Developing a Peer Assessment of Lecturing Instrument: Lessons Learned

Lori R. Newman, MEd, Beth A. Lown, MD, Richard N. Jones, ScD, Anna Johansson, PhD, and Richard M. Schwartzstein, MD

Abstract

Peer assessment of teaching can improve the quality of instruction and contribute to summative evaluation of teaching effectiveness integral to high-stakes decision making. There is, however, a paucity of validated, criterion-based peer assessment instruments. The authors describe development and pilot testing of one such instrument and share lessons learned. The report provides a description of how a task force of the Shapiro Institute for Education and Research at Harvard Medical School and Beth Israel Deaconess Medical Center used the Delphi method to engage academic faculty leaders to develop a new instrument for peer assessment of medical lecturing. The authors describe how they used consensus building to determine the criteria, scoring rubric, and behavioral anchors for the rating scale. To pilot test the instrument, participants assessed a series of medical school lectures. Statistical analysis revealed high internal consistency of the instrument's scores (alpha = 0.87, 95% bootstrap confidence interval [BCI] = 0.80 to 0.91), yet low interrater agreement across all criteria and the global measure (intraclass correlation coefficient = 0.27, 95% BCI = −0.08 to 0.44).

The authors describe the importance of faculty involvement in determining a cohesive set of criteria to assess lectures. They discuss how providing evidence that a peer assessment instrument is credible and reliable increases the faculty's trust in feedback. The authors point to the need for proper peer rater training to obtain high interrater agreement measures, and posit that once such measures are obtained, reliable and accurate peer assessment of teaching could be used to inform the academic promotion process.

Acad Med. 2009; 84:1104–1110.

Traditionally, clinician–educators' teaching has been assessed by students.[1,2] There is, however, growing agreement among medical school administrators and educational researchers that effective assessment of teaching must include evidence from multiple sources.[3–6] Peer review of teaching, combined with student evaluation, can provide essential data to evaluate and improve medical school and clinical teaching.[7] Peer review engages faculty in a discussion about their teaching skills, provides formative assessment of specific instructional techniques, and may be included as a component of summative assessment for academic promotional purposes. Effective peer assessment of teaching should be criterion-based, emphasize teaching excellence, and use instruments that produce highly reliable measures.[1,8–10]

Please see the end of this article for information about the authors.

Correspondence should be addressed to Ms. Newman, Shapiro Institute for Education and Research at Harvard Medical School and Beth Israel Deaconess Medical Center, 330 Brookline Avenue, E/ES-204, Boston, MA 02215; telephone: (617) 667-4742; fax: (617) 667-9122; e-mail: (lnewman@bidmc.harvard.edu).

Program Description

Background and goals

In 2007, the Shapiro Institute for Education and Research at Harvard Medical School (HMS) and Beth Israel Deaconess Medical Center (BIDMC) initiated a program of peer assessment of faculty teaching. The goals of the program are to provide faculty with feedback on their teaching abilities and deficiencies and to inform them of resources available to enhance their teaching performance. At the program's inception, a Shapiro Institute task force (made up of the authors and two members of the institute staff) began developing instruments for the assessment program. The goals of the task force were to design instruments based on validated criteria of effective clinical instruction and to train peer observers to reliably assess teaching performance. The resultant measurements would serve as credible and trusted bases for the formative assessment of faculty's teaching abilities, thereby promoting teaching excellence. In addition, reliable data collected from peer assessments could be used as part of a multisource summative evaluation process to inform clinician–educator promotions.

The task force began its work by developing a peer assessment instrument on medical lecturing. The lecture remains the most commonly used instructional method in the first two years of medical education[11,12] and, thereby, offers fertile ground to assess faculty. Done well, a lecture can be an efficient, effective, and dynamic method to introduce new topics or concepts, organize complex ideas, promote critical thinking skills, and generate enthusiasm for a subject.[13,14] Peer review of lectures, supplemented by student ratings, provides faculty lecturers with a comprehensive appraisal of their teaching skills in this context. Peers are able to judge the appropriateness of the content delivered, the lecturer's expertise, and the quality of studies presented during the lecture.[2,15,16]

We conducted an extensive review of the literature and were unable to identify a validated peer rating instrument to assess the quality of medical school faculty lecturing. We therefore undertook development of our own assessment instrument. We describe (1) our use of the Delphi method to create an instrument for peer assessment of medical lecturing, (2) an analysis of the reliability of the ratings obtained from

pilot testing the instrument, (3) lessons learned in developing the instrument, and (4) the next steps we are taking to improve the interrater agreement among faculty before the instrument is widely implemented.

