International Journal of Teaching and Learning in Higher Education 2018, Volume 30, Number 1, 80-90
http://www.isetl.org/ijtlhe/ ISSN 1812-9129

Reliability and Validity of the Research Methods Skills Assessment

Tamarah Smith
Cabrini University

Samantha Smith
Temple University

The Research Methods Skills Assessment (RMSA) was created to measure psychology majors' statistics knowledge and skills. The American Psychological Association's Guidelines for the Undergraduate Major in Psychology (APA, 2007, 2013) served as a framework for development. Results from a Rasch analysis with data from n=330 undergraduates showed good fit for 16 of the 21 items created. Validity analysis showed that using the RMSA to measure students' statistics knowledge was an improvement over the assessment recommended by the APA Guidelines and was less time consuming. Together the findings provide a preliminary test that can be further developed to provide a tool for instructors and departments to use when assessing psychology students' statistics knowledge and skills. Future research should expand the RMSA to include items that measure knowledge with confidence intervals as the proposed items were removed due to poor fit. Future studies should also aim to replicate the current findings with additional samples of psychology students.

Statistics is an integral part of the curriculum in psychology departments with an estimated 86% of undergraduate psychology programs requiring statistics and related methods courses (Friedrich, Buday, & Kerr, 2000; McKelvie, 2000). In psychology, students often study statistics within the department, as opposed to the mathematics department, and the course(s) can be part of a sequence that integrates research methods (Friedrich et al., 2000). The inclusion of statistics in the psychology curriculum allows for the students to build knowledge and skills consistent with the way in which statistics are applied in psychology. As such, statistics courses taught in psychology may differ from the statistics courses taught outside of a psychology department. For example, probability is a foundational topic in statistics courses. However, the application of probability, such as using the *p*-value to interpret a statistical test, is an applied skill needed among psychology majors, but could be omitted from a statistics courses not taught within psychology. Other skills, such as distinguishing and applying different specific statistical tests (e.g., *t*-tests and ANOVAs) or using effect size, are also examples of important applied skills that need to be covered in the psychology statistics course. These skills are outlined in the American Psychological Association's *Guidelines for the Undergraduate Major in Psychology* (APA, 2007, 2013; hereafter *Guidelines*) along with other statistics objectives for psychology majors.

The *Guidelines* also emphasize the importance of assessing students' levels of mastery of these skills and provides suggested methods for such assessment. For the statistics related skills the recommendation is to review a research project using a rubric for scoring. Such projects, while containing a wealth of information regarding students' knowledge and skills, may be problematic in the assessment setting given the length of time needed to review a research project, the variability between raters, and the challenge that not all students will have the opportunity to complete such a project during their undergraduate career. Also, research projects, although using statistics, cover a broader range of skills such that using them to assess statistics skills would require teasing out items that are specifically related to statistics.

Other methods exist for measuring statistical knowledge and skills. Three commonly used tests are the Statistical Reasoning Assessment (SRA; Garfield, 2003), the Comprehensive Assessment of Outcomes for a first Statistics Course, the CAOS, (delMas, Garfield, Ooms, & Chance, 2007), and the Statistics Concept Inventory (SCI; Allen, 2006). In general, these tests assess knowledge expected of students after completing an introductory statistics course. The tests are consistent with the standards outlined in the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) endorsed by the American Statistical Association (Everson, Zieffler, & Garfield, 2008). They measure statistical reasoning and common misperceptions (i.e., SRA and CAOS) as well as items designed to measure general student knowledge (i.e., CAOS and SCI). There is ample support for these instruments (Allen, 2006; Garfield, 2003; delMas et al. 2007; Zieffler, Garfield, Alt, Dupuis, Holleque, & Chang, 2008). They were developed by experts in the statistics education field and were subjected to pilot testing, and the final instruments have yielded high reliability estimates and correlated with course outcomes. As such, they are recommended for measuring statistical reasoning and knowledge after completing an introductory statistics course.

