MedEdPORTAL®
The Journal of Teaching and Learning Resources

Original Publication

Measuring Assessment Quality With an Assessment Utility Rubric for Medical Education

Jorie M. Colbert-Getz, PhD*, Michael Ryan, MD, MEHP, Erin Hennessey, MD, Brenessa Lindeman, MD, MEHP, Brian Pitts, MD, MEHP, Kim A. Rutherford, MD, Deborah Schwengel, MD, MEHP, Stephen M. Sozio, MD, Jessica George, MD, MEHP, Julianna Jung, MD

*Corresponding author: jorie.colbert@gmail.com

Citation: Colbert-Getz JM, Ryan M, Hennessey E, et al. Measuring assessment quality with an assessment utility rubric for medical education. *MedEdPORTAL*. 2017;13:10588. https://doi.org/10.15766/mep_2374-8265.10588

Abstract

Introduction: Prior research has identified seven elements of a good assessment, but the elements have not been operationalized in the form of a rubric to rate assessment utility. It would be valuable for medical educators to have a systematic way to evaluate the utility of an assessment in order to determine if the assessment used is optimal for the setting. **Methods:** We developed and refined an assessment utility rubric using a modified Delphi process. Twenty-nine graduate students pilot-tested the rubric in 2016 with hypothetical data from three examinations, and interrater reliability of rubric scores was measured with interclass correlation coefficients (ICCs). **Results:** Consensus for all rubric items was reached after three rounds. The resulting assessment utility rubric includes four elements (equivalence, educational effect, catalytic effect, acceptability) with three items each, one element (validity evidence) with five items, and space to provide four feasibility items relating to time and cost. Rater scores had ICC values greater than .75. **Discussion:** The rubric shows promise in allowing educators to evaluate the utility of an assessment specific to their setting. The medical education field needs to give more consideration to how an assessment drives learning forward, how it motivates trainees, and whether it produces acceptable ranges of scores for all stakeholders.

Keywords

Validity, Assessment, Editor's Choice

Appendices

A. Assessment Utility Rubric .docx

B. Assessment Utility Rubric Instructions.docx

All appendices are peer reviewed as integral parts of the Original Publication.

Educational Objectives

After applying the assessment utility rubric, educators will be able to:

1. Describe ways to measure five elements of assessment utility: equivalence, educational effect, catalytic effect, validity evidence, and acceptability.
2. Evaluate the utility of assessments used locally.

Introduction

Given the growing complexities of modern medical education and an increased emphasis on competency-based education, assessments must keep pace to ensure they effectively measure the knowledge, skills, and attitudes expected of a medical provider. All stakeholders, including accrediting bodies and the general public, expect objective evidence to demonstrate both trainee readiness for advancement and the competence of physicians. Thus, it is important to have high-quality assessment tools across the continuum of medical education. The difficulty for medical educators is knowing what constitutes a high-quality assessment and therefore understanding when to revise, omit, or continue to use a particular assessment.

In 2011, Norcini and colleagues published a consensus statement identifying seven elements of a good assessment: validity, reproducibility, equivalence, feasibility, educational effect, catalytic effect, and acceptability.[1] Validity is the accuracy of assessment scores in measuring what they are intended to measure. Reproducibility is the consistency of assessment scores over time and with different raters.

Equivalence refers to the assessment producing similar scores across different groups. Feasibility refers to the ability to administer the assessment with a reasonable amount of time and resources. Educational effect refers to the motivation of trainees to study for the assessment, while catalytic effect refers to the impact of scores and feedback on "driving future learning forward."[1] Acceptability refers to the extent to which all stakeholders find the range of assessment scores to be appropriate and reasonable. It should be noted that five of these elements were derived from a conceptual model of assessment utility published by Van Der Vleuten.[2]

To date, the seven elements have not been operationalized in the form of a rubric or checklist that can be used to rate an assessment. It would be valuable for medical educators to have a systematic way to evaluate the utility of an assessment in order to determine if it is optimal for their setting. Therefore, we developed a rubric to formally measure assessment utility in medical education.

Methods

Rubric Development—Content Validity Evidence

A literature review of assessment articles informed the draft item list of the assessment utility rubric. This draft rubric included the elements described by Norcini and colleagues.[1] However, validity and reproducibility were collapsed into a single element termed *validity evidence.* There were two to five items for each element. Ten medical school assessment experts reviewed the rubric elements and items. Nine of the experts had completed or were close to completion of a Johns Hopkins University master of education in health professions (MEHP) degree. All had finished a graduate course on assessment, and one expert was an instructor for the assessment course. The 10 experts from various institutions represented the continuum of undergraduate to graduate medical education leadership, including course/clerkship directors, curriculum deans, and residency/fellowship directors. For the review, the experts were asked to suggest additional elements to include and better ways of measuring the existing elements. After this review, six items were revised, and three were added to the rubric, resulting in a total of 21 items across the various elements, including four items for feasibility, two items for educational effect, four items for catalytic effect, five items for validity evidence, and three items for acceptability. Then, the experts participated in a modified Delphi process by rating the 21 rubric items in terms of importance for the utility of an assessment on a scale of 0 to 2 where 0 = *not important/does not need to be considered,* 1 = *somewhat important,* and 2 = *critically important.* Consensus was defined as agreement among 80% or more of participants for importance ratings.

