Improving Participant Feedback to Continuing Medical Education Presenters in Internal Medicine: A Mixed-Methods Study

Christopher M. Wittich, MD, PharmD¹, Karen F. Mauck, MD, MS¹, Jayawant N. Mandrekar, PhD², Karol A. Gluth, BA³, Colin P. West, MD, PhD¹, Scott C. Litin, MD¹, and Thomas J. Beckman, MD¹

¹Department of Internal Medicine, Division of General Internal Medicine Mayo Clinic College of Medicine, Rochester, MN, USA;

²Department of Health Sciences Research, Division of Biomedical Statistics and Informatics Mayo Clinic College of Medicine, Rochester, MN, USA;

³Mayo School of Continuous Professional Development Mayo Clinic, Rochester, MN, USA.

BACKGROUND: Feedback is essential for improving the skills of continuing medical education (CME) presenters. However, there has been little research on improving the quality of feedback to CME presenters.

OBJECTIVES: To validate an instrument for generating balanced and behavior-specific feedback from a national cross-section of participants to presenters at a large internal medicine CME course.

DESIGN, SETTING, AND PARTICIPANTS: A prospective, randomized validation study with qualitative data analysis that included all 317 participants at a Mayo Clinic internal medicine CME course in 2009.

MEASUREMENTS: An 8-item (5-point Likert scales) CME faculty assessment enhanced study form (ESF) was designed based on literature and expert review. Course participants were randomized to a standard form, a generic study form (GSF), or the ESF. The dimensionality of instrument scores was determined using factor analysis to account for clustered data. Internal consistency and interrater reliabilities were calculated. Associations between overall feedback scores and presenter and presentation variables were identified using generalized estimating equations to account for multiple observations within talk and speaker combinations. Two raters reached consensus on qualitative themes and independently analyzed narrative entries for evidence of balanced and behavior-specific comments.

RESULTS: Factor analysis of 5,241 evaluations revealed a uni-dimensional model for measuring CME presenter feedback. Overall internal consistency (Cronbach alpha=0.94) and internal consistency reliability (ICC range 0.88–0.95) were excellent. Feedback scores were associated with presenters' academic ranks (mean score): Instructor (4.12), Assistant Professor (4.38), Associate Professor (4.56), Professor (4.70) ($p=0.046$). Qualitative analysis revealed that the ESF generated the highest numbers of balanced comments (GSF=11,

ESF=26; $p=0.01$) and behavior-specific comments (GSF=64, ESF=104; $p=0.001$).

CONCLUSIONS: We describe a practical and validated method for generating balanced and behavior-specific feedback for CME presenters in internal medicine. Our simple method for prompting course participants to give balanced and behavior-specific comments may ultimately provide CME presenters with feedback for improving their presentations.

KEY WORDS: continuing medical education; feedback; internal medicine; participants.

J Gen Intern Med 27(4):425–31

DOI: 10.1007/s11606-011-1894-3

© Society of General Internal Medicine 2011

BACKGROUND

Meaningful feedback is essential for improving the skills of all teachers,^{1–5} including continuing medical education (CME) presenters. Indeed, research indicates that physician reflection on positive and constructive assessment feedback is important for understanding, accepting and applying that feedback.^{2,6}

Evaluation and feedback are uniquely different: evaluation is summative and involves judgment, whereas feedback is formative and specifically intended to improve effectiveness.^{7,8} It is understood that useful feedback is provided in a timely fashion, behavior-specific, and balanced with both positive and constructive elements.⁷ Behavior-specific feedback is important because, unlike vague or judgmental comments, it identifies tangible actions for learners to improve upon. Feedback that is balanced (e.g., containing both positive and constructive elements) is particularly useful for poor performers, because it makes the overall feedback more acceptable, thus allowing learners to reflect more comfortably on the constructive feedback component. Reflection on feedback is important, because it has been observed that reflection is the critical link between receiving and using assessment feedback.² Unfortunately, the feedback provided to CME presenters often lacks mention of specific behaviors, thus providing presenters with no means for improvement.²

CME presentations are usually delivered from the podium, yet there are surprisingly few studies examining characteristics of lecture-based presentations. Experts have suggested that effective lecture-based presentations are clear, organized,

Electronic supplementary material The online version of this article (doi:10.1007/s11606-011-1894-3) contains supplementary material, which is available to authorized users.

