Practical Assessment, Research, and Evaluation

Volume 14 *Volume 14, 2009*

Article 12

2009

Examining Design and Inter-Rater Reliability of a Rubric Measuring Research Quality across Multiple Disciplines

Marilee J. Bresciani

Megan Oakleaf

Fred Kolkhorst

Camille Nebeker

Jessica Barlow

See next page for additional authors

Recommended Citation

Bresciani, Marilee J.; Oakleaf, Megan; Kolkhorst, Fred; Nebeker, Camille; Barlow, Jessica; Duncan, Kristin; and Hickmott, Jessica (2009) "Examining Design and Inter-Rater Reliability of a Rubric Measuring Research Quality across Multiple Disciplines," *Practical Assessment, Research, and Evaluation*: Vol. 14 , Article 12.
DOI: https://doi.org/10.7275/1w3h-7k62
Available at: https://scholarworks.umass.edu/pare/vol14/iss1/12

Examining Design and Inter-Rater Reliability of a Rubric Measuring Research Quality across Multiple Disciplines

Authors

Marilee J. Bresciani, Megan Oakleaf, Fred Kolkhorst, Camille Nebeker, Jessica Barlow, Kristin Duncan, and Jessica Hickmott

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 14, Number 12, May 2009 ISSN 1531-7714

Examining Design and Inter-Rater Reliability of a Rubric Measuring Research Quality across Multiple Disciplines

Marilee J. Bresciani, *San Diego State University*
Megan Oakleaf, *Syracuse University*
Fred Kolkhorst, Camille Nebeker, Jessica Barlow, Kristin Duncan, *San Diego State University*
Jessica Hickmott, *Weber State University*

The paper presents a rubric to help evaluate the quality of research projects. The rubric was applied in a competition across a variety of disciplines during a two-day research symposium at one institution in the southwest region of the United States of America. It was collaboratively designed by a faculty committee at the institution and was administered to 204 undergraduate, master, and doctoral oral presentations by approximately 167 different evaluators. No training or norming of the rubric was given to 147 of the evaluators prior to the competition. The findings of the inter-rater reliability analysis reveal substantial agreement among the judges, which contradicts literature describing the fact that formal norming must occur prior to seeing substantial levels of inter-rater reliability. By presenting the rubric along with the methodology used in its design and evaluation, it is hoped that others will find this to be a useful tool for evaluating documents and for teaching research methods.

This project stemmed from a very specific, practical need to have a single rubric for evaluating a range of research from a variety of disciples. A two-day research symposium would take place to provide undergraduate, masters, and doctoral students with an opportunity to showcase their research. The students participating in the symposium would represent a wide array of disciplines from the performing arts to bio-physics. In order to determine quality of the research, we needed to develop criteria to identify sound research methodology and to do so in a manner that provided constructive feedback. Thus, we sought to develop a rubric that would detail the specific criteria for identifying quality research methodology. The resultant rubric appears in the Appendices and the methodology behind that rubric is described in this paper.

Rubric Development

In every discipline, there is a body of literature that defines what quality of research is and how it can be taught, practiced, and identified in other peers' work as well as students' work. A twenty member multi-disciplinary team set out to develop a rubric that would reflect how a scholarly presentation might be delivered at a conference in varying students' respective fields. To develop the rubric, the team examined existing rubrics available internally such as those from bio-chemistry, postsecondary education, and psychology. The team also considered guidelines used for the review of manuscripts submitted for publication. In addition, performing arts faculty shared their criteria for evaluating performances, and humanities shared their varying perspectives of excellence in research within their multiple fields.

