REVIEW

How Reliable Are Assessments of Clinical Teaching?

A Review of the Published Instruments

Thomas J. Beckman, MD, Amit K. Ghosh, MD, David A. Cook, MD, Patricia J. Erwin, MLS, Jayawant N. Mandrekar, PhD

BACKGROUND: **Learner feedback is the primary method for evaluating clinical faculty, despite few existing standards for measuring learner assessments.**

OBJECTIVE: **To review the published literature on instruments for evaluating clinical teachers and to summarize themes that will aid in developing universally appealing tools.**

DESIGN: **Searching 5 electronic databases revealed over 330 articles. Excluded were reviews, editorials, and qualitative studies. Twenty-one articles describing instruments designed for evaluating clinical faculty by learners were found. Three investigators studied these papers and tabulated characteristics of the learning environments and validation methods. Salient themes among the evaluation studies were determined.**

MAIN RESULTS: **Many studies combined evaluations from both outpatient and inpatient settings and some authors combined evaluations from different learner levels. Wide ranges in numbers of teachers, evaluators, evaluations, and scale items were observed. The most frequently encountered statistical methods were factor analysis and determining internal consistency reliability with Cronbach's α. Less common methods were the use of test-retest reliability, interrater reliability, and convergent validity between validated instruments. Fourteen domains of teaching were identified and the most frequently studied domains were interpersonal and clinical-teaching skills.**

CONCLUSIONS: **Characteristics of teacher evaluations vary between educational settings and between different learner levels, indicating that future studies should utilize more narrowly defined study populations. A variety of validation methods including temporal stability, interrater reliability, and convergent validity should be considered. Finally, existing data support the validation of instruments comprised solely of interpersonal and clinical-teaching domains.**

Received from the Department of Internal Medicine (TJB, AKG, DAC), Department of Medicine, Mayo Clinic College of Medicine, Mayo Clinic and Mayo Foundation; Plummer Medical Library (PJE), Mayo Clinic College of Medicine; Department of Health Sciences Research (JNM), Division of Biostatistics, Mayo Clinic and Mayo Foundation, Rochester, Minn.

Presented at the SGIM annual meeting, Chicago, Ill, May 14, 2004.

Address correspondence and requests for reprints to Dr. Beckman: Division of General Internal Medicine, Mayo Clinic, 200 First Street SW, Rochester, MN 55905 (e-mail: Beckman.Thomas@mayo.edu).

KEY WORDS: **validity; evaluation studies; medical faculty.**
J GEN INTERN MED 2004;19:971–977.

Learner feedback is one of the major criteria for evaluating clinical faculty at academic medical centers. Despite this, clinical teachers have questioned policies that depend heavily upon assessments by learners,[1] and they have also questioned the reliability of learner evaluations.[2] Considering the major impact of learner evaluations on the careers of medical educators, it is essential that assessments be reliable. Supporting this, Downing reminds us that all evaluations in medical education require evidence of validity to be meaningfully interpreted.[3] Likewise, Crossley et al. argue that assessment tools must not only appear valid, but they should also be empirically tested for reliability and validity.[4]

Numerous studies have described the psychometric characteristics of instruments designed for assessing clinical teaching by learners.[5–25] The authors of these studies utilized diverse numbers of raters, subjects, and evaluations, as well as various learning environments, learner levels, and methodologies. There have also been comprehensive review articles highlighting the principles of evaluating clinical teaching. In particular, Snell et al. discussed the importance of evaluation for clinical teachers and medical education programs, and they emphasized the necessity of obtaining reliable, valid, and feasible assessments.[26] Similarly, Williams et al. extensively reviewed the literature regarding sources of bias in clinical performance ratings, thereby giving useful recommendations on ways to improve the value of clinical ratings.[27] We are unaware, however, of articles that specifically review scales designed for the assessment of clinical teaching, and that focus on the psychometric characteristics of these scales. In light of this need, our objective was to review the published literature on the reliability and validity of instruments designed for assessing clinical teaching, to summarize existing knowledge on the evaluation of clinical teachers by learners, and to identify themes that may aid in developing meaningful assessment tools.

