TOWARD THE DEVELOPMENT AND VALIDATION OF THE REASONING ABOUT *P*-VALUES AND STATISTICAL SIGNIFICANCE SCALE

Sharon J. Lane-Getaz
University of Minnesota
Minneapolis

This paper describes the development and validation of the Reasoning about P-values and Statistical Significance (RPASS) scale. The RPASS was designed to support future research on students' conceptual understanding and misunderstanding of statistical significance and the effects of instructional approaches on this understanding. After expert content validation and testing, the 27-item RPASS-4 was administered across five introductory courses at California Polytechnic State University (N = 224). Respondents answered 16 of 27 items correctly, on average. This paper reports evidence of construct validity, both convergent and discriminant validity evidence (n = 56). However, internal consistency reliability was low (α = .42, N = 224). A subset of 15 items was identified with expected coefficient alpha of .66 by removing items with low corrected item-total correlations. Implications for future development and research are discussed.

INTRODUCTION

Leading statisticians and statistics educators recommend that educators emphasize the conceptual understanding of *P*-values and the logic of inference in the first course (Cobb, 2005; Franklin & Garfield, 2006; Moore, 1997). Literature from education, psychology, statistics, and statistical and mathematics education suggests inferential concepts are commonly misunderstood by students and misinterpreted by some researchers. However, there are no instruments with reported evidence of validity and reliability that assess how people understand and misunderstand this topic. The goal of this research is to develop and validate a new assessment instrument for statistical education, the Reasoning about *P*-values and Statistical Significance (RPASS) scale. The intended use of the RPASS is to facilitate research on students' understanding and misunderstanding of inference and the effect of instructional approaches on this understanding.

BACKGROUND

Studies about understanding and misunderstandings of P-values and statistical significance

Literature documenting the use and misuse of *P*-values and statistical significance is extensive (e.g., Cohen, 1994; Kline, 2004; Nickerson, 2000). In the recent literature seven observational studies offer empirical data supporting claims that misunderstandings are common and persistent (Falk & Greenbaum, 1995; Haller & Kraus, 2002; Mittag & Thompson, 2000; Oakes, 1986; Vallecillos-Jimenez & Holmes, 1994; Wilkerson & Olson, 1997; Williams, 1999). None of the studies using questionnaires, items, tests or surveys reported evidence of score reliability or validity of item content. Many studies used too few items to sufficiently assess the content domain. Fourteen difficulties culled from this literature are summarized in Table 1 and grouped into four categories. These difficulties framed the preliminary test blueprint for the RPASS.

What do we want students to know?

In addition to understanding difficulties people have, it is important to define what students should know. Some statistics education professionals have found it useful to think of instructional outcomes using the taxonomy of statistical literacy, reasoning, and thinking (see Ben-Zvi and Garfield, 2004). Instructional outcomes from the Tools for Teaching and Assessing Statistical Inference website (Garfield, delMas, & Chance, 2005) were mapped to this taxonomy and added to the test blueprint. RPASS items assessing statistical literacy might include recognition of definitions, symbols, and graphical representations of *P*-values and statistical significance. Statistical reasoning items might require interpreting results, making comparisons, and making connections between concepts related to *P*-values and statistical significance. Statistical thinking items might require students to connect significant results to the broader context of a statistical investigation.

Table 1
Classification of Difficulties Understanding P-values & Statistical Significance

Category	Difficulties	Selected references
	Misunderstanding Basic terminology and concepts	
B-1	Confusing basic language and concepts of inference	Batanero, 2000 Williams, 1999
B-2	Believing the *P*-value is always low	Williams, 1999
	Confusing Relationships between inferential concepts	
R-1	Confusing test statistics & *P*-values	Williams, 1999
R-2	Confusing samples and populations	Mittag & Thompson, 2000
R-3	Confusing α and Type I error rate or significance level with the *P*-value	Haller & Krauss, 2002 Mittag & Thompson, 2000 Williams, 1999
R-4	Believing *P*-value is independent of sample size	Mittag & Thompson, 2000 Wilkerson & Olson, 1997
R-5	Believing reliability is *1 – P*-value	Daniel, 1998 Haller & Krauss, 2002 Mittag & Thompson, 2000 Oakes, 1986
	Misapplying the Logic of statistical inference	
L-1	Misusing Boolean logic of contra-positive proof ($a \rightarrow b$ and not-b, then not-a) (deterministic)	Batanero, 2000 Falk & Greenbaum, 1995 Oakes, 1986
L-2	Misusing Boolean logic of converse ($a \rightarrow b$ replaced with $b \rightarrow a$) (logic error)	Batanero, 2000
L-3	Thinking *P*-value is probability chance *caused* results or “probability due to chance”	Daniel, 1998
	Misinterpreting the *P*-value as the probability of the truth or falsity of Hypotheses	
H-1	Misinterpreting the *P*-value as the probability the alternative hypothesis is true	Falk & Greenbaum, 1995 Haller & Krauss, 2002 Oakes, 1986
H-2	Misinterpreting the *P*-value as the probability that accepting the alternative hypothesis is false	Falk & Greenbaum, 1995 Haller & Krauss, 2002 Williams, 1998, 1999
H-3	Misinterpreting the *P*-value as the probability the null hypothesis is true	Falk & Greenbaum, 1995 Haller & Krauss, 2002 Oakes, 1986
H-4	Misinterpreting the *P*-value as the probability the null hypothesis is false	Falk & Greenbaum, 1995 Haller & Krauss, 2002 Oakes, 1986

