Appraising the Quality of Medical Education Research Methods: The Medical Education Research Study Quality Instrument and the Newcastle–Ottawa Scale-Education

David A. Cook, MD, MHPE, and Darcy A. Reed, MD, MPH

Abstract

Purpose
The Medical Education Research Study Quality Instrument (MERSQI) and the Newcastle–Ottawa Scale-Education (NOS-E) were developed to appraise methodological quality in medical education research. The study objective was to evaluate the interrater reliability, normative scores, and between-instrument correlation for these two instruments.

Method
In 2014, the authors searched PubMed and Google for articles using the MERSQI or NOS-E. They obtained or extracted data for interrater reliability—using the intraclass correlation coefficient (ICC)—and normative scores. They calculated between-scale correlation using Spearman rho.

Results
Each instrument contains items concerning sampling, controlling for confounders, and integrity of outcomes. Interrater reliability for overall scores ranged from 0.68 to 0.95. Interrater reliability was "substantial" or better (ICC > 0.60) for nearly all domain-specific items on both instruments. Most instances of low interrater reliability were associated with restriction of range, and raw agreement was usually good. Across 26 studies evaluating published research, the median overall MERSQI score was 11.3 (range 8.9–15.1, of possible 18). Across six studies, the median overall NOS-E score was 3.22 (range 2.08–3.82, of possible 6). Overall MERSQI and NOS-E scores correlated reasonably well (rho 0.49–0.72).

Conclusions
The MERSQI and NOS-E are useful, reliable, complementary tools for appraising methodological quality of medical education research. Interpretation and use of their scores should focus on item-specific codes rather than overall scores. Normative scores should be used for relative rather than absolute judgments because different research questions require different study designs.

As medical education research continues to grow, it is increasingly important to appraise research study quality.[1–4] Instruments to facilitate critical appraisal support at least four distinct activities—namely, (1) evaluation of the quality of published evidence (e.g., in a systematic or critical literature review); (2) peer review (i.e., to help journal editors and reviewers in manuscript appraisal and decisions); (3) study planning and manuscript preparation (i.e., to help investigators design and describe rigorous studies); and (4) identification of areas for improvement in a field. Those using such instruments must understand the definitions of their various items (e.g., questions or appraisal standards), the validity evidence[5] supporting instrument scores, and how to interpret the scores in various contexts.

Many instruments to appraise the quality of clinical research exist,[6,7] but few have targeted medical education research. However, two instruments have been developed and used in recent years to appraise the methodological quality of studies of medical education—namely, the Medical Education Research Study Quality Instrument (MERSQI)[8] and the Newcastle–Ottawa Scale-Education (NOS-E).[9] A description and detailed validation of the MERSQI were published with its initial use.[8] However, no studies have summarized evidence from subsequent reviews using that instrument, such as interrater reliability and normative data. New data on interrater reliability would help to establish how the instrument performs when applied by investigators other than the development team. Normative data, in turn, would help answer questions relating to score interpretations (e.g., how good are scores for a given study or group of studies?) and would permit an exploration of patterns in study quality across different fields of education research. Although the NOS-E items have been reported,[9] a description of the instrument's development and operational criteria has never been published. Normative data and a synthesis of interrater reliability data would also support the use of this instrument's scores. Finally, contrasting the MERSQI and NOS-E might identify strengths and gaps in the design and content of each instrument.

With these issues in mind, our main objectives in the present study were to evaluate the interrater reliability, normative scores, and between-instrument correlations for the MERSQI and NOS-E, and to identify strengths and weaknesses for each instrument. Using

D.A. Cook is professor of medicine and medical education, Department of Medicine; associate director, Mayo Center for Online Learning; and research chair, Mayo Clinic Multidisciplinary Simulation Center, Mayo Clinic College of Medicine, Rochester, Minnesota.

D.A. Reed is associate professor of medicine and medical education, Department of Medicine, and senior associate dean of academic affairs, Mayo Medical School, Mayo Clinic College of Medicine, Rochester, Minnesota.

Correspondence should be addressed to David A. Cook, Division of General Internal Medicine, Mayo Clinic College of Medicine, Mayo 17, 200 First St. SW, Rochester, MN 55905; telephone: (507) 266-4156; e-mail: cook.david33@mayo.edu.

Acad Med. 2015;90:1067–1076.
First published online June 23, 2015
doi: 10.1097/ACM.0000000000000786

Messick's validity framework,[5] these analyses provide evidence regarding the content, internal structure, relationships with other variables, and potential consequences of scores derived from these instruments. Our secondary objective was to use this information to identify patterns of study quality across different fields of education research.

Method

Instrument development and description: MERSQI

The MERSQI was developed in 2007 as part of a study examining associations between funding for and quality of medical education research, and was "designed to measure the [methodological] quality of experimental, quasi-experimental, and observational studies."[8] Content domains and specific items were developed from literature on study quality and then iteratively revised. The final MERSQI included 10 items clustered in six domains (see Table 1). Further evidence to support score validity included very good interrater reliability, excellent intrarater reliability (same rater on different occasions), and favorable correlations with global quality ratings from two independent experts, the study's three-year citation rate, and the impact factor of the journal of publication. In a follow-up study, MERSQI scores demonstrated a significant favorable association with decisions of acceptance or rejection in a peer-reviewed journal.[10]

Instrument development and description: NOS-E

The original Newcastle–Ottawa Scale (NOS) was developed to evaluate the quality of nonrandomized comparative studies included in a meta-analysis of clinical research.[11] This instrument has become quite popular in clinical research (as of this writing, a PubMed search for "Newcastle–Ottawa Scale" revealed 299 citations, nearly all of which refer to the scale's application in appraising a body of evidence). The original scale for cohort studies included eight items clustered into the following three broad domains:

- *Selection of the study groups:* representativeness of the exposed cohort, selection of the nonexposed cohort, ascertainment of exposure, and demonstration that the outcome of interest was not present at the start of the study
- *Comparability of the groups:* comparability on the basis of the design or analysis
- *Ascertainment of the outcome of interest:* independent or blind assessment, follow-up sufficiently long for outcome to occur, and adequacy of follow-up.