Received May 29, 2011

Revised August 9, 2011

Accepted September 12, 2011

Published online September 27, 2011

relevant, use patient cases and high-quality audiovisuals, and provide supporting evidence.⁹⁻¹³ Likewise, successful CME presenters are observed to be prepared, organized, respect time limits, identify key points, and facilitate interactions with course participants.⁹⁻¹³

Validated feedback instruments for assessing CME faculty have been reported. One questionnaire focused on characteristics of the presenter, presentation, and content of CME presentations,⁹ yet this questionnaire did not include behavior-specific feedback. While other studies have explored peer feedback¹⁴ and the attributes of an effective medical lecture,^{10,11} these studies did not incorporate the attributes of an effective lecture into an evaluation instrument. We are unaware of validated instruments that are designed to improve feedback to CME faculty by generating balanced and behavior-specific comments from course participants.

The goals of this study are 1) to develop and validate an instrument for generating balanced and behavior-specific feedback from a national cross-section of participants to presenters at a Mayo Clinic CME course, 2) to determine whether CME teaching effectiveness scores are associated with characteristics such as presenter experience and the use of interactive cases; and 3) to qualitatively review participants' narrative comments for evidence of meaningful feedback that results from a balanced and behavior-specific CME faculty assessment instrument.

METHODS

Study Design and Participants

This was a prospective, randomized validation study with qualitative data analysis that included all participants (N=317) at the Mayo Clinic Selected Topics in Internal Medicine (STIM) 2009 course. STIM is an annual, week-long course that is accredited by the Mayo School of Continuous Professional Development (MSCPD) and qualifies participants to receive 24.25 hours of CME credit. The format consists of 30- to 45-minute long podium presentations. Presenters are encouraged (but not required) to use clinical cases and the audience response system. Course faculty members from the Mayo Clinic College of Medicine are

selected by the course directors to give one or two presentations. Course participants typically represent a broad geographic cross-section including the United States, Canada, and Europe. The MSCPD requires that 20% of participants are given a standard feedback form (SF) used by all courses accredited by the school. The SF consisted of two items (Practical Value/Content and Presentation Skills) using a 5-point Likert scale (1=very poor, 2=poor, 3=average, 4=good, 5=excellent). In this investigation, a new form designed to elicit balanced and behavior-specific feedback was developed and analyzed. This study was deemed exempt by the Mayo Clinic Institutional Review Board.

Feedback Instrument Development

The new CME faculty assessment instrument (individual items in Table 1, entire form in [online Web Appendix](#)) was designed to gather behavior-specific and balanced feedback for CME faculty presenters. The content for this new assessment instrument was derived from existing literature^{7,9-12,14} and input from a panel of Mayo experts with experience in evaluating CME presenters. Specifically, the Likert-scaled items and open-ended questions for this instrument were developed based on the principles of effective presentations (clear [item 1], organized [item 1], relevant [item 2], use of patient cases [item 3], high-quality audiovisuals [item 4], and providing supporting evidence [item 6]) and successful CME presenters (preparation [item 6], organization [item 1], following time limits [item 7], identifying key points [item 8], and facilitating interactions with course participants [item 5]).⁹⁻¹³ The instrument was refined by eliminating one item ("speaker was engaging") based on further review and consensus among the panel of Mayo experts (authors CMW, KFM, JNM, and TJB) who had experience in assessment research and directing CME courses. In order to optimize instrument accuracy and usability, we ultimately agreed upon 8 items based on 5-point Likert scales (1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree).

