Topical Article

SOCIETY FOR THE TEACHING OF PSYCHOLOGY

Teaching of Psychology
39(1) 17-23

DOI: 10.1177/0098628311430643
http://top.sagepub.com

SAGE

A Reliable and Valid Weighted Scoring Instrument for Use in Grading APA-Style Empirical Research Report

Kathleen Puglisi Greenberg[1]

Abstract

The scoring instrument described in this article is based on a deconstruction of the seven sections of an American Psychological Association (APA)-style empirical research report into a set of learning outcomes divided into content-, expression-, and format-related categories. A double-weighting scheme used to score the report yields a final grade that reflects the relative importance of outcomes in each category and the differential contribution of each section of the report to the report as a whole. The scores produced by the instrument are reliable between and within raters and significantly correlated with students' cumulative grade point averages. The author hopes the instrument can provide a useful framework for scoring any set of learning outcomes an instructor defines as the essential elements of an APA-style research report.

Keywords

grading, scoring, scoring rubrics

One of the first topics addressed in many introductory research methods texts is the notion of objectivity and the integral role it plays in the process of scientific inquiry (e.g., Beins, 2009; Christensen, Johnson, & Turner, 2011; Goodwin, 2008; Neuman, 2011). However, the importance of objectivity extends beyond the realm of science and into the domain of the research methods classroom itself when those of us who teach this course take on the task of grading the American Psychological Association (APA)-style research reports that we often require our students to write. In light of this interconnectedness between what we tell our students about the importance of objectivity and what we actually do when we grade their papers, it may be said that those of us who teach research methods have a special obligation to "practice what we preach" and grade our students' reports as fairly and objectively as possible.

One way to approach the issue of objectivity in grading is through the use of grading rubrics (Moskal, 2000; Peat, 2006). However, despite their widespread use in both K-12 and higher education settings, it is just recently that we have begun to see efforts to create rubrics for the specific purpose of grading students' empirical research reports. Gottfried, Vosmik, and Johnson (2009) described the process they used to develop a rubric of this kind. Despite their very thoughtful approach, they did not find high levels of interrater reliability in the scores the rubric produced. Nonetheless, they argued persuasively that rubric development should continue in light of the many benefits that rubrics provide (see Malini, 2010, for a review of rubric use in higher education).

More recently, there have been two reports of rubrics developed for grading APA-style research papers. Stellmack, Konheim-Kalkstein, Manor, Massey, and Schmitz (2009) created a rubric for grading APA-style introductions in which eight dimensions are rated on a 4-point scale of achievement.[1] Thaler, Kazemi, and Huscher (2009) designed one for grading an entire research report, using a 6-point Likert scale to gauge the achievement of 10 learning outcomes based directly on the guidelines provided in the fifth edition of the *Publication Manual of the American Psychological Association* (APA, 2001).[2] Measures of convergent validity were strong for both rubrics; however, only the Thaler et al. rubric was found to have significant interrater reliability (as measured by the correlation between the scores of different raters).

Comparison of the two rubrics suggests that one possible source of the discrepancy in findings is a difference in the complexity of the rating tasks. The rating scale used by Stellmack et al. (2009) consists of four qualitatively different levels of achievement, each defined by a set of specific criteria. In this case, the rater's task is one of deciding which level, or category,

[1] State University of New York College at Old Westbury, Old Westbury, NY, USA

Corresponding Author:
Kathleen P. Greenberg, Department of Psychology, State University of New York College at Old Westbury, P.O. Box 210, Old Westbury, NY, 11568
Email: greenbergk@oldwestbury.edu

best describes the extent to which an outcome has been achieved. When an achievement category is defined by one or two criteria, the task is fairly straightforward. For example, the ''sources'' learning outcome in the Stellmack et al. rubric is given the highest achievement rating based on (a) the number of sources referenced and (b) the extent to which these sources are relevant to the topic and cited in the paper. However, when an achievement category is based on several criteria—as is often the case for higher levels of proficiency—the rater's task is complicated by the need to acquire a holistic understanding of the set of criteria that defines each of the four levels of achievement, and also by the need to decide how many criteria within a category must be met for that level of achievement to be considered fair and appropriate. In the Thaler et al. (2009) rubric, the levels of the rating scale vary (essentially) on a single continuum of ''sufficiency.'' This use of a single continuum gives the scale quantitative properties that allow the rater to make what seems, on judgment, like a simpler decision about the extent to which the outcome has been achieved. With a simpler task, differences between raters based on proficiency with the rubric would be minimized, as would any differences associated with the interpretation of the scoring categories, particularly when a category is defined by multiple criteria that may have been only partially met.