METHODS

Electronic databases including MEDLINE, EMBASE, PsycINFO, ERIC, and Social Science Citation/Science

Citation indices were searched using the terms *validity*, *medical faculty*, *medical education*, *evaluation studies*, *instrument*, and the text word *reliability*. Included were studies in the English language dating from 1966 to fall of 2003. Excluded were review articles, editorials, qualitative studies, and case discussions. Authors TJB and PJE performed independent literature searches using the above criteria, yielding over 330 articles. Furthermore, by extracting citations from the bibliographies of these articles and by consulting colleagues with expertise in medical education, additional articles were found. After applying the above search criteria and reviewing all titles and abstracts, author TJB identified 21 relevant studies describing instruments designed for evaluating clinical faculty by learners. Three investigators (TJB, AKG, DAC) subsequently reviewed these studies using data abstraction sheets, which aided in identifying categories of validity evidence, statistical methods, and essential discussion points. Characteristics of the learning environments and validation methods were tabulated. After comparing completed abstraction sheets and discussing their observations, the three investigators determined salient themes among the evaluation studies. Although the authors generally agreed on the abstracted findings, when disagreement occurred, the final decision rested with author TJB.

RESULTS

Details regarding the studies' educational settings, evaluators, and subjects are summarized in Table 1. Among the studies with identifiable learning settings, many combined evaluations from inpatient and outpatient settings, and some authors combined evaluations from different learner levels. Evaluators included students, residents, and fellows, and only one study determined the reliability of peer ratings.[5] The most common subjects of evaluation were faculty in the specialties of internal medicine, and the least common subjects of evaluation taught in the specialties of family medicine, surgery, and emergency medicine. Wide ranges in numbers of teachers (10 to 711) and evaluators (3 to 374) were observed.

Characteristics of the validated instruments are summarized in Table 2. Disparate numbers of evaluations (30 to 7,845), items (1 to 43), domains (0 to 14), and Likert scale points (4 to 10) were found. Although studies presented a variety of validity evidence, the most commonly used statistics were factor analysis and determining internal consistency reliability with Cronbach's α. Less common validation methods were determining test-retest reliability, interrater reliability, and demonstrating convergent validity between new instruments and previously validated ones. Other measures supporting the validity of instruments, utilized in various ways, included analysis of variance (ANOVA), interitem and intraclass correlation coefficients, Pearson correlation coefficients, and the Spearman Brown formula. Additionally, numerous authors attempted to demonstrate adequate sampling of the content domain by consulting experts and established assessment methods.[5,8,10–12,14,18] It is noteworthy that applying previous instruments and established educational frameworks to new

Table 1. Validation Studies: Educational Setting, Teachers, and Evaluators

Author	Instrument	Setting	Teachers (*N*)	Evaluators (*N*)	Evaluators (Category)
Beckman, 2003	MTEF	I	10	3	P
Benbassat, 1981	—	—	—	83	S
Cohen, 1996	TES	I, O	43	—	S
Copeland, 2000	CTEI	I, O	711	—	S, R, F
Donnelly, 1989	—	I	90	100	S
Donner, 2003	—	—	—	80	R
Guyatt, 1993	—	I	41	—	R
Hayward, 1995	—	O	15	—	R
Irby, 1981	CTAF	—	230	320	S
James, 1999	MedIQ	—	—	131	S
Litzelman, 1998	SFDP	I	178	374	S
Litzelman, 1999	SFDP	I, O	38	36	R
Macgill, 1986	—	O	19	24	R
McLeod, 1993	CTEQ	I, O	37(S), 15(R)	—	S, R
Ramsbottom, 1994	CTAF	I, O	29	—	R
Risucci, 1992	—	—	62	23	R
Shellenberger,1982	PEQ	—	—	197	S
Solomon, 1997	—	I, O	147	—	S
Steiner, 2000	ERS	ED	29	18	R
Tortolani, 1991	—	—	62	23	R
Williams, 2002	GRS	I, O	96	98	R

MTEF, Mayo Teaching Evaluation Form; TES, Teaching Effectiveness Scores; CTEI, Clinical Teaching Effectiveness Instrument; MedIQ, Medical Instructional Quality; CTAF, Clinical Teaching Assessment Form; PEQ, Preceptor Evaluation Questionnaire; SFDP, Stanford Faculty Development Program; CTEQ, Clinical Tutor Evaluation Questionnaire; ERS, Emergency Rotation Scale; GRS, Global Rating Scale; I, inpatient; O, outpatient; ED, emergency department; P, peer; S, student; R, resident; F, fellow; —, information unavailable.