The SRA, CAOS and SCI are targeted specifically at the experiences students will have in an introductory statistics course, but their questions are limited in their coverage of content that is emphasized in psychology statistics courses (Friedrich et al., 2000) and outlined in the APA *Guidelines*. Among these limitations are the skills outlined above: applying probability by using a *p*-

value, choosing appropriate statistics tests for different scenarios, and using effect sizes. We aimed to develop a test that would be more inclusive of knowledge and skills taught in the psychology statistics course, but less time consuming than reviewing a research project. We used the *Guidelines* to provide a framework from which to begin to develop this test.

Measuring Statistical Skills Amongst Psychology Students

In 2007 the APA established the first version of the *Guidelines*, which includes goals, curriculum and assessment planning aimed at helping the construction of undergraduate programs (APA, 2007). In 2013, the second version of the *Guidelines* was published in which the number of goals was reduced from ten to five and detailed indicators were specified within each goal for both the two- and four-year levels. The goals include a range of topics such as general knowledge in psychology, communication and professional skills, and critical and scientific thinking. Goal 2.0, entitled "Scientific Inquiry and Critical Thinking," includes 24 indicators in which research skills are listed. Among these are specific statistics skill indicators, including the ability to interpret basic descriptive statistics, to identify statistical and meaningful significance (i.e., *p*-value vs. effect size estimates) and to choose appropriate statistical tests. It is these specific statistics indicators that we aimed to measure with the RMSA. Although this first version of the test measured only content related to statistics, the instrument was titled the Research Methods Skills Assessment (RMSA) to be consistent with the broader skills identified in Goal 2.0 and allow for future versions of the test to include methods questions as well as statistics questions.

Goals of the Current Project

Recognizing the need for an efficient way to assess students' skills in statistics, we created a 21-items test with the goal of assessing the research skill indicators focused specifically on statistics in Goal 2.0 of the *Guidelines*. As such, the focus of the questions on the RMSA surrounded knowledge and skills in interpreting descriptives, significance, effect size, and confidence intervals, as well as choosing appropriate tests for various scenarios common in psychological research.

The purpose of this phase of our project was to preliminarily examine the quality of the RMSA items for measuring performance with the indicators related specifically to statistical practices outlined in the *Guidelines*. We examined this by first conducting a Rasch analysis to determine well-fitting items. Our intention was to establish reliable and valid questions for each of the indicators in the *Guidelines* pertaining to statistics content Following this, we tested incremental validity to determine the ability of the RMSA to measure statistics content knowledge and skills above and beyond the use of a scored research project, the suggested method in the *Guidelines*. We expected that the RMSA, given its more direct measure of statistical skills, would provide a better measure than the rubric score for this specific content. Although other tests exist to measure statistical skills, their lack of content specifically related to the APA Guidelines makes them fundamentally different from the RMSA items. As such, we chose not to compare the RMSA to other statistics assessments.

Method

Participants

The participants in this study were recruited from four different Northeast institutions (n=330; 73.9% female; mean age=22.63(6.09); 71.4% Caucasian, 12.4% Black/African American, 14.3% Hispanic/Latino, 1.9% Asian). The institutions ranged in size but were primarily private with the exception of one large public institution. The primary focus was to examine performance when administering the RMSA utilizing a paper-pencil "closed-book" classroom setting; however, two instructors requested an online version of the assessment for their students to complete in class.

Inclusion criteria for participating in the study dictated that students had to be enrolled in, or have already completed, a course taught within the psychology department that had the specific purpose of instructing them on statistics. The inclusion criteria was set this way given that psychology departments vary in the courses they provide to meet the need of statistics instruction (Friedrich et al., 2000). For example, some programs provide a stand-alone statistics course and others provide a course that integrates statistics and methods. The subject pool for this study included those completing stand-alone statistics courses (n=155) and those completing courses that integrated statistics and methods (n=175). This resulted in a sample (n=330) for the Rasch analysis that exceeded the general sample size guidelines of 200 participants (Thorpe & Favia, 2012).

