

Evaluation of Student Learning: A Continuum from Classroom to Clerkship: A Webcast Audioseminar Series for Spring 2004

Carol Whitfield, Ph.D., Phyllis Blumberg, Ph.D., Byron Crawford, M.D., Debra DaRosa, Ph.D., Rebecca Henry, Ph.D., Brian Mavis, Ph.D. and Sebastian Uijtdehaage, Ph.D.

Pennsylvania State University College of Medicine
500 University Drive
Hershey, Pennsylvania 17033-0850 U.S.A.

Phone: (+)1-717-531-8570

Fax: (+)1-717-531-4786

Email: cwhitfield@psu.edu

ABSTRACT

In the spring of 2004, IAMSE sponsored a webcast audioseminar series titled "Evaluation of Student Learning: A Continuum from Classroom to Clerkship". Six nationally recognized experts in evaluation of student learning presented seminars that described various ways to develop and use evaluation methods in settings generally found across the medical curriculum. Our audience included members of institutional faculty development programs and individual faculty members from many countries across the world. Our webcast series allowed registrants to listen to the presentation in real time while viewing the presenter's slides on their computer web browser. The presentations were interactive, allowing the audience to ask questions or provide information from their own experiences. Audio recordings of the seminars, accompanied by the slides were archived on the International Association of Medical Science Educators (IAMSE) website, and are available to registrants who want to review the seminars. Evaluation of student learning proved to be a very popular topic, and the audience numbered well over 100 for each of the six seminars. We urge educators to carefully read the following philosophical and practical approaches to evaluation of student learning. Use these white papers to convince colleagues, Chairs and Deans that there must be a solid evaluation plan for their institution. It is important for educators to measure the return values on education and make them a part of annual reports. Each seminar speaker provided a summary of content and major points of discussion following their presentation. These summaries are reproduced below.

Fundamentals of Evaluation in Medical Education

Brian Mavis, Ph.D.

Associate Professor

Office of Medical Education Research & Development
Michigan State University College of Human Medicine
April 6, 2004

Feedback is a key feature of any system that promotes learning. This is true whether we are talking about an individual student's efforts to learn new knowledge or skills, or an organization's efforts to improve its process or product. It is in this context that evaluation was discussed as it applies to medical education. Fundamentally, evaluation is the systematic collection of information for decision-making. It is a key component of a process of action, reflection and planning. Evaluation questions and strategies can range from a focus on learner's experiences and abilities to larger organizational concerns characterized by questions about the curriculum, students, faculty, institutional

processes or organizational mission. Regardless of the focus of a specific evaluation effort, the purpose of an evaluation is quality improvement.

The first part of the presentation focused on student assessment and its relationship to determining competency. Learners vary in their level of competency from novice to expert; the challenge is choosing assessment strategies appropriate for the level of competency. Assessment strategies vary in the extent to which they are objective or subjective and quantitative or qualitative, thus each requires specific implementation considerations to assure reliability, validity, efficiency and acceptability. Since each assessment strategy has strengths and weakness, a system of assessment that uses multiple strategies will provide the most accurate reflection of learner competency. In basic science education, the multiple choice question (MCQ) is the most frequently used method of student assessment, most likely because of their objective quantitative format as well as their

familiarity to both learners and faculty. However, since patients don't present with five choices during a medical encounter, MCQs have their limitations too. A number of questions were provided to help educators think through decisions about which student assessment methods to choose.

The second part of the presentation focused on program evaluation. Essentially, while the process of designing a program evaluation is similar to designing a student assessment, there are differences in terms of scale as well as the types of questions that frame the data gathering. The program evaluation model by Kirkpatrick was used, indicating that evaluations can focus on participant reactions, learning, behavior change or real world impact. Again, evaluation strategies were discussed in terms of the various levels of Kirkpatrick's model, with idea that each has strengths and weaknesses and that multiple measures provide the more data for decision-making purposes. When deciding on an evaluation strategy, the question of resources, stakeholders, mission and values need to be considered. The discussion following the presentation focused on different methods of collecting information and their appropriateness to different needs or situations. In addition, there was discussion of the strategies for disseminating evaluation information to faculty, as a means of involving faculty in on-going planning and decision-making.

Evaluating Student Learning in the Didactic Setting

Byron E. Crawford, II, M.D.
Associate Professor Pathology & Laboratory Medicine
Tulane University Health Sciences Center
April 22, 2004

The seminar on "Evaluating Student Learning in the Didactic Setting" presented different methods to assess student learning both objective and subjectively. However, before assessing student learning, one must determine curricular expectations through development of specific objectives for each contact hour in a course. This also allows for appropriate exam development in which all written exam questions match or correspond to an objective. With excellent course objectives and exam development, one may use multiple objective and subjective methods to evaluate student learning; both short-term learning and long-term learning.