According to the developers,[11]

> content validity of the NOS has been established based on a critical review of the items by several experts in the field who evaluated its clarity and completeness for the specific task of assessing the quality of studies to be used in a meta-analysis. Also, the NOS has been refined based on experience using it in several projects.

However, a recent empiric study of the NOS in clinical medicine found variable interrater reliability,[12] and this and one other critique[13] identified problems with vague and possibly arbitrary operational definitions.

In adapting this scale for our use in a meta-analysis of Internet-based education for health professionals,[9] we deleted three items (ascertainment of exposure, prestudy outcome assessment, and sufficiently long follow-up) because exposure was uniformly present and the outcomes (i.e., knowledge or skill) were almost-uniformly deficient at baseline, and because such outcomes do not lag in their appearance following intervention. For the remaining five items, we developed operational criteria and definitions to enable their use in education research (see Table 1) after careful review of the original NOS coding manual and iterative application, discussion, and revision based on a subsample of studies. We refer to this revised instrument as the Newcastle–Ottawa Scale-Education (NOS-E). The NOS-E has five items, each of which we consider a separate domain (see Table 1).

Interrater reliability of each instrument

In 2014 we obtained already-collected raw data from previously published work and used the intraclass correlation coefficient (ICC) to calculate interrater reliability for the 10 MERSQI items and 5 NOS-E items. We then pooled all available raw data into a single dataset and calculated the interrater reliability across all studies. We also calculated the raw agreement. For studies for which we did not have raw data, we obtained interrater reliability estimates from published reports.[8,14] We used thresholds proposed by Landis and Koch[15] to classify interrater reliability (0.21–0.4 = fair, 0.41–0.6 = moderate, 0.61–0.8 = substantial, and 0.81–1 = almost perfect).

Normative scores

To identify normative data in addition to those found in our own published work, we searched for studies using the MERSQI and NOS-E to appraise methodological quality. We searched PubMed using search terms "MERSQI OR Medical Education Research Study Quality Instrument" and used Google to identify all publications citing the original description of the MERSQI.[8] From the articles identified in these searches, one of us (D.A.C.) abstracted information on the education topic, total number of raters involved, number of original studies reviewed, mean overall and domain-specific scores, and overall score range. We also searched PubMed and Google for "Newcastle–Ottawa Scale AND education." We found only one additional potentially relevant study,[16] which used a very different variant of the NOS; thus, we did not include that study.

Correlation between instruments

We used Spearman rho to estimate correlation among overall scores from each instrument. We used SAS 9.3 (SAS, Cary, North Carolina) for all analyses. Because no human participants were involved in this research, ethical approval was not required.

Results

Comparison of the instruments

Table 1 contains a detailed description of each instrument and its operational criteria. Whereas each instrument contains items related to sampling, controlling for confounders, and integrity of outcomes, there are several salient differences. Most important, the MERSQI was designed for use with any nonqualitative research report, whereas the NOS-E was designed for quantitative comparative studies (i.e., a separate comparison group or a one-group pretest–posttest comparison). As such, the MERSQI applies more broadly but also omits potentially important issues related to comparative studies.

Second, the MERSQI definitions are more specific, whereas the NOS-E items require more rater judgment. For example,

Table 1

Characteristics of the MERSQI and the NOS-E[a]

Domain: item	Response options: scores[b]	Operational definitions
MERSQI[c]		
Study design	• Single-group cross-sectional or single-group posttest only: 1 • Single-group pretest and posttest: 1.5 • Nonrandomized, 2 group: 2 • Randomized controlled trial: 3	• Survey studies are cross-sectional. • Case–control and cohort studies (2 or more defined cohorts) are considered 2-group nonrandomized.
Sampling: institutions	• 1 institution: 0.5 • 2 institutions: 1 • 3 or more institutions: 1.5	• Number of institutions refers to origin of study participants (not study authors).
Sampling: response rate	• Not applicable • < 50% or not reported: 0.5 • 50%–74%: 1 • ≥ 75%: 1.5	• Response rate is the proportion of those eligible who completed the posttest or survey. For intervention studies, this is the proportion of those enrolled who completed the intervention evaluation. • Use "not applicable" only if a response rate truly does not apply (e.g., data obtained from a medical record or professional organization database).
Type of data	• Assessment by study participant: 1 • Objective: 3	• Observer ratings are considered objective.
Validity evidence for evaluation instrument scores	• Not applicable • Content: 1 • Internal structure: 1 • Relationships to other variables: 1	• Relevant content evidence would include using theory, guidelines, experts, and existing instruments to identify or refine the instrument. • Relevant internal structure evidence would include all reliability (internal consistency, interrater, interstation, and test–retest) and factor analysis. • Relevant evidence of relationships to other variables would include expert–novice comparisons and concurrent or predictive correlation with other variables. • Use "not applicable" only if the study does not measure a psychological construct *and* there is no instrument to rate (e.g., gender as the sole outcome); should be used very rarely.
Data analysis: sophistication	• Descriptive analysis only: 1 • Beyond descriptive analysis: 2	• Descriptive analyses include frequency, mean, and median. • Any test of statistical inference is considered "beyond descriptive."
Data analysis: appropriate	• Data analysis appropriate for study design and type of data: 1	• Considered "no" if there is a statistical error or if authors failed to analyze data at all.
Outcome	• Satisfaction, attitudes, perceptions, opinions, general facts: 1 • Knowledge, skills: 1.5 • Behaviors: 2 • Patient/health care outcome: 3	• General facts include participant demographics. • Knowledge/skills are in a test setting (paper, computer, simulation, or patients in a nonauthentic setting). • Behaviors are physician actions with real patients in a clinical context, or other activities in a real context. • Patient/health care outcomes are actual effects on real patients, programs, or society.
NOS-E[d]		
Representativeness of intervention group	• Not representative: 0 • Very or somewhat representative of average learner in community: 1	• Representativeness is judged in relation to the community of eligible learners (e.g., the entire school year class, training program, or faculty). • "Very" representative indicates that all or a random sample of eligible learners enrolled. • "Somewhat" representative indicates that 75%–99% of eligible learners enrolled, or eligible but unenrolled learners are compared with those enrolled and found to be similar. • "Not representative" indicates either that sampling is not described or < 75% of eligible learners enrolled. • An intervention group defined by completion of the intervention (e.g., completers versus noncompleters) is "not representative" regardless of proportion.
Selection of comparison group	• No separate comparison group (e.g., single-group pretest–posttest): 0 • Drawn from a different community: 0 • Drawn from the same community: 1	• "Same" community indicates that there is no obvious reason to suspect systematic difference between intervention and comparison group. • A "different" community would include a different training program, a historical cohort with different exposures, or subjects substantially different in characteristics such as age, gender, performance, or desire to participate (e.g., volunteers versus nonvolunteers, experts versus novices).