Table 2. Characteristics of Validated Instruments and Study Methods

Author	Evaluations (*N*)	Domains* (*N*)	Likert Scale†	Items (*N*)	Methods
Beckman, 2003	30	7	5	28	IRR (KCC), ICR (α)
Benbassat, 1981	—	1	10	9	Factor analysis, Pearson
Cohen, 1996	3,750	—	5	4	IRR (ICC), test-retest
Copeland, 2000	7,845	1	5	15	Factor analysis, ICR (α), IRR (g-coefficient), Pearson
Donnelly, 1989	218	2	7	12	ICR (α), IRR (ICC)
Donner, 2003	—	—	4	43	Test-retest (Pearson)
Guyatt, 1993	—	14	5	14	Factor analysis, Pearson, ANOVA
Hayward, 1995	142	4	5	18	Factor analysis, ICR (α), IRR (Spearman, ANOVA), generalizability coefficient
Irby, 1981	1,567	4	5	9	Factor analysis, ICR (Split-half), IRR (Spearman, ANOVA)
James, 1999	156	4	4–6	25	Factor analysis, ICR (α), correlation coefficients
Litzelman, 1998	1,581	7	5	25	Factor analysis, ICR (α, Pearson, IIC)
Litzelman, 1999	360	7	5	25	Factor analysis, ICR (α, IIC)
Macgill, 1986	195	—	—	—	—
McLeod, 1993	—	2	6	25	Factor analysis, ANOVA, Pearson, Spearman
Ramsbottom, 1994	639	8	6	9	IRR (ICC, Spearman), ANOVA
Risucci, 1992	—	1	5	10	Factor analysis, ICR (α), Test-retest, IRR (Pearson)
Shellenberger, 1982	—	6	4	34	Factor analysis, ICR (α)
Solomon, 1997	1,570	—	4	13	IRR (Ebel method, Spearman)
Steiner, 2000	48	4	5	4	ICR (α), Spearman, multitrait-multimethod matrix
Tortolani, 1991	—	2	5	10	Factor analysis, test-retest (Pearson)
Williams, 2002	—	—	5	1	Multivariate Random Intercept Model

* *Domains identified in evaluation instruments.*

† *Number of points on instrument's Likert scales.*

ANOVA, analysis of variance; α, Cronbach's coefficient α; ICC, intraclass correlation coefficient; ICR, internal consistency reliability; IIC, interitem correlation coefficient; IRR, interrater reliability; KCC, Kendall's coefficient of concordance; Pearson, Pearson's correlation coefficient; Spearman, Spearman-Brown prophecy formula; —, information unavailable.

educational settings was a theme among only a handful of studies.[5,11,13,15,16,18,19] Finally, 14 domains of teaching were identified (see Fig. 1), with the most common domains being interpersonal and clinical-teaching skills.

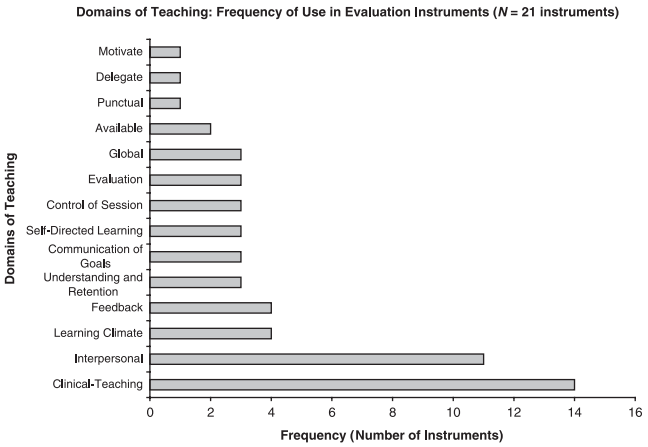

FIGURE 1. Domains of teaching: frequency of use in evaluation instruments (*N* = 21 instruments).