Feedback Instrument Data Collection

Participants at the 2009 STIM course were randomized using a computerized random number generator to receive either 1)

Table 1. CME Faculty Assessment Instrument: Mean Scores and Reliability

Items	Mean	Standard Deviation	ICC*	95% CI**	Cronbach's alpha***
Speaker presented information in a clear and organized manner	4.62	0.61	0.93	(0.90, 0.96)	
Content was relevant	4.61	0.63	0.92	(0.88, 0.95)	
Examples or cases were given to facilitate my understanding	4.51	0.76	0.94	(0.91, 0.97)	
The slides added to the effectiveness of the presentation	4.50	0.75	0.95	(0.93, 0.97)	
Speaker included opportunities to learn interactively	4.37	0.88	0.94	(0.91, 0.97)	
Speaker provided adequate evidence to support the content, presented neither too much nor too little	4.49	0.72	0.88	(0.82, 0.93)	
An appropriate amount of information (content) was provided, neither too much nor too little	4.49	0.75	0.92	(0.87, 0.95)	
Speaker summarized key points	4.56	0.68	0.88	(0.82, 0.93)	
Overall (all 8 items)	4.52	0.61			0.94

*ICC=Intraclass correlation coefficient, a measure of interrater reliability

**CI=Confidence interval

***Cronbach's alpha represents internal consistency reliability for all eight items

the Standard Form (SF), which is required by the MSCPD to be distributed to 20% of participants for accreditation, 2) the Generic Study Form (GSF) consisting of eight items on 5-point scales and a generic request for comments, or 3) the Enhanced Study Form (ESF) consisting of the same eight items on 5-point scales with a request for balanced and behavior specific feedback, along with brief descriptions of what is meant by “balanced” and “behavior specific.” That is, comments on the GSF were solicited with the phrase: “Comments.” Alternatively, comments on the ESF were solicited with the phrases: 1) “What did the speaker do well? (Please describe specific behaviors and provide examples)”, and 2) “What could the speaker improve on? (Please describe specific behaviors and provide examples).” The SF was randomly distributed to 20% of participants, the GSF to 40% of participants, and the ESF to 40% of participants. The evaluation forms were included in each registration packet and participants completed only one type of form throughout the CME course. Participants were asked to complete evaluations on all course presentations and place their completed packets in a box at the registration desk at the conclusion of the course. Lastly, all of the data, including both Likert-scale responses to each instrument item and participants’ hand-written narrative comments, were entered into an electronic database to facilitate quantitative and qualitative data analysis.

Quantitative Data Analysis

The Likert-scaled items for the Generic and Enhanced Study Forms, which were identical, were combined for the factor analysis. In order to account for correlation due to multiple ratings within a speaker and talk combination, we first generated an adjusted correlation matrix using a generalized estimating equations approach. This adjusted correlation matrix was then used to perform factor analysis. Additionally, for a sensitivity analysis, we performed factor analysis using unadjusted correlation matrix and within speaker and talk combinations separately. Factors were extracted using the minimal proportion criteria. The final model was confirmed by inspecting a scree plot. Items with factor loadings ≥ 0.50 were retained. Interrater reliability for each item was calculated using the intra-class correlation coefficient with corresponding 95% confidence intervals. Intraclass correlations were interpreted according to Landis and Koch as follows: coefficients < 0.4 poor, 0.4 to 0.75 fair to good, and > 0.75 excellent.¹⁵ Internal consistency reliability for items comprising each factor and overall were calculated using Cronbach’s coefficient alpha, where alpha > 0.7 is considered acceptable.^{16,17}

Associations between the feedback instrument item and overall scores and presenter and presentation variables were determined. Overall feedback scores were reported as the mean and standard deviation of all eight instrument items. Presenters’ academic ranks (no academic rank, instructor, assistant professor, associate professor, professor), presenters’ practice types (internal medicine, internal medicine specialty, non-internal medicine specialty), use of cases during the presentation (yes, no), and use of the audience response system during the presentation (yes, no) were treated as categorical variables. Overall feedback scores and

presenter and presentation variables were compared using generalized estimating equations to account for multiple observations within various talk and speaker combinations. Statistical significance was set at $p < 0.05$. Statistical analyses were conducted using SAS (SAS Institute Inc., Cary, NC).

