QUALITY OF MEDICAL EDUCATION SCHOLARSHIP

Predictive Validity Evidence for Medical Education Research Study Quality Instrument Scores: Quality of Submissions to *JGIM*'s Medical Education Special Issue

Darcy A. Reed, MD, MPH[1], Thomas J. Beckman, MD[2], Scott M. Wright, MD[3], Rachel B. Levine, MD, MPH[3], David E. Kern, MD, MPH[3], and David A. Cook, MD, MHPE[2]

[1]Division of Primary Care Internal Medicine, Mayo Clinic College of Medicine, Rochester, MN, USA; [2]Division of General Internal Medicine, Mayo Clinic College of Medicine, Rochester, MN, USA; [3]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

BACKGROUND: Deficiencies in medical education research quality are widely acknowledged. Content, internal structure, and criterion validity evidence support the use of the Medical Education Research Study Quality Instrument (MERSQI) to measure education research quality, but predictive validity evidence has not been explored.

OBJECTIVE: To describe the quality of manuscripts submitted to the 2008 *Journal of General Internal Medicine* (*JGIM*) medical education issue and determine whether MERSQI scores predict editorial decisions.

DESIGN AND PARTICIPANTS: Cross-sectional study of original, quantitative research studies submitted for publication.

MEASUREMENTS: Study quality measured by MERSQI scores (possible range 5–18).

RESULTS: Of 131 submitted manuscripts, 100 met inclusion criteria. The mean (SD) total MERSQI score was 9.6 (2.6), range 5–15.5. Most studies used single-group cross-sectional (54%) or pre-post designs (32%), were conducted at one institution (78%), and reported satisfaction or opinion outcomes (56%). Few (36%) reported validity evidence for evaluation instruments. A one-point increase in MERSQI score was associated with editorial decisions to send manuscripts for peer review versus reject without review (OR 1.31, 95%CI 1.07–1.61, p=0.009) and to invite revisions after review versus reject after review (OR 1.29, 95%CI 1.05–1.58, p=0.02). MERSQI scores predicted final acceptance versus rejection (OR 1.32; 95% CI 1.10–1.58, p= 0.003). The mean total MERSQI score of accepted manuscripts was significantly higher than rejected manuscripts (10.7 [2.5] versus 9.0 [2.4], p=0.003).

CONCLUSIONS: MERSQI scores predicted editorial decisions and identified areas of methodological strengths and weaknesses in submitted manuscripts. Researchers, reviewers, and editors might use this instrument as a measure of methodological quality.

Electronic supplementary material The online version of this article (doi: 10.1007/s11606-008-0664-3) contains supplementary material, which is available to authorized users.

KEY WORDS: medical education research; research quality; research methods.
J Gen Intern Med 23(7):903–7
DOI: 10.1007/s11606-008-0664-3

Deficiencies in the quality of medical education research are widely acknowledged.[1–4] Medical education leaders have appealed for increased methodological rigor, including larger multi-institutional studies,[5,6] greater attention to validity and reliability of assessments,[7] and examination of clinically relevant outcomes.[8,9] Nonetheless, the quality of the current body of published education research remains suboptimal, with the majority of articles reporting single institution studies[10] and less rigorous outcomes, such as learner satisfaction or acquisition of knowledge and skills.[10,11]

An instrument measuring the quality of education research studies could be useful to investigators designing studies and to journal editors reviewing submitted manuscripts. We have developed a Medical Education Research Study Quality Instrument (MERSQI)[10] to measure the methodological quality of experimental, quasi-experimental, and observational studies in medical education. In a previous study we demonstrated content, internal structure, and criterion validity evidence for MERSQI scores, including the relationship of one factor, funding, to study quality.[10] However, predictive validity evidence has not been established for MERSQI scores.

In this study, we examined whether MERSQI scores predicted editorial decisions for the 2008 medical education special issue of the *Journal of General Internal Medicine (JGIM)*. *JGIM* regularly publishes a special issue containing medical education research pertinent to general internal medicine. We hypothesized that submitted manuscripts with higher MERSQI scores would be more likely to be sent for peer review, have revisions invited after review, and ultimately be accepted for publication compared to manuscripts with lower MERSQI scores.

METHOD

Study Design

We conducted a cross-sectional assessment of the quality of manuscripts submitted to the 2008 *JGIM* medical education

special issue. Submitting authors were given the opportunity to decline to include their manuscript in this study. *JGIM* editors were not aware of authors' study participation status; all editorial decisions were made independent of study participation. The Mayo Foundation Institutional Review Board deemed this study exempt from review.

