INTRODUCTORY AND INTERMEDIATE STUDENTS' UNDERSTANDING AND MISUNDERSTANDING OF *P*-VALUES AND STATISTICAL SIGNIFICANCE

SHARON LANE-GETAZ
Macalester College and Saint Olaf College
lanegetaz@macalester.edu

ABSTRACT

This study examines reliability and validity evidence for the Reasoning about P-values and Statistical Significance (RPASS) scale and reports evidence of introductory and intermediate students' understanding and misunderstanding of inference. RPASS is being developed to facilitate research on students' inferential understanding and the effects of instructional approaches on this understanding. RPASS-6 was constructed by combining the previous RPASS with items from the Assessment Resource Tools for Improving Statistical Thinking: Test of Significance Topic Scale (ARTIST TOS). Expert ratings were reported. The 23-item RPASS-6 was administered in four introductory and three intermediate level courses. Respondents answered 74% correctly, on average. A reliability analysis identified 20 items with sufficient internal consistency to conduct research. Implications for future development and research are discussed.

Keywords: *Statistics education research; P-values; Statistical Significance; Inference; Conceptual understanding; Online Assessment*

1. INTRODUCTION

Statistics and mathematics education professionals have called for assessments with sufficient reliability and validity to be used for education research (e.g., Ben-Zvi & Garfield, 2004; Garfield, 2006; Shaughnessy, 1992). To accomplish this, the development of test instruments with an adequate assessment of their psychometric properties (validity, reliability, and bias) is critical (Schaeffer & Smith, 2007). The purpose of this study is to gather reliability and validity evidence as well as evidence of introductory and intermediate students' understanding and misunderstanding of *P*-values and statistical significance. The instrument used is the Reasoning about *P*-values and Statistical Significance (RPASS) scale. RPASS was developed to facilitate research and evaluation in the teaching and learning of inference and to evaluate the effect of different teaching approaches on understanding inference.

Previous versions of the RPASS were field tested at the University of Minnesota in fall 2004 with 17 items to 333 students, and at California Polytechnic State University (Cal Poly) in the spring of 2006 with 27 items to 224 students. Even though statistics education experts validated the content of 28 RPASS items, RPASS scores for 27 of those items yielded low reliability for the Cal Poly respondents (Lane-Getaz, 2007a). Low reliability constrains correlations used as evidence of construct-related validity. Whether one should be concerned with construct-related validity in education research is debatable (Embretson, 2007; Kane, 2008). While construct-related validity evidence was gathered in this study per the current *Standards* (AERA, APA, & NCME, 1999), the primary concern was to improve RPASS reliability. Some researchers hold that reliability and content-related validity evidence should take precedence in instrument development

for education research (Lissitz & Samuelson, 2007; Mislevy, 2007). Nevertheless, some evidence of RPASS content-related validity was reported in previous research (Lane-Getaz, 2007a). By improving reliability, one might expect stronger evidence of construct-related validity as well. To this end, two questions were posed:

Question 1. Can a selection of the RPASS and ARTIST Test of Significance topic scale items be combined in one scale to produce sufficiently reliable scores to conduct research in students' understanding and difficulties with reasoning about *P*-values and statistical significance?

Question 2. How does the proposed instrument reflect introductory and intermediate level statistics students' understanding and difficulties with reasoning about *P*-values and statistical significance for different courses and levels of statistical preparation?

2. METHODS

2.1. STUDY 1—SCALE DEVELOPMENT ITEM ANALYSIS

Subjects and setting. A sample of 82 students in two introductory statistics courses at Cal Poly who completed both the 27-item RPASS-4 and the 10-item ARTIST TOS as separate scales during the spring quarter of 2006.

Instruments. The 10-item Assessment Resource Tools for Improving Statistical Thinking: Test of Significance Topic Scale (ARTIST TOS) (delMas, Ooms, Garfield, & Chance, 2006) and the 27-item RPASS-4 described in previous research (Lane-Getaz, 2007b). Both instruments have reported content-related validity evidence.

