on statistical significance.	.16
3a-2	Understands the *p*-value as a rareness measure.	.14
4a-3	Believes causal conclusion can be drawn from small *p*-values regardless of study design.	.14
1-1	Recognizes a formal textbook definition of the *p*-value without context.	.13
3b-3	Believes *p*-value is always a low number (or always desired to be a low).	.13
3a-3	Belief *p*-values are always a low value or are always desired to be a low value	.12
6-7	Differentiates between concepts of Type I and Type II error.	.12

Note. [a]The Pretest proportion correct for this item was less than .50.

4. DISCUSSION

4.1 Above the band

Of the 23 items above the band the students improved in statistical literacy and improved their inferential reasoning about 14 correct conceptions. For five of these improvements less than half of the students had answered the items correctly on the Pretest: item 6-1 definition of the *p*-value, item 3b-1 assessing significance graphically, item 3c-2 reasoning about variation (as shown in Appendix A.), item 1-3 reasoning about the impact of the alternative hypothesis on the *p*-value, and item 6-7 reasoning about Type I versus Type II errors. In terms of statistical literacy, students recognized definitions of the *p*-value as evidenced from items 1-2 and 6-1. They were better able to link the concept of a *p*-value with the concept of sampling variation (item 2-1). A greater proportion of the students recognized the *p*-value as a rareness measure (item 3a-2). As for inferential reasoning, students were able to assess statistical significance graphically (item 3b-1). They showed improved reasoning about variation (item 3c-2). Students showed improved reasoning about the impact of the alternative hypothesis on the magnitude of the *p*-value (items 1-2, 4b-21). Students were better able to differentiate between the concepts of *p*-values, Type I and Type II errors (items 6-2, 6-7) and showed improved reasoning about the impact of sample size on the *p*-value and who the strength of evidence against the null hypothesis is measured by the *p*-value (items 2-2, 4a-2, 6-3).

Table 2. "Above the Band" concepts that provide evidence of improved statistical literacy and reasoning about inference-related concepts.

Improved Statistical Literacy	Items
Recognize textbook definitions of *p*-value	1-1, 6-1
Link *p*-value to sampling variation	2-1
Understand *p*-value as a rareness measure	3a-2
Improved Statistical Reasoning	
Assess significance graphically	3b-1
Employ reasoning about variation	3c-2
Reason about strength of evidence vs. *p*-value	2-2, 4a-1, 5-2, 6-3
Assess impact of alternative hypothesis on *p*-value	1-3, 4b-1
Interpret small *p*-values and Type I and II errors	6-2, 6-7
Reason about small sample size impact on *p*-value	6-4

Of the 23 items above the band, students showed improved statistical reasoning about nine common misconceptions or difficulties with inferential reasoning. They were better able to state conclusions within the confined of the scope of inference. They reasoned that a random sample is needed to generalize a pattern found in a sample to a population (item 5-4). They also reasoned that random assignment of subjects to groups is needed to draw a causal conclusion (item 4-3). Not only did students develop better conceptions of what the *p*-value is, they also better understood what the *p*-value is not. They recognized that the *p*-value is not always small or always desired to be a low value (items3a-3, 3b-3). A greater proportion of students no longer thought that the *p*-value was the probability that the null hypothesis is true or false (items 5-1, 5-2). Students also were able to differentiate the *p*-value from the significance level, α (item 4a-1). A greater proportion of students no longer interpreted a small *p*-value to mean that chance caused the results observed and they did not misinterpret a small *p*-value as evidence of definitive, contrapositive proof. Most importantly, students improved their reasoning on three items where less than half addressed the misconceptions correctly on the Pretest. First, students were better able to recognize that in order to generalize results to a population one must have a random sample (item 5-4). Second, students were better able to interpret small *p*-values (item 3b-3). Third, student recognized that the *p*-value is not to be interpreted as the probability that the null hypothesis is false (item 5-1).

Table 3. "Above the Band" concepts that provide evidence of more students with overturned or suppressed common misconceptions or difficulties.

Improved recognizing what the *p*-value is not.	Item
Not the probability that the Null Hypothesis is false or true	5-1, 5-2
Not always small or always desired to be a low value	3a-3, 3b-3
Not equivalent to Type I error or significance level (α)	4a-1
Improved recognizing that a small *p*-value does not mean:	
That chance caused the results observed	2-4
That there is definitive, contrapositive proof	3a-1
Improved interpreting statistical significance with caution:	
Attend to study design before drawing causal conclusion	4a-3
Recognize statistical significance is necessary, but not sufficient, to generalize from a sample to a population	5-4

