Developing and Testing an Instrument to Measure the Effectiveness of Clinical Teaching in an Academic Medical Center

H. Liesel Copeland, PhD, and Mariana G. Hewson, PhD

ABSTRACT

Purpose. Instruments that rate teaching effectiveness provide both positive and negative feedback to clinician–educators, helping them improve their teaching. The authors developed the Clinical Teaching Effectiveness Instrument, which was theory-based and generic across their entire academic medical center, The Cleveland Clinic Foundation. They tested it for reliability, validity, and usability.

Method. In 1997, using an iterative qualitative development process involving key stakeholders, the authors developed an institution-wide instrument to routinely evaluate clinical faculty. The resulting instrument has 15 questions that use a five-point evaluation scale. The instrument, which was administered to medical students, residents, and fellows over a 20-month period, produced data that were rigorously tested for instrument characteristics, reliability, criterion-related and content validity, and usability.

Results. This instrument, implemented in all departments across the institution, produced data on a total of 711 clinician–educators. Correlation coefficients among the items were high (.57 to .77). The scores were reliable (*g* coefficient of 0.935), and the instrument had both content and criterion-related validity.

Conclusions. The Cleveland Clinic's Clinical Teaching Effectiveness Instrument is reliable and valid, as well as usable. It can be used as an evaluation tool for a wide variety of clinical teaching settings.

Acad. Med. 2000;75:161–166.

Ratings from students are commonly an essential component of teaching-evaluation systems in tertiary and professional educational institutions. A review of the literature[1] confirmed that scores from instruments in which students rate their teachers' effectiveness are reliable; correlatable with measures such as student learning, instructor self-evaluations, and peer ratings; and generalizable across different teaching situations. Such instruments offer quality assurance measures by providing empirical data to all levels of administrative educators. More important, clinician–educators can receive direct feedback from their trainees (medical students, residents, and fellows) on their teaching performances in diverse settings. These data can be used to reward good teachers, improve average teachers, and help all teachers who wish to enhance their abilities by giving them knowledge of their own scores on specific teaching skills. Faculty development programs can use the feedback by concentrating on those skills identified as lacking. The feedback may also be useful in making decisions about academic promotions and allocating teaching responsibilities within clinical departments. In addition, an institution-wide instrument allows medical educators to research variables that may affect teaching effectiveness, such as programmatic, teaching, and demographic differences.

The Cleveland Clinic Foundation, a large Midwestern academic medical center, previously evaluated the effectiveness of clinical teaching with diverse, department-specific instruments that lacked comparability. We needed a new instrument that was theory-based and generic across the entire institution to compare teaching competencies among faculty, divisions, and depart-

Dr. Copeland *is a postdoctoral fellow in medical education research and evaluation, and* ***Dr. Hewson*** *is director, both at the Center for Medical Education Research and Development, Division of Education, The Cleveland Clinic Foundation, Cleveland, Ohio.*

This article was first presented at the American Educational Research Association Conference, Montreal, Quebec, Canada, April 1999.

Correspondence and requests for reprints should be addressed to Dr. Hewson, Education NA-25, The Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195; e-mail: ⟨hewsonm@ccf.org⟩.

ments (which are subordinate to divisions). The key requirements for the new instrument were that it be practical and convenient to use (e.g., short, visually appealing, and scannable), useful for clinician–educators in motivating self-improvement and for the annual performance review, clinically credible for all divisions, valid, and reliable. Our goal was to develop and test an instrument that fulfills these needs.

METHOD

Development of the Instrument

We determined that the instrument should be based on the clinical education literature (so that it would be valid), that it should have broad-based support within the institution (so that it would be accepted), and that it should be useful. Therefore, we developed the Clinical Teaching Effectiveness Instrument in conjunction with current literature[2–8] and through collecting qualitative data from a series of interviews with all relevant stakeholders.

The first draft of the instrument was based on an inventory of effective clinical teaching behaviors,[3] which was based on a model of tailored clinical teaching.[4] This first draft was reviewed and modified by a committee in the department of medicine (the residency program director, the education administrator, the chief resident, a community physician, and a medical educator). This draft instrument was then modified with feedback from at least 20 meetings with representatives from each of the stakeholder groups (residency and medical student program directors, department and division chairs, educational administrators, clinician–educators, residents, and medical students) and from the major clinical teaching divisions (medicine, pediatrics, surgery, anesthesiology, radiology, and pathology). This iterative process involved asking for opinions about the important qualities of teaching to identify key items needed for the instrument. We invited feedback about the specific instrument items on each draft.