Procedures. A new RPASS scale was constructed for Study 2. The new RPASS-6 scale was constructed by conducting an item analysis combining results from the administration of the 27-item RPASS-4 scale with the 10-item ARTIST TOS scale administered at Cal Poly in spring 2006. The objective of this preliminary reliability study was to identify a new scale with stronger internal consistency reliability than the RPASS-4 ($\alpha = .42$). The reliability analysis of these 37 items was conducted as if the two scales were administered as one scale. Items with low corrected-item-total-correlations (corrected-$r_{pb} < .10$) were iteratively removed until estimated internal consistency reliability was greater than .70. The remaining items were cross-referenced with the Test Content Blueprint developed in previous research (Lane-Getaz, 2007a, 2007b).

2.2. STUDY 2—RPASS-6 RELIABILITY AND VALIDITY STUDY

Study 2 collected a new set of data with the 23-item RPASS-6 at the end of the fall semester of 2007 to evaluate reliability for a new large group. Human subjects consent forms were approved at the three participating institutions.

Subjects and setting. Five statistics education professionals were asked to provide content–related validity evidence for the 23-item RPASS-6 during the fall semester of 2007. Subsequently, a sample of 177 students from seven introductory and intermediate statistics and probability courses completed the 23-item RPASS at the end of the semester (see Table 1). Four introductory level statistics courses participated from two undergraduate liberal arts colleges in the Midwest United States. Each of the institutions had an introductory statistics course with an algebra prerequisite (denoted AP1 and AP2, for institution-1 and -2 respectively). Each institution also had an introductory statistics course with a calculus prerequisite (denoted CP1 and CP2). The two intermediate level regression courses invited to participate were offered at liberal arts institution-2 (denoted RG2) and at an east coast university (denoted RG3). The seventh course was an intermediate level probability course at liberal arts institution-2 (denoted PR2). The columns of Table 1 list the seven participating courses. The rows of Table 1 list the class

standing of the respondents. The participation rate of those invited who completed the RPASS-6 in its entirety was 60%. The sample used to examine construct-related validity was a respondent subset (39 AP2 students) who took three additional assessments.

Table 1. Number of RPASS Respondents by Class Standing and Course[a]

	Introductory level statistics				Intermediate level statistics			
	Institution-1		Institution -2		Institution-2		Institution-3	
	AP1	CP1	AP2	CP2	RG2	PR2	RG3	Total
Freshman	0	1	19	5	1	0	0	26
Sophomore	12	12	10	19	3	5	0	61
Junior	8	7	2	11	3	12	0	43
Senior	3	5	9	7	3	7	0	34
Master's	0	0	0	0	0	0	12	12
Other	0	0	0	1	0	0	0	1
Sum/Invited	23/62	25/72	40/47	43/61	10/10[b]	24/29	12/14	177/295
% Response	37%	35%	85%	70%	100%	83%	86%	60%

Note. [a]Sample consists of respondents from seven courses across three institutions. [b]Of 13 students in the RG2 course, 3 were also enrolled in CP2 and took the RPASS with that class.

Instruments. The 23-item RPASS-6 developed in the preliminary reliability study was completed by the 177 students described as the sample. Furthermore, three additional instruments were administered to 39 AP2 students to gather evidence of construct-related validity: the ARTIST Bivariate Quantitative Data Topic Scale (to examine discriminant evidence), an inferential open-ended free response item selected and then modified from the ARTIST Assessment Builder data base (to examine convergent evidence), and the ARTIST Comprehensive Assessment of Outcomes in a first Statistics course (CAOS) post test (delMas et al., 2007). The CAOS test should also provide convergent evidence.

Procedures. Five statistics education professionals rated the 23 RPASS-6 items on a scale of 1-4 for each item and for the overall scale. A rating of 1 was *strongly disagreed* and a 4 was *strongly agreed* that the item or scale assessed the stated misconceptions or learning objectives.