Data Collection

A team of investigators (DAR, TJB, SMW, and RBL) who were not involved in *JGIM* editorial decisions used the Medical Education Research Study Quality Instrument (MERSQI)[10] to measure the quality of studies submitted to *JGIM*'s medical education issue. All studies were de-identified using the procedures described below. Although high interrater reliability of MERSQI scores has already been established,[10] two investigators independently scored a subset of studies (55 of 100, 55%) to confirm reliability in this sample. After confirming high interrater reliability, the remaining 45 articles were scored by one investigator. Disagreements were resolved by consensus.

Information on manuscript type (educational innovation, original article, brief report, perspective, review, resource paper, and recommendations/guidelines), initial publication decision (reject without peer review, reject after review, or revise after review), and final decision (reject or accept) was provided by the *JGIM* editorial office.

De-identification of Studies

The *JGIM* editorial office removed author names and affiliations from manuscripts and then sent them to an administrative assistant who removed all other identifying information from the manuscript, including acknowledgments, institution names in manuscript text, and references. After MERSQI scoring was complete, the *JGIM* editorial office provided provisional and final publication decisions using manuscript unique identifiers.

JGIM editorial decisions were made without knowledge of MERSQI scores or other study results. As Co-Editor for the *JGIM* medical education special issue, investigator DAC was not involved in data collection (grading studies) or data analysis and had no knowledge of individual manuscripts' MERSQI scores.

Quality Assessment Instrument

The MERSQI is a ten-item instrument designed to assess the methodological quality of experimental, quasi-experimental, and observational medical education research studies.[10] The ten items reflect six domains of study quality: study design, sampling, data type (subjective or objective), validity of assessments, data analysis, and outcomes. The maximum score for each domain is 3. A total MERSQI score is calculated as the sum of domain scores with appropriate reductions in the denominator for "not applicable" responses. Thus, possible total MERSQI scores range from 5 to 18. Total MERSQI scores are adjusted to a denominator of 18 to allow for comparison of scores across studies. The MERSQI instrument and scoring algorithm is available online (Appendix).

We have previously demonstrated strong validity evidence for MERSQI scores including: (1) content evidence based on expert consensus and published literature supporting instrument items, (2) internal structure evidence based on factor analysis and excellent interrater, intrarater, and internal consistency reliability, and (3) criterion validity evidence (relationships to other variables) demonstrated by strong correlations between MERSQI scores and the impact factor of the journal in which the study was published, the number of times the study was cited in the 3 years after publication, and global quality ratings by independent experts.[10]

Data Analysis

Total, domain, and item MERSQI scores for submitted and published studies were summarized using descriptive statistics and compared using Wilcoxon rank sum test. We used logistic regression to examine associations between total MERSQI scores and initial (reject without peer review, reject after review, or revise after review) and final (reject or accept) editorial decisions. Interrater reliability was determined using intraclass correlation coefficients (ICC). We considered a two-tailed $p < 0.05$ statistically significant for all analyses. Data were analyzed using STATA 8.0 (STATA Corp., College Station, TX).

RESULTS

Characteristics of Submitted Manuscripts

One hundred thirty-one manuscripts were submitted to the 2008 *JGIM* medical education special issue. Thirty-one were excluded (16 used qualitative methods exclusively, 14 were not original research, and 1 author declined to include his or her manuscript), leaving 100 quantitative, original research manuscripts for analysis.

Of the remaining 100 manuscripts, 58 were submitted as original articles, 35 were submitted as educational innovations, and 7 were submitted as brief reports. Almost half of studies (46%) involved residents as study participants, while 37% involved medical students, and just 7% included faculty. Ten percent of studies included a combination of students, residents, and faculty as study participants.

Quality of Submitted Studies

The interrater reliability of MERSQI scores was excellent with ICCs for individual items ranging from 0.76 (95% CI 0.67–0.83) to 0.98 (95% CI 0.97–0.99) (Table 1).

Table 1. Interrater Reliability of MERSQI Scores

MERSQI Item	Interrater Reliability ICC (95% CI)*
1. Study design	0.98 (0.97–0.99)
2. Institutions	0.97 (0.96–0.99)
3. Response rate	0.82 (0.69–0.89)
4. Type of data	0.78 (0.63–0.87)
5. Validity: Internal structure	0.90 (0.82–0.94)
6. Validity: Content	0.93 (0.88–0.96)
7. Validity: Relationships to variables	0.91 (0.85–0.95)
8. Appropriateness of analysis	0.76 (0.67–0.83)
9. Sophistication of analysis	0.96 (0.92–0.97)
10. Outcome	0.83 (0.68–0.89)

