An Assessment of Reliability and Validity of a Rubric for Grading APA-Style Introductions

Mark A. Stellmack
University of Minnesota

Yasmine L. Konheim-Kalkstein
Department of Educational Psychology, University of Minnesota

Julia E. Manor, Abigail R. Massey, and Julie Ann P. Schmitz
University of Minnesota

This article describes the empirical evaluation of the reliability and validity of a grading rubric for grading APA-style introductions of undergraduate students. Levels of interrater agreement and intrarater agreement were not extremely high but were similar to values reported in the literature for comparably structured rubrics. Rank-order correlations between graders who used the rubric and an experienced instructor who ranked the papers separately and holistically provided evidence for the rubric’s validity. Although this rubric has utility as an instructional tool, the data underscore the seemingly unavoidable subjectivity inherent in grading student writing. Instructors are cautioned that merely using an explicit, carefully developed rubric does not guarantee high reliability.

Rubrics are tools for evaluating and providing guidance for students’ writing. Andrade (2005) claimed that rubrics significantly enhance the learning process by providing both students and instructors with a clear understanding of the goals of the writing assignment and the scoring criteria. Stevens and Levi (2004) noted that rubrics facilitate timely and meaningful feedback to students. Peat (2006) suggested that, because of their explicitly defined criteria, rubrics lead to increased objectivity in the assessment of writing. Thus, different instructors might use a common rubric across courses and course sections to ensure consistent measurement of students’ performance. To this end, this article describes the assessment of the reliability and validity of a rubric that was designed for use in grading APA-style introductions in multiple sections of an introductory research methods course.

Rubrics used in many subject areas in higher education generally include two elements: (a) a statement of criteria to be evaluated, and (b) an appropriate and relevant scoring system (Peat, 2006). Rubrics can be classified as either holistic or analytic (Moskal, 2000). Holistic rubrics award a single score based on the student’s overall performance, whereas analytic rubrics give multiple scores along several dimensions. In analytic rubrics, the scores for each dimension can be summed for the final grade. Although an advantage of the holistic rubric is that papers can be scored quickly, the analytic rubric provides more detailed feedback for the student and increases consistency between graders (Zimmero, 2004).

Regardless of its format, when used as the basis of evaluating student performance, a rubric is a type of measurement instrument and, as such, it is important that the rubric exhibits reliability (i.e., consistency of scores across repeated measurements) and validity (i.e., the extent to which scores truly reflect the underlying variable of interest). Although reliability and validity have been noted as issues of concern in rubric development (Moskal & Leydens, 2000; Thaler, Kazemi, & Huscher, 2009), the reliability and validity of grading rubrics seldom have been assessed, most likely due to the effort and time commitment that is required to do so.

Two types of reliability that can be evaluated for a grading rubric are interrater agreement, the extent to
which scores resulting from use of the rubric are consistent across graders, and interrater reliability, the correlation between the scores of different graders (Tinsley & Weiss, 2000). When the task of grading students’ papers is divided among multiple graders, as is the case in our course and in many undergraduate courses at large universities, high interrater agreement is particularly important to achieve uniform grading across sections of the course. (Thaler et al., 2009, in contrast to this study, evaluated their rubric in terms of interrater reliability.) For the remainder of this article, when we refer to the general concept of reliability, we do so with respect to the computation of interrater agreement.

One way in which the interrater agreement of a rubric can be enhanced is through the careful description of criteria on which grades will be based. Without specific grading instructions, one grader might emphasize grammar, for example, whereas another might emphasize content. A well-developed rubric guides graders in placing the desired emphasis on specific, uniform criteria, so that the role of subjective opinions is minimized (Newell, Dahm, & Newell, 2002). Training of graders usually is necessary to maximize interrater agreement (Zimmaro, 2004). A reliable measuring instrument also is one that yields the same score in repeated measurements. With respect to a grading rubric, this task amounts to showing that the same grader would assign the same grade to a given paper if the grader were to grade the paper again. The repeatability of grades within graders is called intrarater agreement (Moskal & Leydens, 2000).

To say that a rubric exhibits validity means that it measures the underlying variable of interest. In this case, the variable of interest is “quality of psychological writing.” One way to demonstrate validity is to provide evidence that different measures of the same variable are correlated with one another (i.e., convergent validity; Crocker & Algina, 1986). In this case, we assessed validity by comparing measures of writing quality obtained with the rubric to judgments of writing quality made by an independent evaluator who has experience in teaching and grading student writing but who did not use the rubric.

