Journal of Surgical Research **142,** 304–307 (2007)
doi:10.1016/j.jss.2007.03.060

Developing a Scoring Rubric for Resident Research Presentations: A Pilot Study

Joseph L. Musial, Ph.D.,[1] Ilan S. Rubinfeld, M.D., Alton O. Parker, M.D., Craig A. Reickert, M.D., Sarah A. Adams, B.A., Sishir Rao, B.A., and Alexander D. Shepard, M.D.

Department of Surgery, Henry Ford Hospital, Detroit, Michigan

Submitted for publication January 8, 2007

Background. **A requirement for all Accreditation Council for Graduate Medical Education (ACGME) approved residencies is the provision of "an opportunity for residents to participate in research." To comply with this requirement, most training programs encourage their residents to conduct research and to report their results. Few guidelines exist, however, for assessing the efficacy of the presentations. The goal of this pilot study was to develop a valid, one-page scoring rubric to be used during oral resident research presentations. Such a scoring rubric will facilitate acceptable agreement among faculty raters.**

Methods. **Content validity was addressed by adhering to the Standards for Educational and Psychological Testing. A one-page, five-domain, behaviorally worded scoring rubric was developed. Inter-rater reliability was derived and three ACGME General Competencies were also addressed within the rubric.**

Results. **The initial scoring rubric was tested with 11 resident oral presentations. The inter-rater reliability was 0.56 using Cronbach's alpha. The rubric was modified and the scale restricted to a 3-point scale. It was then tested with 17 additional presentations, which were independently rated by two general surgery faculty members. Cronbach's Alpha increased to 0.61.**

Conclusions. **An objective method to evaluate a resident's oral research presentation has been successfully piloted. This content valid rubric possesses *good* inter-rater reliability according to established guidelines. Clearly defined behaviors have been outlined within the rubric. Program directors will have psychometrically sound evidence for the ACGME. Future research will address generalizability and concurrent validity using other types of resident assessment data.**

Key Words: **rubric; resident assessment; ACGME competencies; inter-rater reliability.**

INTRODUCTION

The Accreditation Council for Graduate Medical Education (ACGME) has established a set of *Key Considerations for Selecting Assessment Instruments and Implementing Assessment Systems*. These considerations state that an elected assessment approach should provide valid and reliable data, the approach should be feasible, provide useful information, and possess generalizability or external validity [1]. Among the six General Competencies, Interpersonal and Communication Skills and Professionalism are perhaps the most challenging competencies to evaluate. There is a paucity of standardized methods by which to assess resident performance in these two areas. According to the ACGME Tool Box for Assessments, raters who judge resident performance in ACGME competency areas retrospectively often face trouble with reliability. Some concerns highlighted by the toolbox include untrained raters scoring subjectively irrespective of resident performance, as well as rater's evaluating with too severe or lenient ratings. In addition, studies have been mixed about the ability to discriminate between different individuals, as well as the reliability of ratings across physician evaluators [2].

Regardless of intent, every assessment type has both strengths and weaknesses. In general, there are four types of rating errors: (1) errors of severity; (2) errors of leniency; (3) errors of central tendency, or avoiding extreme judgments; and (4) the halo effect, which involves giving biased ratings due to an overall impression of the person [3]. Classical measurement theory

[1] To whom correspondence and reprint requests should be addressed at Department of Surgery, Henry Ford Health System, 2799 West Grand Blvd., CFP-106, Detroit, MI, 48202. E-mail: JMUSIAL1@hfhs.org.

0022-4804/07 $32.00

identifies two forms of error associated with measurement: systematic and unsystematic. Systematic errors, also referred to as constant errors, occur during repeated measurements and unsystematic errors, or random errors, including those that vary unpredictably during repeated measurements. An example of a systematic error would occur if an expert rater used the wrong metric score consistently across all measurements. This would impact validity. An example of an unsystematic error might include inconsistent measurements as a result of distractions that occur from one measurement to another. This would negatively impact score reliability [4].

Despite the numerous challenges facing faculty assessments of resident performance, scoring rubrics are an ideal method by which to make assessments more objective. Scoring rubrics are a uniform method, with precise criteria by which to rate a students performance or artifact. Ideally, two independent raters should be able to derive the same score. Recently, scoring rubrics have begun to appear in graduate medical education settings; they have been used to evaluate both norm-referenced, peer-benchmarked global evaluations [5] as well as resident presentations at the University at Buffalo [6]. In addition, many physician evaluators are accustomed to scoring rubrics from clinical practice, i.e., APGAR, Sepsis, Trauma Score, Glasgow Comma Score, to name a few.

METHODS

A one-page scoring rubric was developed and piloted to assess a resident's oral research presentation. The aim of the rubric was to provide a transparent, objective, feasible, and user-friendly method that would facilitate acceptable agreement among faculty raters who judge resident research presentations. The scoring rubric adhered to the *Standards for Educational and Psychological Testing* that was jointly developed by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [7]. The research plan had IRB approval. The reliability analysis was performed using the statistical software SPSS, version 14.0.

