What is the Validity Evidence for Assessments of Clinical Teaching?

Thomas J. Beckman, MD,¹ David A. Cook, MD, MHPE,¹ Jayawant N. Mandrekar, PhD²

¹Division of General Internal Medicine, Department of Internal Medicine, Mayo Clinic College of Medicine, Mayo Clinic and Mayo Foundation, Rochester, Minn, USA; ²Division of Biostatistics, Department of Health Sciences Research, Mayo Clinic College of Medicine, Mayo Clinic and Mayo Foundation, Rochester, Minn, USA.

BACKGROUND: Although a variety of validity evidence should be utilized when evaluating assessment tools, a review of teaching assessments suggested that authors pursue a limited range of validity evidence.

OBJECTIVES: To develop a method for rating validity evidence and to quantify the evidence supporting scores from existing clinical teaching assessment instruments.

DESIGN: A comprehensive search yielded 22 articles on clinical teaching assessments. Using standards outlined by the American Psychological and Education Research Associations, we developed a method for rating the 5 categories of validity evidence reported in each article. We then quantified the validity evidence by summing the ratings for each category. We also calculated weighted κ coefficients to determine interrater reliabilities for each category of validity evidence.

MAIN RESULTS: Content and Internal Structure evidence received the highest ratings (27 and 32, respectively, of 44 possible). Relation to Other Variables, Consequences, and Response Process received the lowest ratings (9, 2, and 2, respectively). Interrater reliability was good for Content, Internal Structure, and Relation to Other Variables (κ range 0.52 to 0.96, all P values $<.01$), but poor for Consequences and Response Process.

CONCLUSIONS: Content and Internal Structure evidence is well represented among published assessments of clinical teaching. Evidence for Relation to Other Variables, Consequences, and Response Process receive little attention, and future research should emphasize these categories. The low interrater reliability for Response Process and Consequences likely reflects the scarcity of reported evidence. With further development, our method for rating the validity evidence should prove useful in various settings.

KEY WORDS: validity; clinical teaching; evaluation studies.

DOI: 10.1111/j.1525-1497.2005.0258.x

J GEN INTERN MED 2005; 20:1159-1164.

Experts stress the need for reliable and valid teaching assessments.^{1,2} Despite this, medical educators have not used consistent validity criteria when developing and evaluating instruments to assess clinical teaching. For example, we recently reviewed the literature on the psychometric characteristics of instruments for assessing clinical teachers.³ In our analysis of these studies,⁴⁻²⁵ we found that authors usually pursue a limited variety of validity evidence, and authors' interpretations of validity substantially differ in terms of the rel-

ative importance placed on different categories of validity evidence.

These findings led us to reflect more deeply on the definition and measurement of validity in assessing clinical teaching. According to modern theory, validity is a hypothesis, and all sources of validity evidence contribute to accepting or rejecting this hypothesis.¹ For this reason, scores from teaching assessment instruments should be supported by a variety of validity evidence. The American Psychological and Education Research Associations published standards that identify 5 sources of validity evidence: (1) Content, (2) Response Process, (3) Internal Structure, (4) Relation to Other Variables, and (5) Consequences²⁶ (see Table 1). Notably, this 5-category validity framework, articulated by Messick²⁷ over 10 years ago, has been increasingly regarded by education and psychology researchers as the most comprehensive conceptualization of validity. Moreover, experts emphasize the importance of incorporating these sources of evidence into clinical teaching assessments.^{1,3} Experts also assert that validity is a property of scores and score interpretations, and not a property of the instrument itself.^{1,26,27}

In the present study, we critically evaluated the published literature for evidence supporting the validity of clinical teaching assessments. Our aims were: (1) develop a reliable and systematic method by which medical educators can rate the validity of scores from teaching assessment instruments, (2) evaluate the quantity and quality of validity evidence for scores from published teaching assessment instruments, and (3) identify areas warranting further research.