After extensive discussion between committee members and their departmental colleagues, consultation with the NRC website, and Creswell (2008), four main areas or constructs were developed for evaluation of the content (Organization, Originality, Significance, Discussion and Summary). A fifth area was developed for evaluation of the Delivery of the presentation. Within each of the rubric areas, five different ranked levels (numbered 1 through 5) were developed and described according to the worst possible and best possible scenarios. For example, for the category of Organization, a score of "1" had an associated descriptor of, "Audience could not understand/follow presentation because there was no sequence of information," while the descriptor for a score of "5" was, "Student presented information in logical order, which audience easily followed." Descriptors for intermediate scores of 2 through 4 then were compromises between the two extremes. In its earliest stages, the rubric was very much geared toward a science-based presentation, with associated expectations of a literature review, description of methods, explanation of the results, and conclusions. Of course, this could not apply to all the disciplines, such as some humanities and fine and performing arts fields. Thus, the rubric was subjected to numerous revisions of categories and descriptors to make the constructs more generalizable across the disciplines.

The final version of the rubric was the result of numerous reviews and revisions by the team of faculty from a range of academic disciplines that included the fine arts, education, humanities, engineering, psychology, and physical and life sciences. Regardless of the variances in research quality within each discipline, all team members could agree that a quality research presentation needed to include a strong rationale for and a clear purpose of the research or project, a brief description of the procedures, the primary results or outcome, and a discussion that demonstrated a thorough understanding of the conclusions or outcomes and their implications. Because judges of the student work would only be given a minute or so between student research presentations to complete the rubric, the number of performance levels in the rubric was limited to five. The final rubric categories or constructs that were selected reflected the agreed upon research quality across the various disciplines. Perhaps the strength of the rubric was that considerable time was devoted to ensuring that levels for each performance standard were unambiguous and that the vocabulary was such that it would be unmistakably understood by judges from all academic areas.

Application of the Rubric

The research methodology rubric that was used at the research symposium was designed by approximately 20 faculty who represented various disciplines across the university. Under the direction of two professors from different disciplines (i.e., Life Sciences and Linguistics), the feedback from these various faculty was combined into one rubric, which was posted on the university's research symposium website one week prior to the research competition. An education professor provided oversight to ensure that the rubric feedback would indeed remain instructive to students for their future improvement of their research presentations. See Appendix A for the rubric.

It is uncertain how many of the symposium's evaluators examined the rubric prior to the two-day research symposium; however, the only training that was given the evaluators was a written set of instructions for using the rubric that directed that their evaluations be "*fair, consistent, and that standards/expectations are appropriate for the academic level (i.e., undergraduate, master's, doctoral).*" See Appendix B for the instruction sheet. Even though instructions were provided to judges, it is unknown how many of the judges read the instructions.

The student research presentations were organized into 41 sessions, which grouped presentations of similar disciplines into one session. Evaluators were assigned sessions to judge student research based on their scheduled availability. In some cases, session planners were able to assign evaluators to judge research within their area of discipline expertise. However, in other cases faculty were asked to judge research that resided outside their own areas of discipline expertise due to evaluators' time constraints.

Students presented their research in 10 minute oral presentations within their assigned sessions, and the judges had an opportunity to ask clarifying questions. In most cases, only one clarifying question was asked. The judges then had approximately 5-10 minutes to complete the scoring of the student's research using the research methodology rubrics.

Reliability

There is "nearly universal" agreement that reliability is an important property in educational measurement (Colton, Gao, Harris, Kolen, Martinovich-Barhite, Wang, & Welch, 1997, p. 3). Many assessment methods require raters to judge or quantify some aspect of student behavior (Stemler, 2004), and Johnson, Penny, and Gordon (2000) challenge those who design and implement assessments to strive to achieve high levels of inter-rater reliability. Because the rubric in this study was used by a variety of disciplinary faculty to score student research quality, inter-rater reliability measures are worth investigating.

Raters are often used when student products or performances cannot be scored objectively as right or wrong but require a rating of degree. The use of raters results in the subjectivity that comes hand in hand with an interpretation of the product or performance (Stemler, 2004). In order to combat subjectivity and unfairness, many assessors develop rubrics to improve inter-rater reliability. Moskal and Leydens (2000) state that rubrics respond to concerns of subjectivity and unfairness by formalizing the criteria for scoring a student product or performance. They write, "The descriptions of the score levels are used to guide the evaluation process. Although scoring rubrics do not completely eliminate variations between raters, a well-designed scoring rubric can reduce the occurrence of these discrepancies." In fact, Colton et al. (1997, p. 9-10) state, "Generally it has been found that it is possible to define rubrics so well that raters can be trained to score reliably."