DISCUSSION

A review of the published literature revealed 21 evaluation studies on instruments designed for assessing clinical teaching. Each of these studies demonstrated at least some evidence of construct validity. Current standards published by the American Psychological and American Educational Research Associations identify 5 categories of validity evidence: 1) content, 2) responses, 3) internal structure, 4) relationship to other variables, and 5) consequences.[28] When considering categories of validity, it is understood that validity evidence exists to various degrees, but there is no threshold at which an assessment is said to be valid. It is also understood that while historically there are numerous categories of validity, all validity is construct validity.[3] Furthermore, not all types of validity evidence are required for every assessment, although evidence of an instrument's validity should be obtained from a variety of sources.[3,28]

Among the 21 studies reviewed, the most commonly observed sources of validity evidence were in the category of internal structure, which relates to the psychometric characteristics of an instrument.[3] One important element of

an instrument's internal structure is reliability (see Table 3). The most frequent measures among the reviewed studies were exploring scale dimensionality with factor analysis and determining the internal consistency reliability of teaching domains. These findings were expected, considering the importance of demonstrating an instrument's internal consistency reliability with Cronbach's coefficient α when assessing the validity of a new evaluation tool.[29,30] Likewise, the frequent use of factor analysis reflects the importance of showing that items represent a common latent variable,

Table 3. Types of Reliability: Descriptions, Indices, and Definitions

Reliability	Description	Indices	Definitions	Comments
Internal consistency	Do all the items on a test* measure the same trait? We would expect high correlation between items measuring a single trait. Internal consistency is widely reported, in part because it can be calculated after administering a single test form once.	Split-half reliability	Divide a test into equal halves, and calculate the correlation between the halves.	One drawback is that the "effective" test is only half as long as the actual test; the Spearman-Brown† formula can adjust this result.
		Kuder-Richardson	Similar concept to split-half, but accounts for all items (several different formulas for specific situations).	Assumes all items are equivalent, measure a single trait, and have dichotomous responses.
		Cronbach's α	A generalized form of the Kuder-Richardson formulas.	Assumes all items are equivalent and measure a single trait; can be used with dichotomous or continuous data.
		Factor analysis	A statistical method to find clusters (factors) of related items.	It is useful in this setting to determine whether the factors the test identified are those it was intended to assess.
Temporal stability	Does the test produce similar results when administered a second time?	Test-retest	Administer the same form to the same person at different times.	If a trait is stable, we would expect similar results on the same test at different times.
Equivalence	Do two different tests intended to measure the same trait produce similar results?	Alternate forms	Administer different forms to the same individual at the same or different times.	If the tests are truly similar, we would expect similar results.
Agreement (interrater reliability)	Is one rater's score similar to another's? This may be the most important source of reliability when assessing clinical performance.	Percent agreement	Percent of identical responses.	Does not account for agreement that would occur by chance.
		Phi	Simple correlation.	Does not account for chance.
		Kappa	Percent agreement corrected for chance.	
		Kendall's tau	Agreement on ranked data.	
		Intraclass correlation coefficient	Uses analysis of variance (ANOVA) to estimate how well ratings from different raters coincide.	
Generalizability theory	How much of the error in measurement is due to each factor (item, item grouping, subject, rater, day of test, etc.) involved in the measurement process?	Generalizability theory	Complex model that allows estimation of multiple sources of error.	This method can be used in any setting where reliability is assessed. For example, it can determine the relative contribution of internal consistency and interrater reliability to the overall reliability of a given test.