4.2 Within the band

Of the 13 items within the band, students showed virtually no improvement reasoning about eight correct conceptions. They showed no improvement in reasoning about variation given comparative boxplots (item 3c-1 as shown in Appendix A) and making the correct rejection decision (item 4b-3). For these two items less than half of the students had answered the items correctly on the Pretest and Posttest. Interestingly, the informal inferential reasoning that students applied when comparing boxplot centers and variation was tenuous (compare results for items 3c-1 within the band and 3c-2 above the band). Students also showed no appreciable improvement recognizing an informal definition of the *p*-value (item 1-2) or that the *p*-value is a conditional probability, conditioned on the null being true (item 2-3). Students showed no improved reasoning that confidence intervals can be used to determine statistical significance (item 2-5) and have difficulty differentiating the concept of *p*-values from effects (item 4a-2). They showed no improvement interpreting a large *p*-value (item 4b-2). They also do not improve their reasoning about the impact of sample size on *p*-values (4b-4, 6-4).

Also among the 13 items within the band, students showed virtually no improvement reasoning about four misconceptions or difficulties. Students showed no improvement in beliefs that *p*-values are always low values (item 3b-2). They also had similar responses on the Pretest and Posttest when asked to differentiate statistical significance from practical importance (4b-5, 6-5). Students were confused increasing replications with increasing sample size as having the effect of decreasing variation in a sampling distribution (4b-6). Finally, students were no more likely to select the option to check conditions before making an inference (6-6).

Table 4. "Within the Band" concepts that show no improvement in statistical reasoning from Pretest to Posttest

No improvement on correct conceptions within the band	Items
Recognize an informal definition of *p*-value	1-2
Recognize *p*-value as a conditional probability	2-3
Use Confidence Intervals for statistical significance	2-5
Employ reasoning about variation	3c-1
Interpret a large *p*-value	4b-2
Differentiate *p*-values from effects	4a-2
Make the correct rejection decision	4b-3
Consider impact of sample size on *p*-values	4b-4, 6-4
No improvement on misconceptions and difficulties within the band	
Belief *p*-values are always a low value or desired to be low	3b-2
Differentiate statistical vs. practical significance	4b-5, 6-5
Belief increased replications will decrease variability in a sampling distribution (equivalent effect as increased sample size)	4b-6
Check conditions before making an inference	6-6

4.3 Below the band

The one item below the 95% confidence band (item 3b-4) suggests students empoyed better reasoning on the Pretest than on the Posttest. When asked to choose the correct direction to shade the *p*-value in the sampling distribution of means, students tend to select the option that "shading to the right" is the valid statement, even though the alternative hypothesis suggests that one should shade the larger left tail of the distribution. The wording of the scenario that precedes the item, the figure shown along with the scenario, and the two options available to the student appear in Appendix B. It

seems that more students think the *p*-value must always be less than 50% after taking the course.

5. RECOMMENDATIONS

Some introductory statistics students will eventually become statistical consulting clients. This research was intended to be presented to consulting statisticians who may have clients who have taken a statistics class but may not remember much of what they once knew. Some statistics clients may not have had a statistics class at all. The difficulties that students have will likely be compounded in clients after many years of little to no exposure to statistical concepts at all. There are undoubtedly similar confusions among introductory statistics students and clients encountered by statistical consultants. Statistics clients like students, need to understand what the key statistical terms mean and what they do not mean (statistical literacy). They also they need to understand some of the caveats and limitations of the statistical inference process (statistical reasoning/thinking).

Three key points are enumerated that should be emphasized—whether teaching introductory statistics students or dealing with clients who need a refresher on statistical concepts. First, the *p*-value is not a magical number; it is an integrated part of the larger statistical process. Second, to interpret the *p*-values and statistical significance properly, one must tend to both the logic of inference and the scope of inference. The logic of inference determines how we interpret results (statistical reasoning). This inferential logic requires attending to sample sizes, effect sizes and, most importantly, whether the necessary conditions for inference were met (or how badly they were violated). The scope of inference determines what we can conclude. The scope of the inference that can be made from the sample depends on how randomness was used in the study design; how the data were gathered. Attending to the scope of inference requires a broader understanding of the entire statistical process (statistical thinking).

Furthermore, the confusions that persist about confidence intervals (CI) should also be reinforced. In terms of statistical literacy, the CI estimates population parameters or true effects, given the sample data observed. From a statistical reasoning perspective, the CI also provides complementary information that *p*-values do not provide alone; namely, the bounds for the effect. The CI can be used to assess statistical significance. When interpreting a confidence interval, one should note whether a given null hypothesis value is contained in the confidence interval or not. For example, if the null hypothesis is that there is no effect, is zero in the confidence interval of the difference? One can determine the direction of the effect, by noting if the interval is all positive or all negative.

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.

