LINKING THE RANDOMIZATION TEST TO REASONING ABOUT *P*-VALUES AND STATISTICAL SIGNIFICANCE

Sharon J. Lane-Getaz
Departments of Mathematics, Statistics and Computer Science, and Education,
Saint Olaf College, United States of America
lanegeta@stolaf.edu

This quasi-experiment compares student learning outcomes from three college statistics courses to investigate whether greater randomization test content explains gains in conceptual understanding of inference, adjusting for prior knowledge and mathematical ability. The study uses a 34-item Reasoning about P-values and Statistical Significance (RPASS) scale to measure gains in students' inferential understanding. Of two introductory courses, one has limited randomization content ($n_1 = 55$). The second emphasizes randomization, simulation, and P-values throughout ($n_2 = 26$). The third is a second course in statistics that reviews randomization tests at the beginning of the course ($n_3 = 24$). Comparative results, score reliability, and the changes in respondents' correct conceptions and misconceptions are reported. Directions for future research are discussed.

INTRODUCTION

Teaching inference using randomization or permutation tests is believed to deepen students' conceptual understanding of inference (Cobb, 2007; May & Hunter, 1992; Rossman, 2008). Psychologists May & Hunter (1993) asserted that teaching inference using permutation tests for comparing two sample groups could serve three pedagogical benefits. They assert that using randomization simulations or permutation tests to teach inference can:

- Clarify whether hypotheses being tested relate to sample statistics or population parameters
- Offer a more appropriate analysis option, if conditions for normal theory are not met
- Differentiate randomization distributions—used to make inferences about group associations—from normal distributions—used to make inferences about populations from random samples

Cobb (2007) further claims randomization tests can bring the "logic of inference" to the center of the introductory course. Randomization simulations can be used to explicitly highlight the "Three R's of inference…[to] randomize data production; repeat by simulation to see what's typical and what's not; and reject any model that puts your data in its tail" (Cobb, 2007). Rossman (2008) also suggests that "simulation of the randomization test provides an informal and effective way to introduce students to the logic of statistical inference." There is no evidence of a link between teaching inference using randomization distributions and students' inferential understanding. The goal of this study is to explore whether and to what extent greater exposure to randomization tests and simulations explain gains in students' inferential understanding as measured by the *Reasoning about P-values and Statistical Significance* (RPASS) scale (Lane-Getaz, 2007).

METHODS

This quasi-experimental study compares student learning outcomes from three college statistics courses to investigate whether greater randomization test content explains gains in conceptual understanding of inference. The design is similar to a group comparison experiment but without randomization of subjects to groups (Pedhazur & Schmelkin, 1991); therefore, gain scores were adjusted for prior knowledge and mathematics ability. Concepts for items with a large pretest to posttest change in the proportion of students' answering the item correctly ($\geq .30$) are reported.

Instrument

The 34-item *Reasoning about P-values and Statistical Significance* (RPASS-7) scale was administered as a pretest and posttest to measure the effects of different courses on students' conceptual understanding and misunderstanding of inference (Lane-Getaz, 2007, 2008). Internal consistency reliability of RPASS scores was estimated using Cronbach's coefficient alpha. Student explanations were requested on eight selected RPASS items as insight to item functioning.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

Subjects and setting

During spring 2009 students enrolled in three sections of Principles of Statistics (Course-1), three sections of Statistics for Science (Course-2), and two sections of Statistical Modeling (Course-3) at a small US liberal arts college were invited to participate. Of the 215 students enrolled in the courses, 132 completed the RPASS as a pretest and posttest. The sample includes 105 respondents who answered every item. Table 1 details their class year and gender by course.

Table 1. Number of Sample Respondents by Class Year, Gender, and Course (N = 105)

	Class year				Gender		
Course	Freshman	Sophomore	Junior	Senior	Female	Male	Total
Course-1	23	22	5	5	45	10	55
Course-2	9	10	2	5	19	7	26
Course-3	5	5	11	3	12	12	24

Course descriptions

All three courses expose students to randomization tests, use real data, and require completion of a final research project. Course-1 is a terminal, introductory course for the liberal arts with an Algebra prerequisite. The text is *Statistics: Concepts and Controversies* (Moore & Notz, 2002). Tools include *SPSS, Fathom,* and *Minitab.* Course-2, an introductory course for students in the sciences with a Calculus prerequisite, introduces randomization content early and repeatedly throughout the course. The text is *Investigating Statistical Concepts, Applications, and Methods* (Chance & Rossman, 2006). Tools used include *Minitab* and applets that accompany the textbook. Course-3, a second course in statistics, reviews randomization tests and *t*-tests for the first third of the course, then builds toward multiple regression, ANOVA, and multiple logistic regression using *R*. The text is *The Statistical Sleuth* (Ramsey & Schafer, 2002).