When the process of continual modification and refinement reached the "point of redundancy"[9] (i.e., the meetings no longer resulted in new ideas or disagreements), we concluded that consensus had been attained, and we finalized the instrument. Then the graduate medical education council, the medical student education council, and the educational governance committee reviewed the instrument and voted to accept it. The new instrument was then formally presented to an institution-wide meeting of all program directors. This iterative process allowed us to gain support from all areas within the institution and helped us inform people of the impending changes in the institutional evaluation system.

The Clinical Teaching Effectiveness Instrument (see boxed text) has 15 items on clinical teaching behaviors, one general item, and space for written comments. Each item is rated on a five-point scale. Resident and student raters are guaranteed anonymity. Trainees are asked to specify the length of time spent with each clinician–educator, their residency programs, and their levels of training. We use this information to study the effects of these modifying variables on measures of clinical teaching effectiveness.

Testing the Instrument

Beginning in the 1997 academic year, we implemented the new instrument across all divisions and collected ratings from medical students, residents, and fellows. Data from the instrument are collated and summarized into formal reports that are systematically fed back to individual clinician–educators, program directors, and department and division chairs. All data for each clinician–educator are explicitly reviewed as part of the formal institutional annual performance-review process. We assessed the psychometric properties of the new instrument in terms of its general characteristics, reliability, validity, and usefulness.

Statistical Analysis

We summarized the instrument characteristics using descriptive statistics (means, standard deviations, and response rates). To test for the relationships between items, we computed Pearson correlation coefficients between all 15 items and conducted a factor analysis.

To test the psychometric quality of the instrument, we estimated reliability and validity. We estimated reliability through conducting a generalizability analysis (a method of estimating, from the basis of analysis of variance, the amounts of variance added by different components of the study) and computation of a *g* coefficient (reliability-type coefficient). We assessed the internal consistency of the instrument by calculating the Cronbach alpha.

We assessed the validity of the new instrument through application of modified content (face) and criterion-related validation studies. To determine whether the items on the instrument adequately represent the domain of interest (another modified content validity study), we conducted a content analysis of a systematic random selection of 440 completed instruments (20% of the respondents' written comments). We also performed a modified content validation study by comparing our instrument with several published clinical-teaching-evaluation instruments. We conducted a criterion-related validation study to assess the relationship between scores on the instrument and a selected criterion measure. We did this by computing correlation coefficients between scores on an old teaching-effectiveness instrument (a "retrospective" criterion) with the overall scores on our new instrument.

Items on The Cleveland Clinic's Clinical Teaching Effectiveness Instrument

Rating Scales

DK/NA Don't Know/Not Applicable	1 Never/ Poor	2 Seldom/ Mediocre	3 Sometimes/ Good	4 Often/ Very good	5 Always/ Superb

Items

1. Establishes a good learning environment (approachable, nonthreatening, enthusiastic, etc.)
2. Stimulates me to learn independently
3. Allows me autonomy appropriate to my level/experience/competence
4. Organizes time to allow for both teaching and care giving
5. Offers regular feedback (both positive and negative)
6. Clearly specifies what I am expected to know and do during this training period
7. Adjusts teaching to my needs (experience, competence, interest, etc.)
8. Asks questions that promote learning (clarifications, probes, Socratic questions, reflective questions, etc.)
9. Gives clear explanations/reasons for opinions, advice, actions, etc.
10. Adjusts teaching to diverse settings (bedside, view box, OR, exam room, microscope, etc.)
11. Coaches me on my clinical/technical skills (interview, diagnostic, examination, procedural, lab, etc.)
12. Incorporates research data and/or practice guidelines into teaching
13. Teaches diagnostic skills (clinical reasoning, selection/interpretation of tests, etc.)
14. Teaches effective patient and/or family communication skills
15. Teaches principles of cost-appropriate care (resource utilization, etc.)

Statistical computations were obtained through SPSS for Windows (SPSS Inc., Chicago, IL) and generalizability analysis was conducted using Genova.[10]

Results

Instrument Characteristics

From October 1997 to March 1999, we collected a total of 8,048 completed instruments, which included 424 (5.3%) instruments left blank due to self-reported insufficient time with a faculty member and 203 (2.5%) instruments with unidentified trainee levels. A median of eight instruments per faculty member was collected for 711 educators. Each trainee could rate more than one faculty member, so trainees may be represented multiple times in a count. The total numbers of instruments by staff department and evaluator status are given in Table 1.