In cases where raters using rubrics produce inconsistent scores, several problems may exist. Inter-rater reliability scores can be influenced by several factors, including "the objectivity of the task/item/scoring, the difficulty of the task/item, the group homogeneity of the examinees/raters, speededness, number of tasks/items/raters, and the domain coverage" (Colton et al., 1997, p. 4). Rater inconsistency can also be due to inadequate training of raters, inadequate detail of rubrics, or "the inability of raters to internalize the rubrics" (Colton et al., 1997, p. 9). Furthermore, Wolfe, Koa, and Ranney (1998, p. 465) suggest that "scorers with different levels of scoring ability do not focus on different [product or performance] features, but probably have different levels of understanding about the scoring criteria."

Moskal and Leydens (2000) note that discussing rater scoring differences and making appropriate changes to rubrics is worthwhile and ultimately helps improve assessment reliability. A second way to improve rater consistency is to make rubric criteria and performance descriptions more specific. However, some researchers warn that including rigid definitions in rubrics might limit their generalizability (Colton et al., 1997). Finally, Stemler (2004) cautions assessors that inter-rater reliability is a function of an assessment situation, not of an assessment tool itself.

In any study, a researcher should decide which type of inter-rater reliability best suits the purpose of the assessment and then be sure to treat and summarize the assessment data in ways that are consistent with that type of inter-rater reliability. The three types of inter-rater reliability are consensus estimates, consistency estimates, and measurement estimates (Stemler, 2004). Consistency estimates, used in this study, are based on the assumption that "it is not really necessary for two judges to share a common meaning of the rating scale, so long as each judge is consistent in classifying the phenomenon according to his or her own definition of the scale" (Stemler, 2004). In this study, it was assumed that the faculty raters would be able to come to agreement on how to apply the rubric, even though the faculty all came from various disciplines and received no training on how to apply the rubric. The statistical analysis was designed to test these assumptions. The analysis further emphasized the case by variable data structure where each rater was an independent case and that each observation of the student's research was a variable (MacLennan, 1993).

Intra-class correlation coefficients, of which Cronbach's alpha is one form, are used as measures of consistency when evaluating multiple raters on ordered category scales. The interpretation of Cronbach's alpha is that it is the expected correlation between pairs of student scores if we were to choose two random samples of k judges and compute two different scores for each student each based on the k judges. Though some authors discourage the assignment of a strength of reliability scale to this statistic as it is dependent on the number of judges (Cortina, 1993), 0.7 is generally considered a satisfactory value of alpha (Nunnally, 1978). Once the rubric scores were entered into spreadsheets and loaded into SPSS, Cronbach's alpha was used to evaluate the inter-rater reliability of the judges within each judging session.

Findings

Correlations between the five rubric categories are given in Table 1. Category averages were computed for each student using the scores from all judges who evaluated the student. Then category averages were used to compute correlations. The correlations range from a low of 0.459 between Significance/Authenticity and Delivery to a high of 0.770 between Organization and Discussion/Summary.

Table 1. Rubric Category Correlations

	Organ.	Origin.	Signif.	Discuss.
Originality	0.612			
Significance	0.544	0.729		
Discussion	0.770	0.636	0.626	
Delivery	0.730	0.550	0.459	0.660

Relative frequencies of the scores by rubric category are given in Table 2. Half-point scores did not have descriptors on the rubric and were used less frequently than integer valued scores. The mode for all five categories was a score of 4. The means for the five categories are very similar ranging from a low of 3.61 for Originality to a high of 3.78 for Organization.