**Calculated for 55 studies that were scored by two independent raters*

Table 2. Mean Total MERSQI Scores for Manuscripts Submitted to the 2008 *JGIM* Medical Education Special Issue

	Initial Editorial Decision				Final Editorial Decision	
	All Manuscripts (n=100)	Rejected Without Review (n=25)	Rejected After Review (n=34)	Revision Invited (n=41)	Rejected (n=64)	Accepted (n=26)
All manuscripts	9.6 (2.6)* n=100	8.4 (2.3) n=25	9.2 (2.1) n=34	10.7 (2.7) n=41	9.0 (2.4) n=65	10.7 (2.5) n=35
Type of submission						
Original article or brief report	10.3 (2.2) n=65	9.5 (2.2) n=13	10.0 (1.5) n=21	10.8 (2.5) n=31	9.8 (2.0) n=38	10.8 (2.7) n=27
Educational innovation	8.3 (2.7) n=35	7.2 (1.9) n=12	7.9 (2.3) n=13	10.3 (3.2) n=10	7.8 (2.4) n=27	10.0 (3.1) n=8

**Mean (standard deviation)*

The mean (SD) total MERSQI score of studies was 9.6 (2.6), range 5–15.5. Most studies used single group cross-sectional (54%) or pre-post designs (32%). Fourteen percent of studies included a control or comparison group, and 5% were randomized. Nearly one quarter (22%) of studies were multi-institutional. Nineteen percent failed to report a response rate. Less than half (42%) included objective measurements. Thirty-six percent of manuscripts reported at least one measure of validity evidence for scores from their evaluation instruments: 29% demonstrated content, 20% internal structure, and 9% relationships to other variables (e.g., criterion, concurrent, or predictive validity) evidence. Errors in data analysis were identified in 30% of submitted manuscripts. Most studies (56%) reported satisfaction or opinion outcomes, while a minority reported knowledge or skills (32%), behavior (7%), or patient-related outcomes (5%).

The mean (SD) total MERSQI score of the 35 manuscripts submitted as "educational innovations" was lower than the 65 studies submitted as "original articles" or "brief reports" [8.3 (2.7) versus 10.3 (2.2), p<0.001], (Table 2). Manuscripts submitted as original articles or brief reports had higher MERSQI scores than those submitted as educational innovations in domains of sampling [2.0 (0.6) versus 1.6 (0.5), p=0.002]; validity of evaluation instruments' scores [0.8 (0.9) versus 0.3 (0.7), p=0.003]; and data analysis [2.7 (0.5) versus 2.0 (0.8), p<0.001]. There was no difference in MERSQI scores by submission category in the domains of study design, type of data, and outcomes.

Association Between MERSQI Scores and Editorial Decisions

Of the 100 submitted manuscripts in the analysis, 75 were sent for peer review, and 25 were rejected without peer review. Of the 75 sent for peer review, 41 received an invitation to revise, and 34 were rejected immediately after peer review. Ultimately, 35 manuscritpts were accepted for publication, and 65 were rejected. For logistic reasons, some manuscripts will be published in a regular issue of JGIM and do not appear in the special issue.

MERSQI scores were associated with an initial editorial decision to send a manuscript for peer review versus reject without review [OR 1.31 for a one-point MERSQI score increase; 95% confidence interval (95% CI) 1.07–1.61, p=0.009] and to invite revision after review versus reject after review (OR 1.29; 95% CI 1.05–1.58, p=0.02). MERSQI scores also predicted final acceptance versus rejection (OR 1.32; 95% CI 1.10–1.58, p=0.003). Thus, a one-point increase in MERSQI score was associated with a 1.32 odds of manuscript acceptance.

The mean total MERSQI score of the 35 accepted manuscripts was significantly higher than the 65 rejected manuscripts [10.7 (2.5) versus 9.0 (2.4), p=0.003] (Table 2). Accepted manuscripts received higher mean MERSQI scores than rejected manuscripts in the domains of sampling [2.1 (0.6) versus 1.8 (0.6), p=0.03]; validity of evaluation instruments' scores [0.9 (1.0) versus 0.5 (0.8), p=0.02]; data analysis [2.7 (0.6) versus 2.4 (0.7), p=0.01); and outcomes [1.5 (0.5) versus 1.3 (0.5), p=0.006) (Figure 1). MERSQI scores were similar for accepted and rejected manuscripts in the domains of study design and type of data.