* *The word "test" may apply to any instrument—test, survey, performance rating scale, etc. "Items" are the individual questions on the instrument. The "trait" is what is being measured, such as knowledge, attitude, or skill in a specific area.*

† *The Spearman Brown "prophecy" formula allows one to calculate the reliability of a test when the number of items is increased (or decreased). For more details regarding the concepts in this table, please see references 29, 30, and 36–38.*

and of proving the unidimensional nature of a set of items prior to calculating α.[30] One example of a study determining an instrument's multidimensionality and internal consistency reliability comes from Litzelman et al., who used factor analysis with oblique rotation to support the existence of a 7-category framework.[16] This 7-factor model explained 73% of variation in the study data and the coefficient α's were all acceptably high. Another example of a multidimensional model comes from Irby and Rakestraw.[13] By using factor analysis with a principle component solution to orthogonal factors, these authors identified 4 factors that accounted for 87% of their study data's variance. Notably, the models described by Litzelman and Irby were the most common ones to be reevaluated by different authors, among the studies reviewed. Less frequent measures of reliability among the reviewed studies were temporal stability (test-retest reliability) and interrater reliability. Only 5 of the studies reported test-retest reliabilities, which ranged from 3 weeks to 9 years.[7,10,18,20,24] Nine studies demonstrated interrater reliability.[5,7–9,12,13,19,20,22]

The relationship of an assessment to other variables, such as the convergent validity between new and established instruments, is a powerful source of validity that was used infrequently in the reviewed studies. For example, Steiner et al. gave resident physicians the Irby scale (previously validated) and the Emergency Rotation (ER) Scale for the evaluation of emergency department faculty.[23] Correlations between the Irby and ER scales were high (>0.70). Residents in a study by Williams et al. also completed two faculty evaluation forms, the Global Rating Scale (GRS) and the Stanford Faculty Development Program (SFDP) scale, yielding high correlations between the two scales (range 0.86 to 0.98).[25] What's more, these studies illustrate a disadvantage of testing for convergent validity, which is the time burden on learners who are required to complete dual forms.

Another source of validity evidence that was rarely used in the reviewed studies relates to correlations between assessments and educationally relevant outcomes.[3] Authors James and Osborne utilized this type of evidence in their study of the Medical Instructional Quality (MedIQ) instrument, thereby demonstrating good correlations between MedIQ preceptor scores and resident physicians' clerkship grades and specialty choices.[14] Similarly, Benbassat and Bachar showed that instructors were significant sources of variance in their students' adjusted clinical examination scores.[6] Regarding faculty performance, Tortolani et al. found correlations between resident evaluations of faculty and the amount of faculty involvement in teaching, clinical, and research activities.[24]

There were 14 domains of teaching among the evaluation studies, with clinical-teaching and interpersonal skills being the most common. Furthermore, several studies demonstrated that interpersonal skills are distinguishable from other dimensions such as cognitive and teaching skills. Donnelly and Woolliscroft revealed medical students' abilities to discriminate between cognitive and interpersonal aspects of teaching by resident and attending physicians.[9] Hayward et al. showed that the clinical-teaching scale is separable from availability, respect, and slow-staffing scales.[12] Irby and Rakestraw found that the interpersonal relations factor is distinct from the factors of knowledge, clinical skills, and supervision skills.[13] McLeod et al.'s analysis of student and resident ratings revealed that an instrument could be reduced to personality and pedagogic domains.[18] A final example of a two-dimensional construct comes from Durning et al.[31] Their study involving faculty-on-learner evaluations showed that the 7-category American Board of Internal Medicine Evaluation Form (ABIM-MEF) could be collapsed into the domains of judgment-knowledge-skills and attitude-humanism. In summary, these findings raise the possibility that instruments comprised solely of interpersonal and clinical-teaching domains may adequately assess the proficiency of clinical teachers.

It is noteworthy that common sources of bias in the assessment of behaviors pertains to observers, to the relationship between observers and the subjects of evaluation, and to the environment in which the evaluation takes place.[32] Many of the studies used evaluations from both inpatient and outpatient settings, and some authors even pooled evaluations from different learner levels.[8,18] In light of this, experts have observed that different teaching skills are required for instruction in the outpatient versus inpatient settings, and that various levels of learners perceive clinical teaching differently.[18,19,23,33] Other potential sources of observer-related bias include the tendencies of learners to consistently give high ratings and to exhibit the halo effect when completing faculty evaluations.[2,9,11,18,22] Indeed, bias from raters is recognized as a major source of construct-irrelevant variance in assessments of clinical performance by both learners and teachers.[34] Consequently, combining evaluations from different educational settings and learner levels may have been an overlooked confounder in many previous studies.