The average rating for all 15 items was 4.12 (SD = .772), with mean ratings for individual items ranging from 3.92 to 4.25. Twelve of the 15 items had response rates of 92.6% or higher (7,043 to 7,595 of 7,624 returned instruments with ratings). The most frequently skipped questions had response rates of 89% ("teaches principles of cost-appropriate care"), 88% ("coaches my skills"), and 86% ("teaches communication skills"). Though the distribution of the scores is negatively skewed, parametric and nonparametric tests gave the same findings, and only parametric tests are reported here. We note the lack of data for anesthesiology at the time of this analysis and are currently investigating the reasons.

The Pearson correlation coefficient showed that all the items on the instrument are interrelated, with correlations ranging from .57 to .77. The factor analysis of the 15 rating items resulted in a single component explaining 70.6% of the variance, indicating that we are measuring one core concept. All 15 rating items loaded with at least .79 weight. The four highest loading items were "adjusts teaching to my needs," "adjusts teaching to diverse settings," "teaches diagnostic skills," and "asks questions that promote learning."

Reliability

To check the reliability of the instrument, we calculated a *g* coefficient. This required using the same number of instruments for every faculty member. Thus, we randomly selected five instru-

Table 1

Numbers of Clinical Teaching Effectiveness Instruments Returned by Raters, by Training Level and Division, The Cleveland Clinic, 1997–98

	Trainees		
Department	Medical Students	Residents	Fellows
Anesthesiology	74	14	—
Medicine	704	1,804	1,035
Pathology	16	791	18
Pediatrics	212	1,336	52
Radiology	45	227	85
Surgery	472	706	254
TOTAL*	1,523	4,878	1,444

*Of 7,845 instruments plus 203 received from raters with unidentified trainee levels.

Table 2

Generalizability Study: Variance Component Estimates of the Rater-within-educator-by-item Design for the Teaching Effectiveness Evaluation Instrument Ratings from 1,475 Trainees, The Cleveland Clinic, 1997–98

Source	Degrees of Freedom	Variance Component (Expected Mean Square)
Clinician educator (c)	294	.2609
Item (i)	14	.1919
Rater within educator (r:c)	1,180	.0478
Educator–item interaction (ci)	4,116	.3813
Item–rater interaction (ir:c)	16,520	.7732

ments per faculty member and used a final data set of 295 clinician–educators. In our study design, we acknowledged that differences between trainees on evaluating the same faculty member could come from three sources (effects): (1) the clinician–educators, (2) the items, and (3) the trainees (raters) who were nested within clinician–educators (i.e., every clinician–educator was rated by different trainees). When computing the *g* coefficient, we chose a design where items were fixed at 15 and raters were considered random. This design matched our interest in determining the characteristics of the measurement of teaching effectiveness as defined by only the 15 specific items on our instrument and rated by any trainee similar to the ones used in our study. Table 2 presents the variance component estimates for the generalizability study.

Analysis of these variance estimates for the study effects (which indicate the most likely source contributing to the differences in an individual's rating) indicate that the largest source of error contributing to differences in ratings of one faculty member was due to trainees' interpreting items differently. The *g* coefficient for our design of five random raters (r) nested within clinician–educators (c) crossed with 15 fixed items (i)—[(r:c) × i]—was calculated to be .935. Even if we were to use only one trainee per educator, the *g* coefficient would be .742; with seven trainees per educator, it would rise to .953. Thus, ratings from our instrument are reliable. This also means that the 95% confidence interval for the mean is ± .377 for five trainees (± .752 if one trainee is used). The internal consistency of our instrument was high (Cronbach coefficient alpha = .97).

Validity

Content validity. The content analysis of the written comments indicated that the instrument was not missing any distinctively unique category. The resulting categories of the written comments corresponded to the concepts in all 15 (100%) of the questions on our instrument. Many of the comments supported the importance of establishing a good learning environment, providing regular feedback, using questions effectively, stimulating independent learning, and incorporating current literature into teaching. Additionally, trainees' comments reinforced how valuable it was for clinician–educators to provide appropriate autonomy, organize their time for teaching and care giving, and give explanations for opinions, advice, and actions.