Data from the stand-alone statistics course and integrated statistics and research methods courses were analyzed together. The groups were similar in terms of their college GPA, M=3.21(.47) vs. M=3.14(.42), $t(90)$=.75, p=.45, overall RMSA score M=.61(.15) vs. M=.58(.21), $t(95)$=.82, p=.412, and each subsection score of the RMSA (Section 1, M=.74(.22) vs. M=.76(.22); Section 2, M=.45(.33) vs. M=.43(.33); Section 3, M=.36(.40) vs. M=.34(.38); Section 4, M=.36(.30) vs. M=.32(.28); all ps>.05). The course grade between these

Table 1
Content, Taxonomy and Related APA Objectives for each RMSA Item

Content Area (Question #)	Objectives Measured (Taxonomies Used)	Objective (APA V2)	Total Items
Descriptives	Interpret basic statistical results (knowledge, application, analysis)	2.2e,E	9
Q1	Sample Size		1
Q2	Mean		1
Q3 & Q4	Standard Deviation		2
Q5 – Q9	z scores		5
Significance	Appropriate use of statistical vs. practical significance (knowledge, comprehension, application, evaluation)	2.2e,E	4
Q10 & Q11	p-values		2
Q12 & Q13	Cohen's d		2
Confidence Intervals	Identifying confidence intervals appropriate for sample means as opposed to individual scores; assumes knowledge of Central Limit Theorem (knowledge, application)	2.2e,E	2
Q14	Confidence for individual scores		1
Q15	Confidence intervals for sample means		1
Choosing Tests	Ability to choose an appropriate test for a given research hypothesis (knowledge, comprehension, application)	2.3B; 2.4d, D	6
Q16	One-sample t-test		1
Q17 & Q18	One-way and repeated measures ANOVA		2
Q19 & Q20	Pearson and Spearman correlations		2
Q21	Chi-square		1

Note: APA-V2 objectives reflect foundational level with lowercase letters and baccalaureate level with uppercase letters.

two groups showed a significant difference with the students in the stand-alone statistics course having a final grade that was higher than those in the integrated statistics and research methods course, M=.87(.07) vs. M=.84(.08), $t(102)=2.03$, p=.045. Given that these scores differed only by 3% and that all other measures of ability (i.e., GPA and RMSA scores) did not differ, we combined the data for the groups for all analyses.

A subsample provided course performance data (described below) to aid in the investigation of incremental validity. Access to course performance data was dependent on the instructor having assigned the specific coursework and their willingness to participate, resulting in a convenience sampling. Final statistics course grades were provided for n=116 students, and rubric scores for a final research projects were provided for n=28.

Measures

Research methods skill assessment. A full copy of the RMSA is provided in the Appendix. The RMSA included 21 questions to measure key statistical knowledge and application skills reflected in the objectives in the *Guidelines*. Table 1 provides the objective each item targets, as well as the foundational/2-year and baccalaureate/4-year level. The items were created to span four of the taxonomies proposed by Bloom (Aiken & Groth-Marnat, 2006; Bloom, 1956): general knowledge, comprehension, application, analysis, and evaluation (see Table 1). This resulted in items that go beyond the measurement of rote memory and measure more in-depth comprehension of concepts.

In the first section of the RMSA, a table of descriptive statistics was displayed, and a series of four questions was given to assess students' ability to "interpret basic statistical results" (objective 2.3a, APA, 2007, p. 13). Following this, four z-scores were provided, and students were to identify which of these were statistically significant at the given alpha level. To assess students' skills with objective 2.3b, to "distinguish between statistical significance and

practical significance" (APA, 2007, p. 13), the results of an independent samples *t*-test were provided, and four questions were asked to identify the students' knowledge of whether the results were statistically significant and/or meaningful and which values in the results (i.e., *p* or *d*) revealed this information. Other questions assessed the students' ability to choose an appropriate test for a given research hypothesis (objective 2.4e). True/false questions pertaining to confidence intervals (objective 2.3c) and derived from Garfield, delMas, and Chance (n.d.) were also included.

Completion time was typically 15 minutes with some students finishing more quickly and few taking up to 20 minutes. Directions for completion encouraged this quick pace by specifying, "If you know the answer to a question, please write it down. If you do not know an answer, that is okay, simply move on to the next question." Each question on the RMSA is recorded as correct (1) or incorrect/not answered (0), and the points are summed and averaged across the number of items. This provides a final score indicating the percentage of items correct.