Participants

In 2007, after receiving institutional review board approval from the BIDMC Committee on Clinical Investigation, the task force invited all members of the BIDMC's Resource Faculty in Medical Education to participate in a study to develop an instrument for peer assessment of lecturing and to measure the reliability of the scores obtained from the instrument. The Resource Faculty consists of HMS physician faculty members, representing all major clinical departments at BIDMC, who have a strong commitment to medical education and experience teaching in a variety of medical school and hospital settings. Resource Faculty members are recognized educational leaders selected by their department chairs to lead professional development activities, faculty and program evaluation, and curriculum development.[17] Resource Faculty members represent the departments of anesthesia, dermatology, emergency medicine, medicine, neonatology, neurology, obstetrics–gynecology, orthopedic surgery, psychiatry, radiology, radiation oncology, and surgery. The majority are graduates of the BIDMC's Rabkin Fellowship in Medical Education[18] or scholars of the Harvard Macy Institute's Educators in the Health Professions program. All are members of the Academy at Harvard Medical School and have participated in intensive faculty development teaching activities. A total of 14 Resource Faculty participated.

Instrument Development

Criteria identification

We used the modified Delphi method[19] to develop our instrument for peer assessment of lecturing. The Delphi method is shown to be an effective consensus building process to use when published information is inadequate or nonexistent.[20] The modified Delphi method is an iterative process designed to establish expert consensus on specific questions or criteria by systematic collection of informed judgments from professionals in the field. Using this method, a researcher first surveys a panel of experts individually about a particular issue or set of criteria. After analyzing and compiling their responses, the researcher resurveys the experts, asking each to indicate agreement or disagreement with the items. Repeated rounds of surveys are carried out until full consensus is reached. For development of the peer assessment of lecturing instrument, the Resource Faculty members served as the expert panelists. We chose to involve the Resource Faculty because of their educational expertise, diverse clinical backgrounds, and experience teaching in a variety of instructional settings. Furthermore, we felt the Resource Faculty would have a strong interest and commitment to the development of this instrument, as their education leadership role involves the peer assessment of teaching.

In preparation for the first survey round, we generated an initial list of effective lecturing behaviors, skills, and characteristics. To compile the list, we spoke with faculty members with extensive expertise in lecturing and reviewed the medical literature for observable, effective lecturing behaviors[11–14,21–25] (Figure 1, Delphi Round 1). We constructed and distributed a listing of 19 possible criteria to the panelists and asked them to rate the importance of including each item in an instrument to assess medical lecturing. We based the ratings on a four-point scale: 1 = very important; 2 = important; 3 = not important; 4 = eliminate. We also asked panelists to suggest different wording, note redundancies, or propose additional items for the instrument. All 14 Resource Faculty experts responded to the first Delphi survey round.

We used measures of central tendency and dispersion to analyze the data collected from the first survey round. Calculating these measures allowed us to determine the level of group consensus for inclusion or exclusion of each criterion. The mean value of 2.5 (the midpoint of our four-point scale) was chosen as the numerical indicator of group consensus. Those criteria with mean values less than 2.5 were included. Standard deviation (SD) was used to measure the dispersion of responses for each criterion and provide further evidence of group consensus. The smaller the SDs, the greater the consensus. Those criteria with an SD of less than 1 were included. Seventeen of the 19 criteria had means between 1.0 and 2.2 and SDs between 0.00 and 0.96. Two of the criteria had means of 2.6 and 2.9, with SDs of 1.1 and 1.2, and were eliminated.

In addition, we edited the criteria according to the panelists' suggestions for rewording. Five items were reworded to describe explicit, observable behaviors. For example, the original criterion "Captures and keeps the audience's attention," became "Captures attention by explaining or demonstrating need, importance, or relevance of topic." Several panelists noted redundancies among six of the criteria. We therefore eliminated three of these criteria. The outcome from the first Delphi survey round resulted in a listing of 14 criteria. We summarized and distributed to the panel of experts the data from the first survey round and the resulting list of criteria, along with a written request for a second round of review (Figure 1, Delphi Round 2).