If there is some merit to this conceptualization of the two types of rubrics, it suggests that the utility of a rubric as an objective and reliable means of scoring written work may be limited by the degree to which it is designed with largely qualitative or quantitative scaling properties. A rubric designed with scoring categories defined by an amalgam of learning criteria is inherently a qualitative measuring instrument; as such, it would be expected to yield scores that differ more between raters than would be the case with a rubric in which outcome achievement is rated on a single dimension.

In addition to issues with weak reliability, the ''qualitative'' type of rubric does not provide students with feedback as to which criteria within a scoring category they have met; nor, of course, does it provide information about the extent to which each of those criteria has been mastered. It is also more difficult for students to use this type of rubric proscriptively because the scoring categories (or at least the category that defines the highest level of achievement) must be deconstructed in order for students to be able to identify the criteria that will be used to evaluate their work.

Consideration of these limitations of qualitative rubrics as scoring devices led to the development of the scoring instrument described here. It is referred to as a scoring instrument, and not as a rubric, because each learning criterion is rated separately and on a single dimension of ''achievement.'' The main benefits of the instrument as a way of grading APA-style research reports are that (a) it makes explicit the learning criteria upon which the grade is based; (b) it delineates those criteria in a way that makes it easy for students to use them as a guide when preparing their reports; (c) it provides specific feedback regarding the extent to which each learning criterion has been achieved; (d) it yields subscores for each section of the report so students can see which sections were done well and which may need some work; (e) it produces scores that reflect the relative importance of content-, expression-, and format-related outcomes; (f) it results in a final grade that reflects the differential contribution of each section of the report to the report as a whole; and (g) it standardizes the scoring process so that the subscores and final grade it produces are reliable both between and within raters, whether the scoring is done by an experienced instructor or an advanced undergraduate student.

The purpose of this article is to describe the instrument and present data demonstrating significant interrater and intrarater reliability and significant criterion validity as well. I hope the description can serve as a framework that can be adapted to fit whatever set of learning outcomes an instructor defines as the essential elements of an APA-style research report.

The Scoring Instrument

The version of the instrument described here[3] consists of a set of 60 learning outcomes that collectively represent what I expect my students to include or show evidence of in the research reports they write. The outcomes associated with each section of the report (i.e., title page through references section) fall into three superordinate categories—Content, Expression, and Format—based on whether they respectively pertain to what was said, how it was said, or whether the text was formatted properly. (See Table 1 for a model of the introduction section of the instrument.) Each outcome is rated on a 4-point scale from 0 (*absent/not at all*) to 3 (*completely*) based on the extent to which the work provides evidence that the outcome has been achieved. The ratings can be written on a two-page score sheet given to students, or entered directly into an Excel spreadsheet used to calculate the scores. After the outcomes are rated, the ratings within each category are summed and converted to a percentage of possible points earned. This percentage is then weighted to reflect the relative importance of each category. For example, one might assign of weight of 0.5 for content-related outcomes, 0.3 for those that are expression-related, and 0.2 for those associated with formatting. The weighted outcomes are calculated separately for each section of the report, yielding a set of Section Subscores that provide students with a measure of their performance for each section. The final scoring is done by weighting each Subscore to reflect the impact it is to have on the final grade (with the body of the report presumably being given the most weight). Copies of the instrument, the two-page score sheet, and the Excel file are available at http://goo.gl/ckb5n.