RESULTS

The 105 sample respondents answered 25.5 of 34 RPASS posttest items correctly, 75% on average. RPASS posttest results were unimodal and somewhat left skewed (M = 25.5, SD = 4.6, Mdn = 26, IQR = 6). A preliminary two-way ANOVA was conducted to compare RPASS pretest scores included in the sample to those excluded, by course. There was no statistically significant interaction effect ($F_{(2,168)} = .23$, $p = .79$). Similarly, no statistically significant interaction was found between included and excluded posttest scores by course ($F_{(2,134)} = .37$, $p = .69$). Course-3 posttests included in the sample did have scores that were 2.3 items higher, on average, than posttests excluded from the sample (n = 24, n = 17, respectively; $t_{(39)} = 2.56$, $p = .02$). Nevertheless, the sample was sufficiently representative of all respondents taking the RPASS pretests and posttests.

Figure 1 displays boxplots for RPASS pretest scores. Figure 2 depicts boxplots for RPASS posttest scores, clustered by instructor. The clustered boxplots suggest no instructor effect, which is of import since Instructor 3 is the researcher. Course-1 and Course-2 had similar RPASS pretest distributions with Course-3 having higher initial scores, on average. Course-2 and Course-3 had similar RPASS posttest distributions. Thus, boxplots of gain scores by course (Figure 3) show Course-2 achieved the greatest gains in inferential understanding. Table 2 enumerates RPASS pretest and posttest means, mean gains, and standard deviations by course.

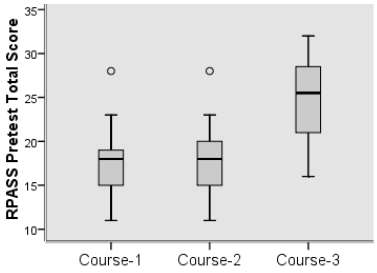

Figure 1. Boxplots of RPASS pretests, N = 105

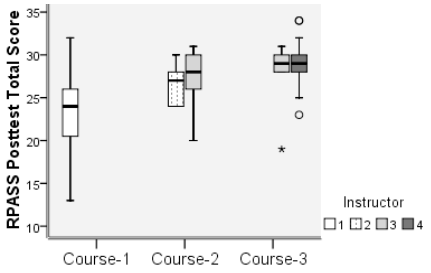

Figure 2. Clustered boxplots of RPASS posttests by instructor within course, N = 105

Table 2. Sample RPASS Pretest and Posttest Means, Gains, and SDs by Course (N = 105)

Course	*N*	RPASS pretest mean (*SD*)	RPASS posttest mean (*SD*)	Mean gains (*SD*)
Course-1	$n_1 = 55$	17.4 (3.0)	23.2 (4.5)	5.8 (5.0)
Course-2	$n_2 = 26$	17.8 (4.0)	27.4 (2.9)	9.6 (4.5)
Course-3	$n_3 = 24$	25.0 (4.8)	28.7 (3.1)	3.7 (4.4)

To facilitate course comparisons given the absence of randomization, scores were adjusted using RPASS pretests (for prior knowledge) and ACT Mathematics (for math ability), likely confounds. The effect of the course was not statistically significant ($F_{(2,83)} = 2.08$, $p = .13$, partial $\eta^2 = 5\%$, $N = 88$: $n_1 = 51$, $n_2 = 15$, $n_3 = 22$). With the sample size reduced to the 88 respondents who took the ACT Math exam, power may have been insufficient to detect an effect. Helmert contrasts indicated the adjusted mean gain for Course-1 was 2 items less than the adjusted mean gain for the combination of Course-2 and Course-3 ($p = .05$ two-tailed, 95% CI ranging from 0 to 3.9 items).

Three additional ANCOVA analyses were run imputing missing values to increase sample size and power. The course taken was statistically significant in all three analyses. In the first and most conservative analysis zero was imputed for missing gain scores, increasing sample size to 108. The course taken explained 7% of the variation in gains, adjusted for RPASS pretests and ACT Math scores ($F_{(2,103)} = 4.0$, $p = .02$, partial $\eta^2 = 7\%$, $N = 108$: $n_1 = 66$, $n_2 = 17$, $n_3 = 25$). Course-2 and Course-3 respondents answered an average of 2.6 additional items correctly compared to those in Course-1 (95% CI from .7 to 4.5 items). ACT Math remained statistically significant in this analysis ($F_{(1,103)} = 15.8$, $p < .001$, partial $\eta^2 = 13\%$, $N = 108$). The RPASS adjusted mean gains, standard errors, and 95% confidence intervals are reported in Table 3 by course. Figure 4 shows the adjusted mean gains plotted by course. For the remaining two analyses sample size increased to 123 (imputing zeros for skipped items) and 129 (imputing zero for skipped items and missing gains). ACT Math was no longer statistically significant. Course explained 13.9% and 7% of variation in gains, respectively. Increasing sample size through imputation also improved the RPASS posttest score reliability estimate from $\alpha = .76$, $N = 105$ to $\alpha = .82$, $N = 175$.