Table 2. Frequencies of Score Assignments

Score	Organ.	Origin	Signif.	Discuss	Delivery
1	1.2%	0.8%	0.3%	0.8%	0.9%
1.5	0.0%	0.0%	0.0%	0.1%	0.1%
2	5.8%	6.6%	4.0%	6.8%	6.1%
2.5	1.5%	1.5%	1.0%	2.3%	3.2%
3	19.4%	30.4%	26.6%	25.6%	22.7%
3.5	7.1%	8.9%	9.5%	7.4%	8.8%
4	43.9%	36.1%	40.9%	39.8%	33.8%
4.5	5.6%	5.0%	5.5%	5.1%	7.0%
5	15.6%	10.8%	12.3%	12.0%	17.4%

The structure of the data, with the number of students per session varying from six to eight, the number of judges per session varying from three to seven, and most judges judging only one session makes it difficult to compute an overall reliability measure. Therefore, measures were computed on a session-by-session basis with missing data handled by case-wise deletion of students who were not judged by all judges in the session. The median Cronbach's alpha value for the 41 sessions is 0.76. The mean alpha value was much lower due to negative alpha values in three sessions. These outlying sessions had only three or four judges, and in each case, we could identify one judge in the session whose exclusion would greatly improve alpha.

Discussion and Conclusions

The researchers were surprised to discover how substantial the level of agreement was on the application of the research rubric across disciplines and across levels of students, particularly when there was no prior training in regards to the understanding of the content of the rubric or the use of the rubric. As such, the researchers desired to examine whether those who participated in the creation of the rubric may have contributed indirectly to the "norming" of other evaluators' expectations (Bresciani, Zelna & Anderson, 2004).

This study did not gather the type of qualitative data that may shed light on the indirect influence these rubric team designers may have had on the informal norming process of the other evaluators. As such, the researchers are unable to determine the extent their influence had on others. However, it is most interesting to note that of the eight judges whose scores could have been removed to improve inter-rater reliability coefficients, three of the eight were on the team that formulated the rubric criteria. Furthermore, in the four sessions where it was unclear which rater's score to remove in order to increase inter-rater reliability coeffecients, a member of the rubric planning team was also judging in each of those four sessions.

When the researchers examined the number of times that evaluators judged sessions to identify whether patterns of agreement existed between number of times the rubric was applied by the same judges, no clear patterns were identified. In other words, increase or decrease in inter-rater reliability coefficients did not appear to be influenced by the number of times each judge evaluated a session.

Thus, the researchers conclude that given no training of evaluators on how to apply the rubrics and given no statistical evidence of patterns of influence by rubric design team members, the rubric design itself must have offered the needed clarity for evaluators to use it with substantial level of agreement. While level of agreement among raters is substantial, it is not clear whether those receiving the evaluators' feedback were in agreement of the marks they received. Additional research would be necessary in order to determine level of agreement among the recipients of the rubric scores.

In closing, the researchers highlight the remarkable level of agreement among judges in the use of one rubric with evaluated undergraduate, masters, and doctorate level research in multiple disciplines. The high level of agreement found in this study contradicts the notion that norming needs to play in order for agreement to be reached (Huba & Freed, 2000). However, given the inclusiveness and the extensiveness of the discussions to create the rubric, informal norming may have occurred and therefore influenced the statistical level of agreement.

REFERENCES

Bresciani, M. J., Zelna, C. L., & Anderson, J. A. (2004). Techniques for assessing student learning and development: A handbook for practitioners. Washington, DC: NASPA.

Colton, D. A., Gao, X., Harris, D. J., Kolen, M. J., Martinovich-Barhite, D., Wang, T., et al. (1997).

DOI: https://doi.org/10.7275/1w3h-7k62

Reliability Issues with Performance Assessments: A Collection of Papers. ACT Research Report Series 97-3.

Cortina J. M. (1993). What is coefficient alpha? An examination of theory and applications, *J App Psychol 78*,98–104.

Creswell, J. W. (2008). Educational research: Planning, conducting, and evaluating quantitative and qualitative research (3rd ed.). Upper Saddle River, NJ: Pearson Education.

Gwet, K. (2001). *Handbook of Inter-Rater Reliability: How to Estimate the Level of Agreement between Two or Multiple Raters.* Gaithersburg, MD: STATAXIS.

Huba, M. E., & Freed, J. E. (2000). Learner-centered assessment on college campuses: Shifting the focus from teaching to learning. Boston: Allyn and Bacon.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and inter-rater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, *13*(2), 121-138.