Twelve experts responded to the second Delphi survey round. Thirteen of the 14 criteria had mean ratings between 1 and 1.3 and SDs between 0.0 and 0.6. One criterion had a mean of 2.5 and an SD of 1.2 and was eliminated from the listing. We again edited and reworded the criteria according to the panelists' suggestions. Most suggestions were recommendations to shorten the criterion's word length, and to add specific behavioral descriptors or anchors to the assessment instrument. The panelists noted redundancy of two criteria, and we therefore eliminated one of these.

We e-mailed a final revised listing of the 12 criteria to the expert panelist for the third Delphi survey round. All 14 experts reached full consensus on this final listing of criteria (Figure 1, Delphi Round 3). Using this listing of 12 criteria of effective lecturing, we constructed our initial peer assessment instrument. We used a three-point scale to rate each criterion: 1 = excellent demonstration, 2 = adequate demonstration, and 3 = does not demonstrate. We also added an option to indicate *unable to assess*, along with a global rating of the lecture.

To differentiate the three levels of lecturer performance, we included behavioral descriptors of each criterion,

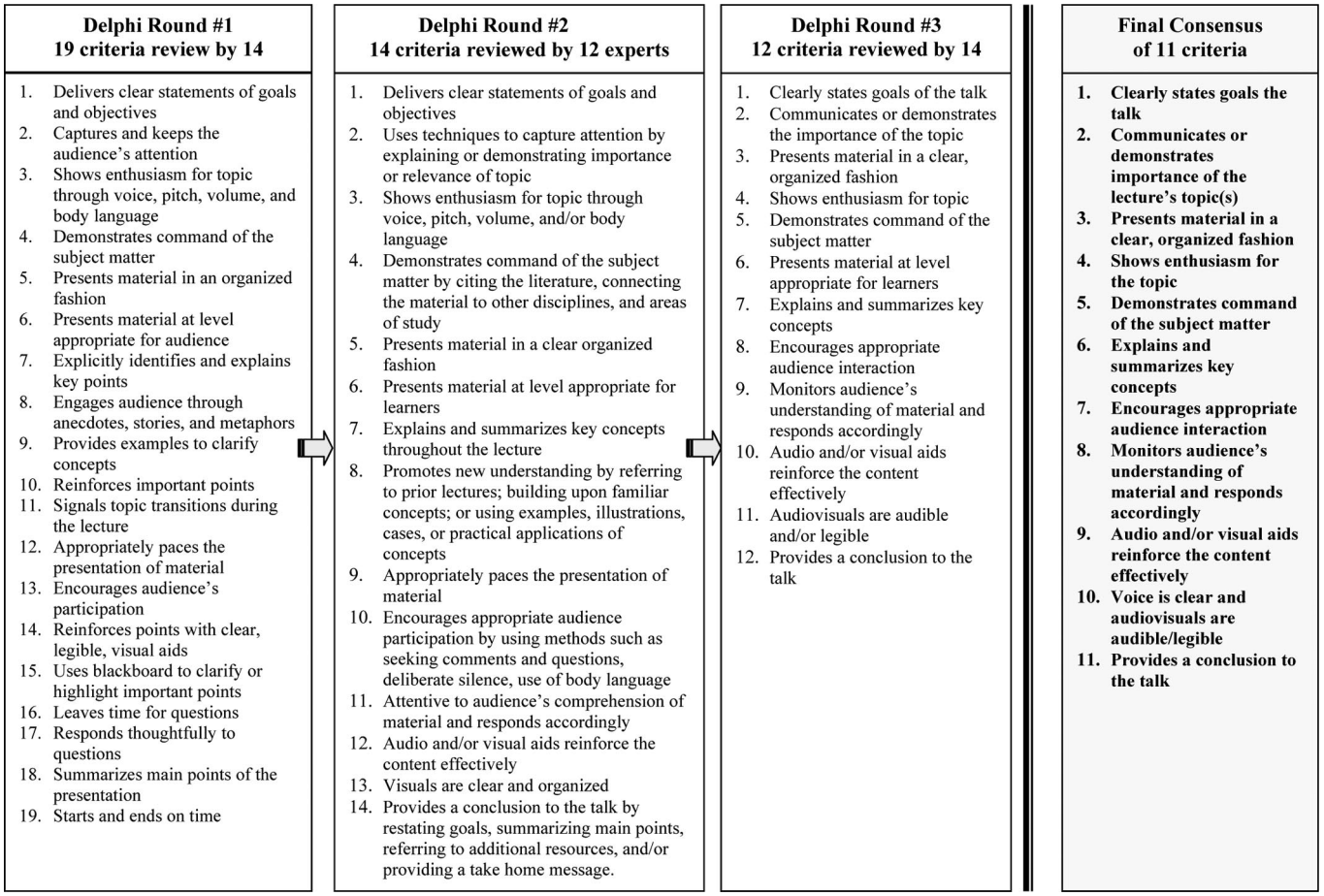

Figure 1 Stages of the modified Delphi process used to determine 11 criteria to include in an instrument for peer assessment of medical lecturing developed at the Beth Israel Deaconess Medical Center, Boston, Massachusetts, 2007.

culled from the literature.[6,26] The behavioral descriptors were placed under the column heading for rating level 1, *excellent demonstration of performance.* For rating level 2, *adequate demonstration of performance,* we used qualifying terms such as "limited in scope." For rating level 3, *does not demonstrate,* we used terms such as "does not present."

We presented the rating scale and criteria to the faculty as a group, who recommended eliminating one additional criterion, "Presents material at level appropriate for learners." The group felt that, to assess this criterion, a peer observer would need to know the learners' opinions regarding the appropriateness of the presentation level. This resulted in identification of 11 criteria of effective lecturing.

Rating scale development

We invited the same Resource Faculty members who participated in the Delphi rounds to consider and review the rating scale and behavioral anchors of the peer assessment instrument to finalize it for pilot testing of interrater reliability. These faculty members met for two, 2-hour sessions to discuss peer observation techniques, consider the behavioral descriptors for each criterion, comment on the sufficiency of the three rating levels, and provide feedback on the overall format of the instrument. To gain experience using the instrument, we asked the group to watch, score, and discuss videotaped lectures filmed during an HMS human physiology course. We showed 10-minute segments from the beginning, middle, and end of each lecture and asked the faculty to rate the elements observed. After rating the lecture segments, the faculty shared their scores and discussed behaviors they saw that persuaded them to choose a particular level of performance. Several faculty made suggestions for minor rewording of the behavioral descriptors.