Course Performance Measures. Course grades and rubric scores for a research project, based on a grading scale of 0-100%, were provided by faculty for a subset of students. Reviewing research projects using a rubric is the recommended form of assessment listed in the *Guidelines,* and as such we compared them to RMSA scores. The rubric scores differed from course grades in important ways. For example, the rubric scores are generated using items that intend to assess learning outcomes for a research project assigned to students. Course grades demonstrate the extent to which a student meets, not only the learning outcomes of one assignment, but many, and they can also include credit for course attendance, participation, and goals of assignments beyond that of the APA *Guidelines'* objectives. For example, assignment goals can include criteria such as properly using APA format or sentence structure.

The rubric scores provide an assessment of students' final research project paper. The paper included a literature review, hypothesis, methods development, data collection, analysis, and conclusion. The faculty who provided the rubric scores developed and used the rubric for departmental assessment to rate each pertinent step of the research process listed above. The rubric utilized a four-point scale that included "does not meet expectation," "partially meets expectation," "meets expectation," and "beyond expectation."

Procedure

All research was approved by the Institutional Review Boards at all schools from which students were sampled. The RMSA was distributed in students' classrooms. Students completed the RMSA in a paper-pencil format or online using a computer in the classroom. The faculty member or researchers monitored the completion of the RMSA.

Results

Rasch Analysis

To assess the appropriateness of the items on the RMSA to measure knowledge and skills with statistics, a Rasch analysis was used. Rasch analysis was developed to examine the individual items on a test that are scored dichotomously, such as the items on the RMSA (0=incorrect, 1=correct; see Thorpe & Favia, 2012). Rasch analysis is a latent variable model that assumes that an underlying latent trait (in this case statistics knowledge and skills) can be explained by responses on a series of measurable items (RMSA items). This approach allows researchers to develop tests that can measure intangible constructs and, as such, is commonly used in educational test construction. The results of a Rasch analysis provide information on the fit of items, that is, their ability to explain the underlying latent trait, as well as the difficulty and discrimination of an item. Difficulty refers to the ability index, or point at which a respondent has a 50% probability of answering an item correctly. Discrimination refers to the ability of an item to separate respondents between those scoring above and below that item's ability index. In our analysis, difficulty and discrimination items are reported as *z*-scores with negative values indicating easier/less discriminate items and positive values indicating more difficult/discriminate items.

We proceeded with the Rasch analysis by first assessing item fit using three criteria (Thorpe & Favia, 2012). After final items were determined, the Rasch model, excluding the removed items, was assessed for fit. Following this, the final model was used to generate difficulty and discrimination estimates for each item.

Item Analysis

Three criteria were used to examine items. Items with less than a moderate (r<0.30) point-biserial correlation (Nandakumar & Ackerman, 2004) were flagged; items that could not, at minimum, moderately discriminate ability (>0.65; Baker, 2001) were flagged; and, items that had a negative effect on the overall Cronbach alpha such that their removal increased the alpha value were also flagged.

Using these criteria, seven items were identified as potentially problematic: I1, I3, I4, I10, I14, I15, and I20 (see Table 2). All seven items made small, if any, negative impact on Cronbach alpha values. The largest increase in the alpha value possible, given removal of a

Table 2
Statistics for Removed and Final Items

Item	Point-Biserial	α if removed	% correct	Difficulty (z)	Discrimination (z)
I1	0.2921	0.7838	0.9534	-5.5023	3.8251
I2	0.3908	0.7794	0.9009	-6.394	4.7073
I3	0.1202	0.766	0.621	0.5137	-0.5164
I4	0.2928	0.75	0.7784	-3.5214	3.3623
I5	0.57	0.7671	0.7872	-9.1354	5.7155
I6	0.6106	0.7634	0.7522	-8.5864	4.7077
I7	0.5428	0.7693	0.7697	-8.665	5.2601
I8	0.5279	0.7706	0.7668	-8.392	6.0223
I9	0.4986	0.7748	0.5219	-0.9675	6.202
I10	0.3764	0.7459	0.6152	-2.9646	3.6427
I11	0.4783	0.7765	0.4082	2.9057	5.0695
I12	0.4159	0.7827	0.4665	1.1487	4.3565
I13	0.5066	0.7729	0.2915	5.2679	5.4785
I14	0.161	0.7606	0.2741	0.2127	0.2128
I15	0.272	0.755	0.449	1.3378	1.8781
I16	0.5054	0.7741	0.4606	1.314	5.5894
I17	0.4885	0.7743	0.2741	5.3062	5.2556
I18	0.5356	0.7711	0.4548	1.515	5.5417
I19	0.5101	0.773	0.3499	4.2676	5.3035
I20	0.425	0.7788	0.2216	5.5805	5.0846
I21	0.4741	0.7764	0.3528	3.9846	4.9344
All**	--	0.78	--	--	--