METHODS

Literature Search

Our method for identifying studies on the assessment of clinical teaching is described in detail elsewhere.³ Electronic databases, including MEDLINE, EMBASE, PsycINFO, ERIC, and Social Science Citation/Science Citation indices were searched for English-language articles published between 1966 and July 2004 using the terms validity, medical faculty, medical education, evaluation studies, instrument, and the text word reliability. This search yielded over 330 articles. Review articles, editorials, qualitative studies, and case discussions were excluded. Additional articles were found by reviewing the bibliographies of retrieved articles and by consulting colleagues with expertise in medical education. After reviewing all titles and abstracts, we identified 22 relevant studies describing instruments designed for assessing clinical faculty by learners.

None of the authors have any conflicts of interest to declare for this paper.

Presented at the Society of General Internal Medicine 28th Annual Meeting, New Orleans, La, May 11-14, 2005. Also presented at the Association for the Study of Medical Education Annual Scientific Meeting, Newcastle upon Tyne, UK, July 11-13, 2005.

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).

Received for publication July 27, 2005

and revised form July 29, 2005

Accepted for publication July 29, 2005

Table 1. Validity: Sources of Evidence, Definitions, and Examples

Evidence Source	Definition ^{1,25}	Examples*
Content	The "relationship between a test's content and the construct it is intended to measure." ²⁶ Refers to themes, wording, and format of items on an assessment instrument. Includes analyses by experts regarding how adequately items represent the content domain. Also includes development strategies to ensure appropriate content representation	Surveying experienced teachers regarding the adequacy and representativeness of proposed instrument items. ¹⁰ Choosing items previously utilized in similar settings. Developing instruments based on established educational theories. ¹⁴
Response process	Analyses of responses, including the actions, strategies, and thought processes of individual respondents or observers. Differences in response processes may reveal sources of variance that are irrelevant to the construct being measured. Also includes instrument security, scoring, and reporting of results	Interviewing and studying learners regarding factors that influence the ratings they assign to teachers. Analyzing varying response patterns among different categories/levels of learners. ^{17,19}
Internal structure	The degree to which individual items within the instrument fit the underlying constructs. Items measuring a unidimensional construct should be homogenous, while items measuring complex constructs should not. Most often reported as measures of internal consistency reliability and factor analysis	Using factor analysis to determine the dimensional structure of an instrument's scores, and determining the reliability of scores. ¹² Studying differential functioning of items among a homogenous group of evaluators
Relations to other variables	The relationship between scores and other variables relevant to the construct being measured. Relationships may be positive (convergent or predictive) or negative (divergent or discriminant) depending on the constructs being measured	How well do teachers' assessment scores predict learners' performance on high-stakes examinations, or their choice of a medical specialty? ¹³ Do scores correlate with other measures of the same construct? ²² Can results of an evaluation be generalized from one setting to another, similar setting?
Consequences	Assessments are intended to have some desired effect (e.g., improve teaching performance), but they also have unintended effects. Evaluating such consequences can support or challenge the validity of score interpretations	Do equally qualified teachers' performances on clinical teaching assessments correlate with factors that are not being measured, such as medical subspecialty, gender, or ethnicity?

*Examples of validity evidence that are from our literature search are followed by citations. Other examples are hypothetical.

Rating Sources of Validity Evidence

We agreed upon operational definitions of the 5 sources of validity evidence based on *Standards*²⁶ published by the American Psychological and Education Research Associations and interpretations by another author.¹ We then developed the following rating scale: N=no discussion of this source of validity evidence and/or no data presented; 0=discussion of this source of validity evidence, but no data presented, or data failed to support the validity of instrument scores; 1=data for this source weakly support the validity of score interpretations; and 2=data for this source strongly support the validity of score interpretations. We adopted this scale after considering multiple alternatives because it awarded numerical points (score=1 or 2) only for articles that provide data supporting the validity of score interpretations, distinguished articles that discussed a category of validity evidence (score=0) from those failing to address a category of validity evidence (score=N), and avoided unnecessary complexity. See Table 2 for specific rating criteria for each category of validity evidence. Authors T.J.B. and D.A.C. independently analyzed the 22 studies using this rating scale. After calculating interrater reliability, authors T.J.B. and D.A.C. discussed their individually assigned ratings for each article until they reached a consensus on the final ratings. These ratings were then summed across all studies to yield a total score for each evidence category.