After the 21 items were rated using the Delphi process, consensus on the importance of each item was reached for 13 of the 21 rubric items; one item was omitted as it was captured in another item. After this, consensus was then reached for four more items. This left three still in need of consensus. We discussed these items via a conference call. For one item, it was decided that the importance weight would vary depending on whether the assessment questions had one correct answer each, such as in multiple-choice examinations, or required a rater's subjective judgment. The other two items in need of consensus dealt with the feasibility element; we grappled with the appropriate amount of time for a coordinator to prepare and administer an assessment and the appropriate time necessary for faculty development. Ultimately, we decided that feasibility information should be provided alongside the utility score so that an institution could determine if the amount of time and resources was commensurate with the utility of an assessment. The final version of the assessment utility rubric is provided in Appendix A. The rubric includes four elements with three items each (i.e., equivalence, educational affect, catalytic effect, and acceptability) and one element with five items (i.e., validity evidence). The rubric also includes space for feasibility items of cost, coordinator time, faculty/standardized patient development time, and complexity of/time for psychometric analysis. Appendix B provides instructions on rubric scoring and examples, with practical advice for each rubric element.

Rubric Pilot—Internal Consistency Validity Evidence

To determine the interrater reliability and discrimination ability of rubric scores, we piloted the rubric with 35 MEHP graduate students in an assessment course not involved in the Delphi process. The majority of

graduate students had a professional degree in medicine (32; 91%), with the remaining three (9%) having a degree in pharmacy, nursing, or a science-related field. The students resided in the United States (26; 74%), Canada (6; 17%), and the Middle East (3; 12%).

The students were provided with the assessment rubric, instructions, and hypothetical assessment data from an objective structured clinical examination (OSCE), a midterm multiple-choice examination (m-MCQ), and a final multiple-choice examination (f-MCQ) for a Year 2 renal course. The examination data generated were theoretical but were loosely based on data typically resulting from internally developed multiple-choice examinations and OSCEs in medical school. Additionally, data were purposely created so that the f-MCQ would have the highest utility score, the m-MCQ would have the lowest utility score, and the utility scores of the OSCE would fall somewhere in between the two multiple-choice exams. Hypothetical data are available on request for faculty development training.

To determine if raters could differentiate among the utility for the three assessments, we compared mean utility scores with a one-way analysis of variance (ANOVA) and Scheffé post hoc tests as appropriate. To assess the interrater reliability of rubric scores, we computed an interclass correlation coefficient (ICC) using a two-way random-effects model with absolute judgment for each assessment. Since the intention is for the rubric to be used by a single rater in the future, we computed single-measure ICCs. An ICC of .60-.74 suggests good agreement, and an ICC greater than .75 suggests excellent agreement.[3] There was evidence of content validity to justify the development of the assessment utility rubric. The Results section below provides details on internal structure validity evidence for assessment utility rubric scores. The Johns Hopkins University School of Medicine Institutional Review Board determined that this study was exempt from review.

Results

Six of the 35 graduate students had incomplete rubric data or did not complete the assignment and were omitted from the analyses, resulting in a final sample size of 29. A one-way ANOVA showed a significant difference among the three assessment utility scores, $F(2,84) = 515.58$, $p \leq .001$, partial $\eta^2 = .93$. Specifically, the average utility score for the f-MCQ was 21% higher (95% CI, 16%-26%) than the average OSCE utility score and 68% higher (95% CI, 63%-74%) than the average m-MCQ utility score ($p \leq .001$ for both comparisons). Additionally, the average OSCE utility score was 47% higher (95% CI, 42%-53%) than the average m-MCQ utility score ($p \leq .001$). The ICC of the rubric ratings for the f-MCQ was .89, for the OSCE was .73, and for the m-MCQ was .74

Discussion

The assessment utility rubric described in this resource expands on prior work[1,2] by providing items to measure the equivalence, educational effect, catalytic effect, validity evidence, and acceptability of an assessment, along with importance weightings for each item. Past medical education literature has focused almost entirely on the psychometrics of assessment scores to ascertain quality.[4] The results of our modified Delphi process suggest that future research needs to provide evidence beyond a high Cronbach's alpha value or generalizability coefficient for assessment quality. Specifically, more consideration needs to be given to how an assessment drives learning forward, how it motivates trainees, and whether it produces acceptable ranges of scores for all stakeholders. Our findings suggest that even assessments with validity evidence for scores may have limited utility when considered within the broader context of educational goals. Assessments that produce qualitative data may offer greater utility, as these assessments may be more useful for motivating trainees and for giving them an understanding of their own learning needs, relative to quantitative assessment, which merely provides an overall score.[4,5]

Although feasibility was one of the elements in the original consensus statement by Norcini and colleagues,[1] we ultimately left it out of the utility equation. Instead, we considered it in the context of local resources. The OSCE can be used to illustrate our rationale. The feasibility of an OSCE is typically measured in terms of cost per student or cost per student per station. However, a recent systematic review demonstrated a lack of consensus in determining what should be included in measuring the

indirect cost of administering OSCEs (faculty time, training, staff support).[6] Thus, feasibility is likely contextual, factors in local resources, and should be considered outside the utility equation of an assessment.