*Item removed
**Based on final items only
Note: Statistics for final items are derived from the model after removing poor performing items

Table 3
Model Fit Statistics

	Model 1 Rasch Model	Model 2 Unconstrained Discrimination	Model 3 Discrimination Varies Across Item
N			
Log Likelihood	-2825	-2819.39	-2782.32
AIC	5682.93	5672.78	5628.64
BIC	5744.34	5738.03	5751.48
χ^2 (df)	--	6.08(1)*	37.07(15)

*p<.05

single item (I3), was 0.01. As such, the point-biserial and discrimination values were used to evaluate item removal. Flagged items were sorted by their correlation values first and then discrimination scores. Point-biserial correlations provide a measurement of montonicity, an underlying assumption of the Rasch model that correct responses increase with ability (Thorpe & Favia, 2012). Discrimination is an important output of the Rasch model that allows researchers to identify items that separate respondents based on ability; however, it is not an underlying assumption of the model. As such, discrimination was considered the secondary priority when determining removal of items.

Five items had point-biserial correlations that were below the cutoff (.30) ranging from 0.12 - 0.29 (I3, I14, I15, I1, I4). Four of these items (I3, I14, I15, I4) also had low discrimination values (<.65) ranging from -0.06 to 0.53. Item one (I1) had a low correlation but high discrimination. The low correlation was likely due to small variability in response where 94% of respondents correctly answered the item. The high rate of correct responses is not surprising given the easy

nature of the question ("how many people were in this study?"). Given this, I1 was not pursued for removal.

One item (I10) was flagged based on a low discrimination value but had a moderate correlation (0.37), and one item (I20) was flagged due to a slight decrease in Cronbach alpha (-0.007) but had a moderate correlation (0.37) and discrimination value (0.94). The removal of I20 was not pursued given that it met the criteria for both the correlation and discrimination and would have a minimal change on the overall Chronbach alpha if removed.

The four items with both low correlations and low discrimination values were removed one at a time based on lowest to highest correlation values. Following this, I10, which had low discrimination but a moderate correlation, was removed. The model was recalibrated after each removal to ensure that additional items flagged for removal continued to meet the point-biserial criteria cutoff. The point-biserial correlation is dependent on the overall test score and as such can change after removal of an item.

After removing five items (I3, I14, I15, I4, I10) sequentially, the overall Cronbach alpha was improved from 0.75 to 0.78. All remaining 16 items had moderate or strong point-biserial correlations and moderate to high discrimination scores.

Model Fit

The fit of the model to the data was first tested for adherence to the assumption that there is a known discrimination parameter fixed at one (Thorpe & Favia, 2012). To check for the fit of the model under this assumption, 200 iterations of a Bootstrap goodness-of-fit test was performed using Pearson's chi-square in the ltm package of R (Rizopoulos, 2006). A non-significant goodness-of-fit test supports the assumption that the model fits the data with a parameter fixed at one; however, the result of the test based on the data was significant (p=.01), suggesting that the discrimination parameter was different from one. As such, a second unconstrained model, allowing a single discrimination parameter to vary, was tested for fit to the data. Models one and two were compared using -2LL with a chi-square test. A significant reduction in the -2LL indicates improvement in model fit. The -2LL for model two decreased (see Table 3) and this change was significant, χ^2=6.08(1), p<.001, indicating Model 2 was an improvement over Model 1.

We examined the two-parameter extension of the Rasch model that allows the discrimination parameter to vary for each item. We used -2LL to test if this third model provided better fit than model two. When allowing the discrimination parameter to vary across items, the fit was improved above using the single parameter, χ^2=37.07(15), p<.001.