DISCUSSION

The quality of manuscripts submitted to the 2008 *JGIM* medical education special issue was modest. Most submissions described single institution studies using cross-sectional designs and reporting satisfaction or opinion outcomes. However, our results indicate that high quality submissions, as measured by MERSQI scores, were ultimately selected for publication. As a result, many of the accepted manuscripts are outstanding examples of methodologically rigorous medical education research.

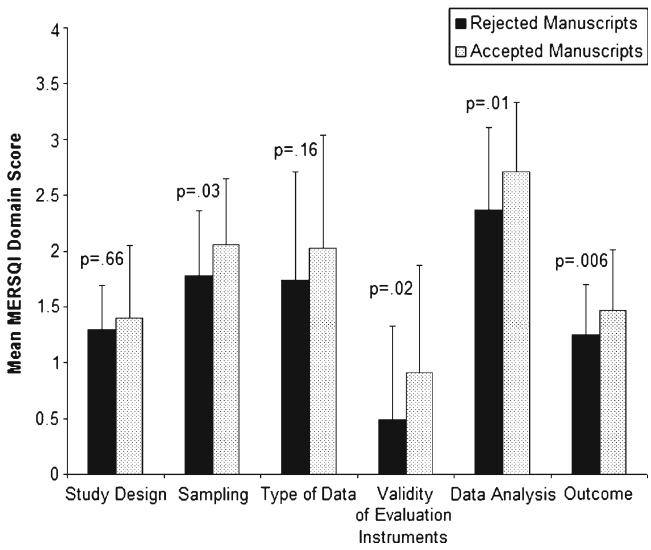

Figure 1. Quality of rejected and accepted manuscripts in the 2008 *JGIM* Medical education special issue. Legend. Calculated for 65 rejected and 35 accepted manuscripts. Columns represent mean domain-specific MERSQI scores. Error bars represent standard deviation of mean domain-specific MERSQI scores. Maximum possible domain-specific MERSQI score is 3.

Few submitted manuscripts reported validity evidence for scores from evaluation instruments. MERSQI scores were lowest in this domain for both accepted and rejected manuscripts. This is consistent with prior observations that many categories of validity evidence are underreported in medical education studies.[12] However, descriptions of validity evidence for evaluation instruments' scores were associated with acceptance for publication, suggesting that reviewers and editors agree that validity evidence is important. Because studies that use measurement instruments with weak or no validity evidence are less likely to be published, authors are advised to gather validity evidence in the beginning stages of study design and implementation. Published frameworks that describe and classify validity evidence may facilitate this effort.[7,13,14]

Less than one-fifth of studies submitted to this issue were multi-institutional, and very few measured learner behaviors (7%) or health care outcomes (5%). These results confirm prior assertions that multi-institutional studies examining clinically relevant outcomes are lacking. Given appeals for greater generalizability[5,6] and clinically relevant education research,[8,9,11] multi-institutional studies measuring higher level outcomes should be prioritized where appropriate for the research question and study aims.

The associations between MERSQI scores and editorial decisions have meaningful implications. First, this finding provides evidence for the predictive validity of MERSQI scores, supporting its role as a measure of education research study quality. Second, the MERSQI may facilitate peer review and editorial decision-making processes. For example, it could be used by editors to screen articles for review versus rejection, or to resolve dissimilar peer reviews. Because peer reviewers frequently disagree[15] and reviews may be influenced by relationships with authors,[16] MERSQI scores could be used to standardize the peer review process and identify important methodological issues. Third, the association between MERSQI scores and editorial decisions authenticates the editorial process by showing that editors' decisions are congruent with established measures of methodological quality.

We acknowledge that the MERSQI focuses solely on the quality domain of methods. Study methods are only one aspect of the multifaceted "quality" of a manuscript. Other important aspects include the quality of the research question,[17,18] accuracy of interpretations drawn from study results,[19] and the quality of reporting.[20] Yet the methods largely determine the confidence one can place in the interpretations drawn from study results. MERSQI scores now have substantial validity evidence supporting their use in assessing the methodological quality of medical education scholarship, and this instrument should thus prove useful to educators, scholars, and editors.