Data Collection and Analysis

For this study, 20 papers were selected randomly from a set of 45 first-and-only drafts submitted by students in two research methods classes at a small public college in suburban New York. Each class was composed primarily of female students

Table 1. Model of the Introduction Section of the Scoring Instrument

III. Introduction → [Goal: provide empirical context and explain study's purpose]	
A. Content [points earned: ____ /out of 12 → percentage: ____%] × 50% =	
1. General orientation	
introduction of topic; definition of variables being studied	____
2. Empirical context	
well-researched, focused literature review using primary source information	____
3. Purpose	
accurate and clear statement of study's purpose	____
4. Hypothesis	
accurate statement of hypothesis(es) being tested, with obvious rationale	____
B. Expression [points earned: ____ /out of 9 → percentage: ____%] × 30% =	
1. Organization of ideas	
ideas/thoughts flow logically from paragraph to paragraph; paragraphs have a topic sentence and supporting details	____
2. Mechanics/clarity	
rules of grammar are followed; punctuation is correct and appropriate; words are spelled correctly	____
3. Voice	
no slang; no informal phrases (e.g., ended up); no connection with the reader (e.g., "you"), etc.	
C. Formatting [points earned: ____ /out of 6 → percentage: ____%] × 20% =	
1. Title	
study title, written in title case; no bold	____
2. Citations	
written in Modified Harvard (name, date) citation format	____

Note: Outcome achievement rating scale ranges from 0 (*absent/not at all*) to 3 (*completely*).

and students in their sophomore or junior year of study. Both were taught by the same instructor, and this instructor served as Rater 1 (R1). Another instructor who taught different sections of the same research methods courses as did R1 served as the second rater (R2). R1 and R2 scored the papers in July 2009. In May 2010, the papers were scored by a third rater (R3) who, at the time, was a graduating senior in the psychology department. The second scoring of the papers by R1 took place in December 2010, roughly 18 months after the original scoring. (For ease of communication, the first and second scoring of the papers by R1 will respectively be referred to as R1 Time-1 and R1 Time-2.) All three raters were familiar with the scoring instrument (R2 used it for two semesters, and R3 used it in an introductory and an advanced research methods course), so very little discussion about how to use it took place other than to briefly review the definitions of each outcome. A person not associated with the research recorded the names of the students whose papers were scored and blackened them out with a permanent marker before distributing the papers for scoring.

Interrater Reliability

Table 2 shows the levels of interrater reliability (as measured with Pearson's *r*) for the Final Score, Section Subscores, and mean[4] Category Scores for each pair of raters (R1-R2, R1-R3, and R2-R3). There is a high level of reliability overall, as indicated by average correlations of $r(18) = .84$ for R1-R2, $r(18) = .64$ for R1-R3, and $r(18) = .74$ for R2-R3. All but three individual coefficients are statistically significant ($p < .05$), and most are higher than .70. Two of the three that did not reach statistical significance were those for the Expression category scores for the two instructor-student pairs of raters (R1-R3 and R2-R3). The lack of statistical significance in this particular area suggests that the scoring of expression-related outcomes may be more subject to individual differences in judgment than those in other categories and that instructors and students may have different standards for what constitutes effective writing. Nonetheless, the high level of interrater reliability between each of the three pairs of raters indicates that the instrument promotes a high level of consistency in grading and that it does so regardless of whether the papers are scored by someone with expertise in the subject matter or someone whose exposure to it is relatively limited.

Table 2. Interrater Reliability for Final Scores, Section Subscores, and Mean Category Scores for Each Rater (R1, R2, R3) Pair

	R1-R2	R1-R3	R2-R3
Final Score	.88*	.72*	.83*
Title Page	.86*	.50**	.43
Abstract	.92*	.76*	.79*
Introduction	.82*	.53**	.79*
Method	.85*	.73*	.87*
Results	.73*	.65*	.62**
Discussion	.84*	.73*	.81*
References	.92*	.79*	.90*
Mean Category Scores			
Content	.82*	.68*	.83*
Expression	.75*	.42	.40
Formatting	.89*	.88*	.89*
MEAN	.84	.64	.74

*$p < .05$. **$p < .01$.