Bakker, A., & Gravemeijer, K. (2004), "Learning to Reason about Distribution," in D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147-168). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

Ben-Zvi, D., Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J Garfield (Eds.), *The challenge of*

developing statistical literacy, reasoning, and thinking (pp. 3-15). Dordrecht, the Netherlands: Kluwer Academic Publishers.

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

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

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

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

de Veaux, D., Velleman, P., & Bock, D. (2006). Intro stats (2nd ed.) Belmont, CA: Brooks/Cole – Thompson Learning.

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

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

Garfield, J., delMas, R., & Chance, B. (2005). Tools for teaching and assessing statistical inference. Retrieved September 25, 1014 from http://www.tc.umn.edu/~delma001/stat_tools/

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

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education, 3(1)*. Retrieved September 25, 1014 from http://www.amstat.org/publications/jse/v3n1/konold.html

Lane-Getaz, S. J. (2014). A graphical approach to examine inferential reasoning development. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics* (ICOTS9, July, 2014), Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute. http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_C101_LANEGETAZ.pdf

Lane-Getaz, S. J. (2013). Development of a reliable measure of students' inferential reasoning ability. *Statistics Education Research Journal (SERJ), 12*(1), 20-47. http://iase-web.org/documents/SERJ/SERJ12(1)_LaneGetaz.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. http://www.stat.auckland.ac.nz/~iase/publications/sat07/Lane-Getaz.pdf

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

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

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

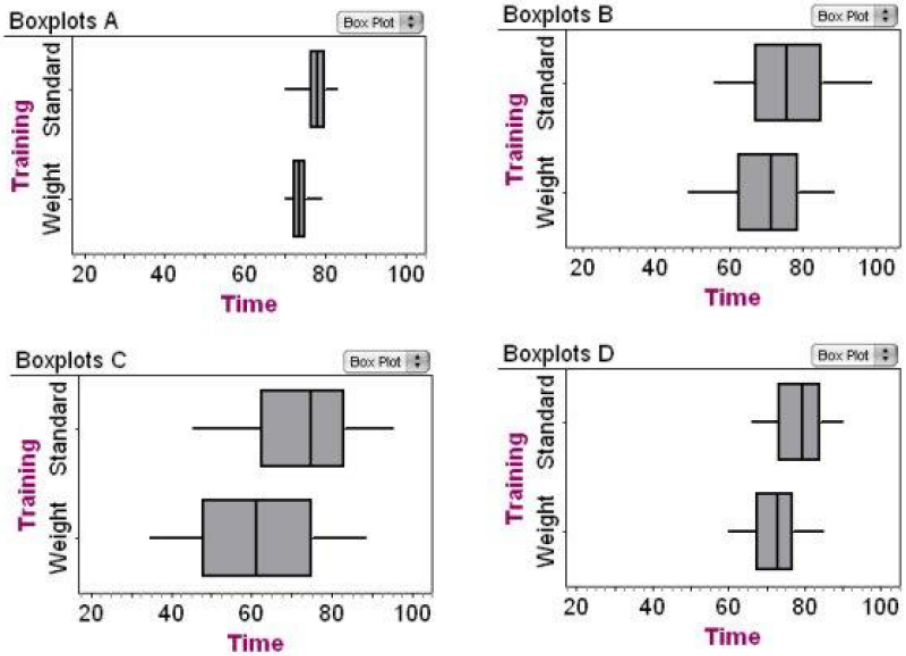
Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008), "A Framework to Support Research on Informal Inferential Reasoning," *Statistics Education Research Journal*, *(7)*2, 40-58.

Appendix A.

Read scenario 3c and interpret the graphical representations presented. Please answer all items. 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.)



3c-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

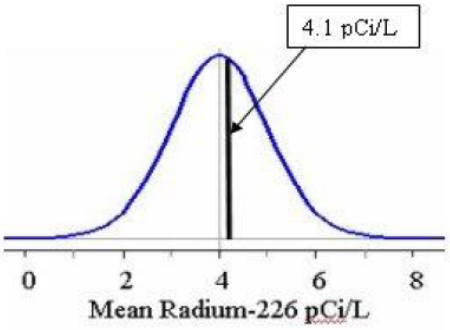
3c-2. Which set of boxplots provides the LEAST convincing evidence that the difference between the two groups of athletes is due to the training program.

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

Appendix B.

Scenario 3b

Radium226 is a naturally occurring radioactive gas. For public safety, the Environmental Protection Agency (EPA) has set the maximum exposure level of Radium226 at a mean of 4 pCi/L (picocuries per liter). Student researchers at a southern Florida university expected to show that Radium226 levels were less than 4 pCi/L. However, these student researchers collected 32 soil specimens with a mean Radium226 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.



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

- o Valid Statement
- o Invalid Statement