(*Table continues*)

Table 1
(Continued)

Domain: item	Response options: scores[b]	Operational definitions
Comparability of comparison group	• No separate comparison group: 0 • Randomized study: o Allocation not concealed: 1 o Allocation concealed: 2 • Nonrandomized study: o Controlled for 1 subject characteristic: 1 o Controlled for 2 or more subject characteristics: 2	• Randomized and nonrandomized studies are coded separately. • Allocation is considered concealed if enrollment, consent, or baseline assessment preceded randomization. • Controlling for subject characteristics requires statistical covariate analysis (e.g., including baseline scores or training level in a multivariate model); directly comparing characteristics between groups (e.g., t-test comparing baseline demographics) is insufficient. • Relevant subject characteristics include (but are not limited to) scores/grades on a pretest, standardized test, or earlier course, and grade point average/class rank.
Study retention[e]	• Poor retention could introduce bias: 0 • Retention unlikely to introduce bias: 1	• High if ≥ 75% of those enrolled provided outcome data, or if authors described those lost to follow-up. • Authors must report the number providing data (percentages or proportions completing the study are insufficient unless the denominator is specified).
Blinding of assessment[e]	• Outcome assessment not blinded: 0 • Outcome assessment blinded: 1	• Blinded if the assessor cannot be influenced by group assignment. • Assessments that do not require human judgment (e.g., multiple-choice tests or computer-scored performance) are considered to be blinded. • One-group studies are not blinded unless scoring does not require judgment or authors describe a plausible method for hiding the timing of assessment. • Participant-reported outcomes are never blinded.

Abbreviations: MERSQI indicates Medical Education Research Study Quality Instrument; NOS-E, Newcastle–Ottawa Scale-Education.
[a]The MERSQI and the NOS-E are instruments used to appraise the methodological quality of medical education original research studies, typically in the process of a literature review of a field or topic in medical education.
[b]Investigators appraise study quality using each domain/item and assign corresponding scores. MERSQI item scores are weighted to reflect features of higher study quality, whereas NOS-E items are weighted equally (1 point each) except for comparability of comparison group (maximum 2 points). Each item (and its score) for either instrument should ideally be reported separately. Scores may also be summed to create a final score (maximum possible 18 for MERSQI, 6 for NOS-E).
[c]The MERSQI has 6 domains and 10 items. The "sampling" and "data analysis" domains have 2 items each. Each of the three sources of validity evidence (content, internal structure, and relationships to other variables) is counted as a separate item.
[d]The NOS-E has five domains, each with one item.
[e]Study retention and blinding are scored separately for each outcome. The maximum score is typically reported, but in certain applications it may be preferable to report outcome-specific NOS-E scores.

the original NOS items regarding the selection of the intervention group and the similarity of the community from which the comparison group is drawn leave much room for judgment regarding degrees of similarity. We developed specific operational criteria during use (see Table 1), but these still require judgment. This is both a strength of the NOS-E (because it allows integration of more data in making judgments) and a weakness (because it requires greater inference from the rater, with concomitantly greater room for error).

Third, in judging the potential for bias in outcome measurement, both instruments use similar coding for participant retention, but the MERSQI evaluates objective (versus subjective) outcome assessment, whereas the NOS-E evaluates blinded (versus unblinded) assessment.

Interrater reliability

Table 2 shows the interrater reliability for both instruments as applied to original research studies in seven topic areas (e.g., Internet-based instruction, virtual patients, simulation-based education, general internal medicine).[8–10,14,17–23] Reliability was "substantial" or better (ICC > 0.60) for nearly all items, and "almost perfect" (ICC > 0.80) for many. For the few instances of low interrater reliability, raw agreement was usually good; we return to this point in the Discussion. Four other studies[24–27] reported interrater reliabilities for overall MERSQI scores ranging from 0.68 to 0.89.