The 23-item RPASS was administered online in seven statistics and probability courses at the end of the fall semester of 2007. The respondents at institution-1 were administered the RPASS outside of class time and participated on a voluntary basis. The respondents at institution-2 and institution-3 were also voluntary participants but were administered the RPASS during class time. Human subjects consent forms were approved at all three institutions.

To examine construct-related validity, three additional tests were administered in the AP2 course in conjunction with other course exams: a free response item scored holistically on a 7-point scale, the Bivariate Quantitative Data Topic Scale, and the CAOS post test. Validity coefficients, Pearson product-moment correlations, are reported in a multitrait-multimethod matrix correlating RPASS-6 scores with scores from similar and dissimilar assessments (Campbell & Fiske, 1959). Furthermore RPASS-6 scores were correlated with reported college entrance test scores: Scholastic Aptitude Test (SAT) scores for Reading and Mathematics, and the American College Testing (ACT) Composite scores. The ACT Composite is the average of student scores across four tests (English, Mathematics, Reading, and Science). Results include distributions of RPASS-6 scores, validity coefficients, the proportion of respondents who answered each item correctly, and internal consistency reliability (computed as Cronbach's coefficient alpha).

3. RESULTS

3.1. STUDY 1—SCALE DEVELOPMENT ITEM ANALYSIS

Seven of the content-validated ARTIST items and 16 of the content-validated RPASS-4 items remained after iteratively removing items with the lowest corrected-item-to-total correlations from the combined scale. This preliminary reliability analysis suggested the 23-item RPASS-6 combined scale would produce a Cronbach's coefficient alpha of .711. These 23 RPASS-6 items were cross-referenced with the Test Content Blueprint by category (Lane-Getaz, 2007a, 2007b). All four of the Test Blueprint major content areas were represented: 14 basic literacy items (denoted B-1 or B-2), 4 relationships between concepts items (denoted R-3, R-4), 3 logic of inference items (denoted L-3, L-4), and 2 tying *P*-values back to hypotheses items (denoted H-1, H-4). These designations are used in the secondary item analysis at the end of Study 2. Nine correct conceptions and seven misconceptions were directly assessed. In addition, seven multiple choice items assessed correct conceptions, with misconceptions as distractors.

3.2. STUDY 2–RPASS-6 SCORES, RELIABILITY AND VALIDITY EVIDENCE

Content-related validity evidence. After the first round of expert review, all but three RPASS-6 items were rated as *agreed* or *strongly agreed* (3 or 4) that the item validly measured the stated objectives or misconception. Even though 20 items were rated 3 or higher, expert-recommended item modifications were implemented to improve item clarity. The three items with a mean rating of less than 3 appear in Table 3 with the recommended modification and the final rating and RPASS-6 item number after the changes were implemented. These ratings remained below the desired level but were not eliminated pending further item information obtained during this study.

Table 3. Mean Validity Rating for Three Items Rated below 3 (agreed) by Expert Raters

Study 1 item number and description	Mean rating	Expert recommended modification	RPASS-6 Item	Mean rating
4. Conclusions and study design	2.4	Move the item from Scenario 1 to a comparative study scenario.	4-2	2.8
8. Caused by chance	2.2	Clarify that there is no random assignment for this item.	2-3	2.6
10. Confusion of the converse	1.8	Clarify that the magnitude of the *P*-value is small.	3-1	2.8

Total scores and reliability evidence. RPASS-6 data were gathered for reporting in the aggregate, by course, level of course, and by institution. Respondents answered 17 of 23 items correctly, 74% correct on average (M = 17.1, SD = 2.7, Mdn = 18, IQR = 4, N = 177). Items were scored with a 1 if the respondent recognized a correct conception items as correct or identified a misconception item as incorrect, 0 otherwise. Question 2-2 (Strong statistical evidence) was answered correctly by all 177 respondents and was removed from the scale to compute reliability. The estimated reliability of RPASS-6 items seemed to have improved (Cronbach's coefficient α = .57, N = 177 with 22 items) compared to the 27-item RPASS-4 (α = .42) but was lower than desired for research.

