

# Deconstructing Quality in Education Research

GAIL M. SULLIVAN, MD, MPH

Past decisions about teaching often were based on the “PHOG” approach: “prejudices, hunches, opinions, and guesses.”<sup>1</sup> In the last decade, major advancements have occurred in the development and understanding of new evidence to guide medical education decisions. The formation of the Best Evidence in Medical Education (BEME) international working groups is an example of this new approach.<sup>2</sup> BEME work groups systematically search for studies to answer key questions, with a rigorous approach to evaluating the quality of evidence. Other groups have examined the quality of methods and of reporting in English-language education research studies.<sup>3–10</sup> Despite these developments, many decisions in medical education are still based on “persuasion and politics.”<sup>11</sup>

One of the primary goals of the *Journal of Graduate Medical Education (JGME)* is to improve the quality of graduate medical education research. Systematic reviews of education research have identified areas of concern.<sup>3–6,10</sup> One of our strategies will be to improve readers’ understanding of these areas. In each issue, the Journal plans to provide a summary about one aspect of research quality. For this issue, the subject is reliability and validity of assessment

Gail M. Sullivan, MD, MPH, is Editor-in-Chief, *Journal of Graduate Medical Education*.

Corresponding author: Gail M. Sullivan, MD, MPH, Editor-in-Chief, *Journal of Graduate Medical Education*, 515 N State, Suite 2000, gsullivan@nsol.uchc.edu

DOI: 10.4300/JGME-D-11-00083.1

instruments used for research outcomes (pp 119–120). This topic is particularly relevant to our readers, as validity and reliability evidence for assessments is routinely underreported in manuscripts submitted to *JGME*.

In this editorial, we introduce areas of concern in the *quantitative methodologies* delineated in systematic reviews of English-language publications and instruments available to examine the quality of education research. These instruments include the Medical Education Research Study Quality Instrument (MERSQI),<sup>5,7</sup> the BEME global scale,<sup>3</sup> and the Newcastle–Ottawa Scale (NOS) for assessing quality of nonrandomized studies.<sup>12</sup> These instruments are based in part on Kirkpatrick’s hierarchy of educational outcomes,<sup>3,13</sup> which provides a valuable conceptual framework for planning and evaluating educational initiatives. Standards are also available to assess the quality of *methods reporting*,<sup>14,15</sup> but they will not be discussed here. Similarly, other topics, such as quality of research questions and overinterpretation of results, will not be addressed in this editorial.<sup>16</sup>

The NOS was developed to rate the quality of nonrandomized studies included in systematic reviews and has data to support its validity.<sup>12</sup> Although the NOS was created for clinical research, it has been modified and used in systematic reviews of educational research.<sup>6,10</sup> Examining one’s own research for the presence or absence of specific items may be instructive (TABLE 1).

In 2 studies, the modified NOS was highly correlated with the MERSQI (TABLE 2) and the BEME global rating scale (TABLE 3).<sup>7,10</sup> Of the 3 scales, the MERSQI may be

TABLE 1 MODIFIED NEWCASTLE–OTTAWA SCALE<sup>12,20</sup>

Category	Description (1 point each)	Maximum Score
Representativeness	Intervention group “truly” or “somewhat” representative of average learner in this community	1
Selection	Comparison group drawn from same community as the exposed cohort	1
Comparability	Nonrandomized 2-cohort studies	2
	Controlled for baseline learning outcome (eg, baseline pretest scores)	
	Randomized studies	Controlled any other baseline characteristic
		Randomized
	Allocation concealed	
Blinding	Blinded outcome assessment <sup>a</sup>	1
Follow-up	Subjects lost to follow-up <sup>a</sup> unlikely to introduce bias: small no. lost (75% or greater follow-up) or description provided for those lost	1
<b>Maximum Total Score</b>		<b>6</b>

<sup>a</sup> Blinding and completeness of follow-up are reported as Yes if this was true for any reported outcome. Modified from Reed et al.<sup>20</sup>

TABLE 2 MEDICAL EDUCATION RESEARCH QUALITY INSTRUMENT<sup>5</sup> FOR QUANTITATIVE STUDIES

Domain	MERSQI Item	Score	Maximum Score
Study design	Single group cross-sectional or single group posttest only	1	3
	Single group pretest and posttest	1.5	
	Nonrandomized, 2 groups	2	
	Randomized controlled trial	3	
Sampling	Institutions studied:		3
	1	0.5	
	2	1	
	3	1.5	
	Response rate (%)		
	Not applicable		
	<50 or not reported	0.5	
	50–74	1	
≥75	1.5		
Type of data	Assessment by participants	1	3
	Objective measurement	3	
Validity of evaluation instrument	Internal structure		3
	Not applicable		
	Not reported	0	
	Reported	1	
	Content		
	Not applicable		
	Not reported	0	
	Reported	1	
	Relationships to other variables		
	Not applicable		
Not reported	0		
Reported	1		
Data analysis	Appropriateness of analysis		3
	Inappropriate for study design or type of data	0	
	Appropriate for study design and type of data	1	
	Complexity of analysis		
	Descriptive analysis only	1	
	Beyond descriptive analysis	2	
Outcomes	Satisfaction, attitudes, perceptions, opinions, general facts	1	3
	Knowledge, skills	1.5	
	Behaviors	2	
	Patient/health care outcome	3	
<b>Total possible score<sup>a</sup></b>			<b>18</b>