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

Research questions

Existing research instruments in statistical education did not address all of the identified content (e.g., Garfield, 2003; Allen, Stone, Rhoads, & Murphy, 2004; delMas, Ooms, Garfield, and Chance, 2006; delMas, Garfield, Ooms, & Chance, in press). A research instrument with reported psychometric properties is needed to assess understanding and misunderstandings about *P*-values and statistical significance. Based on previous research about this topic, and what has been learned about developing research instruments in statistics education, two questions were posed:

Question 1: *Can a research instrument be developed, validated, and piloted to produce sufficiently reliable scores and thereby facilitate future research in students' understanding of and difficulties with reasoning about P-values and statistical significance?*

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

METHODS

Phases I - III: Instrument development and content validation

During Phase I, the preliminary test blueprint was developed based on difficulties culled from the literature. RPASS items were modified from four multiple-choice items selected from the ARTIST (Assessment Resource Tools for Assessing Statistical Thinking) website available from https://app.gen.umn.edu/artist/. The multiple-choice options were converted to multiple true-false item sets to improve reliability and validity (Downing, 1992). The resultant four problem scenarios and 16 true-false items were reviewed by statistics education advisors (n = 5). One item was added, and the 17-item RPASS-1 was piloted at the University of Minnesota the end of fall semester 2004 (N = 333). There was little variation between scores between the four courses tested (Lane-Getaz, 2005). Five correct conceptions and 12 misconceptions were assessed.

In Phase II the blueprint was revised per the ongoing literature review. Learning goals for teaching *P*-values and statistical significance were added from the Tools for Teaching and Assessing Statistical Inference website (Garfield, et al., 2005). Items were added or modified to meet new goals. RPASS content was classified by statistical literacy, reasoning, or thinking. After review with the five statistics education advisors a 25-item RPASS-2 was produced, assessing 7 correct conceptions and 18 misconceptions.

During Phase III the RPASS-2 was administered to students at California Polytechnic State University (Cal Poly) the end of winter quarter 2006. Feedback from testing and 13 student interviews (n = 61) produced the 25-item RPASS-3A. Next, content was validated by 10 subject matter experts from four colleges and universities. Experts recommended that redundant misconception items be removed and more correct conception items were written. After two rounds of feedback and individual interviews with each rater, all ten experts *agreed* or *strongly agreed* that the 28-item RPASS-3C assessed the stated learning objectives or misconceptions. Deleting one additional redundant item produced the 27-item RPASS-4, assessing 13 correct conceptions and 14 misconceptions (Lane-Getaz, 2007).

Phases IV - V: RPASS large scale class testing

Setting and participants

The data in this paper were collected at Cal Poly during spring quarter 2006. A sample of 224 students from five introductory statistics courses completed RPASS-4 (see Tables 2 and 3). Of 56 students who completed two additional instruments to assess construct validity, 37 were AgStat (statistics for agriculture) and 19 LibStat (statistics for liberal arts) students.

Table 2
Number of RPASS-4 Respondents by Class Standing and Statistics Course

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

Table 3
Number of RPASS-4 Respondents by College Major and Statistics Course

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

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

Instruments used to assess construct validity

Convergent and discriminant validity evidence was collected using instruments and items from the ARTIST website. Since no criterion measure existed, a five-part open-ended item related to *P*-values and statistical significance was selected to administer concurrent with the RPASS during finals week. This open-ended item was used to examine convergent validity. A second instrument, the 14-item Bivariate Quantitative Data topic scale was administered during week 9 to examine discriminant validity.

Procedures

RPASS-4 was administered online across five introductory courses over the course of two weeks. Participants were tested in the same 24-station lab. Depending on the instructor, students earned extra credit, homework credit or final exam points for participation. Items were summarized across three dimensions: correct conceptions and misconceptions, the four content areas defined by the blueprint, and the three learning goals for statistics instruction. The mean proportion of correct responses was computed by first computing the mean proportion of correct responses by item, and then computing the means of these proportions for each of the three item groupings.