According to *Standard 1.7*, when a validation rests on the opinion of expert judges or raters, their qualifications and experience as judges should be described. In addition, if information is exchanged, the process should be described [7].

A multi-disciplinary team was assembled to address content validity. The team consisted of a surgical educator who holds a doctorate in educational evaluation and research, 10 years of GME experience, and two years of psychometric experience (J.L.M.), a vascular surgeon who is a program director in general surgery (A.D.S.), two associate program directors in general surgery (I.S.R. and C.A.R.), a PGY-3 surgery resident (A.O.P.), a first year medical student (S.R.), and a senior undergraduate summer intern (S.A.A.). This multidisciplinary team created the narrative descriptions by referring to the descriptors used in the ACGME General Competencies. All of the domains and narrative behaviors were agreed upon by consensus. The rubric went through five iterations during which the rating scale was reduced from a 5-point to a 3-point scale.

Two expert raters were used to establish the inter-rater reliability. Rater one (I.S.R.) is nine years post-residency, board certified in general surgery and critical care, and is an associate program director who has completed over 1000 resident evaluations. Rater two (C.A.R.) is 10 years post-residency, board certified in general surgery, colorectal surgery, and critical care, serves as an associate program director, and has completed over 1000 resident evaluations. Both raters have had significant experience rating oral resident research presentations.

According to *Standard 2.3*, when interpretation emphasizes differences between two rating scores, reliability data should be provided [7].

Cronbach's alpha, or coefficient alpha, was selected for the reliability analysis. Cronbach's alpha, an estimate of internal consistency, is very practical because it requires a single measure, as opposed to repeated measures [8]. In contrast, the test–retest reliability requires two measurements of the same performance [9]. It is also preferable to use two independent raters rather than one rater as single raters lead to lower reliability coefficients [10].

According to *Standard 3.7*, the procedures that were used to develop and review the items should be documented [7].

Initially, both raters had practiced rating presentations using the scoring rubric on resident presentations that were not part of the formal analysis. Both raters discussed how they derived their scores and then agreed upon what constituted their assigned score. Next, a total of 11 oral presentations were independently rated at an institutional resident research forum. The scoring rubric then went through four additional iterations and was reduced from a 5-point to a 3-point scale. The final scoring rubric appears in Fig. 1. A total of 17 oral presentations were then assessed independently by both raters during a regional surgical research conference.

According to *Standard 3.27,* if a test is intended for research purposes only, statement to this effect should be promptly stated [7].

Although we developed an assessment but not a test, the intent of this study was to validate a user-friendly scoring rubric to evaluate resident research presentations. The research proposal clearly states that a pilot approach was used. No norm reference testing occurred nor was any external validity testing conducted at that time.

RESULTS

According to the data from the baseline ($n = 11$) ratings, the mean presentation score was 16.6, with a standard deviation of 3.5 and a standard error of 0.75, based upon a 5-point scale with a maximum possible score of 25 points. The inter-rater reliability derived from Cronbach's alpha was 0.56. The data derived with the revised ($n = 17$) ratings had a mean score of 8.5, a standard deviation of 1.6, and a standard error of 0.28, based upon the final 3-point scale with a maximum score of 15 points. The inter-rater reliability based upon Cronbach's alpha increased to 0.61, which can be interpreted as possessing *good* reliability [11].

DISCUSSION

The preliminary results from this pilot study have lead to the development of a content valid, user-friendly, transparent, scoring rubric that can be used during resident research oral presentations. Three of the ACGME General Competencies, including communication skills, practice-based learning and improvement, as well as professionalism have been assessed

Resident ID# _____

	Practice Based Learning and Improvement			Communications	Professionalism
	Literature Search	**Hypothesis/Aim/Objective**	**Statistical Analysis**	**Presentation Skills**	**Personal Conduct**
1	• No evidence of literature citations included in the talk or slides.	• The hypothesis, aim, or objective, was either not stated or not clearly stated.	• No explanation of the employed statistics.	• Unable to discuss findings in a scientific fashion. • Lacked a logical order. • Materials lack polish. • Presentation lacks polish.	• Arrived late. • Not in professional attire. • Behavior not appropriate for context. • Not prepared.
2	• Basic literature search. • No mention of seminal or landmark studies. • Sufficient number of citations.	• Reasonable explanation of the hypothesis, aim, or objective.	• Reasonable explanation of employed statistics for presenter's level of training.	• Satisfactory presentation skills. • Responses to questions were adequate. • Minor pauses.	• Reasonably prepared. • Appropriate attire. • Behavior appropriate for context.
3	• Thorough literature search. • May have used key words. • May have used search dates e.g. 2000-2005. • May have cited seminal or landmark studies.	• Clearly explained the hypothesis, aim, or objective. • Discussed the hypothesis, aim, or objective, in relation to the data. • May have made a hypothesis decision.	• May have described statistics as it related to design. • May have described sample size and power. • May have described statistical limitations.	• Clear and concise presentation. • Answered technical questions with great ease. • Logical flow. • Reiterated results. • May have recommended future direction.	• Very well prepared. • Professional attire. • Well mannered and respectful towards others. • Arrived early. • Polished presentation.