This was a review of studies based on measurements of Likert-scaled instruments completed by learners for the purpose of evaluating clinical faculty, and there are limitations inherent in this form of data. As noted above, learners at all levels have been shown to give inflated ratings of clinical faculty, which may limit the ability to separate skilled from less skilled teachers. Furthermore, learners' evaluations, while inexpensive and widely available, may be less reliable than alternate forms of faculty assessment. Indeed, these observations have led authors to encourage incorporating peer review into the evaluation of teaching faculty,[2,35] and paying close attention to learners' comments written on faculty evaluations, versus relying solely on learners' responses to Likert-scaled items.[2] A final limitation lies in the generalizability of clinical assessment instruments, regardless of whether such instruments are reliable. This is because, despite the existence of widely acclaimed educational frameworks,[13,16] institutions have their own cultures of teaching, and assessments should be consistent with the philosophy of the institutions in which they are used.[26]

Our review of reliable and validated instruments has some limitations. Although we feel our strategy for searching electronic databases to identify the published validation studies was thorough, it is possible that several studies were overlooked. In response we would point out that, due to the number of relevant databases searched, it is unlikely any significant number of studies were missed. Moreover, many of the observed themes were evident after reviewing only about 15 articles, making it unlikely that the addition of several more articles would change our conclusions. Another limitation derives from our method for cataloging the domains of teaching observed in the reviewed studies. This required the use of judgment when collapsing similar categories into the same domain. For example, domains in the literature such as *interpersonal skills*, *personality*, *interpersonal*, *conduct*, and *interpersonal relations* were all collapsed into the domain of *interpersonal* (Fig. 1). Nevertheless, we feel most people would generally agree with our interpretation of these terms. We also recognize that the use of judgment when interpreting the meaning of construct labels is an unavoidable limitation of evaluating and describing assessment tools.

In conclusion, we identified 21 studies providing a variety of evidence to support the construct validity of instruments designed for assessing clinical teachers. While 5 categories of validity evidence are recognized,[3,28] authors tended to emphasize an instrument's internal structure validity by demonstrating an instrument's dimensionality and internal consistency of teaching domains through the use of factor analysis and Cronbach's coefficient α, respectively. Less frequently used evidence of validity included test-retest and interrater reliabilities. Establishing validity by showing convergence between new and established instruments, and by correlating faculty assessments with educationally relevant outcomes, were also less common methods, suggesting that a broader variety of validity evidence should be considered when planning clinical assessments. Previous authors have recognized the different requirements for teaching in inpatient and outpatient settings, and the varying perceptions of clinical teaching by different learner levels, indicating that future studies should use more narrowly defined populations. Additionally, a review of these articles indicates that future studies should consider developing assessment tools comprised solely of interpersonal and clinical-teaching domains. We found that inflated ratings are a limitation of using learners to evaluate faculty, prompting the need for closer attention to learners' comments written on faculty evaluations. Finally, we propose that the unique cultures of teaching at most institutions may ultimately limit the generalizability of even the most carefully designed instruments.