Table 3. Adjusted Mean Gains, Standard Errors, and Confidence Intervals by Course (N = 108)[a]

Course	Adjusted mean gains	Standard Error	95% Confidence interval
Course-1	3.9[b]	.6	(2.8, 5.0)
Course-2	7.0[b]	1.0	(5.0, 9.1)
Course-3	6.1[b]	1.0	(4.1, 8.1)

Note. [a]Imputed zero for missing gain scores. [b]Covariates: RPASS pretest = 19.26, ACT Math = 27.32.

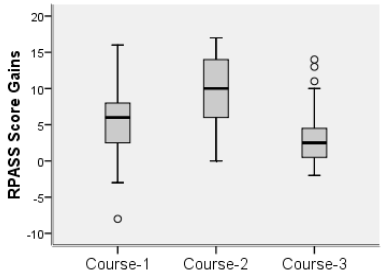

Figure 3. Boxplots RPASS gains, N = 105

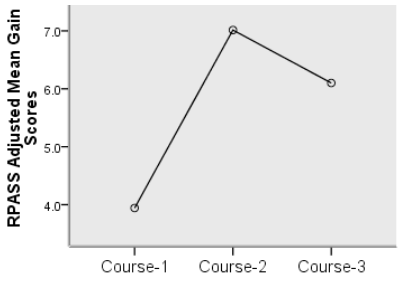

Figure 4. RPASS adjusted mean gains, N = 108

Inferential concepts with the largest pretest to posttest change in the proportion of students answering correctly included: assessing significance graphically (.37 to .77), and understanding that *P*-values are dependent on the direction of the alternative hypothesis (.28 to .65), that smaller *P*-values provide stronger evidence of an effect (.56 to .91), and that *P*-values are related to sampling variation (.59 to .91). Two misconception items showed a large change in the proportion correct: confusing the significance level alpha with the *P*-value (.54 to .86) and ascribing chance as the cause of the observed results (.60 to .90). The proportion change for respondents selecting the

correct *P*-value definition among three common misconceptions increased from .53 to .84. However, written explanations revealed difficulties understanding relationships between one-tailed and two-tailed tests. One item read: "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?" Eleven respondents wrote they would divide the one-tailed *P*-value in half to obtain a two-tailed *P*-value. Another item read: "One student argued that the appropriate *P*-value should be 7/100 or .07 for a one-tailed hypothesis, which was sufficient to reject at the .10 significance level but insufficient to reject at the .05 level." Fourteen students wrote that they needed to conduct a two-tailed test, even though the context called for a one-tailed test. This difficulty had not been previously identified.

DISCUSSION

This quasi-experimental study compares student learning outcomes from three college statistics courses to investigate whether greater randomization content explains gains in inferential understanding. Respondents in the introductory course with greater randomization content (Course-2) attained a statistically higher mean gain in inferential understanding compared to those in the introductory course with less randomization content, even after adjusting for prior knowledge and math ability. The adjusted mean gain for Course-2 was equivalent to Course-3, a second course in statistics. While quasi-experimental designs are insufficient to draw strong causal conclusions, the study does link greater exposure to randomization content to a better understanding of inference. RPASS score reliability was sufficient to support group comparison (Pedhazur & Schmelkin, 1991). However, some potential benefits of randomization content for teaching inference (May & Hunter, 1993; Cobb, 2007) may not be measured by RPASS. Items need to be explicitly identified, validated, and tested to better assess these potential benefits.

This study lays a foundation for future experimental studies to evaluate the effectiveness of randomization methods for teaching inference. Future studies should explore the confusion between one-tailed and two-tailed tests, retention of inferential understanding, and compare courses with and without randomization content. In so doing, evidence can also be used to improve RPASS psychometric properties, so results and conclusions can be transferred across studies.

REFERENCES

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

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

Lane-Getaz, S. J. (2008). Introductory and intermediate students' understanding and misunderstanding of *P*-values and statistical significance. *Proceedings of the 11th International Congress on Mathematical Education* (ICME). Online: http://tsg.icme11.org/document/get/475.

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. Online: http://www.stat.auckland.ac.nz/~iase/publications/sat07/Lane-Getaz.pdf.

May, R. B., & Hunter, M. A. (1993). Some advantages of permutation tests. *Canadian Psychology*, *34*(4), 1-10.

Moore, D. S., & Notz, W. I. (2002). *Statistics: Concepts and controversies* (6th edition). New York: W. H. Freeman.

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

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

Rossman, A. J. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal*, *7*(2), 3-4. Online: www.stat.auckland.ac.nz/~iase/serj/SERJ7%282%29_Rossman.pdf.