Item Difficulty and Discrimination

Item difficulty and discrimination was examined using the final third model that utilized 16 items and a two-parameter extension of the Rasch model. Item difficulty z scores are presented in Table 2. Items earlier on the test were easier for students (I1-I8) with z-scores ranging from -9.1 to -5.5. The negative z-scores indicate that the ability index, or point at which a respondent has a 50% probability of answering correct, is skewed left for each of these items, indicating that they are easier items. The majority of students answered these items correctly (75% to 95% correct). The items were designed to measure knowledge and skills with interpreting descriptive statistics and z-scores.

More difficult items were present in the second half of the test (I9-I20). The z-scores for these items ranged with -.96 to 5.58 from 22% correct to 52% correct. This indicates that the ability index is skewed right for each item, illustrating that they are more difficult items. These items covered topics of statistical significance, effect size, and choice of the best test statistical test for a given scenario.

All items had very high discrimination, ranging from 3.82 to 6.20. This indicates that the items were able to separate respondents between those scoring above and below the ability index of a given item.

Validity

To establish content validity, two instructors with experience teaching statistics for psychology majors at both the undergraduate and graduate level for more than 30 years each reviewed and provided feedback on the items' coverage of the *Guidelines*' indicators. Final adjustments to wording and format were made given that feedback prior to distributing the RMSA to students.

To examine incremental validity, we utilized the technique of Schmidt and Hunter (1998) that compares the overall correlation generated when regressing scores from a current standard for measuring skills alone on course grade compared to the overall correlation when regression the current standard for measuring skills in combination with the RMSA. We used the *Guideline's* suggestion of a rubric score from the evaluation of a research project as the standard existing assessment from which to assess incremental validity. As such, we compared the overall R values obtained when regressing the rubric scores alone on course grade compared to the rubric scores and RMSA regressed on the course grade. When predicting course grade using the rubric scores alone, R^2=.11; when adding the RMSA to the model, the R^2 improved to .18. This indicates a 63% increase in validity and utility of using the RMSA over using the rubric score alone.

Discussion

The purpose of this study was to provide preliminary data for the development of the RMSA. We implemented the RMSA with a large sample of students majoring in psychology across different institutions. Our goal was two-fold: establish questions that pertain to each indicator related to statistics content in the *Guidelines* and determine if the RMSA increased validity when measuring statistical skills compared to using a rubric score alone.

The Rasch analysis provided good fit for a 16-item test. The items cover interpreting descriptive statistics, statistical significance, and effect size, as well as choosing tests appropriate for different scenarios. The items related to confidence intervals were problematic and removed from the test. Future studies should aim to create better fitting items that measure both knowledge and application of confidence intervals. One of the items removed (I3) asked students to assess the normality of data based on descriptive statistics provided in a table. We see this as a critical skill for students in psychology. As such, we would recommend further development of the RMSA to include this question with additional items to help decipher exactly how students are thinking the answer to this question through and better assess their level of skills. For example, do students answer this question incorrectly because they fail to recognize that they can compare the mean, median, and mode to determine skew? Or, are they unaware that the incongruity in these three values suggests skew in the data? Also, there is a need for items that assess students' abilities to interpret graphically displayed descriptives. In this study, descriptives were displayed only in table format.

Despite the need for growth with respect to items that measure confidence intervals and knowledge of descriptive statistics, the incremental validity analysis suggested that the RMSA provides a better indicator of students' statistics skills than rubric scores of a research project. These findings are consistent with our hypothesis. We anticipated that, given the direct emphasis of statistical skills by the current items on the RMSA, the test would provide a better measure of statistics skills than the research project. The research project did include statistical analysis; however, it also included other research knowledge of various designs (quasi, experimental, correlational), independent and dependent variables, and reliability and validity. It will be important, if additional items are created on the RMSA to assess such research skills, to analyze the incremental validity of methods related items on the RMSA compared to using the research project rubric scores.

The data in this study supports the RMSA as a good measure of statistical skills; however, research projects remain a more holistic approach that can allow instructors a context to open dialog with students regarding concepts that they may be struggling to master. If instructors use the RMSA in place of a more holistic project, we would encourage instructors to carefully review the results of the RMSA with students to allow for dialog and further exploration in areas where they may struggle.