During the second rating scale development session, the faculty noted that the three-point rating scale was limiting, as they tended to rate most criteria at the second performance level (adequate demonstration). The group suggested changing the instrument to a five-point scale (1 = excellent demonstration, 2 = very good, 3 = adequate, 4 = poor, and 5 = does not demonstrate criteria) and maintaining descriptive benchmarks for the excellent, adequate, and poor performance rating levels. At a follow-up meeting with the faculty, we distributed the finalized peer assessment of the lecturing instrument consisting of 11 criteria rated on a five-point scale. The group unanimously agreed on this final version (Appendix 1).

Pilot testing reliability of the instrument's measures

We subsequently pilot tested the instrument to measure internal consistency and interrater agreement. We instructed each participant to rate the entirety of four, 1-hour HMS videotaped lectures (not viewed previously)

according to the criteria, and to provide a global rating assessment of the quality of each lecture. Because of faculty time constraints, the number of observers varied in the assessment of the four lectures. We collected a total of 31 peer assessment rating forms for the lectures (the four lectures had 12, 9, 5, and 5 reviewers, respectively).

We analyzed the pilot data to measure reliability of the scores obtained from the instrument. Cronbach alpha was used to assess internal consistency reliability of the ratings.[27] The coefficient alpha was high (a = .87, 95% bootstrap confidence interval [BCI] = 0.80–0.91), indicating that the items on the instrument measure a cohesive set of concepts of lecture effectiveness. Bootstrap resampling approaches were used to obtain interval estimates. Missing data were handled with multiple imputation.[28]

There was some variability in the internal consistency across each of the four lectures (0.92, 0.77, 0.93, 0.87). All but one were close to a minimal threshold of 0.90 for making decisions about individuals, and well above the threshold for making decisions about groups (0.80).[29]

Interrater agreement was assessed by forming all possible pairs of raters who observed the same lecture. The reliability of a randomly selected reviewer's scores was computed using intraclass correlation coefficient (ICC). The measure of ICC for the 31 raters' scores across all criteria and the global measure was fair (0.27, 95% BCI = −0.08 to 0.44). However, there was variability of ICC measures for the individual criteria. For criterion 11 (ICC = 0.69), the magnitude of association across pairs of raters can be described as substantial. For criteria 3 through 7 and 9, the magnitude of association can be described as moderate to fair. The reviewers reached only slight agreement on criteria 1, 2, 8, and 10, and on the global rating of the lectures.[30] Table 1 presents a comparison by criteria of the interrater agreement (as measured by ICC) for all four lectures. The table is arranged in descending order of agreement.