Reliability

For the purpose of determining interrater reliability, the categories of N, 0, 1, and 2 were converted to the numbers 1, 2, 3, and 4, respectively. Weighted κ scores and κ P values were calculated for each category of validity evidence using the weight-

ing scheme proposed by Fleiss and Cohen.^{28,29} Kappa values were interpreted according to Landis and Koch's³⁰ guidelines, where κ values under 0.4 represent poor agreement, values from 0.4 to 0.75 fair to good agreement, and values 0.75 and over represent excellent agreement. For all analyses, κ P values $\leq .05$ were the criteria for concluding that there was significant agreement between observers.

RESULTS

The highest possible score for each category of validity evidence, when summing scores for the 22 articles, is 44. Highest sum scores were given for the categories of Content and Internal Structure (scores 27 and 32, respectively) (Table 3). With the exception of 1 article in the category of Relation to Other Variables, Content and Internal Structure were the only categories with articles having evidence that strongly supported validity (=score of 2) (Table 4).

The lowest sum scores were given for the evidence categories of Relation to Other Variables, Consequences, and Response Process (scores 9, 2, and 2 respectively) (Table 3). Within the categories of Response Process and Consequences, the vast majority of articles provided no or insufficient data to support validity. Table 3 summarizes the scores for each category of validity evidence and Table 4 summarizes validity evidence scores for individual studies of clinical teaching.

Regarding interrater reliability, weighted κ scores were good to excellent for the categories of Content, Internal Structure, and Relation to Other Variables (κ range 0.52 to 0.96, all P values $\leq .01$), but weighted kappa scores were poor for Consequences and Response Process. Table 3 summarizes weighted κ scores and corresponding P values.

Table 2. Criteria for Rating Validity Evidence

Evidence Category	Rating*	Rating Criteria	Illustrative Examples†
Content	N	No discussion of instrument content (includes simply listing items without justification)	Guyatt et al. ¹⁰
	0	Discussion but no data‡	
	1	Listing assessment themes with little or no reference to a theoretical basis, or a poorly defined process for creating and reviewing items	
	2	Well-defined process for developing instrument content, including both an explicit theoretical/conceptual basis for instrument items and systematic item review by experts	
Response process	N	Alternatively, reference to a prior study on an assessment instrument that meets these criteria	McLeod et al. ¹⁷
	0	No discussion. Merely disclosing response rates or numbers of respondents does not constitute evidence	
	1	Discussion but no data. Discussing the impact of response rate on assessment scores, or speculating on the thought processes of learners, does not constitute evidence	
	2	Minimal data regarding thought processes and analysis of responses. Description (without data) of systems that reduce response error, such as computer-scored forms	
Internal structure	N	Multiple sources of supportive data, including critical examination of thought processes, analysis of responses for evidence of halo error or rater leniency, or data demonstrating low response error	Irby and Rakestraw ¹²
	0	No discussion	
	1	Discussion but no data	
	2	Factor analysis incompletely confirming anticipated data structure, or acceptable reliability with a single measure	
Relation to other variables	N	Factor analysis confirming anticipated data structure, or multiple measures of reliability. Variation in responses to specific items among subgroups (differential item functioning) can support or challenge internal structure depending on predictions ²⁶	James and Osborne ¹³
	0	No discussion	
	1	Discussion but no data	
	2	Correlation of assessment scores to outcomes with minimal theoretical importance, or unanticipated score correlations	
Consequences	N	Correlation (convergence) or no correlation (divergence) between assessment scores and theoretically predicted outcomes or measures of the same construct. Such evidence will usually be integral to the study design, and anticipated a priori	Cohen et al. ⁶
	0	No discussion. Speculation on potential applications of the assessment does not constitute evidence	
	1	Discussion but no data. Simply discussing the consequences of assessment (e.g., data regarding usefulness or faculty approval) without linking this to validity does not constitute evidence	
	2	Description of consequences of assessment that could conceivably impact the validity of score interpretations (although these impacts are not explicitly identified by the authors)	
		Description of consequences of assessment that clearly impact on the validity of score interpretations, as supported by data and convincingly argued by the authors. Such evidence will usually be integral to the study design, and anticipated a priori	