REFERENCES

1. Jones RG, Froom JD. Faculty and administration view of problems in faculty evaluations. Acad Med. 1994;69:476–83.
2. Beckman TJ, Lee MC, Mandrekar JN. A comparison of clinical teaching evaluations by resident and peer physicians. Med Teach. 2004;26:321–5.
3. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37:830–7.
4. Crossley J, Humphris G, Jolly B. Assessing health professionals. Med Educ. 2002;36:800–4.
5. Beckman TJ, Lee MC, Rohren CH, Pankratz VS. Evaluating an instrument for the peer review of inpatient teaching. Med Teach. 2003;25:131–5.
6. Benbassat J, Bachar E. Validity of students' ratings of clinical instructors. Med Educ. 1981;15:373–6.
7. Cohen R, McRae H, Jamieson C. Teaching effectiveness of surgeons. Am J Surg. 1996;171:612–4.
8. Copeland HL, Hewson MG. Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical center. Acad Med. 2000;75:161–6.
9. Donnelly MB, Woolliscroft JO. Evaluation of instructors by third year medical students. Acad Med. 1989;64:159–64.
10. Donner-Banzhoff N, Merle H, Baum E, Basler HD. Feedback for general practice trainers: developing and testing a standardized instrument using the importance-quality-score method. Med Educ. 2003;37:772–7.
11. Guyatt GH, Nishikawa J, Willan A, et al. A measurement process for evaluating clinical teachers in internal medicine. Can Med Assoc J. 1993;149:1097–102.
12. Hayward RA, Williams BC, Gruppen LD, Rosenbaum D. Measuring attending physician performance in a general medicine outpatient clinic. J Gen Intern Med. 1995;10:504–10.
13. Irby DM, Rakestraw P. Evaluating clinical teaching in medicine. J Med Educ. 1981;56:181–6.
14. James PA, Osborne JW. A measure of medical instructional quality in ambulatory settings: the MedIQ. Fam Med. 1999;31:263–9.
15. Litzelman DK, Westmorland GR, Skeff KM, Stratos GA. Student and resident evaluations of faculty—how reliable are they? Acad Med. 1999;74(suppl Oct):s25–s27.
16. Litzelman DK, Stratos GA, Marriott DJ, Skeff KM. Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. Acad Med. 1998;73:688–95.
17. Mcgill MK, McClure C, Commerford K. A system for evaluating teaching in the ambulatory setting. Fam Med. 1986;18:173–4.
18. McLeod PJ, James CA, Abrahamowicz M. Clinical tutor evaluation: a 5-year study by students on an in-patient service and residents in an ambulatory care clinic. Med Educ. 1993;27:48–54.
19. Ramsbottom-Lucier MT, Gillmore GM, Irby DM, Ramsey PG. Evaluation of clinical teaching by general internal medicine faculty in outpatient and inpatient settings. Acad Med. 1994;69:152–4.
20. Risucci DA, Lutsky L, Rosati RJ, Tortolani AJ. Reliability and accuracy of resident evaluations of surgical faculty. Eval Health Prof. 1992;15:313–24.
21. Shellenberger S, Mahan JM. A factor analytic study of teaching in off-campus general practice clerkships. Med Educ. 1982;16:151–5.
22. Solomon DJ, Speer AJ, Rosebraugh CJ, DiPette DJ. The reliability of medical student ratings of clinical teaching. Eval Health Prof. 1997;20:343–52.
23. Steiner IP, Franc-Law J, Kelly KD, Rowe BH. Faculty evaluation by residents in an emergency medicine program: a new evaluation instrument. Acad Emerg Med. 2000;7:1015–21.
24. Tortolani AJ, Rissucci DA, Rosati RJ. Resident evaluation of surgical faculty. J Surg Res. 1991;51:186–91.
25. Williams BC, Litzelman DK, Babbott SF, Lubitz RM, Hofer TP. Validation of a global measure of faculty's clinical teaching performance. Acad Med. 2002;77:177–80.
26. Snell L, Tallett S, Haist S, et al. A review of the evaluation of clinical teaching: new perspectives and challenges. Med Educ. 2000;34:862–70.
27. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med. 2003;15:270–92.

28. 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.
29. Nunnally JC, Berstein IH. Psychometric Theory. 3rd ed. New York: McGraw-Hill. 1994:211–54.
30. DeVillis RF. Scale Development: Theory and Applications. London: Sage Publications. 1991;94:102–37.
31. Durning SJ, Cation LJ, Jackson JL. The reliability and validity of the American Board of Internal Medicine Monthly Evaluation Form. Acad Med. 2003;78:1175–82.
32. Schwab DP. Construct validity in organizational behavior. Res Organ Behav. 1980;2:3–43.
33. Perkoff GT. Teaching clinical medicine in the ambulatory setting: an idea whose time may have finally come. N Engl J Med. 1986;314: 27–31.
34. Downing DM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38:327–33.
35. Irby DM. Evaluating instruction in medical education. J Med Educ. 1983;58:844–9.
36. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. In Press.
37. Howell DC. Statistical Methods for Psychology. 5th ed. Pacific Grove, Calif: Duxbury; 2002.
38. McMillan JH, Schumacher S. Research in Education: A Conceptual Introduction. 5th ed. New York: Addison Wesley Longman; 2001.