Qualitative Data Selection and Analysis

The qualitative approach for this study was content analysis, since we had identified the coding categories of balanced (positive and constructive) and behavior-specific comments prior to the qualitative analysis. Initially, five GSF and five ESF forms were selected from the database using a computerized random number generator. In terms of the thematic data analysis, two raters (authors CMW and KFM) – who throughout the process were blinded to the type of form and names of speakers and participants—independently analyzed the forms for hand-written balanced comments (both positive and constructive statements, versus the presence of only one or the other) and hand-written behavior-specific comments (descriptions of actual behaviors [e.g., “Need to enunciate and slow down when speaking; Confusing because there were too many words on the slides”], versus simple judgments [e.g., “nice talk”]). This approach of using the study’s conceptual framework (i.e., elements of effective feedback) as a guide for identifying content within the narrative comments has been described.¹⁸

The raters then discussed their findings on this first round of qualitative data gathering in order to calibrate on their understandings and definitions of the two themes; their level of agreement on this calibration round, even prior to discussion and consensus, was excellent (kappa=1.0 for “balanced;” kappa=0.85, 95% CI 0.56–1.00 for “behavioral;” and kappa=0.80, 95% CI 0.44–1.00 for “judgmental”). After the calibration process, each rater analyzed 50 GSF and 50 ESF, thus generating a total of 200 forms for qualitative data analysis and reporting. The final data were presented as total counts for each of the two themes for both the GSF and ESF, as shown in Table 3. Also, to provide additional information regarding the comparative richness of data for the two forms, the median (minimum, maximum) number of words in the free text response was reported. Fisher’s exact test (for balanced, behavioral, and judgmental comparisons) and Wilcoxon rank sum (for median word count comparison) were used to determine if there were statistically significant differences between the narrative comments on the GSF and ESF.

RESULTS

Characteristics of Participants and Presenters

Among the 317 participants at the STIM 2009 course, 209 (65.9%) were men and 108 (34.1%) were women. The vast majority of participants had an MD degree (267, 84.2%) and specialized in internal medicine (120, 37.9%), family medicine (88, 27.8%), or medical subspecialties (39, 12.3%). The participants were from the across the United States (Midwest 207, 65.3%; West 69, 21.8%; South 19, 6.0%; Northeast 7,

2.2%), Canada (14, 4.4%), and Europe (1, 0.3%). There were 20 course faculty presenters who gave a total of 37 presentations. The majority of presenters practiced internal medicine (8, 40.0%) or a medical subspecialty (7, 35.0%). Presenters included seven (35.0%) assistant professors, six (30.0%) associate professors, and five (25.0%) professors. The majority of presenters elected to incorporate clinical cases (27, 73.0%) and the audience response system (24, 64.9%) into their presentations.

Feedback Instrument Validation

Of the 317 STIM 2009 course participants, 63 (20%) received the SF, 127 (40%) received the GSF, and 127 (40%) received the ESF for the 37 presentations. A total of 6,674 evaluations were completed (out of 11,729 possible forms, response rate=56.9%). For the GSF and ESF, factor analysis revealed a single factor (Eigenvalue=4.60), representing a uni-dimensional model for measuring CME presenter feedback. Individual item loadings ranged from 0.66 to 0.84. Overall, the extracted factors accounted for 100% of the shared variance among the original variables. Notably, factor analysis involving both the adjusted correlation matrix and unadjusted correlation matrix and within speaker and talk combinations separately yielded the same uni-dimensional model.