*Scoring system: N, No data or discussion of this source of evidence; 0, Discussion of this source of evidence, but no data presented, or data corresponding to this source of evidence fail to support the validity of the proposed interpretation; 1, data corresponding to this source of evidence weakly supports validity of proposed interpretation; 2, data corresponding to this source of evidence strongly support validity of proposed.

†Parenthetically listed numbers after authors refer to reference numbers.

‡We did not award a score of 0 to any article in this category. We believe it improbable that a published article would merit this score because any discussion of content evidence would constitute data, and it seems unlikely that data would not at least weakly support the content of any published instrument.

DISCUSSION

Based on our prior observations,³ we anticipated that evidence of Content and Internal Structure validity would be strongly represented in the literature on clinical teaching assessments. After developing and implementing an objective method for scoring the validity evidence, we confirmed that the highest scoring categories are Content and Internal Structure, and the lowest scoring categories are Relation to Other Variables, Consequences, and Response Process. These findings suggest that, contrary to requests for validity evidence from a variety of sources,^{1,3,26,27} most authors report only a limited subset of validity evidence. This raises questions about the validity of interpretations that can be derived from existing teaching assessments. To illustrate how future research could improve upon the past, we will review the sources of validity evidence

found among assessments of clinical teaching and provide examples of articles that utilize validity evidence effectively.

Internal Structure evidence refers to the degree to which individual items fit the underlying construct of interest, and is most often reported as measures of factor analysis or internal consistency reliability.²⁶ Some experts include all reliability in this category.^{1,27} We found that evidence of Internal Structure is commonly demonstrated among studies on clinical teaching assessments, perhaps because such evidence can be sought without prior planning. In other words, performing statistical analyses (e.g., reliability, factor analysis) on preexisting data can generate evidence of Internal Structure. The majority of studies in our review reported at least 1 type of reliability. We are concerned, however, that the types of reliability reported may not be the most important. Interrater reliability is the favored type of reliability when assessing clinical performance³¹;

Table 3. Validity Ratings and Observer Agreement

Category	Sum	κ	P Value
Content	27	0.96	<.0001
Response process	2	−0.10	.56
Internal structure	32	0.52	.01
Relation to other variables	9	0.82	<.0001
Consequences	2	−0.14	.51

Sum reflects the total validity score for all the articles in each category based on agreement between 2 observers. Scores of "0" and "N" were considered to be 0 when summing scores. The maximum sum score is 44. κ refers to weighted κ scores; see Methods for details regarding κ analysis. P value's reflect κ P values.

yet, our review on the psychometric characteristics of clinical teaching assessments revealed that less than half of published studies report this reliability measure.³

As evidence for Internal Structure is common, we were able to recognize patterns that facilitated our rating of this category. Notably, all articles utilized factor analysis and/or some measure of reliability. A noteworthy example of Internal Structure evidence comes from a study by Irby and Rakestraw.¹² In this study, factor analysis with orthogonal rotation revealed 4 distinct factors and satisfactory interitem correlations. Additionally, estimates of interrater reliability for 20 ratings, calculated by the Spearman-Brown prophecy formula, were excellent.

Content evidence refers to themes, wording, and format of items, and includes expert review and other systematic item development strategies. In this study, evidence of Content among clinical teaching assessments ranked second only to Internal Structure. When scoring Content evidence, we utilized 2 criteria: (1) items should represent the construct(s) they intend to measure, that is, the items should have a convincing theoretical basis and (2) instrument development strategies,

including expert review, should be clearly described. Perhaps the best example of content evidence is a study by Guyatt et al.¹⁰ The authors initially defined assessment criteria (domains) using input from faculty, residents, and a careful literature review. Items were then created to represent the defined domains. Finally, the authors articulated a method by which physicians reviewed and modified the ultimate set of items.