A strength of the rubric is that it can be used for a variety of assessment tools (multiple-choice examinations, OSCEs, workplace assessments, etc.) in medical education as opposed to being tailored to only one type of tool. Since most courses and programs use a variety of assessment tools, the rubric allows for easy comparisons of utility. A weakness is that student scores on assessments need to be gathered, stored, and analyzed to use the rubric. Although most medical schools have access to scores from assessments, not all schools have access to a person who can analyze the scores for validity evidence items. In order to meet all elements of assessment utility, it may be necessary for medical schools to increase resources allocated to an assessment program.[7] This weakness may limit the use of the assessment utility rubric.

A few limitations of this resource should be considered. First, our expert panel involved only 10 faculty experts. We purposely selected our experts based on their intimate involvement with medical educational leadership and felt that a smaller panel of highly qualified experts would yield better results than a larger panel of general faculty. Second, we used hypothetical assessment data to measure the discrimination ability and interrater reliability of rubric scores. Although the data were based on common results in medical education, we do not presently know if the rubric would work equally as well with real data.

Applying the assessment utility rubric presented in this resource will allow educators, administrators, and curriculum committees to make evidence-based decisions about designing, selecting, and revising assessments across medical education programs.

Jorie M. Colbert-Getz, PhD: Assistant Dean of Assessment and Evaluation, University of Utah School of Medicine; Assistant Professor, Department of Internal Medicine, University of Utah School of Medicine

Michael Ryan, MD, MEHP: Assistant Dean for Clinical Medical Education, Virginia Commonwealth University School of Medicine; Associate Professor, Department of Pediatrics, Virginia Commonwealth University School of Medicine

Erin Hennessey, MD: Program Director for the Anesthesia Critical Care Medicine Fellowship, Stanford University School of Medicine; Clinical Assistant Professor, Department of Anesthesia and Critical Care Medicine, Stanford University School of Medicine

Brenessa Lindeman, MD, MEHP: Fellow and Associate Surgeon, Department of Surgery, Brigham and Women's Hospital; Instructor of Surgery, Harvard Medical School

Brian Pitts, MD, MEHP: Associate Professor, Department of Anesthesiology, University of California, Davis, School of Medicine

Kim A. Rutherford, MD: Assistant Professor, Departments of Pediatrics and Emergency Medicine, Pennsylvania State University College of Medicine

Deborah Schwengel, MD, MEHP: Program Director for the Anesthesiology Residency Program, Johns Hopkins University School of Medicine; Assistant Professor, Departments of Anesthesiology, Critical Care Medicine, and Pediatrics, Johns Hopkins University School of Medicine

Stephen M. Sozio, MD: Associate Director of the Nephrology Fellowship Program, Johns Hopkins University School of Medicine; Assistant Professor, Department of Medicine, Johns Hopkins University School of Medicine

Jessica George, MD, MEHP: Assistant Professor of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine

Julianna Jung, MD: Associate Director, Johns Hopkins Medicine Simulation Center; Associate Professor, Department of Emergency Medicine, Johns Hopkins University School of Medicine

Disclosures
None to report.

Funding/Support
None to report.

Prior Presentations

Colbert-Getz JM, George J, Jung J, et al. Developing a rubric to assess our assessments: results of a Delphi process with experts from multiple institutions. Poster presented at: Learn Serve Lead 2016: the AAMC Annual Meeting; November 11-25, 2016; Seattle, WA.

Ethical Approval

This publication contains data obtained from human subjects and received ethical approval.

References

1. Norcini J, Anderson B, Bollela V, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach.* 2011;33(3):206-214. https://doi.org/10.3109/0142159X.2011.551559
2. Van Der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract.* 1996;1(1):41-67. https://doi.org/10.1007/BF00596229
3. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess.* 1994;6(4):284-290. https://doi.org/10.1037/1040-3590.6.4.284
4. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach.* 2013;35(7):564-568. http://dx.doi.org/10.3109/0142159X.2013.789134
5. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: validity evidence for qualitative educational assessments. *Acad Med.* 2016;91(10):1359-1369. https://doi.org/10.1097/ACM.0000000000001175
6. Patricio MF, Julião M, Fareleira F, Vaz Carneiro A. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? *Med Teach.* 2013;35(6):503-514. https://doi.org/10.3109/0142159X.2013.774330
7. Eva KW, Bordage G, Campbell C, et al. Towards a program of assessment for health professionals: from training into practice. *Adv Health Sci Educ Theory Pract.* 2016;21(4):897-913. https://doi.org/10.1007/s10459-015-9653-6

Received: January 9, 2017 | **Accepted:** April 28, 2017 | **Published:** May 24, 2017