The SF mean score for practical value/content was 4.54 (standard deviation 0.34) and for presentation skills was 4.49 (standard deviation 0.37). The GSF and ESF combined item mean scores ranged from 4.37 to 4.62 (Table 1). The overall mean score for all eight items was 4.52 (standard deviation 0.61). The item "speaker presented information in a clear and organized manner" received the highest scores, while the item "speaker included opportunities to learn interactively" received the lowest. The interrater reliabilities were excellent, with intra-class correlation coefficients ranging from 0.88 to 0.95 across the eight items. The overall internal consistency reliability was excellent, with the Cronbach's alpha=0.94 for 8 items combined, indicating that the assessment form could

be shortened to include fewer items while potentially maintaining adequate internal consistency reliability.

Associations Between Feedback Instrument Scores and Characteristics of the Presenters and Presentations

Generalized estimating equations indicated that presenter feedback scores were positively associated with the presenter's academic ranks (mean score): Instructor (4.12), assistant professor (4.38), associate professor (4.56), professor (4.70) ($p=0.046$, Table 2). There were no statistically significant associations between presenter feedback scores and presenter practice type, use of clinical cases, or use of the audience response system.

Qualitative Analysis of Participants' Narrative Comments

There was a striking difference between the GSF and ESF with respect to the numbers of balanced and behavior-specific comments, as shown in Table 3. Notably, there were major differences in the numbers of balanced comments (GSF=11, ESF=26; $p=0.01$) and behavior-specific comments (GSF=64, ESF=104; $p=0.001$), thus indicating that simply prompting course participants to provide descriptions of balanced and behavior-specific comments is an effective means for generating meaningful feedback for presenters. The median word count for narrative comments was higher in the ESF (7) versus the GSF (6.5) although this did not reach statistical significance.

DISCUSSION

To our knowledge, this is the first study regarding a practical and validated method for generating balanced

Table 2. Associations between CME Faculty Assessment Instrument Mean Scores and Presenter and Presentation Variables

Variable	Number of Evaluations*	Mean Assessment Score	Standard Deviation	p value**
Presenter Practice Type				0.307
Non-internal medicine	1319	4.39	0.77	
Internal medicine	1828	4.49	0.57	
Internal medicine specialty	2063	4.62	0.48	
Presenter academic rank				0.046
No academic rank	144	4.40	0.59	
Instructor	274	4.12	0.73	
Assistant professor	1646	4.38	0.73	
Associate professor	1704	4.56	0.51	
Professor	1442	4.70	0.44	
Clinical cases used during presentation				0.321
No	1389	4.29	0.77	
Yes	3821	4.60	0.51	
Audience response system used during presentation				0.266
No	1787	4.28	0.75	
Yes	3423	4.64	0.47	

*Total number of evaluations is 5210 which represents the generic and enhanced study forms only

**Overall feedback scores and presenter and presentation variables were compared using generalized estimating equations to account for multiple observations within various talk and speaker combinations

Table 3. Balanced and Behavior-Specific Feedback in Generic and Enhanced Assessment Instruments

	Generic Study Form	Enhanced Study Form	p value**
Forms evaluated	100	100	
Forms with balanced comments	11 (11.0%)	26 (26.0%)	0.01
Number of behavioral comments*	64/152 (42.1%)	104/172 (60.5%)	0.001
Number of judgmental comments*	88/152 (57.9%)	68/172 (39.5%)	0.001
Median (minimum, maximum) number of written words per form	6.5 (1, 29)	7 (0, 57)	0.36
Examples of balanced comments (Included both positive and constructive comments)	The speaker kept my attention by using cases and humor. However, please speak louder The drugs you chose to present were relevant to my practice. The talk was too long Provided examples from the current literature. Please provide summary slides or tables of the most important points		
Examples of behavioral comments (Based on things the presenter actually said or did)	Need to enunciate and slow down when speaking Confusing because there were too many words on the slides The outline and tables made the presentation easy to follow Just read the slides You did not state the objectives of your talk		
Examples of judgmental comments (Opinions, views, or beliefs not supported by observed behaviors)	Excellent presentation Nice talk Funny Informative Terrible lecture Learned a ton The presenter needs to lose weight		