Relation to Other Variables evidence refers to relationships (convergent or discriminating) between scores and other variables relevant to the construct being measured. Relation to Other Variables is a powerful yet underutilized source of evidence among assessments of clinical teaching. We suspect that Relation to Other Variables is underutilized because, in most cases, studies must be specifically designed to evaluate predicted associations or hypotheses. As discussed below, it seems that studies of clinical teaching assessments are often not hypothesis-driven or designed to confirm anticipated associations. Examples of meaningful relationships between faculty assessments and other variables could include correlations (either positive or negative as predicted by theory) between scores and outcomes such as teaching awards or academic promotion, and demonstrating predicted correlations between scores from 2 different instruments designed to assess the same teaching behaviors.²² Unfortunately, we identified only 1 study with data strongly supporting Relation to Other Variables. James and Osborne¹³ showed that student-on-teacher assessment scores (a proxy for instructional quality) were significantly related to certain theoretically predicted outcomes, such as clerkship grades and choice of medical specialty, but were not related to other outcomes, such as National Board of Medical Examiners scores.

The categories of Consequences and Response Process were the least represented sources of validity evidence among published clinical teaching assessments, and the most difficult

Table 4. Validity Ratings on Individual Assessments of Clinical Teaching

Author	Year	Content	Response Process	Internal Structure	Other Variables	Consequences
Beckman	2003	1	0	1	N	0
Benbasset	1981	N	0	1	1	0
Cohen	1996	N	0	1	0	1
Copeland	2000	2	0	2	1	1
Donnelly	1989	N	0	2	N	N
Donner-Banzhoff	2003	2	0	1	1	N
Guyatt	1993	2	0	1	0	0
Hayward	1995	2	0	2	N	0
Irby	1981	2	N	2	N	N
James	1999	2	N	1	2	N
Litzelman	1998	2	N	2	N	N
Litzelman	1999	2	0	2	N	N
Magill	1986	1	0	1	N	N
McLeod	1993	2	1	2	N	0
Ramsbottom-Lucier	1994	2	0	1	N	N
Risucci	1992	N	1	1	0	N
Shellenberger	1982	2	0	2	N	N
Smith	2004	2	N	2	1	0
Solomon	1997	N	0	1	N	0
Steiner	2000	1	0	1	1	0
Tortolani	1991	N	0	2	1	0
Williams	2002	N	0	1	1	0

Rating system: N, no discussion of this source of validity evidence and/or no data presented; 0, discussion of this source of validity evidence, but no data presented, or data for this source fail to support the validity of score interpretations; 1, data for this source weakly support the validity of score interpretations; and, 2, data for this source strongly support the validity of score interpretations.

to rate as demonstrated by the low κ scores. Just as the prevalence of Internal Structure evidence facilitated our ability to assign accurate ratings, we suspect that the dearth of examples of Consequences and Response Process evidence challenged our ability to recognize patterns, and thus to formalize criteria for rating these categories.

Assessments are intended to have some desired effect (consequence), but may also have unintended effects. Therefore, analyzing Consequences of assessments may support the validity of score interpretations, or reveal unrecognized threats to validity. Yet, simply demonstrating consequences, even significant and impressive consequences, does not constitute evidence for validity unless investigators explicitly demonstrate that these consequences have an impact on score interpretations (validity).²⁶ We identified only 2 studies whose data weakly supported evidence of consequences validity. In a study by Cohen et al.⁶ teachers were given assessment scores as formative feedback. All but one of the teachers with low scores improved, as compared with moderate to outstanding teachers, whose scores generally remained the same. Although these data do not show causality, they imply that the assessment process may have influenced teaching behaviors, which in turn constitutes evidence of consequences. In another study, anecdotal reports indicated that an assessment raised awareness of effective clinical teaching behaviors.⁷ Again, this observation implies that the consequences of faculty assessment may impact on the validity of assessment interpretations.