Rater Name: _____ **Total Score** (circle): 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

FIG. 1. Competency-based resident research scoring rubric.

within the five prescribed domains. This rubric should allow residents to better understand the criteria by which they are evaluated. The psychometric data, including the inter-rater reliability and content validity procedures, are aligned with the ACGME Timeline-Working Guidelines, Phase 3, advising programs to focus on more standardized measures.

Concomitantly, the development phase referenced the Standards for Educational and Psychological Testing. Although we did not develop a commercial test, such as the ACT or SAT, we did subscribe to the intent of these important standard setting procedures which, if followed, will lead to both valid and reliable scores.

A recent study from the University of Ohio Hospitals has established a set of criteria to evaluate residency education based on the ACGME competencies. They suggest that proper assessment tools must be "reliable, reproducible, and valid" as well as "open and clearly defined" [12]. In addition, the ACGME has published a set of key considerations for selecting assessment instruments. These elements require that the assessment approach be valid, valuable, reliable, feasible, and demonstrate external validity.

Presently, there is no gold standard for setting a desirable coefficient of agreement, and various coefficient scales are contradictory [13–15]. For example, when working on predictor tests or hypothesized construct measures, Nunnally considers a reliability coefficient of 0.70 or higher as being sufficient but argues that basic research that increases the reliability coefficients to 0.80 as being "often wasteful of time and funds" [16]. Landis and Koch suggest that reliability coefficients between 0.40 and 0.75 represent *fair* to *good* agreement [17]. A brief oral presentation that is delivered by a resident does not constitute a high stake decision, other than conferring a trophy or a nominal cash prize. Therefore, the authors could tolerate a certain degree of error as suggested by Pedhazur and Schmelkin [4]. By consensus, the reliability coefficient obtained of 0.61 was considered *good* for judging low stake oral research presentations.

In summary, the preliminary results from this small sample pilot study should be interpreted with caution. Namely, this scoring rubric does not possess external validity (generalizability) or concurrent validity. To address these shortcomings, the second phase of this study will include testing the scoring rubric in multiple surgical and nonsurgical resident research settings. Concurrent validity will require obtaining similar types of resident assessment data points that had occurred at approximately the same time, such as an end of month resident evaluation metric, and correlate the two sets of scores [18].

ACKNOWLEDGMENTS

The authors formally acknowledge Jack Butler of Butler Graphics for his graphical support, Bruce Fay of Wayne County RESA, and Nicole Shamey of the Plymouth-Canton Community Schools for their valuable editorial assistance.

REFERENCES

1. Lynch, DC, Swing, SR. Accreditation Council for Graduate Medical Education, 2001. Accessed: January 5, 2007. Outcome Project Key Considerations for Selecting Assessment Instruments; Available at: http://www.acgme.org/outcome/assess/keyConsider.asp.
2. Accreditation Council for Graduate Medical Education, Version 1.1, 2000 September. Accessed: January 5, 2007. Toolbox of Assessment Methods; Available at: http://www.acgme.org/outcome/assess/toolbox.asp.
3. Erwin TD. Assessing student learning and development. San Francisco: Jossey-Bass Publishers, 1991:83–84.
4. Pedhazur EJ, Schmelkin LP. Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991:82.
5. Oetting TA, Lee AG, Beaver HA, et al. Teaching and assessing surgical competency in ophthalmology training programs. Ophthalmic Surg Lasers Imaging 2006;37:384.
6. Wings.buffalo.edu/. University at Buffalo, The State University of New York, 2004. Cited January 5, 2007. University at Buffalo Graduate Medical Education; Available from: http://wings.buffalo.edu/smbs/GME/evaluation.htmUniversity.
7. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington DC: American Educational Research Association, 1999.
8. Cronbach LJ. Coefficient α and the internal structure of tests. Psychometrika 1951;16:197.
9. Thompson B, ed. In Score reliability: Contemporary thinking on reliability issues. Thousand Oaks, CA: Sage Publications; 2003:3.
10. SPSS Inc. SPSS Base 10.0 Applications Guide. Chicago, IL: Prentice Hall, 1999.
11. Fleiss JL. Statistical methods for rates and proportions, 2nd ed. New York, NY: John Wiley and Sons, 1981:218.
12. Lee AG, Carter KD. Managing the new mandate in resident education: A blueprint for translating a national mandate into local compliance. Ophthalmology 2004;111:1807.
13. Thorndike RL, Hagen E. Measurement and evaluation in psychology and education. New York, NY: John Wiley and Sons, 1961:189.
14. Pedhazur ES, Schmelkin LP. Measurement design and analysis. Hillsdale, N: Lawrence Erlbaum Associates, 1991:109.
15. von Eye A, Mun EY. Analyzing rater agreement: Manifest variable methods. Mahwah, NJ: Lawrence Erlbaum Associates, 2005:5.
16. Nunnally JC. Psychometric theory. New York: McGraw-Hill, 1978: 245.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159.
18. Mehrens WA, Lehmann IJ. Measurement and evaluation in education and psychology, 2nd ed. New York, NY: Holt, 1978: 292.