Normative scores

Tables 3 and 4 report the results of 28 reviews whose authors applied the MERSQI (n = 26)[8,10,14,17,19–40] and NOS-E (n = 6)[9,17,18,20–22] to appraise study quality, along with means and medians. As explained later, these normative data should not be used to classify studies in absolute terms.

Correlation between MERSQI and NOS-E scores

All projects using the NOS-E also used the MERSQI to code at least half the studies. For studies coded with both instruments, we found score correlations that were statistically significant (all $P < .0001$) and relatively strong: rho = 0.72 for studies of Internet-based instruction, rho = 0.49 for studies of virtual patients, and rho = 0.60 for studies of simulation-based training.

Discussion

In this report, we provide information on two instruments for appraising the

Table 2

Interrater Agreement for MERSQI and NOS-E Scores[a]

Domain: item	Interrater agreement[b] reported in published reviews							
	Medical education[8] (n = 210), 6 raters[c]	General internal medicine[10] (n = 100), 6 raters[c]	Internet-based instruction[9,18] (n = 266), 4 raters[d]	Virtual patients[17] (n = 45), 2 raters	Internet-based instruction[19] (n = 128), 3 raters	Simulation-based education[20–23] (n = 748), 8 raters	Physical exam[14] (n = 14), 3 raters	Combined[e] (N = 1,106)
MERSQI								
Study design	0.86	0.98	0.95 (95%)	0.95 (96%)	0.92 (93%)	0.94 (94%)	1	0.95 (94%)
Sampling: institutions	0.98	0.97		0.72 (93%)	0.96 (98%)	0.63 (91%)	0.4[f]	0.71 (93%)
Sampling: response rate	0.89	0.82		0.61 (82%)	0.95 (96%)	0.37 (74%)	0.4[f]	0.48 (79%)
Type of data	0.90	0.78	0.89 (97%)	0.0 (93%)	0.92 (98%)	0.71 (94%)	1	0.78 (95%)
Validity evidence: content	0.89	0.93		0.68 (84%)	0.90 (95%)	0.46 (75%)	0	0.54 (79%)
Validity evidence: internal structure	0.91	0.90		0.51 (87%)	0.88 (95%)	0.76 (91%)	1	0.77 (92%)
Validity evidence: relationships with other variables	0.72	0.91		0.23 (89%)	0.76 (98%)	0.32 (80%)	1	0.38 (84%)
Data analysis: appropriate	0.94	0.76		0.55 (88%)	0.53 (90%)	0.34 (94%)	1	0.50 (93%)
Data analysis: sophistication	0.86	0.96		1.0 (100%)	0.87 (98%)	0.68 (96%)	1	0.74 (97%)
Outcome	0.83	0.83	0.91 (97%)	1.0 (100%)	0.93 (98%)	0.76 (90%)	1	0.79 (92%)
Overall score				0.77	0.95	0.75		0.79
NOS-E								
Representativeness of intervention group			0.64 (82%)	0.87 (93%)		0.62 (88%)		0.66 (86%)
Selection of comparison group			0.76 (92%)	0.37 (85%)		0.29 (86%)		0.48 (88%)
Comparability of comparison group			0.77 (79%)	0.58 (71%)		0.72 (71%)		0.75 (74%)
Study retention			0.62 (85%)	0.65 (84%)		0.36 (78%)		0.44 (80%)
Blinding			0.75 (90%)	0.75 (91%)		0.58 (80%)		0.62 (83%)
Overall score			0.84	0.85		0.80		0.82

Abbreviations: MERSQI indicates Medical Education Research Study Quality Instrument; NOS-E, Newcastle–Ottawa Scale-Education.

[a]The MERSQI and the NOS-E are instruments used to appraise the methodological quality of medical education original research studies, typically in the process of a literature review of a field or topic in medical education.

[b]Column headings indicate the review topic, number of articles included, and total number of raters involved; two raters assessed each article. In the data cells, numbers indicate the interrater reliability (intraclass correlation or kappa coefficient) of quality assessments as reported in the original reviews. Raw agreement is shown as a percentage in parentheses, when available. Blank cells indicate no data available (i.e., the review team did not code original studies using this instrument or item).

[c]All raters were part of the team that developed the MERSQI.

[d]All raters were part of the team that developed the NOS-E.

[e]Combined analysis includes all of the studies of Internet-based instruction, virtual patients, and simulation-based education involving both instruments, and 54 studies from the general internal medicine review. After merging articles present in > 1 dataset, the maximum N = 1,106 for MERSQI scores and the maximum N = 1,058 for NOS-E scores. Four other studies[24–27] (not included in this table) reported interrater reliabilities for overall MERSQI scores ranging from 0.68 to 0.89.

[f]Reported 0.4 for sampling domain as a whole; did not break down institutions and response rate separately.

Table 3

Normative Data From 26 Studies That Applied the MERSQI[a]