From the comments, we identified a few additional areas not fully captured in any of the items: the willingness to teach and initiate discussions rather than simply being available to answer questions, actively involving trainees in decision making or allowing trainees to express their opinions first, teaching with an organized approach that role-modeled or demonstrated through questioning a logical thinking method, providing practical clinical information or medical "pearls," and incorporating the current clinical experience to highlight key points and emphasize teaching. Additionally, some trainees praised the "fund of knowledge" displayed by their faculty, but, after deliberating with the

Table 3

Criterion-related Validity Coefficients Generated by Correlating the New Instrument with the Former Instrument and the Corresponding Descriptive Statistics for Each Criterion

	No.	Mean (SD)	Correlation with New Instrument (Pearson *r*)	*p* Value
Average of all former instrument items	351	4.08 (0.559)	.428	<.01
Former instrument "overall" item average	350*	4.09 (0.649)	.433	<.01
Average of new instrument items	351	4.11 (0.517)	—	—

*The number for the former instrument is different due to one faculty member's not receiving a score on the overall item.

stakeholders, we decided that it was not justifiable for trainees to rate the knowledge levels of the faculty within this instrument.

Our instrument has good content validity as assessed through comparisons with other instruments. Of the 15 items on an unpublished 1993 University of Toronto's faculty teaching instrument,[11] 11 (73%) were represented on our instrument. Of the 18 items on Westberg and Jason's sample instrument,[7] nine (50%) were embodied in ours. Of the 23 items on one of our original instruments (used for criterion validity), 11 (48%) are represented on the new instrument.

We also assessed validity and comprehensiveness by analyzing, during item development, the extents to which concepts represented by our items were congruent with concepts expressed in the literature and by our faculty and trainees. Complete congruence was obtained for five concepts: (1) offers feedback, (2) establishes a good learning climate, (3) coaches my clinical/technical skills, (4) teaches medical knowledge (diagnostic skills, research data and practice guidelines, communication skills, cost-appropriate care), and (5) stimulates independent learning. Concepts that were common in the literature[2–8] but infrequently mentioned by our respondents were: (1) adjusts teaching to the learner's needs, (2) asks questions to actively involve learners, (3) specifies expectations, and (4) gives clear explanations and answers questions. Concepts mentioned by our respondents but less commonly in the literature were: (1) provides autonomy, (2) organizes time for teaching and care giving, and (3) adjusts teaching to diverse settings (bedside, exam room, operating room, view box).

Criterion-related validity. To calculate criterion-related validity coefficients, we used scores for 351 faculty from the most frequently used former teaching instrument (used by all trainees except those in anesthesiology and internal medicine). This instrument was composed of 22 specific short items on various teaching concepts and a single item that asked for an overall evaluation. We used both the average of the 22 items and the overall score as criterion measures. Validity coefficients were calculated by computing Pearson correlation coefficients between the average score from all items on our new instrument with the average score of the 22 items on the former instrument and the overall item score. Table 3 gives the mean scores, number of staff used in the calculations for each measure, and the correlation values. Faculty scores used in the calculation were mean scores from an average of eight trainees (range of one to 38 trainees rating each faculty member). This validity coefficient suggests that a fundamental criterion of teaching is being assessed and indicates that the instrument is valid.[1]

Usability

The new instrument is highly usable, though we are not using a scanner due to incompatibility in saving written comments. Currently, reports are generated throughout the year, once for each department, to be used during each clinician–educator's annual performance review. These reports include summary data that are distributed to the department and division chairs, program directors, and individual clinician–educators. This high usability is demonstrated by its adoption as a measure in all of the clinical departments. This, along with the congruence to the clinical teaching literature, suggests that this instrument is generalizable. Usability also incorporates the idea of individual clinician–educator's being able to apply their ratings to their own specific behaviors and improve their teaching. Anecdotal reports from clinician–educators suggest that the new Clinical Teaching Effectiveness Instrument is raising awareness of effective clinical teaching behaviors and that more people are seeking help with their teaching.

Discussion

Based on a needs assessment of relevant stakeholders (faculty, trainees, program directors, and chairs) and using specific educational principles, we developed the Clinical Teaching Effectiveness Instrument, which is reliable, valid, and us-

able. We are now able to give consistent regular feedback to all the program directors who make decisions about teaching assignments. We are also able to conduct studies of variables that may impact teaching effectiveness. For example, we intend to investigate the effects of trainee level (medical student, resident, or fellow) and the differences between ambulatory and inpatient teaching.