Table 4 reports the mean, standard deviation, median, and interquartile range of the RPASS-6 score distribution by course. There was little difference in the RPASS-6 score distribution for introductory and intermediate level courses (Introductory: M = 17.15, SD = 2.7, Mdn = 18, IQR = 4, n = 131; Intermediate: M = 17, SD = 2.9, Mdn = 18, IQR = 4, n = 46). Figure 1 depicts boxplots of the RPASS-6 score distributions by course. Table 5 reports summary statistics by institution.

Table 4. Mean (SD), Median (IQR) for RPASS-6 Total Scores by Course (N = 177)

	Introductory				Intermediate		
	AP1	AP2	CP1	CP2	PR2	RG2	RG3
	n = 23	*n* = 40	*n* = 25	*n* = 43	*n* = 24	*n* = 10	*n* = 12
Mean (*SD*)	16.8 (3.1)	17.2 (2.9)	18.9 (1.2)	16.5 (2.5)	16.6 (3)	16.9 (1.8)	16.8 (3.2)
Median (*IQR*)	17 (5)	18 (5)	19 (2)	16 (4)	17.5 (4)	18.5 (3)	17.5 (3)

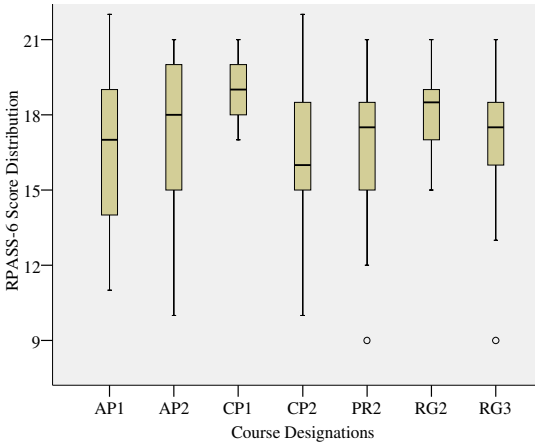

Figure 1. Boxplots of the RPASS score distributions by course, N= 177

Table 5. Mean (SD), Median (IQR) for RPASS-6 Scores by Institution (N = 177)

	Institution-1	Institution-2	Institution-3
	n = 117	*n* = 48	*n* = 12
Mean (*SD*)	17.7 (2.6)	16.9 (2.7)	16.8 (1.2)
Median (*IQR*)	18.5 (2)	17 (4)	17 5 (3)

Construct-related validity evidence. Table 6 reports validity coefficients correlating RPASS-6 scores with scores from similar and dissimilar assessments. The sample consisted of 39 AP2 respondents who took all four assessments. There was a weak but statistically significant correlation of RPASS-6 with the free response item ratings scored holistically on a 7-point scale. All of these RPASS correlations were constrained by the measurement error (low reliability) of the measures. Maximum validity was determined by taking the square root of RPASS reliability. For these 39 respondents RPASS reliability was $\alpha = .66$, so the maximum validity was .81. RPASS-6 scores had a moderate, statistically significant correlation with the ARTIST Bivariate Quantitative Data Topic Scale scores. RPASS-6 had a moderate, statistically significant correlation with the ARTIST CAOS post test. RPASS-6 also had a statistically significant correlation with a subset of 14 inference-related CAOS post test items ($r = .50$, $p < .05$) and the subset of 26 remaining items indirectly related to inference ($r = .64$, $p < .01$).

Table 7 reports Pearson product-moment correlations of RPASS-6 scores with respondent-reported scores on three standardized tests: SAT Reading, SAT Mathematics, and the ACT Composite score. The correlation evidence showed that RPASS-6 had a weak but statistically significant correlation with SAT Reading and the ACT Composite scores as reported by the respondents. There was no statistically significant correlation of the RPASS-6 total scores with the reported SAT Mathematics scores.