Examining the reasoning and thought processes of learners, or systems that reduce the likelihood of response error, can provide evidence of Response Process. We identified 2 articles with Response Process data that weakly supported validity. McLeod et al.¹⁷ showed that students and resident physicians, utilizing the same forms, provide significantly different scores, although it was not entirely clear whether students and residents were assessing the same teachers. Risucci et al.¹⁹ used interitem correlations to estimate halo error, and found that more advanced trainees have lower interitem correlations (lower halo error). Findings from both these studies imply that assessments completed by different learner levels could impact on the validity of these interpretations. Unfortunately, neither author discussed this implication, nor did they discuss the reasons why different levels of learners provide different scores.

We recognize limitations to our study method. As noted above, developing and applying rating methods for Consequences and Response Process was challenging, likely because of the scarcity of evidence in the literature. Another source of variance when assigning ratings was the multitude of criteria contributing to each evidence category. For example, factor analysis, differential item functioning, and various types of reliability all provide evidence for Internal Structure validity.²⁶ Nevertheless, operational definitions became clearer as our review progressed, and by the conclusion of our rating process, we had arrived at a system that accommodated these sources of variance. An additional limitation of our scoring system was that each article was given equal weighting, despite significant heterogeneity of study methods. For example, not all articles disclosed important details regarding the assessment setting or utilized methods designed to discover anticipated outcomes. Therefore, there remains a need to evaluate the quality of the *methods* used in studies on the as-

essment of clinical teaching, separate from the evidence supporting the validity of these assessments.

Another potential limitation of our study is that ratings were assigned to categories of validity evidence on the assumption that each category has equal importance in the assessment of clinical teaching. One might argue that the lesser-used categories of validity evidence (e.g., Relation to Other Variables, Consequences, and Response Process) are underrepresented because they are less important in clinical teaching assessment. However, we are inclined to believe that these categories are underrepresented because seeking evidence from these categories requires studies designed to demonstrate these sources of evidence, and also because these categories tend to be misunderstood. Additional challenges include identifying meaningful correlates and outcomes (for evidence of Relation to Other Variables and Consequences, respectively), and utilizing samples sufficiently large to demonstrate such correlates and outcomes. While evidence for Content and Internal Structure may have primary importance during the initial development of an instrument, we urge authors to seek additional evidence of Relation to Other Variables, Consequences, and Response Process in subsequent studies. Finally, we acknowledge that accurately rating validity evidence from categories of Content, Consequences, and Response Process can be especially challenging because data from these categories are often qualitative.

Our findings have important implications for research in assessing clinical teaching in particular, and for research involving psychometric instruments in general. First, future studies should seek a broader variety of validity evidence with greater attention to the categories of Relation to Other Variables, Consequences, and Response Process. All the same, when reviewing the scores assigned to individual articles in the current study, readers are cautioned against comparing articles solely based on the sum of their evidence scores (indeed, we did not report sum scores for individual articles for this reason). Such scores are only crude estimates of validity. The study by James and Osborne,¹³ for example, did not receive the highest overall score; yet, this was the only study to receive a perfect score in the category of Relation to Other Variables.

A second implication of our findings for future research is that authors of instruments using multiple observers should more frequently report interrater reliability.³ A third implication is that future studies should clearly state hypothesized outcomes and the theoretical bases for these outcomes. Many investigators seemed to analyze existing data, and then attempt to explain the results a posteriori. Of course, not all studies of clinical teaching assessments need to be prospective. Medical educators have been encouraged to use traditional epidemiologic approaches, including retrospective cohort and case-control studies.^{32,33} Nonetheless, we stress that even retrospective studies should be theory-based and hypothesis-driven. Lastly, our study provides a reliable method for rating validity evidence from the categories of Content, Internal Structure, and Relation to Other Variables. Rating the categories of Consequences and Response Process, while not reliable in this study, may improve as authors become more comfortable with these categories and as evidence is presented more frequently in the literature. Our method would be best utilized (as we have performed in this article) to identify weaknesses in a body of literature. We encourage ed-

ucators to expand upon our method for rating the validity evidence, and to apply this or similar methods when analyzing the validity of their assessments. We anticipate that systematic methods for rating the validity evidence will benefit researchers and critical appraisers of the literature in all fields where validity is utilized.