Landis, J. R., & Koch, G. G. (1977). The measure of observer agreement for categorical data. *Biometrics,* 33.

MacLennan, R. N. (1993). Interrater Reliability With SPSS for Windows 5.0. *The American Statistician, 47*(4), 292-296.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubrics development: Validity and reliability. *Practical Assessment, Research, and Evaluation, 7*(10). Available online: http://pareonline.net/getvn.asp?v=7&n=10

Nunnally, J. (1978). Psychometric theory. New York: McGraw-Hill.

SPSS FAQ. (2008) *UCLA: Academic Technology Services, Statistical Consulting Group,* Retrieved April 17, 2008, from http://www.ats.ucla.edu/stat/spss/faq/alpha.html

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating inter-rater reliability. *Practical Assessment, Research, and Evaluation*, *9*(4). Available online: http://pareonline.net/getvn.asp?v=9&n=4

Wolfe, E. W., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*(4).

Citation

Bresciani, Marilee J., Oakleaf, Megan, Kolkhorst, Fred, Nebeker, Camille, Barlow, Jessica, Duncan, Kristin, and Hickmott, Jessica (2009). Examining Design and Inter-Rater Reliability of a Rubric Measuring Research Quality across Multiple Disciplines. *Practical Assessment, Research & Evaluation*, 14(12). Available online: http://pareonline.net/getvn.asp?v=14&n=12.

Corresponding Author

Marilee J. Bresciani,
San Diego State University
3590 Camino del Rio Norte
San Diego, CA 92108
Phone: 619-594-8318
Fax: 619-594-4208
http://interwork.sdsu.edu/student_affairs/

Marilee.Bresciani [at] mail.sdsu.edu

Appendix A

2008 SDSU Student Research Symposium Oral Presentation Rubric

<table>
<tr><th colspan="4">Undergraduate | Master's | Doctoral</th><th>Judge Code:</th><th colspan="2">Abstract number:</th></tr>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>Score</th></tr>
<tr><td>1 Organization</td><td>Lacked sequence in presentation or missing information. Presented too little/much material for allotted time.</td><td>Poor sequence or illogical presentation of information. Some relevant information not presented. Presentation not well timed.</td><td>Some information presented out of sequence. Had some pacing and timing problems.</td><td>Information presented nearly complete and relevant and presented in logical sequence. Pace and timing appropriate.</td><td>Information presented was complete and in logical order. Easy to follow. Very well-timed and well-paced.</td><td></td></tr>
<tr><td>2 Originality</td><td>Problem/purpose lacked creativity or not new. Duplication of previous work. Design/approach inappropriate and/or ignored previous well-established work in area.</td><td>Problem/purpose limited in originality and creativity. Design/approach only marginally appropriate or innovative.</td><td>Problem/purpose moderately original or creative. Design/approach moderately appropriate or innovative.</td><td>Problem/purpose fairly original or creative. Design/approach appropriate or innovative.</td><td>Problem/purpose very creative or original with new and innovative ideas. Explored original topic and discovered new outcomes. Design/approach introduced new or expanded on established ideas.</td><td></td></tr>
<tr><td>3 Significance/ Authenticity</td><td>Project has no significance/authenticity to field and will make no contribution.</td><td>Project has little relevance or significance/authenticity to field and will make little contribution</td><td>Project only moderate relevance or significance/authenticity to field and will make a nominal contribution.</td><td>Project has fair relevance or significance/authenticity to field and will make good contribution.</td><td>Project extremely relevant or has significant importance/authenticity to field and will make an important contribution.</td><td></td></tr>
<tr><td>4 Discussion and summary</td><td>Little or no discussion of project findings/outcomes. Displayed poor grasp of material. Conclusion/summary not supported by findings/outcomes.</td><td>Major topics or concepts inaccurately described. Considerable relevant discussion missing. Conclusions/summary not entirely supported by findings/outcomes.</td><td>Few inaccuracies and omissions. Conclusions/summary generally supported by findings/outcomes.</td><td>Discussion sufficient and with few errors. Greater foundation needed from past work in area. Conclusions/summary based on outcomes and appropriate, included no recommendations.</td><td>Discussion was superior, accurate, engaging, and thought-provoking. Conclusions/summaries and recommendations appropriate and clearly based on outcomes.</td><td></td></tr>
<tr><td>5 Delivery</td><td>Presenter unsettled, uninterested, and unenthused. Presentation was read. Inappropriate voice mannerisms, body language, and poor communication skills. Poor quality of slides/presentation materials; did not enhance presentation/performance.</td><td>Presenter unenthused, monotonous and relied extensively on notes. Voice mannerisms, body language, and communication skills sometimes inappropriate. Poor quality of slides/presentation material; poor enhancement of presentation/performance.</td><td>Displayed interest and enthusiasm. Read small parts of material. Occasionally struggled to find words. Generally appropriate voice mannerisms, body language, and communication skills. Moderate quality of slides/presentation materials.</td><td>Relied little on notes. Displayed interest and enthusiasm. Good voice mannerisms, body language, and communication skills. Good quality of slides/presentation materials; enhanced presentation/performance.</td><td>Relied little on notes. Expressed ideas fluently in own words. Genuinely interested and enthusiastic. Exceptional voice mannerisms, body language, and communication skills. Exceptional slides/presentation quality materials; greatly enhanced presentation/performance.</td><td></td></tr>
<tr><td colspan="7">Comments</td></tr>
</table>