Lessons Learned About Instrument Development and Peer Assessment of Lecturing

Peer review of teaching is a valuable process that engages faculty in discussing and improving the skills of teaching, provides formative assessment to enhance clinician–educator performance, and may be used as part of a multisource, summative assessment to inform high-stakes decisions making, such as academic promotion. Providing feedback to faculty members clarifies good performance, facilitates self-reflection of teaching practice, encourages discussion about effective instruction, and closes the gap between current performance levels and desired goals.[31] Peer assessment of teaching, therefore, can build a community of educators while fostering continuous quality improvement. This report describes how we approached our goal of developing valid instruments for peer assessment to evaluate reliably teaching performance. We feel it is important to share what we have learned during this process with members of the educational community who may seek to implement peer assessment of teaching for formative and summative evaluation.

Lesson 1: Consensus building fosters instrument coherence and self-reflection

The time and effort the Resource Faculty dedicated to the development of the assessment instrument was vital to establishing cohesive measures of effective lecturing. The effort expended likely contributed to the high measurement of internal consistency when we tested the reliability of the instrument. One lesson learned, and noted in the literature, is that collaboration of faculty in the development of an assessment instrument can create a shared definition of good performance.[32] Resource Faculty also noted that the work of establishing the criteria of effective lecturing stimulated self-reflection and consideration of how well they met these standards when giving their own lectures.

Lesson 2: Faculty members must trust the validity and reliability of the evaluation process

For peer assessment to be used as evidence of effective teaching, the process requires a high degree of objectivity to produce credible, reliable, and defensible evaluations.[9] Faculty undergoing peer review need to trust that the ratings are not idiosyncratic scores of their performance. We therefore felt it was critical to test the reliability of our assessment instrument through measuring interrater agreement,[33] as faculty would be more likely to trust the feedback. The instrument itself could then be used as instructional material in faculty development. Conversely, low interrater agreement of the instrument's scores would be a significant threat to its usefulness in a comprehensive assessment program or inclusion in high-stakes decision making.

There was considerable variability in our instrument's interrater agreement measures. There are several possible explanations for this variability. The most significant factor is that we did not provide proper rater training. In our two, 2-hour faculty development sessions, the Resource Faculty discussed peer observation techniques, offered comments on the instrument, and practiced using the assessment tool. However, these were not formal training sessions (see Lesson 3). A second factor contributing to the low interrater agreement measure may be that the faculty raters used predetermined, internal standards in judging the quality of a lecturer's performance. Braskamp and Ory[34] note that, at times, raters compare a person's performance or contribution against those of others, or against some a priori standard derived from previous experience. In our study, the faculty might have approached the peer observation event with an internal bias about how the lecture should be presented. This may have been the case, in particular, if the topic was of interest to the faculty or within the faculty's own discipline. Therefore, the faculty's idiosyncratic perceptions may have superseded more objective appraisal of the lecturing performance. In Lesson 3, we explore how best to address this phenomenon.

Lesson 3: Peer rater training is essential for high-stakes evaluation

Careful attention to rater training has been singled out as the most effective strategy for increasing accuracy and consistency of performance assessment ratings.[32] During training, raters learn to avoid common rater errors (such as halo, leniency, and central tendency) and discuss behaviors indicative of each performance dimension until individual perceptions are brought into closer

Table 1

Comparison of Interrater Agreement Among 31 Beth Israel Deaconess Medical Center Faculty in Rating the Criteria of Lecture Effectiveness of Four Harvard Medical School Lectures, Boston, Massachusetts, 2007

Criterion*	Intraclass correlation coefficient
11 (Provides a conclusion)	0.69
3 (Presents material in a clear, organized fashion)	0.60
4 (Shows enthusiasm for the topic)	0.56
6 (Explains and summarizes key concepts)	0.46
5 (Demonstrates command of the subject matter)	0.45
9 (Audio and/or visual aids reinforce the content effectively)	0.38
7 (Encourages appropriate audience interaction)	0.22
8 (Monitors audience's understanding and responds accordingly)	0.20
2 (Communicates or demonstrates importance of lecture topic)	0.14
1 (Clearly states goals of the talk)	0.07
10 (Voice is clear and audiovisuals are audible/legible)	−0.06
Global rating (Overall, how would you rate this lecture?)	0.19

* Criteria are listed in descending order of agreement.

congruence with those held by the group.[35] To increase proficiency at discriminating between performance dimensions, raters view and discuss samples of each performance level included on the rating scale. Most important, raters practice scoring performances and receive feedback from a training facilitator on the accuracy of their scores.