Table 3. Final Scores and Mean Category Scores by Rater (R1, R2, R3)

	R1 Time-1		R2		R3	
	M	*SD*	*M*	*SD*	*M*	*SD*
Final Score	77.9	13.0	76.0	12.7	81.0	12.1
Mean Category Scores						
Content	78.7	13.1	79.8	10.9	77.0	12.8
Expression	80.2	15.9	74.1	16.2	88.5*	14.3
Formatting	86.5	11.0	84.5	9.3	84.5	9.3

Note: Univariate test results for each dependent measure are as follows: Final Score: $F(2, 57) = .790$, $p > .05$; mean Content score: $F(2, 57) = .261$, $p > .05$; mean Expression score: $F(2, 57) = 4.298$, $p < .05$; mean Formatting score: $F(2,57) = .277$, $p > .05$.
*Significantly different ($p < .05$) from the mean for R2 as determined by Tukey post hoc comparisons; all remaining comparisons revealed no significant differences.

Intrarater Reliability

An intrarater reliability analysis for the Final Score, Section Subscores, and mean Category Scores (again, as measured with Pearson's *r*) showed that the scores generated by the instrument are highly consistent from Time-1 to Time-2. All coefficients are statistically significant ($p < .01$) and, with the exception of the Expression category scores, greater than or equal to $r(18) = .84$.

Criterion Validity

Criterion validity was assessed by measuring the relationship between the Final Scores and the cumulative grade point averages (GPAs) of 10 students whose GPA information could be obtained. Although this clearly is a very limited sample, the correlation between the two measures is significant ($p < .05$) for the two instructor raters, and marginally significant ($p = .084$) in the case of the student rater. Specifically, for R Time-1, $r(8) = .75$; for R1 Time-2, $r(8) = .76$; for R2, $r(8) = .80$; and for R3, $r(8) = .57$. Although the small sample makes it impossible to say with confidence that the scores generated by the instrument are predictive of overall academic performance, the findings nonetheless are consistent with this conclusion.

Subscores and Final Grades

Although the inter- and intrarater reliability data indicate that the instrument promotes relative consistency in scoring both across and within raters, the question remains as to whether the Final Scores generated by the instrument are of the same magnitude regardless of who has scored the paper. The results of a MANOVA conducted on the Final Scores and the mean Category Scores show that the Time-2 scores for R1 are not significantly different from the scores for R2 or R3 and, furthermore, that the scores for R2 and R3—with the exception of the mean Expression score—are not significantly different from each other. Specifically, the MANOVA provides evidence for an overall mean difference among the raters, $F(8, 108) = 4.00$, $p < .01$, that is traceable (with follow-up univariate analyses and Tukey post hoc comparisons) to a sole difference between the mean Expression Category Scores for Raters 2 and 3, $F(2, 57) = 4.30$, $p < .05$. All the remaining scores—including the Final Scores—are not significantly different across raters (see Table 3). Thus, it does appear that the instrument yields scores that are neither statistically nor practically different regardless of whether the scoring is done by the course instructor, another instructor, or an upper-division student who has used the instrument in class.

One final question is whether the scores generated by the instrument differ, on average, when the papers are graded by the same instructor at two different points in time. Table 4 shows the Final Score and mean Category Scores for the first rater at Time-1 and at Time-2. A series of paired *t* tests revealed that the scores at Time-1 are significantly lower than the scores at Time-2 (all ps $< .01$) by about 15 points (translating into a difference of about 1.5 letter grades on a standard 100% scale). Thus, although the instrument produces scores that are reliable over time (as indicated by the intrarater reliability data), it clearly does not serve to eliminate all sources of subjectivity in the grading process.