Table 6. Reliability and Validity Coefficients for AP2 Respondents[a]

	Written	Online		
	Convergent: Free response	Discriminant: Bivariate	Convergent: CAOS	RPASS-6
Free Response Item Ratings	n/a			
Bivariate Quantitative Data Topic Scale	.41 *	.41 (32)[b]		
CAOS Post Test	.25	.53 ***	.65 (47)[b]	
RPASS-6	.39 *	.60 ***	.68 ***	.66 (39)[b]

Note. [a]Off-diagonal elements are validity, $n = 39$. [b]Internal consistency reliability, Cronbach's coefficient alpha, sample size in parenthesis. * $p < 0.05$, 2-tailed *** $p < 0.001$, 2-tailed

Table 7. Correlations of RPASS-6 Scores with Reported Standardized Test Scores

	SAT Mathematics	SAT Reading	ACT Composite	RPASS-6
SAT Mathematics	1.0 (113)			
SAT Reading	.09 (110)	1.0 (110)		
ACT Composite	.29 (39)	.63 (37) **	1.0 (88)	
RPASS-6	.11 (113)	.24 (110) *	.31 (88) **	1.0 (177)

Note. Sample sizes in parentheses. * $p < 0.05$, 2-tailed ** $p < 0.01$, 2-tailed

Secondary reliability analysis. A subsequent reliability analysis of RPASS-6 produced a 20-item subset with improved reliability ($\alpha = .64$). Item 2-2 was previously removed from the scale since all respondents answered it correctly. Items 2-3 and 3-1 had negative point biserial correlations and were also removed from the scale. The analysis was stopped when further item removal would reduce reliability (see Table 8). Item 4-2 had a mean validity rating less than 3 but unlike Items 2-3 and 3-1 was not removed, since the corrected item-to-total correlation was sufficiently high (corrected-$r_{pb} = .272$).

Table 8. Reliability Analysis of 20-Item Subset of the RPASS-6 Scale (N = 177)

RPASS-6 item number	C/M[a]	Blueprint category	Proportion correct	*SD*	Item-total correlation	α-if-deleted
1-1. Textbook definition	C	B-1	.84	.37	.283	.618
1-2. Dependence on alternative	C	B-1	.58	.50	.305	.613
1-3. Lay definition	C	B-1	.78	.42	.250	.622
1-4. Smaller the *P*-value	C	B-1[b]	.89	.32	.478	.600
2-1. *P*-value in sampling variation	C	B-1	.84	.37	.354	.610
3-2. *P*-value as rareness measure	C	B-1	.85	.36	.252	.622
3-3. *P*-value as always low	M	B-2	.99	.11	.048	.638
4-1. Type I / α and *P*-value	M	R-3[b]	.87	.34	.206	.638
4-2. Conclusions and study design	M	L-4	.76	.43	.272	.627
4-3. Large difference or effect	C	B-1	.85	.36	.233	.619
5-1. Probability: null is false	M	H-4	.67	.47	.277	.624
5-2. Probability: alternative is true	M	H-1	.85	.35	.132	.618
5-3. Sample size and significance	C	R-4[b]	.51	.50	.100	.635
6-1. *P*-value definition	C/M	B-1[b]	.82	.39	.240	.645
6-2. Smaller *P*-value for significance	C/M	B-1[b]	.94	.23	.215	.623
6-3. Strong evidence	C/M	B-1[b]	.83	.38	.268	.628
6-4. Sample size	C/M	R-4[b]	.63	.49	.220	.620
6-5. Practical significance	C/M	B-1[b]	.66	.48	.090	.626
6-6. Necessary conditions	C/M	B-1[b]	.82	.39	.123	.637
6-7. Type I vs. II error	C/M	R-3[b]	.76	.43	.241	.623

Note. [a]C: 8 Correct conceptions, M: 5 Misconceptions, C/M: 7 multiple choice with misconception distracters. [b]10 items have three or more options (the remainder are two-option items).