The success of rater training programs requires that participants commit to the time and effort necessary to internalize the standards of the system and become consistent in their use of the ratings. The need for a high level of commitment among all faculty participants can make training a large group of peer raters problematic. One solution might be to establish a small cadre of faculty who undergo intensive rater training together. Reliable appraisal data obtained from this cadre of peer raters could then be used in summative, high-stakes assessment of lecturing effectiveness.

Before implementing the Instrument for Peer Assessment of Medical Lecturing on a hospital- and medical-school-wide scale, we must address the process issues described in Lessons Learned. Our first step is to increase accuracy of the ratings and achieve acceptable interrater agreement of the instrument's scores. To address this issue, we have initiated a rater training program at BIDMC.

Conclusions

Our report demonstrates that the modified Delphi method can be used to determine a cohesive set of agreed-on criteria for an instrument to be used in the peer assessment of lecturing. Furthermore, through group consensus building, faculty can successfully establish an appropriate scoring rubric and identify behavioral descriptors for the instrument. The instrument, presented in full for reader use and future research (Appendix 1), can be used in its current state to provide formative assessment and instruction on lecturing performance. Because of the variability of interrater agreement regarding the instrument's scores, further study is needed, along with appropriate rater training, for this tool to be used in summative, high-stakes assessment of lecturing effectiveness.

Ms. Newman is acting director, Faculty Programs in Medical Education, and codirector, Rabkin Fellowship in Medical Education, Shapiro Institute for Education and Research, Harvard Medical School and Beth Israel Deaconess Medical Center; and associate in medicine, Harvard Medical School, Boston, Massachusetts.

Dr. Lown is director of faculty development, Department of Medicine, Mount Auburn Hospital; codirector, Rabkin Fellowship in Medical Education, Shapiro Institute for Education and Research, Harvard Medical School and Beth Israel Deaconess Medical Center, The Mount Auburn Fellowship in Medical Education, and The Harvard Medical School Academy Fellowship in Medical Education; and assistant professor of medicine, Harvard Medical School, Boston, Massachusetts.

Dr. Jones is associate director, Social and Health Policy Research, Institute for Aging Research, Hebrew SeniorLife, Harvard Medical School; and assistant professor of medicine, Harvard Medical School, Boston, Massachusetts.

Dr. Johansson is assistant director, Office of Educational Research, Shapiro Institute for Education and Research at Harvard Medical School and Beth Israel Deaconess Medical Center; and instructor in medicine, Harvard Medical School, Boston, Massachusetts.

Dr. Schwartzstein is vice president for education, Beth Israel Deaconess Medical Center; faculty associate dean for medical education, Harvard Medical School; executive director, Shapiro Institute for Education and Research at Harvard Medical School and Beth Israel Deaconess Medical Center; associate chief, Division of Pulmonary and Critical Care Medicine, Beth Israel Deaconess Medical Center; and professor of medicine, Harvard Medical School, Boston, Massachusetts.

Acknowledgments

The authors gratefully acknowledge Dr. Charles J. Hatem for his vital contributions to the study and to faculty development. The authors also wish to thank the faculty who participated in the study and development of the assessment instrument.

Dr. Jones is supported in part by NIH Grant AG008812, "Biostatistics and Evaluation Core, Harvard Older Americans Independence Center."