Demographic comparisons of student's responses to the RMSA are needed. Upon determining a more comprehensive test that addresses the need for the items listed above, comparisons should be examined for potential gender, racial, and ethnic differences. Also important is the consideration of appropriate overall RMSA scores. Given that students may study statistics throughout a sequence of courses (e.g., Statistics I & II or Statistics followed by Research Methods/Capstone), comparison of item fit and discrimination should be conducted for these varying levels of completion with the course sequence.

The findings in this study, and especially those related to incremental validity, need to be replicated in a large diverse sample. Only one institution participating in this study assessed student research projects with a rubric, which limited the available data. The data collected in this study provided a foundation from which a relatively brief test can be developed for use in assessing psychology major's statistical knowledge and skills. The implication of such a test is vast, as it would allow for faculty to quickly assess their curriculum's effectiveness in meeting *Guidelines* indicators in this area. Unlike other goals and indicators in the *Guidelines*, Goal 2.0 has few standard tests listed for assessing the indicators in that goal. The RMSA could fill this gap by providing a quick and easily administered test to provide an indicator of students' statistical skills and knowledge.

References

Aiken, L., & Groth-Marnat, G. (2006). *Psychological testing and assessment.* Boston, MA: Pearson Education Group.

Allen, K. (2006). *The Statistics Concept Inventory: Development and analysis of a cognitive assessment instrument in statistics*. Retrieved from https://dx.doi.org/10.2139/ssrn.2130143

American Psychological Association, Task Force on Psychology Major Competencies. (2007). *APA guidelines for the undergraduate psyc hology major.* Washington, DC: Author.

American Psychological Association, APA Board of Educational Affairs Task Force on Psychology Major Competencies. (2013). *APA guidelines for the undergraduate psychology major*. Washington, DC: Author.

Baker, F. B. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Original work published in 1985. Retrieved from http://echo.edres.org:8080/irt/baker/

Bloom, B. S. (1956). *Taxonomy of education objectives, handbook 1: Cognitive domain.* New York, NY: Longman.

delMas, R. G., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*, 28-58.

Everson, M., Zieffler, A., & Garfield, J. (2008). Implementing new reform guidelines in teaching introductory college statistics courses. *Teaching Statistics*, *30*(3), 66–70. doi: 10.1111/j.1467-9639.2008.00331.x

Friedrich, J., Buday, E., & Kerr, D. (2000). Statistical training in psychology: A national survey and community on undergraduate programs. *Teaching of Psychology, 27*, 248-257.

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*, 22-38.

Garfield, J., delMas, B., & Chance, B. (n.d.) *Tools for teaching and assessing statistical inference.* Retrieved from https://apps3.cehd.umn.edu/artist/

McKelvie, S. J. (2000). Psychological testing in the undergraduate curriculum. *Canadian Psychology/Psychologie canadienne, 41*(3), 141-148. doi: 10.1037/h0086863

Nandakumar, R., & Ackerman, T. (2004). Test modeling. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology in the social sciences* (pp. 93-105). Thousand Oaks, CA: Sage.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analysis. *Journal of Statistical Software, 17*(5), 2-25.

Schmidt, F. L., & Hunter, J. (1998). The validity and utility of selection methods in personnel psychology: Practical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Thorpe, G. L., & Favia, A. (2012). Data analysis using Item Response Theory Methodology: An introduction to selected programs and applications. Unpublished manuscript, Department of Psychology, University of Maine, Orono, ME. Retrieved from http://digitalcommons.library.umaine.edu/psy_facpub/20

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. *Journal of Statistics Education, 16*(2) Retrieved from http://www.amstat.org/publications/jse/v16n2/zieffler.html.

TAMARAH SMITH, Ph.D.. is an assistant professor of psychology at Cabrini University. Her research focuses on the impact of statistics apprehensions, that is, the mindset, anxieties, attitudes, and other motivational factors that impact student learning in statistics. Her work is funded by the National Science Foundation and has been implemented in psychology classrooms, as well as STEM programs and teacher and faculty development programs.

SAMANTHA SMITH, M.S., completed her graduate studies in applied behavior analysis at Temple University and is a Board Certified Behavior Analyst. She is a member of PMABA, APBA, and ABAI. Her most recent publication evaluated the effectiveness of small group work with children diagnosed with autism.

Acknowledgements

This project was funded by a joint grant from Psi Chi and the Society for Teaching Psychology.