Two possible sources of such subjectivity are an instructor's expectations and standards for defining acceptable performance. The nature of the discrepancy between the scores at Time-1 and Time-2 suggests that R1 may have used a stricter standard when scoring the papers for the first time than when scoring them 18 months later. Given that R1 was the instructor for the course from which the papers had been selected, and that the papers were first scored not long after they had been submitted, it is possible that the stricter standard reflects a tendency to have high expectations when papers are initially graded because one so readily remembers what students were asked to do and what was done to help them do it. (How many times have we found ourselves perplexed and even a bit frustrated when a student provides an incorrect answer to a question on a topic that we know for certain was well-covered in class?) Thus, it follows that when there is no such recollection (as presumably would be the case 18 months after the papers were originally scored), expectations "return" to a level that is consistent with those of other instructors who lack this personal point of reference. Admittedly, this line of reasoning is highly speculative (and circular); however, it is supported by the data in Table 3, showing no differences between the

Table 4. Final Scores and Mean Category Scores for Rater 1 (R1) at Time-1 and Time-2

	R1 Time-1		R1 Time-2			
	M	*SD*	*M*	*SD*	*t*	*p Value*
Final Score	63.2	17.9	77.9	13.0	–8.808	.001
Mean Category Scores						
Content	67.3	15.0	78.7	13.1	–7.397	.001
Expression	59.5	26.7	80.2	15.9	–4.949	.001
Formatting	82.8	13.4	86.5	11.0	–3.764	.001

Time-2 scores of R1 and the scores obtained by the other two raters. Nonetheless, it will be important in future research on the instrument to explore the role of instructor expectations and standards as possible sources of score variability, even when attempts are made (as with the current instrument) to adopt a standardized approach to generating them.

Summary and Conclusions

The weighted scoring instrument described in this article was developed in an effort to promote objectivity in the scoring of APA-style research reports. An analysis of 20 papers scored by three raters—the course instructor, an instructor not associated with the course, and an upper-division psychology major—showed that the instrument has a high degree of inter- and intrarater reliability, and that the scores it yields are significantly correlated with students' cumulative GPA. Moreover, an interrater analysis of final scores suggests that the instrument does produce scores that are not significantly different across raters, (presumably) as long as the expectations and standards of those raters are aligned.

A comparison of the current scoring instrument with the rubrics designed by Stellmack et al. (2009) and Thaler et al. (2009) suggests that the reliability of a scoring rubric may, in part, be a function of whether the rating scale it uses is essentially qualitative or quantitative in nature. When outcome achievement is scored categorically on the basis of multiple criteria, the rating task is more qualitative than quantitative; when the learning criteria are scored separately, as is the case with the current instrument, the task is more quantitative. In this sense, the Stellmack et al. rubric, the Thaler et al. rubric, and the current scoring instrument represent three places along a continuum of measurement, with the Stellmack et al. rubric closest to the qualitative end, the current instrument closest to the quantitative end, and the Thaler et al. rubric at some point in between. Given that quantitative measures are more reliable than qualitative measures, it is perhaps not surprising to see gradations in reliability across the three measurement devices, with a lack of reliability reported by Stellmack et al., moderately strong reliability reported by Thaler et al., and strong reliability in the case of the current instrument.

If there is validity to this conceptualization of rubrics as scoring devices that vary along a qualitative-quantitative measurement continuum, it may not be appropriate to think of holistic and analytical rubrics as representing mutually exclusive categories. In the literature, rubrics are often characterized as being either holistic or analytical depending on whether they respectively yield a single score or multiple scores (e.g., Kan, 2007; Mertler, 2001). However, it could be argued that what differentiates one from the other is not whether the work is evaluated as a whole or in parts, but the extent to which individual learning criteria are combined to form more complex, or higher-order, learning outcomes. In a holistic rubric, there is one such outcome, defined on the basis of the entire set of established learning criteria. However, an analytical rubric can have any number of outcomes, from just two (if the set is described by only two dimensions) to as many outcomes as there are in the entire set (at which point, it may not be considered a rubric at all). Thus, in the characterization of a rubric, it may not be a matter of deciding whether it is holistic or analytical, but rather, if it is not holistic (i.e., qualitative), then how analytic (quantitative) is it?