References

1 Beckman TJ, Lee MC, Rohren CH, Pankratz VS. Evaluating an instrument for the peer review of inpatient teaching. Med Teach. 2003;25:131–135.
2 Leamon MH, Servis MH, Canning RD, Searles RC. A comparison of student evaluations and faculty peer evaluations of faculty lecturers. Acad Med. 1999;74(suppl): 22–24.
3 Simpson D, Fincher RM, Hafler JP, et al. Advancing educators and education by defining the components and evidence associated with educational scholarship. Med Educ. 2007;41:1002–1009.
4 Wilkerson L, Irby DM. Strategies for improving teaching practices: A comprehensive approach to faculty development. Acad Med. 1998;73:387–396.
5 Arreloa RA. Developing a Comprehensive Faculty Evaluation System. Bolton, Mass: Anker Publishing Company; 1995.
6 Centra JA. Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness. San Francisco, Calif: Jossey Bass; 1993.
7 Irby DM. Peer review of teaching in medicine. J Med Educ. 1983;58:457–461.
8 Van Note Chism N. Peer Review of Teaching: A Sourcebook. Bolton, Mass: Anker Publishing; 1999.
9 Kohut GF, Burnap C, Yon MG. Peer observation of teaching: Perceptions of the observer and the observed. Coll Teach. 2007; 55:19–25.
10 DeZure D. Evaluating teaching through peer classroom observation. In: Seldin P. Changing Practices in Evaluating Teaching:

A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions. Bolton, Mass: Anker Publishing; 1999:70–96.
11 Gelula MH. Effective lecture presentation skills. Surg Neurol. 1997;47:201–204.
12 Laidlaw JM. Twelve tips for lecturers. Med Teach. 1988;10:13–17.
13 Steinert Y, Snell LS. Interactive lecturing: Strategies for increasing participation in large group presentation. Med Teach. 1999;21:37–42.
14 Cantillon P. ABC of learning and teaching in medicine: Teaching large groups. BMJ. 2003; 326:437–440.
15 Irby DM, DeMers J, Scher M, Matthew D. A model for the improvement of medical faculty lecturing. J Med Educ. 1976;51:403–409.
16 Nelson MS. Peer evaluation of teaching: An approach whose time has come. Acad Med. 1998;73:4–5.
17 Schwartzstein R, Huang G, Coughlin C. Development and implementation of a comprehensive strategic plan for medical education at an academic medical center. Acad Med. 2008;83:550–559.
18 Hatem CJ, Lown BA, Newman LR. The academic health center coming of age: Helping faculty become better teachers and agents of educational change. Acad Med. 2006;81:941–944.
19 Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs. 2000;32:1008–1015.
20 Jones J, Hunter D. Qualitative research: Consensus methods for medical and health services research. BMJ. 1995;311:376–380.
21 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: 366–371.
22 Nierenberg DW. The challenge of "teaching" large groups of learners: Strategies to increase active participation and learning. Int J Psychiatry Med. 1998;28:115–122.
23 Casteel CP, Mortillaro NA, Taylor AE. Teaching effectiveness analysis plan applied to lectures in medical physiology. Am J Physiol. 1989;256(suppl):3–8.
24 Brown G, Manogue M. AMEE Medical Education Guide No. 22: Refreshing lecturing: A guide for lecturers. Med Teach. 2001;23:231–244.
25 Sullivan RL, McIntosh N. Delivering Effective Lectures. Available at: (http://www.reproline.jhu.edu/english/6read/6training/lecture/delivering_lecture.htm). Accessed April 3, 2009.
26 Seldin P. Changing Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions. Bolton, Mass: Anker Publishing; 1999.
27 Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297–334.
28 Royston P. Multiple imputation of missing values: Update. Stata J. 2005;5:118–201.
29 Nunnally JC, Bernstein IH. Psychometric Theory. 3rd ed. New York, NY: McGraw-Hill College Division; 1994.
30 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–174.
31 Nicole DJ, Macfarlane-Dick D. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Stud High Educ. 2006;31:199–218.
32 Williams RG, Klamen DA, McGaghie WC. Cognitive, scholar and environmental sources of bias in clinical performance ratings. Teach Learn Med. 2003;15:270–292.
33 Shea JA, Fortna GS. Psychometric methods. In: Norman G, van der Vleuten CPM, Newble DI, eds. International Handbook of Research in Medical Education: Part One. Dordrecht, the Netherlands: Kluwer Academic Publishers; 2002.
34 Braskamp LA, Ory JC. Assessing Faculty Work: Enhancing Individual and Instructional Performance. San Francisco, Calif: Jossey-Bass; 1994.
35 Govaerts MJB, van der Vleuten CPM, Schuwirth LWT, Muijtjens AMM. Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. Adv Health Sci Educ. 2007;12:239–260.