*Values reflect numbers of comments for a given category (behavioral or judgmental) in the numerator, over the total numbers of discrete comments reviewed on the forms in the denominator

**Fisher's exact test used for balanced, behavioral, and judgmental comparisons; Wilcoxon rank sum used for word count comparisons

and behavior-specific feedback from course participants to presenters of CME in any specialty, including internal medicine.

We found that CME course participants who were randomized to an assessment form with simple prompts were able to provide robust and meaningful feedback, as compared to participants who were simply asked to "comment." That is, when participants were given simple descriptions (e.g., give both positive and constructive comments and list specific behaviors), they were willing and able to generate meaningful balanced and behavior-specific feedback, which is regarded as essential for improving both teaching in general and CME in particular.¹⁻⁵ This study finding has improved our method for collecting narrative comments at Mayo Clinic CME meetings, and shows some potential for arming CME presenters with information to improve their skills. However, our study findings may be useful only to the extent that CME presenters seriously reflect on the feedback that they receive, as it has been shown that reflection on assessment feedback is an important step towards understanding, accepting and applying feedback among physicians.²

In this study, feedback scores by CME course participants were positively associated with the presenters' academic ranks. Likewise, other studies have shown that students' ratings of lecture-based presentations were positively correlated with their teachers' academic ranks.¹⁹⁻²² Studies of clinical teachers, however, have demonstrated an *inverse* correlation between teaching scores and academic rank.²³⁻²⁵ This may be because higher-ranking faculty members in clinical settings spend more time on research and administrative activities, perhaps at the cost of time and energy that would otherwise be spent developing interpersonal relationships with their learners.²³ The conflicting observations in

these studies could reflect the different skills required to give a CME presentation versus teaching in a clinical setting. Most classroom lectures and CME presentations are delivered to relatively larger audiences, which constrains the presenters' opportunities to generate discussions and facilitate interactions with students or course participants. Additionally, CME participants may be more inclined to equate presenter excellence with advanced experience as reflected by academic rank. On the other hand, teaching in clinical settings usually involves smaller groups of learners and a greater opportunity for group-based discussion, which, in turn, is enhanced by teachers who are approachable and by minimal psychological size differentials between teachers and learners.

Systematic reviews of CME program evaluations have demonstrated the widespread use of methods with poor validity evidence,^{26,27} a focus on learner outcomes,²⁸ and neglect of the higher Kirkpatrick outcomes of behavior and results.²⁷ Our study addresses these limitations by providing a measure of feedback for CME presenters that is based on observed behaviors and is supported by strong validity evidence. An accepted validity paradigm in medical education states that construct validity is upheld by evidence from the following sources: content, response process, internal structure, relations to other variables (criterion), and consequences.²⁸⁻³³ In this study, *content evidence* draws from items that reflect literature on qualities of CME presenters and presentations^{7,9-12,14} and revision by a panel with expertise in teaching CME. *Internal structure* evidence is supported by factor analysis showing a unidimensional assessment of feedback to CME presenters and by excellent overall internal consistency reliability. *Relations to other variables* (criterion) evidence is established by a positive association between CME feedback

scores and presenters' academic ranks. Our previous review of the literature indicated that these sources of validity evidence are the categories most commonly reported in medical education studies.³³ Our CME feedback instrument should add to the limited armamentarium of validated methods for assessing CME presenters. Nonetheless, in the future it will be important to establish *Consequences* evidence by determining whether our method actually improves presentations, and ultimately, better patient care. For example, a next step will be to determine whether concerted efforts to improve presenter-participant interactions at CME courses (e.g., polling the audience to raise their hands and encouraging audience comments amidst presentations) will increase feedback scores for presenters. Additionally, in the future it will be important to establish *Consequences* evidence by determining whether CME presenters reflect on the feedback that is generated by our study method, and whether this reflection actually results in improved presentations.