4. DISCUSSION

4.1. QUESTION 1—RELIABLITY AND VALIDITY

The estimated reliability of the RPASS-6 items was higher (Cronbach's coefficient α = .57, N = 177 with 22 items) than the 27-item RPASS-4, α = .42 (Lane-Getaz, 2007a, 2007b). Item 2-2 was removed from RPASS-6 since all students answered the item correctly. The reliability of the 22 remaining items was lower than desired for research but was sufficiently high that correlations did not require corrections for attenuation.

The expert ratings of the 23-item RPASS-6 provided sufficient evidence of content-related validity. Experts identified three potentially problematic items before the data were collected. Two of these items were also identified during the secondary realibility analysis, after the data were collected. With two independent sources identifying Items 2-3 and 3-1 as problematic, these two items were removed from the developing scale. The remaining 20 items have an estimated reliablity of α = .64, N = 177 using existing data.

The construct-related validity evidence provided mixed results. RPASS-6 had statistically significant, positive correlations with all of the statistically-related tests. RPASS-6 correlated weakly with the 7-point free response item ratings (r = .39). This result is consistent with the statistically significant but weak results obtained in previous research (Lane-Getaz, 2007a, 2007b). The ostensibly discriminant construct-related validity evidence was not consistent with previous research. Scores from the 14-item ARTIST Bivariate Quantitative Data Topic Scale had a moderate correlation with the RPASS-6 scores (r = .60). The RPASS to Bivariate Scale correlation was essentially zero in previous research (Lane-Getaz, 2007a, 2007b). The highest validity coefficient correlated the RPASS-6 scores with the 40-item CAOS post test scores (r = .68). Some of this correlation is attributed to the 14 inference-related CAOS items (r = .50, $p < .05$). However, the RPASS-6 correlates higher with the other 26 CAOS items (r = .64, $p <$ .01). These correlations of CAOS and RPASS-6 may suggest that students cannot understand inference without a good understanding of foundational concepts: standard deviation, mean, median, IQR, variation from center, understanding histograms, et al.

Results suggest RPASS-6, a content-validated measure of inferential understanding, also assesses a students' ability to learn introductory statistics, in general. This interpretation is consistent with the correlation between the RPASS-6 and the ARTIST Bivariate Scale. The ARTIST Bivariate Scale also had a weak but statistically significant correlation with free response item ratings (r = .41). Interrelationships between statistical inference and bivariate statistical concepts may be too intertwined for a bivariate scale to provide discriminant construct-related validity evidence. Both can be modeled at a higher level by the general linear model. Another explanation for this pattern of correlations may be the method of testing. There were higher correlations among the online tests than between the written free response item ratings and online tests (see Table 6).

There was evidence of RPASS-6 discrimination. RPASS-6 correlations with college entrance scores were weaker than with the statistical measures. RPASS-6 had its weakest and statistically insignificant correlation with reported SAT Mathematics scores. The lack of correlation with SAT Math suggests RPASS-6 discriminates from mathematics ability. Correlational results with the SAT Reading and ACT Composite suggest reading ability is needed to take the RPASS but RPASS performance is not dependent on reading ability.

4.2. QUESTION 2—STUDENTS' INFERENTIAL UNDERSTANDING

While the RPASS undergoes this continuous improvement process, the results offer insight into students' understanding and misunderstanding of P-values and statistical

significance. Respondents answered 74% of the 23 RPASS-6 items correctly, on average. Consistent with previous RPASS results, respondents linked statistical significance to smaller *P*-values (Item 1-4 and Item 6-2). Respondents also seemed to understand how to interpret large *P*-values (Item 3-3). The two most difficult items for these respondents were Items 1-2 and 5-2. For Item 1-2, understanding that the *P*-value depends on the alternative hypothesis, the proportion of respondents who answered the item correctly was *p-hat* = .58. For Item 5-2, misunderstanding the impact of sample size on statistical significance, the proportion answering correctly was *p-hat* = .51. The item assessing misunderstanding sample size impact on statistical significance was also the least understood concept in previous research (Lane-Getaz, 2007a). Understanding the impact of sample size on statistical significance and understanding practical versus statistical significance (see Item 6-5, *p-hat* = .66) are topics that may need greater emphasis in introductory and intermediate courses.