Construct validity evidence was gathered in two of the five courses (n = 56). Pearson product-moment correlations were computed between the open-ended item ratings and the RPASS to provide convergent validity evidence. Correlating the Bivariate Quantitative Data topic scale with the RPASS provided discriminant validity evidence.

RESULTS AND ANALYSIS

RPASS-4 results

Respondents answered 16 of 27 items correctly, on average, with standard deviation of 3 items (N = 224). Table 4 summarizes the mean proportion of correct responses across three item grouping and the number of items per item grouping. Table 5 reports the mean proportion of correct responses (RPASS-4 item difficulties) and the corrected item-total correlation by item. The learning goals and correct conception or misconception assessed are also identified. Items are sorted by difficulty within blueprint category.

Table 4

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

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

Table 5

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

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

Note. RPASS-4 mean difficulty 16 correct / 27 items = .60, *SD* = 3 items; assessed 13 correct conceptions, 14 misconceptions. [a]Corrected item-total correlation removes item contribution from total. [b]Three-option item.

Reliability

The RPASS-4 reliability across the five introductory courses was low (Cronbach's coefficient $\alpha = .42$, $N = 224$). Thus, 42% of the variation in RPASS scores could be attributed to true score variation. The remainder of the variation could be attributed to measurement error.

Validity

Pearson's *r* was used to compute correlations to examine construct-related validity. The instrument used to assess discriminant validity had low reliability for this subgroup; therefore, the discriminant correlation was corrected for attenuation. In addition, both convergent and discriminant correlations were further corrected for attenuation due to the low reliability of RPASS-4. Table 6 presents these uncorrected and corrected validity coefficients as off-diagonal elements. The on-diagonal elements are the instrument reliabilities for the Bivariate Quantitative Data topic scale and RPASS-4. The proportion of rater agreement is reported for the open-ended item.

The uncorrected convergent correlation was positive and statistically significant but weak. The discriminant correlation was very weak, and not statistically significant. Correcting the correlations for attenuation, yielded a more moderate correlation, suggesting the low reliability of RPASS-4 scores constrained the convergent comparison measure. However, the discriminant correlation remained weak, even after correction. The lack of correlation with the discriminant measure discredits plausible rival interpretations – such as general statistics knowledge or general intelligence – to explain relationships found. Furthermore, testing methods do not explain correlations (or lack of correlation) found. That is, the dissimilar Bivariate Quantitative Data topic scale was online as was RPASS-4; whereas the open-ended item with more similar content was administered via paper and pencil (Campbell & Fiske, 1959). The pattern of validity coefficients provides some evidence that RPASS-4 measures the desired construct.

Table 6
RPASS-4 Reliability and Validity Coefficients for AgStat and LibStat Respondents[a]

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

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

Investigating RPASS-5 reliability and validity

An item analysis was conducted to identify a subset of items that might have higher internal consistency reliability. Using Phase IV data, 12 items were iteratively removed from the scale with corrected item-total correlations less than .15. Coefficient alpha was estimated as .66 for the remaining 15 items (RPASS-5) using existing data. After correcting for attenuation, the convergent validity coefficient for RPASS-5 was moderate (corrected $r = .49$). The discriminant correlation remained very weak (corrected $r = .15$). The pattern of validity coefficients provides some evidence that the RPASS-5 item subset measures the desired construct. RPASS-5 assessed 7 correct conceptions and 8 misconceptions.

DISCUSSION AND CONCLUSIONS

Limitations

The 76% participation rate suggests results are representative of the five targeted courses, the population. Generalizations to other populations should be made with caution. There are four factors that may have limited the construct-related validity evidence obtained. First, no criterion measure existed to provide an adequate comparison. Second, the instrument used to examine convergent validity did not sample across the same content domain as the RPASS. Third, the instrument used to examine convergent validity was not content-validated in the form used. Fourth, correcting correlations for attenuation yielded moderate concurrent evidence and weak discriminant evidence, suggesting low reliability of the RPASS attenuated the convergent validity correlation. Two limitations may have impacted the reliability evidence obtained. The stratified structure of the RPASS (e.g., RPASS correct conception and misconception item specifications) may have constrained reliability as measured by Cronbach's coefficient alpha (Cronbach and Shavelson, 2004). Internal consistency reliability may also have been constrained by students' inconsistent reasoning on these kinds of items as discussed by Konold (1995). If inconsistent student reasoning limits the internal consistency reliability of the scores, a better measure of reliability might be a test-retest correlation (stability) rather than internal consistency.