Appendix

Research Methods Skills Assessment

DIRECTIONS: The following questions ask varying questions pertaining to statistical concepts. Please read each question carefully and provide your best answer.

Questions 1 – 9 are based on the data in Table 1 which represents the results of a hypothetical administration of the SAT Quantitative Test.

Table 1: Results for SAT Quantitative Test

Statistic	Value
n	100
Mean	700
Median	500
Mode	500
Standard Deviation	300

Q1: How many people were sampled?

Q2: What was the mathematical average for the SAT score?

Q3*: Does it appear that the data are normally distributed? *(this item removed)*

Q4*: What measure listed in the table provides information about the spread of the data? *(this item removed)*

A series of tests were run on the SAT data presented in Table 1. First, z-scores were calculated for each student to determine any outliers. An outlier was defined as having a score more than two standard deviations from the mean.

Use the following information to determine which students are outliers. Circle the correct response on the right.

Q5: Student # 1 has a z-score of 1.64 Outlier Not an outlier
Q6: Student # 2 has a z-score of 2.35 Outlier Not an outlier
Q7: Student # 3 has a z-score of 0 Outlier Not an outlier
Q8: Student # 4 has a z-score of -2.21 Outlier Not an outlier

Q9: What score on the SAT Quantitative test did Student # 3 obtain?

The next four questions are based on the following hypothetical example: A clinical psychologist was interested in testing the effects of a new treatment for anxiety. He randomly assigned 30 subjects to two groups: Group A received the treatment, which lasted four weeks; Group B was assigned to a waiting list control. A standardized test of anxiety was given to all subjects at the end of the four weeks. This test has a maximum score of 30 where a higher score indicates a greater amount of anxiety. The psychologist obtained the following data:

Table 2: Results of the Experiment on Anxiety

	Mean	**Standard Deviation**	**Value of t-test**	**Value of p**	**Value of d**
GROUP A	17.80	4.23	2.24	.033	.85
GROUP B	20.93	3.39			

Q10*: Did the treatment significantly affect anxiety? *(this item removed)*

Q11: What statistic did you use to determine if the treatment affected anxiety?

Q12: Is this a meaningful difference?

Q13: What statistic did you use to determine if this is a meaningful difference?

Questions 14 and 15 are based on the following:

A 95% confidence interval is calculated for a set of weights and the resulting confidence interval is 42 to 48 pounds. Indicate whether the following two statements are true or false.

Q14*: A total of 95% of the individual weights are between 42 and 48 pounds. *(this item removed)* True False

Q15*: If 200 confidence intervals were generated using the same process, about 10 of the confidence intervals would not include the population mean (μ). *(this item removed)* True False

Questions 16 through 21 are based on the following:

Researchers at the National Institute of Health have developed a new depression scale. The test is scored on a scale of 0-50 with higher scores indicating higher levels of depression. The scale was given to a large national sample and it was determined that the mean of the test is 25 with a standard deviation of 5 (these values, therefore, are considered to be the population mean and standard deviation).

Please match the appropriate statistical test from the list below that would be used to answer each research question related to the scenario above.

a. One-way between subjects ANOVA
b. One-sample t-test
c. Spearman correlation *(this item removed)*
d. Repeated measures ANOVA
e. Pearson correlation
f. Chi-square test

Q16: A professor gives the test to his class of students and finds that the mean for this group of students is 35. Which test would he use to determine if his students are significantly more depressed than the population on which the test was normed? _____

Q17: The test was given to a sample of 15 women and 10 men. The mean for women was 24 and the mean for men was 21. Which test would he use to determine if the two means were significantly different from each other?

_____

Q18: A teacher of statistics gives the test before and after the midterm exam in her class. Which statistical test would be used to decide if there is a significant difference between these two means?

_____

Q19: Which test can be used to determine if there is a relationship between income (in dollars) and scores on the depression test? _____

Q20: What test can be used to determine if there is a relationship between ethnicity (African American, Caucasian, Hispanic) and scores on the depression test? _____

Q21: In reviewing the scoring protocols for the test, it was discovered that some of the test takers did not complete all of the items. To analyze this, the tests were coded as "completed" or "not completed". Which test would be used to determine if a higher percentage of males completed the test as compared to females?