First author (year)	Topic	No. of articles	No. of raters	Overall scores		Domain-specific mean scores[b]					
				Mean	Range	Study design	Sampling	Data type	Validity	Analysis	Outcome
Reed (2007)[8]	All medical education	210	5	10	5–16	1.3	1.9	1.9	0.7	2.6	1.4
Reed (2008)[10]	General medicine education	100	4	9.6	5–15.5	1.3	1.9	1.8	0.6	2.5	1.3
Reed (2009)[28]	Surgery education	19	3	11	7.5–15	1.4	2.2	2.3	1.1	2.7	1.4
Windish (2009)[24]	Quality improvement education	18		9.9	5–14	1.5	1.8	2	0.3	2.3	1.7
Cook (2010)[17]	Virtual patients	45	2	12.1	9.5–16	2.3	1.9	3	0.6	2.8	1.5
Lie (2011)[29]	Cultural competency education	7	4		5.5–12						
Reed (2010)[25]	Duty hours	64	3	11.9	6–17.5	2.2	1.6	24	0.9	2.7	2.2
Cook (2011)[19]	Internet-based education	133	3	11.7	6–16	2.2	1.9	2.6	0.7	2.8	1.5
Cook (2011)[20]	Simulation-based education (comparison with no intervention)	627	8	11.5	6–17	1.9	1.8	2.7	0.7	2.8	1.6
Fletcher (2011)[30]	Duty hours	27	3	15.1	12–16.7						
Kothari (2011)[26]	Substance abuse education	31	5	10.4	6.3–14.8	1.6	1.7	2.3	1	2.6	1.3
Ma (2011)[31]	Central venous catheter training	20	2	12.6	9–15.6	1.9	1.7	2.3	1.4	2.9	2.1
Yucha (2011)[32]	All nursing education	133	4	9.8	6–14.5	1.4	1.6	1.7	1	2.6	1.2
Cook (2012)[21]	Simulation-based education (comparison with nonsimulation intervention)	97	8	12.8	6–16.5	2.7	1.9	2.9	0.8	2.9	1.7
Johnson (2012)[33]	Lesbian/gay nursing research	40	1	9.4	7–14.4	1.1	1.9	1.1	1.1	1.9	1.1
Quartey (2012)[34]	Complementary medicine education	12	2	10.8	8.5–13.5	1.8	0.9	1.8	0.4	1.4	1.6
van der Leeuw (2012)[35]	Residency education	97	1	12.9	9–15.6						
Batt-Rawden (2013)[36]	Empathy education	15	2	10.1	6.5–14						
Brennan (2013)[37]	Prescribing behaviors	64	2	13.3[c]	6–18	1.7		3		2.8	2
Cheston (2013)[27]	Social media	14	2	8.9	5–15.5	1.4			0.9	2	
Cook (2013)[22]	Simulation-based education (comparison with other simulation intervention)	303	8	12.3	6.5–16	2.7	1.8	2.7	0.7	2.8	1.5
Cook (2013)[23]	Simulation-based assessment	417	6	12.3	5–17	1.4	1.9	2.9	1.8	2.9	1.5
de Jong (2013)[38]	Work-based learning	22	2	11.8[d]	8–14.5						
Eaton (2013)[39]	Internal medicine residency training	223	5	11.1		1.4	1.8	2.2	1.3	2.7	1.6
Mookherjee (2013)[14]	Physical exam education	14	3	9	6.5–11						
Schneider (2013)[40]	Nursing education	37	2			1.6	1.8	1.6			1.3
Weighted average				*11.5*		*1.8*	*1.8*	*2.5*	*1.0*	*2.7*	*1.5*
Median				*11.3*	*8.9–15.1*	*1.6*	*1.8*	*2.3*	*0.9*	*2.7*	*1.5*

Abbreviation: MERSQI indicates Medical Education Research Study Quality Instrument.

[a]The MERSQI is an instrument used to appraise the methodological quality of medical education original research studies, typically in the process of a literature review of a field or topic in medical education. Throughout the table, blank cells indicate that insufficient information was reported.

[b]Each domain could score a maximum of three points.

[c]Raw scores based on a maximum of 13.5 (excluding response rate and validity codes) and then standardized to a maximum of 18.

[d]Median score.

Table 4
Normative Data From Six Studies That Applied the NOS-E[a]

Author (year)	Topic	No. of articles	No. of raters	Overall scores		Domain-specific mean scores[b]				
				Mean	Range	Representativeness, intervention group	Selection, comparison group[c]	Comparability[c] (score by study design[d])	Follow-up	Blinding
Cook (2008, 2010)[9,18]	Internet-based education	254	4	2.83	0–6	0.43	0.75	0.7 (R = 1.39, N = 0.23)	0.72	0.68
Cook (2010)[17]	Virtual patients	45	2	3.22	0–6	0.49	0.88	0.82 (R = 1.19, N = 0.23)	0.71	0.75
Cook (2011)[20]	Simulation-based education (comparison with no intervention)	627	8	2.08	0–6	0.22	0.93	1.09 (R = 1.53, N = 0.16)	0.70	0.49
Cook (2012)[21]	Simulation-based education (comparison with nonsimulation intervention)	97	8	3.82	1–6	0.35	0.95	1.04 (R = 1.39, N = 0.04)	0.82	0.66
Cook (2013)[22]	Simulation-based education (comparison with other simulation intervention)	303	8	3.58	0–6	0.19	0.93	1.12 (R = 1.51, N = 0.10)	0.72	0.61
Weighted average				*2.70*		*0.28*	*0.89*	*0.99 (R = 1.47, N = 0.15)*	*0.71*	*0.57*
Median				*3.22*	*2.08–3.82*	*0.35*	*0.93*	*1.04*	*0.72*	*0.66*

Abbreviation: NOS-E indicates Newcastle–Ottawa Scale-Education.
[a]The NOS-E is an instrument used to appraise the methodological quality of medical education original research studies, typically in the process of a literature review of a field or topic in medical education.
[b]Each domain could score a maximum of one point except Comparability, which could score a maximum of two points (see Table 1 for scoring rubric).
[c]Includes only studies with a separate comparison group.
[d]Score stratified by study design: R = randomized trials; N = nonrandomized two-group studies (maximum two points for either design).

methodological quality of education research studies and offer the first detailed description of the NOS-E. Our findings indicate that interrater reliability is generally very good, although restriction of range attenuates the reliability for some items. We also report normative data for these instruments' scores.