The purpose of this study was to develop a rubric for grading student writing of the introduction section of an APA-style manuscript and to evaluate the rubric’s reliability and validity. We chose to focus on the introduction section because those involved in teaching our research methods course frequently identify it as the most difficult assignment in the course for instructors to grade and for students to write. In addition, we chose to analyze a rubric for only the introduction rather than for an entire manuscript in an attempt to simplify our task of maximizing interrater and intrarater agreement. In this article, we briefly describe the development of an analytic grading rubric, present data regarding interrater agreement and intrarater agreement in the use of the rubric, and assess the rubric’s validity.

Method

Writing Assignment and Graders

In a section of the research methods course at the University of Minnesota, students designed and conducted a research project of their own choosing in groups of approximately 4 students. The students were given a writing assignment in which they were to compose a well-written APA-style introduction, including a literature review of at least five peer-reviewed sources, in order to build an organized argument leading to a statement of the hypothesis of the students’ research project. Each student wrote his or her own paper (rather than a common group paper).

The researchers (the five authors) who developed and evaluated the rubric in this study included two instructors who had taught the research methods course during four previous semesters, a graduate student who had acted as a teaching assistant in the research methods course during one semester, and two advanced undergraduates who had taken the research methods course and who had been identified by their instructors as exceptional students in the course. In the research methods course at the University of Minnesota, student papers are usually graded by first- or second-year graduate student teaching assistants. As such, we felt that it was appropriate to include instructors as well as advanced undergraduate students who had demonstrated high levels of writing ability.

Rubric Development

In the early stages of developing a rubric, we began with an extremely detailed rubric that had been used in the past and that contained more than 30 specific statements that could be evaluated with respect to the student’s paper (e.g., “You did not indicate how your sources relate to your study” and “You did not integrate your sources into a coherent whole argument”). The specificity of the statements coupled with their inherent subjectivity led to frequent disagreements among graders as to whether a particular criterion was satisfied.
Table 1. Brief Descriptions of the Eight Dimensions in the Rubric

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Quality Assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>APA formatting</td>
<td>How well the paper follows the rules of APA formatting</td>
</tr>
<tr>
<td>Literature review and argument support</td>
<td>Description of previous literature and its application to the current study</td>
</tr>
<tr>
<td>Purpose of study</td>
<td>How well the student clarified why the topic should be researched, as well as how student's study follows from previous research</td>
</tr>
<tr>
<td>Study description and hypothesis</td>
<td>Student's brief description of what would be done in his or her study and the hypothesis</td>
</tr>
<tr>
<td>Overall organization and logical flow</td>
<td>How well the paper builds a coherent, smoothly flowing argument that culminates in the hypothesis</td>
</tr>
<tr>
<td>Sources</td>
<td>The quality of sources used, including whether the student used at least five peer-reviewed sources that are relevant to the topic of the study</td>
</tr>
<tr>
<td>Scientific writing style</td>
<td>Whether the paper uses appropriate scientific language and style</td>
</tr>
<tr>
<td>Composition/grammar/word choice</td>
<td>How well the paper is written in terms of grammar, sentence structure, and phrasing</td>
</tr>
</tbody>
</table>

As a result, we discovered, as have others (Peat, 2006; Stevens & Levi, 2004; Thaler et al., 2009), that a smaller number of criteria was more practical.

We reduced the grading criteria to eight broad dimensions of content that were of particular importance for this type of writing assignment (see Table 1), each containing four possible scoring levels (0–3 points). It is generally considered to be advantageous to have fewer scoring levels with meaningful distinctions than to have more scoring levels where it might be difficult to distinguish between categories (Moskal, 2000; Stevens & Levi, 2004). We decided that four scoring levels for each dimension gave the desired resolution to distinguish between levels of achievement. (See Table 2 for an example of the scoring levels for one dimension.)

Each week, during the course of about 10 weeks, we independently graded several introductions written by research methods students and convened to examine our scores. We discussed disagreements and made adjustments to the rubric in a way that we felt would prevent those disagreements in the future. Adjustments involved altering the wording that defined the dimensions and that distinguished between scoring levels within a dimension. We repeatedly refined the rubric in this way until we believed that we achieved maximum interrater agreement. (The complete, final version of the rubric that was evaluated in this article is available online at http://www.psych.umn.edu/psylabs/acoustic/rubrics.htm or via e-mail from the first author.)