What is relevant about this discussion is the suggestion that achieving a level of objectivity in grading students' APA-style research reports with the use of a rubric (or scoring instrument) may require a highly analytic—even molecular—approach in which the assessment criteria/learning outcomes are minimally dimensionalized, if at all. In addition to bringing a level of objectivity to the report grading process, a scoring instrument of this kind has the additional advantage of providing students with both detailed guidance and specific feedback to help them produce a high-quality report the first time they write one and an even better one the next time around. Tutors and instructors also benefit from the detailed feedback, as it serves to pinpoint areas of weakness and make the remediation process potentially more efficient and effective. Specific advantages offered by the current instrument include the "dimensional" feedback students receive in the form of the Section Subscores and the scores within the Content, Expression, and Format achievement categories, along with the assurance that the grade they earn is a fair and appropriate reflection of the relative importance of these three types of outcomes and the degree to which each section of the report contributes to the quality of the report as a whole.

Although the instrument was designed to assess the achievement of outcomes comprising this type of report, if the ability to write one is a key learning outcome of the research methods

course, the scores produced by the instrument also can be used to provide a holistic assessment of the degree to which this broader outcome has been achieved. This can be done by establishing categories of achievement each defined by a range of scores and calculating the percentage of students whose grades fall within each category. These data can then be used as the basis for a comparative assessment across instructors, within instructors, and across introductory and advanced research methods courses with an APA-style report course requirement.

Limitations

In addition to limitations associated with the small sample size, there are limitations specific to the instrument as well. Clearly, the content of an APA-style research report is indicative of a student's ability to demonstrate a number of higher-order skills, including those related to critical thinking, quantitative reasoning, and effective writing. As such, it may not be possible to reduce these complex skills into a number of discrete elements without losing the essence of what these skills represent (Sadler & Good, 2006). On the other hand, from a strictly behavioral point of view, it could be argued that it is possible to teach complex behavior by identifying and shaping its individual components, much like it is possible to teach someone to play the piano by first having them learn to play individual notes. Nonetheless, to the extent that a high-quality APA-style report, like a piano concerto, is more than the sum of its parts, it is acknowledged that the instrument may be better suited for scoring fundamental outcomes than for scoring outcomes that reflect higher-order, and more abstract, cognitive processes.

Another potential limitation concerns the issue of how time-consuming and possibly tedious it is to rate a large number of outcomes. In actuality, with a bit of practice (and presumably less than what is required to become proficient in the use of a qualitatively scored rubric), the scoring process becomes fairly automatic, partly because the criteria to be scored are precisely the elements the instructor is looking for and expects to see in the report. For example, in the current instrument, there are 15 outcomes associated with the formatting of the references section. They reflect every feature (however minor) of a properly formatted reference, from the use of hanging indents and double-spacing to the proper formatting of the authors' names. (Even the use of an ampersand between the penultimate and last author's name is included because experience has shown that it is not unusual to see students leave it out.) Despite the large number of outcomes, the rating task is fairly easy for someone familiar with APA-reference format, as it is possible to quickly scan the reference list and identify the elements that may be problematic. Nonetheless, another option is to reduce the number of criteria by combining them into a single outcome; however, because too much consolidation changes the nature of the instrument from one that is intended to be more quantitative than qualitative, a more consolidated instrument may yield scores that are not as reliable across and within raters, as is the case when the outcomes are narrowly defined. A final point on the issue of scoring concerns the calculation of the Final Score and the Section Subscores. Although this is accomplished easily with the use of an Excel program, it remains necessary to input each of the ratings. It must be said, however, that this too becomes fairly automatic with practice and that it has the additional advantage of providing a permanent record of students' scores and grades.

Green and Emerson (2007, p. 498) define a ''good grading system'' on the basis of eight criteria: It (a) is reliable across graders, (b) provides consistent feedback, (c) specifies the criteria upon which the grade is based, (d) differentiates among levels of achievement, (e) is stable over time, (f) aggregates performance data in a meaningful way, (g) acknowledges some level of subjectivity in the grading process, and (h) is practical to implement. It could be argued that the scoring instrument described here meets all of these to at least some degree; as such, it may offer promise as a way of maximizing objectivity in the grading of the APA-style reports that are often such a significant part of the curriculum of a research methods course.