REFERENCES

1. **Downing SM.** Validity: on the meaningful interpretation of assessment data. *Med Educ.* 2003;37:830-7.
2. **Crossley J, Humphris G, Jolly B.** Assessing health professionals. *Med Educ.* 2002;36:800-4.
3. **Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN.** How reliable are assessments of clinical teaching? A review of the published instruments. *J Gen Intern Med.* 2004;19:971-7.
4. **Beckman TJ, Lee MC, Rohren CH.** Evaluating an instrument for the peer review of inpatient teaching. *Med Teach.* 2003;25:131-5.
5. **Benbassat J, Bachar E.** Validity of students' ratings of clinical instructors. *Med Educ.* 1981;15:373-6.
6. **Cohen R, McRae H, Jamieson C.** Teaching effectiveness of surgeons. *Am J Surg.* 1996;171:612-4.
7. **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.
8. **Donnelly MB, Woolliscroft JO.** Evaluation of clinical instructors by third year medical students. *Acad Med.* 1989;64:159-64.
9. **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.
10. **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.
11. **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.
12. **Irby DM, Rakestraw P.** Evaluating clinical teaching in medicine. *J Med Educ.* 1981;56:181-6.
13. **James PA, Osborne JW.** A measure of medical instructional quality in ambulatory settings: the MedIQ. *Fam Med.* 1999;31:263-9.
14. **Litzelman DK, Westmorland GR, Skeff KM, Stratos GA.** Student and resident evaluations of faculty—how reliable are they? *Acad Med.* 1999;74(suppl):s25-7.
15. **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.
16. **McGill MK, McClure C, Commerford K.** A system for evaluating teaching in the ambulatory setting. *Fam Med.* 1986;18:173-4.
17. **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;27:48-53.
18. **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.
19. **Risucci DA, Lutsky L, Rosati RJ, Tortolani AJ.** Reliability and accuracy of resident evaluations of surgical faculty. *Eval Health Prof.* 1992;15:313-24.
20. **Shellenberger S, Mahan JM.** A factor analytic study of teaching in off-campus general practice clerkships. *Med Educ.* 1982;16:151-5.
21. **Solomon DJ, Speer AJ, Rosebraugh CJ, DiPette DJ.** The reliability of medical student ratings of teaching. *Eval Health Prof.* 1997;20:343-52.
22. **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.
23. **Tortolani AJ, Risucci DA, Rosati RJ.** Resident evaluation of surgical faculty. *J Surg Res.* 1991;51:186-91.
24. **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.
25. **Smith CA, Varkey AB, Evans AT, Reilly BM.** Evaluating the performance of inpatient attending physicians: a new instrument for today's teaching hospitals. *J Gen Intern Med.* 2004;19:766-71.
26. **American Education Research Association and American Psychological Association.** Standards for Educational and Psychological Testing. Washington, DC: American Education Research Association; 1999.
27. **Messick S.** Validity. In: Linn RL, ed. *Educational Measurement*. 3rd ed. Phoenix, Ariz: Oryx Press; 1993.
28. **Fleiss JL, Cohen J.** The equivalence of weighted kappa and the intra-class correlation coefficient as measures of reliability. *Educ Psychol Meas.* 1973;33:613-9.
29. **Fleiss JL, Cohen J, Everitt BS.** Large sample standard errors of kappa and weighted kappa. *Psychol Bull.* 1969;72:323-7.
30. **Landis JR, Koch GG.** The measure of observer agreement for categorical data. *Biometrics.* 1977;33:159-74.
31. **Downing SM.** Reliability: on the reproducibility of assessment data. *Med Educ.* 2004;38:1006-12.
32. **Carney PA, Neirenberg DW, Pipas CF, Brooks WB, Stukel TA, Keller AM.** Educational epidemiology: applying population-based design and analytic approaches to study medical education. *JAMA.* 2004;292:1044-50.
33. **Beckman TJ, Cook DA.** Educational epidemiology (letter). *JAMA.* 2004;292:1969.