Evaluating Intrarater Agreement

Forty papers were selected randomly from a pool of papers written by students in the research methods class. All identifying information was removed, and none of the researchers knew the final scores that were assigned to the papers in the course. Each researcher graded 24 of the 40 papers, such that three researchers graded each paper. Scores were then compared and interrater agreement was calculated based on the number of agreements for each dimension for each paper. Thus, there were 320 potential opportunities for agreement across all papers (40 papers × 8 dimensions).

Two estimates of agreement, conservative and liberal, were calculated. Agreement was defined conservatively as all scores assigned for a dimension being equal across the three graders of that paper. Agreement was defined liberally as all scores assigned for a dimension by the three graders being within 1 point of one another. These criteria are accepted in the measurement literature (Tinsley & Weiss, 2000) and have been applied in past studies of interrater agreement for grading rubrics (Newell et al., 2002). Agreement on total overall score out of 24 possible points (8 dimensions × 3 points maximum for each) for the 40 papers was also calculated and is described in the results.

Evaluating Interrater Agreement

Approximately 2 weeks after grading the papers, each grader regraded five of the papers that he or she originally graded. All five graders regraded different papers from one another, such that 25 different papers were regraded. To compare the first grading of each paper to the second grading, intrarater agreement was evaluated quantitatively in the same way as interrater agreement described earlier. A total of 25 papers were thus examined to assess intrarater agreement.
Table 2. An Example of the Grade Levels Within a Dimension of the Rubric

<table>
<thead>
<tr>
<th>Study description and hypothesis</th>
<th>3 Points</th>
<th>2 Points</th>
<th>1 Point</th>
<th>0 Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>• The paper gives a general description of what the research will entail (what will be done) without exhaustive methodological details.</td>
<td>The study description and hypothesis are present but one or both are unclear.</td>
<td>Either the study description or hypothesis is missing.</td>
<td>The study description and hypothesis are missing.</td>
<td></td>
</tr>
<tr>
<td>• It is clear what variables are going to be measured or compared (e.g., independent and dependent variables are identified).</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• The hypothesis is testable and contains terms that are operationally defined.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Evaluating Validity

To determine whether the rubric truly measured the intended variable “quality of psychological writing,” we examined whether scores assigned using the rubric aligned with an experienced psychology instructor’s ranking of those same papers. The average score (out of 24 possible points) was computed across the three graders for each of the 40 papers. Ten papers were chosen from approximately equally spaced intervals across the range of scores, based on the average score. An instructor of another methods-related psychology course at the University of Minnesota who requires APA-style writing was asked to rank order the 10 papers based on a holistic judgment of their content and quality with respect to what she would consider well-written APA-style introductions. The instructor was given only the description of the writing assignment that was given to the students, and the instructor was asked to apply whatever criteria she felt was appropriate in rank ordering the papers. The Spearman rank-order correlation coefficient between the independent judge’s rankings and the rankings based on the average scores of the papers using the rubric was computed.

Results

Interrater agreement by the liberal definition (i.e., with all graders scoring within 1 point for a given dimension) was 287/320 = .90 (Cohen’s kappa = .84). Interrater agreement defined conservatively (i.e., all scores assigned by the three graders within a dimension being equal to each other) was 119/320 = .37 (Cohen’s kappa = .33). Chance agreement given four scoring levels and three graders would be .34 and .06 by the liberal and conservative definitions, respectively. Newell et al. (2002) found comparable levels of agreement for three graders using a rubric for grading students’ solutions of chemical engineering problems, a task that was not writing based. The rubric developed by Newell et al. also had four scoring levels within each dimension. Newell et al. reported interrater agreement liberally defined of .93 and conservatively defined of .47.

Analysis of the total score assigned to each paper out of 24 points showed that all graders arrived at the same total score for 5% of the papers. The total scores on 75% of the papers were within 4 points of each other. The maximum range of total scores for a single paper was 8, observed for only one paper.

Intrarater agreement (i.e., when the same papers were graded a second time by the same graders) was 196/200 = .98 by the liberal definition and 156/200 = .78 by the conservative definition. Given that each paper was graded twice, chance agreement would be .62 and .25 by the liberal and conservative definitions, respectively.