Appendix 1

Instrument for Peer Assessment of Medical Lecturing, Beth Israel Deaconess Medical Center, Boston, Massachusetts, 2007

Criteria for Effective Lecturing		Excellent Demonstration of Criteria 1	Very Good Demonstration of Criteria 2	Adequate Demonstration of Criteria 3	Poor Demonstration of Criteria 4	Does not Demonstrate Criteria 5	Rating or Unable to Assess (U/A)	Comments
1	Clearly states goals of the talk	During introduction, communicates purpose of the presentation. For example, may provide an overview of content, state expected learning outcomes, pose rhetorical/challenging questions to be answered.		States the goals, but description is limited in scope (e.g., *only* states topics to be covered or provides the format of talk).		Does not provide overview or communicate the goals of talk.		
2	Communicates or demonstrates importance of the lecture's topic(s)	Clearly explains the topics' and subtopics' relevance, context, applicability, and/or the significance to the audience (e.g., presents compelling information, case, or data; uses a "hook").		States the importance of topic, but provides limited description of why learners need to know the material.		Does not communicate or describe why the topic is of importance.		
3	Presents material in a clear, organized fashion	Uses an explicit, organized framework so that the presentation flows logically (e.g., articulates a structure and sequence to the talk, frames subtopics, links concepts).		Presentation has some organization, but limited in structure, linkage, and/or sequence.		Does not present material in a clear, organized fashion.		
4	Shows enthusiasm for topic	Demonstrates keen enthusiasm for topic through voice, eye contact, energy, movement and/or body language (e.g., varies pitch, inflection, tempo, and volume; gestures to emphasize importance).		Shows some enthusiasm for topic, but limited in display.		Does not show enthusiasm for the topic.		
5	Demonstrates command of the subject matter	Demonstrates strong understanding of subject matter (e.g., cites the literature, refers to overarching subject area, draws upon personal experiences, speaks to advances or current controversies in the field, provides informative answers).		Demonstrates some command of subject, but breadth of understanding is limited (e.g., unable to elaborate with greater detail or information).		Does not demonstrate a command of subject matter.		
6	Explains and summarizes key concepts	Defines new terms/principles, synthesizes information (e.g. identifies important points; uses examples, analogies, metaphors; thinks out loud).		Explains some key concepts, or provides vague explanations.		Does not explain or summarize key concepts.		
7	Encourages appropriate audience interaction	Stimulates active participation (e.g., makes eye contact, solicits comments and questions, polls the audience, uses deliberate silence, poses open-ended questions, invites learners to interact with each other; manages flow of discussion).		Encourages some interaction or uses less effective strategies (e.g., asks close-ended questions, offers little wait time, often turns back to audience, and reads from slides).		Does not engage or encourage interaction (e.g., reads all slides without looking at audience; defers questions, yet does not answer them).		
8	Monitors audience's understanding of material and responds accordingly	At appropriate intervals assesses and responds to audience's understanding of material (e.g., asks probing questions or polls audience; asks if material is clear, then tailors response by rephrasing or providing alternative examples; adjusts the pace of lecture to accommodate learners).		Pays some attention to the audience's understanding of topic, but tailoring of response is limited.		Does not pay attention to the audience's understanding of material, or checks in but doesn't respond accordingly.		
9	Audio and/or visual aids reinforce the content effectively	Appropriately chooses and designs instructional material to emphasize key points, demonstrate relevance of material, or stimulate thought.		Some of the audio and/or visual aids reinforce content, or material is less than effective (e.g., slides are cluttered).		Audio and/or visual aids do not reinforce content.		
10	Voice is clear and audiovisuals are audible/ legible	Sensitive to the setting and tailors audio and visual aids so all can see and hear (e.g., checks if audience can hear/see material; talks to audience not to blackboard, laptop, or screen; visual material is well organized, text is legible, and graphics are clear).		At times voice is unclear or audiovisuals are inaudible/illegible.		Voice is unclear and audiovisuals are inaudible/ illegible.		
11	Provides a conclusion to the talk	Concludes presentation by synthesizing information, summarizing main points, and inviting/responding to questions (e.g., repeats or rephrases questions as needed). Open to hearing learners' perspectives/opinions.		Provides synthesis and/or summary of talk, but limited in scope. Invites few questions and/or provides limited or ambiguous responses.		Does not synthesize or summarize information.		

Overall, how would you rate this lecture (*please circle*):

___1__ Excellent ___2__ Very Good ___3__Good ___4__Fair ___5__Poor