This study has limitations. It involved a CME course from one institution, which could limit generalization of the findings to other settings. However, the presenting faculty reflected a range of academic ranks within Mayo Clinic, and our study sample included a broad cross-section of national and international course participants. While the participation rate was 56.9%, this is fairly typical for physician survey research studies.^{34,35} Although there were twice as many balanced comments on the ESF (26) than the GSF (11), this number still remains relatively low, possibly indicating that it is difficult to generate balanced comments by any method. We acknowledge that although our study intervention generated more balanced and behavior specific feedback than the standard assessment method, we have yet to demonstrate whether CME presenters actually reflect on this feedback in order to improve their presentations. CME presenter evaluations were clustered within the participants; yet, we accounted for this by employing conservative statistical methods including factor analysis based on adjusted correlation matrixes to account for within speaker and talk combinations, and calculations of associations using generalized estimating equations to address repeated measures within presenters. Finally, we acknowledge that survey non-respondents may have had different views than respondents regarding the quality of presentations.

CONCLUSIONS

We describe a new, practical, and validated method for improving balanced and behavior-specific feedback for CME presenters in internal medicine, which should add to the dearth of existing validated methods for assessing CME presenters. The observed association between feedback assessment scores and presenter academic rank may indicate that CME course participants believe that more experienced presenters give higher quality presentations. We encourage future research on improving the skills of junior CME faculty and on the contribution of teacher-learner interactivity to the quality of CME presentations.

Our simple method for prompting course participants to give balanced and behavior-specific comments was highly effective, should be easily replicated in future CME courses, and should ultimately provide presenters with feedback for improving their presentation skills.

Acknowledgments: We acknowledge Ms. Gail Ludens for her administrative support.

Funding: None.

Conflict of Interest: None disclosed.

Ethical Approval: This study was deemed exempt by the Mayo Clinic Institutional Review Board.

Corresponding Author: Christopher M. Wittich, MD, PharmD; Department of Internal Medicine and College of Medicine Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA (e-mail: Wittich.christopher@mayo.edu).

REFERENCES

1. Beckman TJ. Lessons learned from a peer review of bedside teaching. *Acad Med.* 2004;79(4):343-6.
2. Sargeant JM, Mann KV, van der Vleuten CP, Metsemakers JF. Reflection: a link between receiving and using assessment feedback. *Adv Heal Sci Educ.* 2009;14(3):399-410.
3. Irby DM. What clinical teachers in medicine need to know. *Acad Med.* 1994;69(5):333-42.
4. Skeff KM. Enhancing teaching effectiveness and vitality in the ambulatory setting. *J Gen Intern Med.* 1988;3(2 Suppl):S26-33.
5. Branch WT Jr, Paranjape A. Feedback and reflection: teaching methods for clinical settings. *Acad Med.* 2002;77(12 Pt 1):1185-8.
6. McGee SR, Irby DM. Teaching in the outpatient clinic. Practical tips. *J Gen Intern Med.* 1997;12(Suppl 2):S34-40.
7. Beckman TJ, Lee MC. Proposal for a collaborative approach to clinical teaching. *Mayo Clin Proc.* 2009;84(4):339-44.
8. Ende J. Feedback in clinical medical education. *JAMA.* 1983;250(6):777-81.
9. Wood TJ, Marks M, Jabbour M. The development of a participant questionnaire to assess continuing medical education presentations. *Medical Education.* 2005;39(6):568-72.
10. Copeland HL, Longworth DL, Hewson MG, Stoller JK. Successful lecturing: a prospective study to validate attributes of the effective medical lecture. *J Gen Intern Med.* 2000;15(6):366-71.
11. Collins J, Mullan B, Holbert J. Evaluation of speakers at a national radiology continuing medical education course. *Med Educ Online.* 2002;7(17):1-8.
12. Litin SC, Ende J. The lecture: tips to make your next presentation go better than your last. In: Sa S, ed. *Methods of teaching medicine.* Philadelphia: ACP Press; 2010:49-60.
13. Brown G, Manogue M. AMEE medical education guide No. 22: refreshing lecturing: a guide for lecturers. *Med Teach.* 2001;23(3):231-44.
14. Newman LR, Lown BA, Jones RN, Johansson A, Schwartzstein RM. Developing a peer assessment of lecturing instrument: lessons learned. *Acad Med.* 2009;84(8):1104-10.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159-74.
16. DeVellis RF. *Scale Development: Theory and Applications.* Newbury Park, CA: Sage; 1991.
17. Beckman TJ. Determining the validity and reliability of clinical assessment scores. In: Henderson M, ed. *A Textbook for Internal Medicine Education Programs.* Washington, D.C.: Association of Program Directors in Internal Medicine and Association of Specialty Professors; 2007:139-46.
18. Miles MB, Huberman AM. *Qualitative data analysis.* Thousand Oaks: Sage Publications; 1994:58.