Appendix B

Instructions for Using Oral Presentation Rubric

Judge Code: As a judge you have an assigned identification code. Record this code number on every presentation you evaluate. If you do not remember your number, ask the Student Ambassador in your session to locate your number.

Abstract Number: Every student is assigned an abstract number. Record the abstract number, which identifies the specific presentation, on every presentation you evaluate. This number can be found in the SRS program and will be on the abstracts sent to you in advance of the SRS.

Scoring: Presentations are judged on a 5-point scale. Record a score (1 through 5) for each of the five categories. You can assign scores ONLY in full or half-point increments. Any score of less than a half-point increment will be given the next lowest half-point score.

Scoring Guidance: It is extremely important that your evaluations are fair, consistent, and that standards/expectations are appropriate for the academic level (i.e., undergraduate, master's, doctoral). Be very discriminating with awarding a 5. This score should be reserved for only truly exceptional presentations.

Please add constructive comments in the designated box following each presentation. Scoring sheets will be returned to students after the symposium to help them in future presentations.

After each presentation, the Student Ambassador will collect your sheets for tabulation of the scores.

Scoring Categories: Standards and expectations for the five rubric categories are described below:

- ***Organization*** refers to the quality and completeness of information presented. Students are allowed only 10 minutes to deliver their presentation (and 5 minutes for questions), thus only the most relevant information should be presented. Moreover, the presentation should be well-paced and make use of the entire time allotment.
- ***Originality*** refers to the research problem or project purpose and to the design or approach. The problem/purpose should be original and imaginative and display independent thought and/or creative skill and imagination. The design/approach should expand on established ideas or introduce new ideas.
- ***Significance/Authenticity*** refers to the importance or worth of the project purpose. This category addresses the question of whether it was a meaningful project to conduct and would make a worthwhile contribution to the discipline. Significance, or authenticity, for the performing arts is realized though a meaningful relationship of a particular work to the re-creative artist.
- ***Discussion and Conclusion*** is the discourse of the findings, the experiential process or outcomes of the project. It should focus on the project and explain the reasoning for the selection of the process or outcomes. It should also describe the process/outcomes in the context of past findings or performance(s) or works and be accurate, engaging, and though-provoking. A hypothesis-based project should include a conclusion based on the results. Non-hypothesis-based projects should include a summary that condenses the findings, process or outcomes.
- ***Delivery*** refers to the style of the presenter and the quality of the presentation. The presentation should be given in a manner (e.g., voice mannerisms, body language, communication skills) that shows the enthusiasm, skill, and interest of the student and be made with little reliance on notes. The delivery also considers the quality of slides or other presentation materials, which should enhance the presentation/performance.