This study has several limitations. First, we assigned MERSQI scores to manuscripts at the time of initial submission, but did not re-score manuscripts after revisions were made. Thus, although many MERSQI items are unlikely to change with revisions (i.e., study design, number of institutions, response rate, outcomes), errors in data analysis and reporting of validity evidence may be identified in the peer review process and corrected prior to publication. Therefore, MERSQI scores of published studies may be higher than initial submissions. Second, we excluded qualitative studies from this analysis because fundamental differences in study design, sampling, evaluation instruments, and analysis preclude summative comparison to other study types.[21,22] Although we were unable to assess the quality of qualitative manuscripts using the MERSQI, we observed that a similar percentage of qualitative and quantitative submissions were accepted for publication (31% and 35%, respectively), suggesting that editors value both approaches to education research. Finally, we examined the quality of studies submitted to a single journal, which limits generalization of our findings to a broader range of journals. However, the mean total MERSQI score in this sample [9.6 (SD 2.6)] is similar to that of a sample of published studies from 13 peer-reviewed journals, including general medicine, subspecialty medicine, and medical education journals [9.9 (2.3)].[10]

Limitations notwithstanding, this study characterizes the quality of a sample of submitted medical education manuscripts and identifies their methodological strengths and limitations. The results also provide predictive validity evidence for MERSQI scores as a measure of the quality of medical education research. The MERSQI may be a useful tool for education researchers, reviewers, and journal editors to gauge the quality of submitted and published education scholarship.

Acknowledgments: *This study was presented in abstract form at the Society of General Internal Medicine 32nd Annual Meeting in April 2008. No funding organization, or sponsor had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.*

Dr. Wright is an Arnold P. Gold Foundation Associate Professor of Medicine, and he is also a Miller-Coulson Family Scholar. No funding organization or sponsor had any role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript. Dr. Reed had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. The authors would like to thank Kathryn Trana for assistance with manuscript blinding.

Conflict of Interest: *Dr. Cook is a Deputy Editor for the Journal of General Internal Medicine and was a Co-Editor for the 2008 medical education special issue. No other conflicts of interest exist for any of the authors.*

Corresponding Author: *Darcy A. Reed, MD, MPH; Division of Primary Care Internal Medicine, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55901, USA (e-mail: reed.darcy@mayo.edu).*

REFERENCES

1. **Dauphinee WD, Wood-Dauphinee S.** The need for evidence in medical education: the development of best evidence medical education as an opportunity to inform, guide, and sustain medical education research. Acad Med. 2004;79(10):925–30.
2. **Wartman SA.** Research in medical education: the challenge for the next decade. Acad Med. 1994;69(8):608–14.
3. **Shea JA, Arnold L, Mann KV.** A RIME perspective on the quality and relevance of current and future medical education research. Acad Med. 2004;79(10):931–8.
4. **Lurie SJ.** Raising the passing grade for studies of medical education. JAMA. 2003;290(9):1210–2.
5. **Carline JD.** Funding medical education research: opportunities and issues. Acad Med. 2004;79(10):918–24.
6. **Regehr G.** Trends in medical education research. Acad Med. 2004;79 (10):939–47.
7. **Cook DA, Beckman TJ.** Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119 (2):166.e7–16.

8. **Chen FM, Bauchner H, Burstin H.** A call for outcomes research in medical education. Acad Med. 2004;79(10):955–60.
9. **Whitcomb ME.** Using clinical outcomes data to reform medical education. Acad Med. 2005;80(2):117.
10. **Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM.** Association between funding and quality of published medical education research. JAMA. 2007;298(9):1002–9.
11. **Prystowsky JB, Bordage G.** An outcomes research perspective on medical education: the predominance of trainee assessment and satisfaction. Med Educ. 2001;35(4):331–3.
12. **Beckman TJ, Cook DA, Mandrekar JN.** What is the validity evidence for assessments of clinical teaching? J Gen Intern Med. 2005;20(12): 1159–64.
13. **Downing SM.** Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37(9):830–7.
14. 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; 1999:9–24.
15. **Cullen DJ, Macaulay A.** Consistency between peer reviewers for a clinical specialty journal. Acad Med. 1992;67(12):856–59.
16. **Schroter S, Tite L, Hutchings A, Black N.** Differences in review quality and recommendations for publication between peer reviewers suggested by authors or by editors. JAMA. 2006;295(3):314–17.
17. **Prideaux D.** Researching the outcomes of educational interventions: a matter of design. BMJ. 2002;324:126–7.
18. **Cook DA, Bordage G, Schmidt HG.** Description, justification and clarification: a framework for classifying the purposes of research in medical education. Med Educ. 2008;42(2):128–33.
19. **Colliver JA.** The research enterprise in medical education. Teach Learn Med. 2003;15(3):154–5.
20. **Cook DA, Beckman TJ, Bordage G.** Quality of reporting of experimental studies in medical education: a systematic review. Med Educ. 2007;41 (8):737–45.
21. **Stacy R, Spencer J.** Assessing the evidence in qualitative medical education research. Med Educ. 2000;34(7):498–500.
22. **Cote L, Turgeon J.** Appraising qualitative research articles in medicine and medical education. Med Teach. 2005;27(1):71–5.