19. **Nation JR, LeUnes AD, Gray M.** Student evaluations of teachers of psychology as a function of academic rank. *Teach Psychol.* 1976;3(4).
20. **Aleamoni LM, Yimer M.** An investigation of the relationship between colleague rating, student rating, research productivity, and academic rank in rating instructional effectiveness. *J Educ Psychol.* 1973;64(3).
21. **Linsky AS, Straus MA.** Student evaluations, research productivity, and eminence of college faculty. *J High Educ.* 1975;46(1).
22. **Levinson W, Rubenstein A.** Integrating clinician-educators into Academic Medical Centers: challenges and potential solutions. *Acad Med.* 2000;75(9):906–12.
23. **Beckman TJ, Mandrekar JN.** The interpersonal, cognitive and efficiency domains of clinical teaching: construct validity of a multi-dimensional scale. *Med Educ.* 2005;39(12):1221–9.
24. **Donnelly MB, Woolliscroft JO.** Evaluation of clinical instructors by third-year medical students. *Acad Med.* 1989;64(3):159–64.
25. **Cohen R, MacRae H, Jamieson C.** Teaching effectiveness of surgeons. *Am J Surg.* 1996;171(6):612–4.
26. **Ratanawongsa N, Thomas PA, Marinopoulos SS, et al.** The reported validity and reliability of methods for evaluating continuing medical education: a systematic review. *Acad Med.* 2008;83(3):274–83.
27. **Tian J, Atkinson NL, Portnoy B, Gold RS.** A systematic review of evaluation in formal continuing medical education. *J Contin Educ Health Prof.* 2007;27(1):16–27.
28. **Cook DA, Beckman TJ.** Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med.* 2006;119(2):166. e7–16.
29. **Messick S.** Validity. In: **Linn RL, ed.** *Educational Measurement.* Phoenix, AZ: Onyx Press; 1993.
30. **Messick S.** Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol.* 1995;50(9):741–9.
31. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U. S.). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association; 1999.
32. **Downing SM.** Validity: On meaningful interpretation of assessment data. *Med Educ.* 2003;37(9):830–7.
33. **Beckman TJ, Cook DA, Mandrekar JN.** What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med.* 2005;20(12):1159–64.
34. **Asch DA, Jedrzejewski MK, Christakis NA.** Response rates to mail surveys published in medical journals. *J Clin Epidemiol.* 1997;50(10):1129–36.
35. **Kellerman SE, Herold J.** Physician response to surveys. A review of the literature. *Am J Prev Med.* 2001;20(1):61–7.