Implications for future research

Assessing inferential topics that are most commonly taught across courses should reduce guessing and improve reliability of scores. Even though omitting the twelve low or negatively correlating items from RPASS-4 may compromise content validity, the elimination of noisy items improved internal consistency. The content of the fifteen RPASS-5 items seems to be a more appropriate content domain for assessing introductory students' understanding. Lengthening RPASS-5 with additional items that cover the same content should increase score variation and thereby improve reliability (Cronbach, 1951). Development of RPASS-6 might include the 15 items from RPASS-5, pus content-validated items from the ARTIST Test of Significance Topic scale (delMas, Ooms et al., 2006) and/or inference-related items from the Comprehensive Assessment of Outcomes in a first Statistics course (CAOS) (delMas, Garfield et al., in press). RPASS-6 would need to be revisited by experts to confirm items sufficiently sample the content domain. Other item improvements might include altering 2-option items to include a third option, where appropriate, to lessen guessing effects (Rodriguez, 2005). RPASS items with low corrected item-total correlations might be further developed using student interviews to explore alternative item wording.

Some reform-based courses have integrated the *P*-value and statistical inference topics throughout the introductory course in order to improve students' inferential reasoning (e.g., Chance & Rossman, 2006; Lane-Getaz & Zieffler, 2006). After additional development, pre- and posttest administration of a future RPASS version in courses with and without instructional interventions may facilitate evaluating the effectiveness of new teaching approaches on inferential understanding. Since random assignment of teaching methods is rarely feasible, results from a broader standardized test could provide a statistical control for a comparative study. Future research questions about inferential reasoning might include: exploring how repeated administration of the RPASS impacts student learning; investigating the development of inferential reasoning as reflected in RPASS scores obtained before, during, and at the end of an introductory course; examining how students' and instructors' correct conceptions and misconceptions compare; or exploring what connections exist, if any, between students' understanding of random sampling and random allocation and their RPASS responses.

Conclusions

Question 1: Can a research instrument be developed, validated, and piloted to produce sufficiently reliable, valid scores and thereby facilitate future research in students' understanding of and difficulties with reasoning about P-values and statistical significance?

This research provided content-related and some construct-related validity evidence. However, reliability of the RPASS-4 total score was low. Deleting items to improve reliability reduces some content coverage but the 15-item RPASS-5 does sample from all four major content areas defined in the item blueprint and RPASS-5 provides a more reliable starting point for future RPASS development.

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

Most respondents seemed to attain statistical literacy. Evidence of statistical reasoning or thinking was less apparent. RPASS results support the relationships between statistical literacy, reasoning and thinking as described by delMas (2002). Misconceptions appeared to be commonly held and respondents exhibited contradictory conceptions as theorized by Konold (1995). Educational and cognitive psychologists have questioned whether targeted instruction and assessment can overturn prior misconceptions about probability and statistics concepts (e.g., delMas and Bart, 1989; Konold, 1995). These respondents learned basic inferential concepts but continued to harbor contradictory misconceptions after instruction. Targeted assessment and instruction may be warranted. With further development the RPASS may be useful for examining inferential reasoning, identifying misconceptions, and designing and evaluating alternative methods for teaching this topic.

ACKNOWLEDGMENTS

The author wishes to thank Bob delMas and Andy Zieffler for feedback on this paper, and Bob delMas and Joan Garfield for their guidance on the dissertation this paper summarizes.

REFERENCES

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

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 and J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-15). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

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

Cobb, G. (2005, May). *Introductory statistics: A saber tooth curriculum?* Presentation at the first United States Conference on Teaching Statistics (USCOTS). Retrieved from http://www.causeweb.org/uscots/uscots05/plenary/

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

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

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

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

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

delMas, R. C., Ooms, A., Garfield, J. B., & Chance, B. (2006). Assessing students' statistical reasoning. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*.

Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/17/6D3_DELM.pdf

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

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

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

Franklin, C., & Garfield, J. (2006). The guidelines for assessment and instruction in statistics education (GAISE) project: Developing statistics education guidelines for pre K-12 and college courses. In G. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth yearbook* (pp. 345-375). Reston, VA: National Council of Teachers of Mathematics.

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

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

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

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, D. C.: 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

Lane-Getaz, S. J. (2007). *Development and validation of a research-based assessment: Reasoning about P-values and statistical significance*. Unpublished doctoral dissertation. University of Minnesota, Minneapolis. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/dissertations/07.Lane-Getaz.Dissertation.pdf

Lane-Getaz, S. J. (2005). Reasoning about *P*-values and statistical significance: Development and pilot of a research-based survey. *Proceedings of the Fourth International Forum on Statistical Reasoning, Thinking, and Literacy* [CD-ROM], Auckland, NZ.

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

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

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

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

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

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

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

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

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

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