<sup>a</sup> Scores range from 5 to 18. Adapted from Reed et al.<sup>5</sup>

TABLE 3 **BEST EVIDENCE IN MEDICAL EDUCATION GLOBAL SCALE<sup>3</sup>**

Strength of Evidence	
1	No clear conclusions can be drawn
2	Results ambiguous, may be a trend
3	Conclusions can probably be based on the results
4	Results are clear and very likely to be true
5	Results are unequivocal
Outcomes <sup>a</sup>	
Level 1	<b>PARTICIPATION</b> Learner feedback on the learning experience (eg, organization, presentation, content, teaching materials, quality of instruction)
Level 2a	<b>ATTITUDES or PERCEPTIONS</b> Changes in attitudes towards intervention or simulation
Level 2a	<b>KNOWLEDGE and SKILLS</b> <i>Knowledge:</i> acquisition of concepts, procedures, or principles <i>Skills:</i> acquisition of thinking and problem solving, psychomotor, or social skills
Level 3	<b>BEHAVIORAL CHANGE</b> Transfer of learning to the workplace or willingness to apply new knowledge and skills
Level 4a	<b>ORGANIZATION PRACTICE</b> Wider changes in organization or delivery of care, attributable to educational program
Level 4b	<b>PATIENT BENEFITS</b> Improvement in health or well-being of patients as a direct result of educational program

<sup>a</sup> Hierarchy of increasing importance. Modified from Littlewood et al.<sup>3</sup>

most useful for researchers wishing to examine their work for methodologic rigor, as it includes a comprehensive list of review items and also has a growing body of validity evidence.<sup>5,7</sup> Less evidence is available for the BEME global rating scale, which includes a modified version of Kirkpatrick's hierarchy<sup>17</sup> of the outcomes of educational interventions. Kirkpatrick's hierarchy of levels is also included in the MERSQI scale, with higher points assigned to higher levels of outcomes. Kirkpatrick's hierarchy, also termed Kirkpatrick's pyramid (FIGURE), is employed widely by education experts to characterize the level of outcomes in an educational intervention. Authors could enhance the quality of their papers by including a discussion of their work in relation to the BEME global or Kirkpatrick frameworks. To date, these discussions rarely occur in *JGME* submissions.

The levels of Kirkpatrick's outcomes include (1) participation rates or learner satisfaction; (2) changes in attitudes, knowledge, and skills; (3) changes in behaviors; and (4) changes to the care system or patient outcomes. For

example, in a study comparing an interactive web-based program with readings and lectures for teaching residents techniques for smoking cessation, potential outcomes could be classified as follows:

Level 1: The percentage of residents completing each intervention; resident satisfaction with the interventions

Level 2a: Residents' attitudes about smoking cessation counseling

Level 2b: Residents' performance of smoking cessation counseling with standardized patients

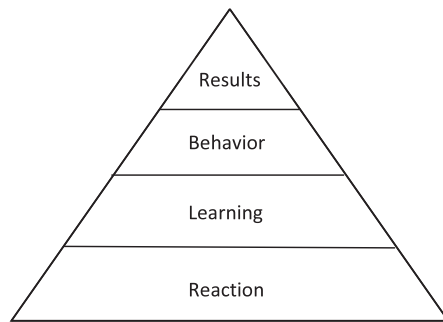
Level 3: Documentation of smoking cessation counseling in clinic charts; clinic patients' reports of smoking cessation counseling

Level 4b: Number of patients who quit smoking in residents' clinics.

In systematic reviews of education research, the majority of studies reported outcomes at Kirkpatrick levels 1 and 2.<sup>5,7</sup> Although undoubtedly easier to study, achievement of outcomes at these levels may not translate into effective, sustained changes in behaviors or improved patient outcomes. In general, outcomes reported were more often subjective rather than objective. Of greater concern is that outcomes are entirely absent in many studies: 19% in one 2008 review.<sup>7</sup> On average, the data analysis portion of reviewed papers received the highest quality ratings, while validation of assessment instruments received the lowest quality ratings.<sup>5,7,9,10</sup>

Other areas of methodologic concern found in literature reviews include (1) predominance of single-site studies; (2) small studies that are underpowered to find a difference between intervention and comparison groups; (3) lack of a comparison group or lack of description of the intervention for the comparison group (eg, description of usual teaching); (4) inadequate description of multifactorial interventions; (5) overconfidence in randomization to eliminate the influence of confounding variables (ie, bias); and (6) overuse of the single-group pretest/posttest strategy to assess differences, with resulting potential overestimates of the magnitude of the effect of the intervention.<sup>16,18</sup> In future issues of *JGME*, we will examine some of these issues in greater detail.