The high mean and slight skewness of our data imply that we have a tendency toward a ceiling effect. We interpret this to indicate that while we can differentiate between teachers of high and low ability, we are less able to differentiate among highly competent teachers. This is not a serious concern, since our aim is to ensure that all faculty teach effectively and, through faculty development, help those who have not achieved this level. Although response rates for the individual items varied, none was so low as to jeopardize interpretation. The high intercorrelations and the factor analysis show that this instrument has a high internal consistency, signifying that all the items are related and each trainee tends to give consistent ratings across items. Additionally, we believe that we did not get a higher criterion-related validity coefficient because our new instrument is providing more specific information about clinical teaching behaviors than did the former instrument.

Our instrument is generalizable to clinical teaching in a variety of settings, as evidenced by its adoption by different clinical departments and its use in both inpatient and outpatient settings. We did not design the instrument to assess the full range of teaching skills such as lecture and problem-based learning instructional skills. Therefore, its generalizability is limited to the measurement of clinical teaching; it may not cover some of the non–clinical-teaching activities that occur in academic medical centers.

Further limitations arise from differences among the departments. Though they use the same instrument, divisions and departments differ in how much time they give residents and fellows to complete the instrument and in the frequencies throughout the year at which instruments are collected. Departments also differ in their numbers of faculty. Trainees, who always complete instruments for more than one faculty member, may lose focus when they must rate a large number of teachers. Additionally, because we wanted to generalize across departments, the instrument may lack some specificity for individual departments. Despite this, we believe that the high validity, good reliability, and specific behavior-based items result in a usable instrument.

Though trainees' ratings of the faculty are a highly valued component of teaching evaluation, it is advisable also to gather other types of data for a complete evaluation of teaching effectiveness. Alternative types include peer evaluations, self evaluations, and observations.

Conclusion

All the divisions at The Cleveland Clinic now use the Clinical Teaching Effectiveness Instrument, and they all (except anesthesiology) routinely collect and score the data. The instrument is reliable and valid, as well as usable. Based on our preliminary data, the instrument is useful in measuring improvement among our faculty. Results from the instrument can also easily be applied to promote self-improvement among our faculty. The items represent specific theoretical constructs important to clinical teaching and therefore are generalizable. The strength of our instrument lies in the qualitative development process of iterative meetings with key stakeholders and informants and the ability to provide a thorough explanation of, and justification for, our measure of teaching effectiveness. We can now compare the teaching of individuals, different departments, and divisions throughout the institution, and we can address research questions concerning variables affecting clinical teaching. By providing a psychometrically sound and theory-based instrument, we can not only improve the teaching at this academic medical center but also promote the importance of clinical teaching and demonstrate the value the institution places on such efforts through appropriate responses.

The authors thank all the program directors and administrators within the medical student program and the 144 residency and fellowship programs, in every division at The Cleveland Clinic, who make the application of this teaching-effectiveness instrument a reality. Special thanks to Andrew Fishleder, MD, Garron Weiker, MD, and Jeffery Hutzler, MD, for their support of this venture and to Wilma Doyle, MA, and Patricia Chapek, MBA, for their administrative support.

References

1. Pritchard RD, Watson MD, Kelly K, Paquin AR. Helping Teachers Teach Well. San Francisco, CA: New Lexington Press, 1998.
2. Anderson DC, Harris IB, Allen S, et al. Comparing students' feedback about clinical instruction with their performances. Acad Med. 1991;66:29–34.
3. Hewson MG, Jensen NM. An inventory to improve clinical teaching in the general internal medicine clinic. Med Educ. 1990;24:518–27.
4. Hewson M. Clinical teaching in the ambulatory setting. J Gen Intern Med. 1992;7:76–82.
5. Irby DM. What clinical teachers in medicine need to know. Acad Med. 1994;69:333–42.
6. Skeff KM, Stratos GA, Bergen MR. Evaluation of a medical faculty development program: a comparison of traditional pre/post and retrospective pre/post self assessment ratings. Eval Health Prof. 1992;15:350–66.
7. Westberg J, Jason H. Collaborative Clinical Education: The Foundation of Effective Health Care. New York: Springer Publishing, 1993.
8. Whitman N, Lawrence P. Surgical Teaching: Practice Makes Perfect. Salt Lake City, UT: Department of Family and Preventive Medicine, University of Utah School of Medicine, 1991.
9. Lincoln YS, Guba EG. Naturalistic Inquiry. Beverly Hills, CA: Sage Publications, 1985.
10. Brennan RL. Elements of Generalizability Theory. Iowa City, IA: ACT publications, 1992.
11. Slade S. Evaluation of Faculty Teaching. University of Toronto, Department of Family and Community Medicine, Toronto, ON, Canada. Unpublished instrument, 1993.