In evaluating the validity of the rubric, the Spearman rank-order correlation coefficient between the independent judge’s rankings and those obtained with the rubric was .49. For comparison, the mean Spearman rank-order correlation coefficient between rankings of the five graders who scored papers with the rubric was .54. Thus, the correlation between rankings of those who used the rubric and the judge who applied her own independent criteria is comparable to the correlation between graders who used the rubric and the same explicitly stated criteria. This result provides support for the validity of the rubric.
Discussion

Although rubrics are used frequently in evaluating student writing, little research has focused on assessing the quality of rubrics as measurement instruments (i.e., their reliability and validity). Clearly, it is desirable to establish that a rubric is a valid measure of the variable that one is attempting to evaluate. The rubric’s reliability is of particular concern to instructors of large psychology classes that use multiple graders.

We undertook development of the present rubric with the optimistic view that near-perfect interrater agreement could be obtained through careful and diligent refinement of the rubric. That turned out not to be the case, with graders reaching perfect agreement only 37% of the time, and graders agreeing perfectly with themselves (intrarater agreement) only 78% of the time. Note that development and evaluation of the rubric spanned much of a semester, which probably represents a greater degree of effort than most instructors typically devote to rubric development. Although interrater agreement and intrarater agreement were not as high as we might have hoped, the rubric exhibited a reasonable degree of reliability in that three graders agreed with each other within 1 point 90% of the time. In addition, as indicated earlier, the introduction is frequently identified as the most problematic writing assignment for both writers and graders in the research methods course. Therefore, we might have expected it to be particularly difficult to attain high reliability for a rubric designed for use in grading introductions. As a whole, these data establish important benchmarks in the evaluation of grading rubrics, primarily because of the lack of such data.

Interpreting reliability and validity coefficients as “high” or “low” is a subjective proposition that is dependent on the way in which the information provided by the coefficients will be used and, as such, is of limited practical use. As long as the coefficients exceed chance, as in this study, reliability and validity exist to some degree. For our purposes, the results of this study suggest ways in which the numerical scores obtained using the rubric can be converted to letter grades. For example, given that numerical scores spanned the entire range of possible scores from 0 to 24, it is reasonable for the letter grade categories to be well distributed across the range of possible scores. Furthermore, given that the rubric used here resulted in graders assigning point totals within 4 points of one another on 75% of the papers, one might consider making the range of each letter grade for this writing assignment at least 4 points. For example, in our application of the grading rubric, after considering the overall quality of writing that existed in different ranges of scores, we intend to assign a grade of A for scores from 21 to 24 and a grade of F for scores from 0 to 3 (out of 24).

The fact that the graded introductions were based on student-designed studies might have influenced the reliability of the grading. In many research methods courses, students write their first APA-style manuscripts describing a research project that the instructor provides. It might be expected that an introduction describing a canned experiment would produce greater reliability because the ideal content and structure of the introduction could be more specifically defined for the graders.

The rubric used here was developed over the course of approximately 10 weeks during which the graders became increasingly familiar with its use and with each other’s grading tendencies. In effect, the development period also served as a training period for the graders. Even after such an extensive period of training, the interrater agreement remained imperfect. The training period in most real-world situations likely would amount to one or two sessions at the beginning of a semester. As a result, one might expect interrater agreement to be lower in practice than that obtained here, particularly when there is substantial turnover among graders from semester to semester as in our research methods course.

The results of this study underscore the inherent subjectivity of evaluating student writing. This subjectivity is problematic if one desires a grading rubric that can produce objective assessments across graders and course sections. An understanding of the reliability of a rubric can aid an instructor in converting scores obtained with the rubric into letter grades by revealing the potential variability associated with assigning a score to any particular paper. The fact that our rubric displayed unexpectedly low interrater and intrarater agreement will lead us to reconsider the way in which student writing is scored in our research methods course and the way in which those scores are incorporated into final grades. At the very least, we hope that this study will lead others to more rigorously assess grading rubrics as measurement instruments.

References


Notes

1. We thank Dr. Gail Peterson and Jamie Peterson for their assistance with this project. We also thank Dr. Randolph Smith and three anonymous reviewers for providing valuable feedback on an earlier version of this manuscript.

2. Yasmine L. Konheim-Kalkstein is now a Psychology faculty member at North Hennepin Community College.

3. The rubric that we evaluated in this article is available online at http://www.psych.umn.edu/psylabs/acoustic/rubrics.htm or via e-mail from the first author.

4. Send correspondence to Mark A. Stellmack, Department of Psychology, University of Minnesota, 75 East River Pkwy., Minneapolis, MN 55455; e-mail: stell006@umn.edu.
Copyright of Teaching of Psychology is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.