*JGME* editors suggest that authors consider evaluating their planned and ongoing work with the above-described instruments, the MERSQI, NOS, and BEME global scale, and other quality scales developed for specific interventions, such as online teaching modules.<sup>19</sup> In addition, authors should consider Kirkpatrick's hierarchy when formulating studies and considering outcome measures. These additional steps in reflection may produce a study and eventual manuscript that requires fewer revision cycles and is ultimately of greater value to consumers of medical education research.



4. Results = Change in patients or the system/organization's practices
3. Behavior = Change in behaviors or practice
2. Learning = Change in attitudes, knowledge, or skills
1. Reaction = Satisfaction

Adapted from BEME Guide No. 8<sup>13</sup>

FIGURE

KIRKPATRICK'S LEVELS OF LEARNING<sup>13</sup>

#### References

- 1 Harden RM, Lilley PM. Best evidence medical education: the simple truth. *Med Teach*. 2000;22(2):117–119.
- 2 Harden RM, Grant J, Buckley G, Hart IR. BEME guide no. 1: best evidence medical education. *Med Teach*. 1999;21(6):553–562.
- 3 Littlewood S, Ypinazar V, Margolis SA, Scherpbier A, Spencer J, Dornan T. Early practical experience and the social responsiveness of clinical education: systematic review. *BMJ*. 2005;331(7513):387–391.
- 4 Price EG, Beach MC, Gary TL, et al. A systematic review of the methodological rigor of studies evaluating cultural competence training of health professionals. *Acad Med*. 2005;80(6):578–586.
- 5 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):1002–1009.
- 6 Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. *JAMA*. 2008;300(10):1181–1196.
- 7 Reed DA, Beckman TJ, Wright SM, Levine RB, Kern DE, Cook DA. Predictive validity evidence for medical education research study quality instrument scores: quality of submissions to JGIM's medical education special issue. *J Gen Intern Med*. 2008;23(7):903–907.
- 8 Reed DA, Beckman TJ, Wright SM. An assessment of the methodologic quality of medical education research studies published in *The American Journal of Surgery*. *Am J Surg*. 2009;198(3):442–444.
- 9 Windish DM, Reed DA, Boonyasai RT, Chakraborti C, Bass EB. Methodologic rigor or quality improvement curricula for physician trainees: a systematic review and recommendations for change. *Acad Med*. 2009;84(12):1677–1692.
- 10 Cook DA, Levinson AJ, Garside S. Method and reporting quality in health professions education research: a systematic review. *Med Educ*. 2011;45(3):227–238.
- 11 Norman G. Research in medical education: three decades of progress. *BMJ*. 2002;324(7353):1560–1562.
- 12 Wells GA, Shea B, O'Connell D, et al. The Newcastle–Ottawa Scale (NOS) for assessing the quality of non-randomised studies in meta-analyses. Available at: [http://www.ohri.ca/programs/clinical\\_epidemiology/oxford.asp](http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp). Accessed February 22, 2011.
- 13 Best Evidence in Medical Education. BEME Guide No. 8: A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education. Steinert Y, Mann K, Naismith L, Chin K (reviewers). Available at: <http://www2.warwick.ac.uk/fac/med/beme/reviews/published/steinert/>. Accessed March 3, 2011.
- 14 von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, for the STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med*. 2007;147(8):573–577.
- 15 American Educational Research Association. Standards for reporting on empirical social science research in AERA publications. Available at: [http://www.aera.net/uploadedFiles/Opportunities/StandardsforReportingEmpiricalSocialScience\\_PDF.pdf](http://www.aera.net/uploadedFiles/Opportunities/StandardsforReportingEmpiricalSocialScience_PDF.pdf). Accessed April 8, 2011.
- 16 Colliver JA, McGaghie WC. The reputation of medical education research: quasi-experimentation and unresolved threats to validity. *Teach Learn Med*. 2008;20(2):101–103.
- 17 Kirkpatrick D. Revisiting Kirkpatrick's four-level model. *Training and Development*. 1996;50(1):54–59.
- 18 Cook DA, Beckman TJ. Reflections on experimental research in medical education. *Adv Health Sci Educ Theory Pract*. 2010;15(3):455–464.
- 19 Shortt SED, Guillemette J, Duncan AM, Kirby F. Defining quality criteria for online continuing medical education modules using modified nominal group technique. *J Contin Educ Health Prof*. 2010;30(4):246–250.
- 20 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):1002–1009.