Future research is planned to determine whether students who use the instrument write higher quality reports than students who do not. Also of interest is the question of whether having students calculate their own scores has any pedagogical value, as would be suggested by the work of Sadler and Good (2006) who, using a test-retest learning paradigm, found that self-grading produces gains in test performance larger than would be expected by simple repetition alone. The issue of instructor expectations as a source of subjectivity and score variability needs to be explored as well. Of course, it will also be important to formally assess students' opinions of the instrument to determine whether they find it helpful and how they think it might be improved. I welcome and encourage others who might wish to use a version of the instrument in their courses to further investigate its utility and pedagogical value.

Acknowledgments

Special thanks are due to Nancy Bray and Tressa Cincotta for scoring the papers, B. Runi Mukherji for her valuable comments on an earlier version of the scoring instrument, and Coleman Paul and William S. Altman for their thoughtful comments on an earlier draft of the article.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Notes

1. A copy of the Stellmack, Konheim-Kalkstein, Manor, Massey, and Schmitz (2009) rubric can be found at http://www.psych.umn.edu/psylabs/acoustic/rubrics.htm.
2. A copy of the Thaler, Kazemi, and Huscher (2009) rubric can be found at http://docs.google.com/Doc?id=df6b863n_0dw8dm3gj.
3. The instrument described here is based on the guidelines provided in the fifth edition of the *Publication Manual of the American Psychological Association* (American Psychological Association,

2001). The instrument I currently use has been modified to reflect the revisions in the sixth edition of the *Publication Manual* (American Psychological Association, 2010) and changes in my thinking about the importance and operationalization of certain outcomes.

4. The mean was computed by averaging the scores within each achievement category across all sections of the report.

References

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Beins, B. C. (2009). *Research methods: A tool for life* (2nd ed.). New York, NY: Pearson.

Christensen, L. B., Johnson, R. B., & Turner, L. A. (2011). *Research methods: Design, and analysis* (11th ed.). New York, NY: Pearson.

Goodwin, C. J. (2008). *Research in psychology: Methods and design* (6th ed.). Hoboken, NJ: Wiley.

Gottfried, G. M., Vosmik, J. F., & Johnson, K. E. (2009). *Assessing student learning: A collection of evaluation tools*. Retrieved from http://teachpsych.org/otrp/resources/gottfried09.pdf

Green, K. H., & Emerson, A. (2007). A new framework for grading. *Assessment and Evaluation in Higher Education*, *32*(4), 495–511. doi:10.1080/02602930600896571

Kan, A. (2007). An alternative method in the new educational program from the point of performance-based assessment: Rubric scoring scales. *Educational Sciences: Theory and Practice*, *7*(12) 144–152.

Malini, R. Y. (2010). A review of rubric use in higher education. *Assessment and Evaluation in Higher Education*, *35*(4), 435–448. doi:10.1080/02602930902862859

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research, and Evaluation*, *7*(25). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=25

Moskal, B. M. (2000). Scoring rubrics: What, when, and how? *Practical Assessment, Research and Evaluation*, *7*(3). Retrieved from http://pareonline.net/getvn.asp?v=7&n=3

Neuman, W. L. (2011). *Social research methods: Qualitative and quantitative approaches* (7th ed.). New York, NY: Pearson.

Peat, B. (2006). Integrating writing and research skills: Development and testing of a rubric to measure student outcomes. *Journal of Public Affairs Education*, *12*(3), 295–311.

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, *11*(1), 1–31. doi:10.1207/s15326977ea1101_1

Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P. (2009). An assessment of reliability and validity of a rubric for grading APA-style introductions. *Teaching of Psychology*, *36*, 102–107. doi:10.1080/00986280902739776

Thaler, N., Kazemi, E., & Huscher, C. (2009). Developing a rubric to assess student learning outcomes using a class assignment. *Teaching of Psychology*, *36*, 113–116. doi:10/1080/00986280902739305