5. CONCLUSIONS

5.1. SUMMARY

The process of using a preliminary reliability analysis to combine items from the RPASS and ARTIST Test of Significance scales produced a scale with improved reliability. The expert ratings of the 23-item RPASS-6 provided evidence of content-related validity for this new combined scale. The construct-related validity evidence (correlations with other scales) suggests the RPASS-6 scale not only measures inferential understadning but also measures general introductory statistical knowledge. The scale also discriminates from mathematics ability.

5.2. FUTURE RESEARCH

While there is little difference in this study when comparing students' RPASS scores from introductory and intermediate courses, the introductory course with a Calculus prerequisite (CP1) did have a higher mean and median score, on average, and a smaller standard deviation than any of the other course distributions. The CP1 course had a relatively low voluntary response rate of 35%. The lower response rate may be attributed to the fact that these students took the RPASS outside of class time. It is possible that more able and motivated students self-selected to participate. Nevertheless, this artifact is consistent with results observed in previous research with the RPASS-4 at Cal Poly. The course with the highest mean and median and the smallest variation in scores used the same textbook and teaching approach as the CP1 course in this study (Chance & Rossman, 2006), which emphasizes *P*-values throughout the course. It is worth noting that this textbook emphasizes randomization simulations as an approach to informal inference. This phenomenon warrants further analysis.

If construct-related validity is to be further explored, results suggest that the Bivariate Quantitative Data Topic Scale is not conceptually dissimilar enough to justify its use for gathering discriminant construct-related validity evidence. Correlation of RPASS scores with college entrance test scores, and particularly the SAT Mathematics scores, may be more appropriate for discriminant evidence.

The ongoing literature review should examine extensions of the RPASS content domain to include reasoning about inferential topics, such as confidence intervals and the relationship between power and statistical significance. RPASS could be further expanded by reiterating the methods used in this study. One could append 14 inference-related CAOS items to the RPASS-6 scale and conduct a preliminary reliability analysis to produce an expanded item set for administering to a new cross-institutional sample.

ACKNOWLEDGEMENTS

The author wishes to thank Robert delMas for feedback on an early draft of this paper, the five expert raters who contributed time and advice concerning the item content, and the six course instructors who allowed their classes to participate in this study.

REFERENCES

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

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.

delMas, R. C., Garfield, J. B., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28-58. Retrieved from http://www.stat.auckland.ac.nz/serj

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

Embretson, S. E. (2007). Construct validity: A universal validity system of just another test evaluation procedure? *Educational Researcher, 36*(8), 449-455.

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

Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher, 37*(2), 76-82.

Lane-Getaz, S. J. (2007a). Toward the development and validation of the Reasoning about *P*-values and Statistical Significance scale. In B. Phillips & L. Weldon (Eds.), *Proceedings of the International Statistical Institute / International Association of Statistical Education Satellite Conference on Assessing Student Learning in Statistics*, Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://www.swinburne.edu.au/lss/statistics/IASE/CD_Assessment/papers/IASE_SAT_07_Lane-Getaz.pdf

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

Lissitz, R. W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*(8), 463-469.

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

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

Other resources:
ARTIST website is available from https://app.gen.umn.edu/artist/
A sample of the RPASS instrument is online and available from:
https://www.surveymonkey.com/s.aspx?sm=OS_2baZG9m_2fi9bRKDYP3POBQ_3d_3d

SHARON LANE-GETAZ
Assistant Professor of Statistics and Education
Saint Olaf College
Northfield, MN 55057