Limitations

We did not collect new data for this study but, rather, synthesized and reanalyzed data from previously published reviews. However, in many cases the interrater reliabilities had not previously been reported, and our study is the first known synthesis of normative data for the instruments studied. Neither the MERSQI nor the NOS-E is perfect, as outlined below, but each plays a valuable and potentially complementary role in the appraisal of study quality. Our search for studies using the MERSQI and NOS-E could be incomplete, but the studies we identified appear sufficient to provide useful normative data. Only a limited number of unique raters have used either instrument, and all of the NOS-E studies involved at least one of the original developers. As such, the performance of these instruments in others' hands remains incompletely understood.

Integration with other literature

The original NOS has been criticized for low interrater reliability[12] and limitations in the operational definitions.[12,13] Reliability estimates for all NOS-E items were consistently higher than those for the NOS,[12] perhaps because the NOS-E includes more detailed operational definitions.

The quality of research reporting is a separate but related field.[4] Studies in both clinical medicine[41–44] and medical education[19,45–47] document frequent deficiencies in reporting quality. Because poor reporting can impair the appraisal of study methods,[9,46,48] methodological quality scores depend in part on the quality of reporting within published articles. We advocate the use of accepted standards, such as those found at the EQUATOR Network,[49] to ensure complete and transparent reporting.

Some meta-analyses have found significantly different effects for studies of low versus high methodological

quality,[9,20,22] yet differences in study design typically appear to have limited impact on quantitative outcomes.[19,50–53] However, even if quantitative results are similar, studies of higher quality allow stronger inferences (e.g., about causality or the meaningfulness of outcomes), which underscores the importance of evaluating methodological quality and using this information when interpreting study results.

Patterns in methodological quality and implications for new research

The variation in scores across reviews (Tables 3 and 4) reflects both differences in criteria used to select studies for inclusion in a given review and differences in research traditions or tacit standards in a given educational field. Because the MERSQI applies to a broad range of study designs, inclusion criteria may have greater influence on between-review variability for MERSQI scores than for NOS-E scores, because the latter are used only for evaluating comparative studies.

Study designs and outcomes had relatively low median MERSQI scores. However, selection of both design and outcome are dependent on the research question; we discourage investigators from aspiring to an inappropriate design or outcome for the sole purpose of meeting an arbitrary quality score. As noted below, score norms for study design and outcome vary for different research questions.

Reflection on suboptimal quality scores suggests four areas for improvement in planning and/or reporting research methods.

- First, in the studies we evaluated, samples were rarely appraised as representative; this may reflect nonsystematic sampling (a methodological weakness) or failure to describe selection processes (a reporting deficiency). This issue could be addressed by representative sampling and complete reporting.
- Second, comparison groups were nearly always selected from the same community, but comparability was appraised as suboptimal. This could be improved in nonrandomized studies by using covariate analysis to adjust for baseline scores or other demographic features, and in randomized studies by ensuring concealed allocation.
- Third, validity evidence was infrequently reported, even for studies focused on the evaluation of assessment instruments.[23] Rigorous frameworks for evaluating validity evidence are well established.[5,54,55]
- Fourth, although assessments were usually objective, follow-up was incomplete, and blinding (which includes objectively scored tests such as multiple-choice exams) was only done about two-thirds of the time. Both high retention and blinding of assessment are essential in minimizing study bias.[56,57]

Implications for appraising research

As we commented earlier, neither the MERSQI nor the NOS-E is perfect. Although correlation was relatively strong, the scores of one instrument still explained only 24% to 52% of the variance in the other. The MERSQI focuses on design issues and is quite objective, whereas the NOS-E weighs the implications of study procedures and requires more judgment. The MERSQI lacks items on blinding and comparability of cohorts, whereas the NOS-E lacks items on objective assessment, validity evidence, data analysis, and level of outcomes. Because all of these methodological features influence quality, and because increased subjectivity is both a strength and a weakness,[58] we believe that these instruments complement rather than replace one another. Future quality coding efforts might either use both instruments together or modify one instrument to include omitted items present on the other. Also, investigators using the NOS-E to evaluate studies measuring long-term retention or delayed outcomes may wish to restore the item deleted from the original NOS regarding "sufficient delay before measurement."

Interrater reliability was high for most items from both instruments, but ICCs were low in several studies for MERSQI "relationships to other variables" and "appropriateness of analysis" items, and for the NOS-E "selection of comparison group" item. However, the raw agreement for these items was relatively high. In these and several other instances of low interrater reliability, the vast majority of original studies received the same code (e.g., nearly all studies were judged to have appropriate analyses). This reduces the variability of scores, which in turn increases the relative contribution of rater error such that even a few disagreements will substantially lower the ICC (i.e., the psychometric phenomenon of ceiling effect or restriction of range[59]). An extreme example is found in the coding of "data type" for virtual patients, for which three disagreements resulted in ICC = 0. Reliability coefficients should not be ignored, but they are not the sole indicator of interrater agreement, and rote adherence to fixed values may be inappropriate. Rather, interrater reliability should be used to identify items requiring greater rater training and consensus and/or methodological features that tend to be poorly reported (and thus are difficult to discern). Most important, these issues highlight the need to involve more than one reviewer when coding quality, and to come to consensus on final codes.

Two other patterns evident in the interrater reliabilities underscore the need for rigorous rater training. First, the interrater reliability for the MERSQI is generally higher than that for the NOS-E. Because in the studies we evaluated, the same raters coded both instruments (usually concurrently), this most likely reflects the less objective items and less-defined operational criteria of the NOS-E. Second, interrater reliability was highest for studies coded by the original development team and was typically lower for larger groups of raters in which training and standardization might be more challenging. We have found that some degree of training or experience in general principles of study design is required for accurate coding and that iterative pilot testing and discussion are typically required to achieve high interrater agreement. Moreover, investigators must use standardized operational criteria and consultation with experienced users if scores are to be comparable with other reviews (external generalization). We are currently developing materials to facilitate such training and standardization.

Normative data such as those in Tables 3 and 4 provide a benchmark for score interpretations and facilitate comparisons across fields. However, such data should not be used to define absolute standards of high or low research quality, and cross-topic comparisons should be pursued with care, because scores vary greatly depending on topic- and field-specific factors. First,

different research questions inherently require different study designs, and higher-level outcomes are not inherently better.[60] Second, fields at different developmental stages and with different research traditions may command different standards. Raising the bar may be appropriate for some but not all fields. Third, reviews that restrict inclusion to specific study designs or outcomes will naturally have higher (or lower) scores. For example, reviews that include surveys and descriptive reports will have lower study design scores than reviews limited to comparative trials; a meta-analysis limited to randomized trials may have higher scores than a narrative review with broader inclusion criteria; and reviews focused on validity evaluations or restricted to higher outcomes (e.g., behaviors) will naturally have higher scores for validity and outcomes, respectively, than reviews without such restrictions. As such, we discourage the use of the normative data to interpret or classify study quality in absolute terms, but instead encourage their use as a reference point against which to judge study quality in relation to other bodies of similar evidence.

Finally, total quality scores have limited applications. The MERSQI was originally developed to provide a single score with which to explore associations with study funding.[8] Yet, instruments appraising methodological quality are most often employed in systematic and nonsystematic reviews of original research, in which cases it is preferable to focus on individual instrument items rather than a total quality score. Although most of the studies using the MERSQI reported item-specific scores, few accounted for this information when synthesizing study results. We remind those conducting literature reviews that "the degree to which reviewers explore the strengths, weaknesses, heterogeneity and gaps in the evidence determines in large part the value of the review."[61] Domain-specific MERSQI and NOS-E codes are intended to facilitate the critical interpretation of individual studies or groups of similar-quality studies. This can be done quantitatively (e.g., using subgroup meta-analysis) or qualitatively (in critical synthesis narrative review).

Conclusions

The MERSQI and NOS-E are useful, reliable, and complementary tools for appraising the methodological quality of medical education research. The validity evidence we presented herein supports the content (description and contrasting of items), internal structure (interrater reliability), and relationships with other variables (between-instrument score correlation) of inferences drawn from their scores.[5] The discussion of score norms, interpretations, and uses anticipates potential consequences.[55]

The median normative scores do not indicate high/low quality thresholds because different research questions require different study designs; these data should be used for relative rather than absolute judgments. Rater training is essential prior to using these instruments. Interpretations should focus on item-specific rather than overall scores.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: Reported as not applicable.

References

1 Harden RM, Grant J, Buckley G, Hart IR. BEME guide no. 1: Best evidence medical education. Med Teach. 1999;21:553–562.
2 Wolf FM. Methodological quality, evidence, and Research in Medical Education (RIME). Acad Med. 2004;79(10 suppl):S68–S69.
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;79:931–938.
4 Cook DA, Bowen JL, Gerrity MS, et al. Proposed standards for medical education submissions to the Journal of General Internal Medicine. J Gen Intern Med. 2008;23:908–913.
5 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med. 2006;119:166.e7–166.16.
6 Deeks J, Dinnes J, D'Amico R, Sowden A, Sakarovitch C. Evaluating non-randomised intervention studies. Health Technol Assess. 2003;7:186.
7 Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. Int J Epidemiol. 2007;36:666–676.
8 Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association between funding and quality of published medical education research. JAMA. 2007;298:1002–1009.
9 Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: A meta-analysis. JAMA. 2008;300:1181–1196.
10 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;23:903–907.
11 Wells GA, Shea B, O'Connell D, et al. The Newcastle–Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm. Accessed March 19, 2015.
12 Hartling L, Milne A, Hamm MP, et al. Testing the Newcastle–Ottawa Scale showed low reliability between individual reviewers. J Clin Epidemiol. 2013;66:982–693.
13 Stang A. Critical evaluation of the Newcastle–Ottawa Scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol. 2010;25:603–605.
14 Mookherjee S, Pheatt L, Ranji SR, Chou CL. Physical examination education in graduate medical education—a systematic review of the literature. J Gen Intern Med. 2013;28:1090–1099.
15 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–174.
16 Straus SE, Soobiah C, Levinson W. The impact of leadership training programs on physicians in academic medical centers: A systematic review. Acad Med. 2013;88:710–723.
17 Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: A systematic review and meta-analysis. Acad Med. 2010;85:1589–1602.
18 Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Instructional design variations in Internet-based learning for health professions education: A systematic review and meta-analysis. Acad Med. 2010;85:909–922.
19 Cook DA, Levinson AJ, Garside S. Method and reporting quality in health professions education research: A systematic review. Med Educ. 2011;45:227–238.
20 Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: A systematic review and meta-analysis. JAMA. 2011;306:978–988.
21 Cook DA, Brydges R, Hamstra SJ, et al. Comparative effectiveness of technology-enhanced simulation versus other instructional methods: A systematic review and meta-analysis. Simul Healthc. 2012;7:308–320.
22 Cook DA, Hamstra SJ, Brydges R, et al. Comparative effectiveness of instructional design features in simulation-based education: Systematic review and meta-analysis. Med Teach. 2013;35:e867–e898.
23 Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. Acad Med. 2013;88:872–883.
24 Windish DM, Reed DA, Boonyasai RT, Chakraborti C, Bass EB. Methodological rigor of quality improvement curricula for physician trainees: A systematic review and recommendations for change. Acad Med. 2009;84:1677–1692.
25 Reed DA, Fletcher KE, Arora VM. Systematic review: Association of shift length, protected sleep time, and night float with patient care, residents' health, and education. Ann Intern Med. 2010;153:829–842.

26 Kothari D, Gourevitch MN, Lee JD, et al. Undergraduate medical education in substance abuse: A review of the quality of the literature. Acad Med. 2011;86:98–112.

27 Cheston CC, Flickinger TE, Chisolm MS. Social media use in medical education: A systematic review. Acad Med. 2013;88:893–901.

28 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;198:442–444.

29 Lie DA, Lee-Rey E, Gomez A, Bereknyei S, Braddock CH 3rd. Does cultural competency training of health professionals improve patient outcomes? A systematic review and proposed algorithm for future research. J Gen Intern Med. 2011;26:317–325.

30 Fletcher KE, Reed DA, Arora VM. Patient safety, resident education and resident well-being following implementation of the 2003 ACGME duty hour rules. J Gen Intern Med. 2011;26:907–919.

31 Ma IW, Brindle ME, Ronksley PE, Lorenzetti DL, Sauve RS, Ghali WA. Use of simulation-based education to improve outcomes of central venous catheterization: A systematic review and meta-analysis. Acad Med. 2011;86:1137–1147.

32 Yucha CB, Schneider BS, Smyer T, Kowalski S, Stowers E. Methodological quality and scientific impact of quantitative nursing education research over 18 months. Nurs Educ Perspect. 2011;32:362–368.

33 Johnson M, Smyer T, Yucha C. Methodological quality of quantitative lesbian, gay, bisexual, and transgender nursing research from 2000 to 2010. ANS Adv Nurs Sci. 2012;35:154–165.

34 Quartey NK, Ma PH, Chung VC, Griffiths SM. Complementary and alternative medicine education for medical profession: Systematic review. Evid Based Complement Alternat Med. 2012;2012:656812.

35 van der Leeuw RM, Lombarts KM, Arah OA, Heineman MJ. A systematic review of the effects of residency training on patient outcomes. BMC Med. 2012;10:65.

36 Batt-Rawden SA, Chisolm MS, Anton B, Flickinger TE. Teaching empathy to medical students: An updated, systematic review. Acad Med. 2013;88:1171–1177.

37 Brennan N, Mattick K. A systematic review of educational interventions to change behaviour of prescribers in hospital settings, with a particular emphasis on new prescribers. Br J Clin Pharmacol. 2013;75:359–372.

38 de Jong J, Visser M, Van Dijk N, van der Vleuten C, Wieringa-de Waard M. A systematic review of the relationship between patient mix and learning in work-based clinical settings. A BEME systematic review: BEME guide no. 24. Med Teach. 2013;35:e1181–e1196.

39 Eaton JE, Reed DA, Aboff BM, et al. Update in internal medicine residency education: A review of the literature in 2010 and 2011. J Grad Med Educ. 2013;5:203–210.

40 Schneider BS, Nicholas J, Kurrus JE. Comparison of methodologic quality and study/report characteristics between quantitative clinical nursing and nursing education research articles. Nurs Educ Perspect. 2013;34:292–297.

41 Pitkin RM, Branagan MA. Can the accuracy of abstracts be improved by providing specific instructions? A randomized controlled trial. JAMA. 1998;280:267–269.

42 Clarke M, Chalmers I. Discussion sections in reports of controlled trials published in general medical journals: Islands in search of continents? JAMA. 1998;280:280–282.

43 Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. JAMA. 2006;295:1147–1151.

44 Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. Lancet. 2005;365:1159–1162.

45 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;80:578–586.

46 Cook DA, Beckman TJ, Bordage G. Quality of reporting of experimental studies in medical education: A systematic review. Med Educ. 2007;41:737–745.

47 Cook DA, Beckman TJ, Bordage G. A systematic review of titles and abstracts of experimental studies in medical education: Many informative elements missing. Med Educ. 2007;41:1074–1081.

48 Huwiler-Müntener K, Jüni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. JAMA. 2002;287:2801–2804.

49 Enhancing the Quality and Transparency of Health Research (EQUATOR) Network. www.equator-network.org. Accessed March 29, 2015.

50 Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med. 2000;342:1878–1886.

51 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342:1887–1892.

52 Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA. 2001;286:821–830.

53 Wilson DB, Lipsey MW. The role of method in treatment effectiveness research: Evidence from meta-analysis. Psychol Methods. 2001;6:413–429.

54 Kane MT. Validation. In: Brennan RL, ed. Educational Measurement. 4th ed. Westport, Conn: Praeger; 2006:17–64.

55 Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ Theory Pract. 2014;19:233–250.

56 Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ. 2009;338:b2393.

57 Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. BMJ. 2008;336:601–605.

58 Norman GR, Van der Vleuten CP, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. Med Educ. 1991;25:119–126.

59 Feldt LS, Brennan RL. Reliability. In: Linn RL, ed. Educational Measurement. 3rd ed. New York, NY: American Council on Education and Macmillan; 1989:105–146.

60 Cook DA, West CP. Reconsidering the focus on "outcomes research" in medical education: A cautionary note. Acad Med. 2013;88:162–167.

61 Cook DA, West CP. Conducting systematic reviews in medical education: A stepwise approach. Med Educ. 2012;46:943–952.