The objective ones include use of examinations. Internal exams can be used to assess student learning in specific topics, or blocks with comparison studies of previous successful years. Comparison of one academic year to another year that has been deemed successful may allow one to determine academic achievement in the next year by analyzing class averages and class block averages. This use of internal exams is limited because it is based on internal critique only.

National Board of Medical Examiner subject exams also allow a course director to assess learning of specific topics and blocks through the "item analysis" results. One may also compare the class with other medical schools in the United States and Canada, and one may compare one internal class with another. Class percentile ranks and comparison of expected percentile ranking with observed percentile ranking may give data supporting student learning. Use of subject exams may also evaluate long-term learning and knowledge retention of orphan topics, topics not covered extensively in a course, and topics in which there may have been specific problems.

Subjective means discussed included data obtained from peer review, student surveys both current and retrospective, faculty participating in future courses and the student effective index. These may all be used to evaluate student learning. Obtaining adequate response rates to faculty and student surveys may be challenging. Voluntary participation is the preferred method. Students and faculty should feel a professional obligation to participate in a way that may potentially improve a course and student learning. Enticements may be used and include students receiving extra points, temporary delay in receiving student grades, and for faculty, small gift certificates or small financial gifts for their time and opinions.

Assessment of student learning should occur throughout the course via well-designed internal exams, and preferably with an external end of the year exam. There are other times that a course director may need to focus in on specific topics to evaluate student learning including 1) faculty change in lectures, 2) significant content change, 3) change in teaching methodology, and 4) utilization of new faculty members in teaching of the course.

The most common methods used to evaluate student learning, according to many course directors, are internal exams looking at topic – block specific data, yearly comparisons and class means, and student perception obtained from student surveys. Additional methods of assessing student learning, including these may provide additional support that students in a class are learning the material outlined in the course objectives. It is recommended those additional methods besides internal exams and student surveys be used to evaluate student learning. An end of the course external exam is recommended to both evaluate and compare student learning with other schools.

Evaluating Student Learning in the Clinical Setting

Debra DaRosa, Ph.D.
Professor and Vice Chair of Education
Department of Surgery
Northwestern University Feinburg School of Medicine
May 5, 2004

The purpose of this session was threefold: discuss common problems with clinical performance ratings (CPR), explain

steps necessary to judiciously evaluate problem learners, and describe strategies for enhancing CPR.

The quality of performance ratings are determined by their accuracy, reproducibility, generalizability, and validity. The main sources of errors in CPR systems include the raters: - evaluating behaviors they didn't observe, or don't remember observing -not the performance rating system, and the rating form itself. Problems associated with raters vary, but sample problems involving raters are:

- evaluating behaviors they didn't observe, or don't remember observing
- not using the full scale but rather being hawks (rare) or doves (most common)
- not wanting to record negatives
- rating a learner high or low in all categories rather than discriminating among the different categories

Clinical performance rating systems need to be administered with attention to detail. The who, what, when, how, and so what questions associated with any system should be documented and implemented as such.. Examples of problems include:

- tardy forms or no forms completed
- lack of follow up when negative ratings or comments are submitted
- insufficient number of raters to truly generalize performance
- insufficient attention to due process guidelines

And lastly, examples of problems associated with the rating form include:

- too many items on the form
- no indication as to the extent of observation by the faculty member
- no global rating scale to capture "gestalt" judgment of faculty member

These lists are not exhaustive but represent many of the weaknesses in clinical performance rating systems.

Faculty should be educated on how to detect common symptoms among problem learners and how to effectively intervene. An impaired learner can have psychological, substance abuse, or physical illness problems. It is critical that faculty document noted problems and submit their written concerns to the clerkship or program director. If communicated verbally, the education administrator should document the date and time of the conversation. Preventative measures such as having a meaningful mentor/advisor system, a critical incident report system, and clearly documented expectations for the learners are helpful. The key guidelines are to:

- document changes in personality, performance, or physical appearance in a timely way
- provide clear and consistent communication, both verbal and written
- due process must be afforded
- intervene early
- protect the learner's right to confidentiality

- be aware of your institution's policies for addressing problem learners.

Education administrators can enhance their clinical performance ratings by taking several steps. These steps are nicely spelled out in a paper by Dr. Reed Williams and colleagues entitled "Cognitive, social and environment sources of bias in clinical performance ratings" published in *Teaching and Learning in Medicine*, 2003. The authors offer a list of suggestions that should be considered when aiming to hone your clinical performance evaluation system.

It is a difficult but critical responsibility to evaluate our learners in the clinical environment. There are challenges to implementing a fair and accurate performance evaluation system in the busy and complex hospital environment. But we can hone our ability to judiciously and accurately evaluate our learners with adequate attention to: 1) educating our faculty raters so to ensure adequate calibration and cooperation, 2) planning and documenting a sound performance evaluation system, and 3) having in place procedures for appropriately addressing problem learners.

Options for Evaluating Student Learning in PBL Programs

Phyllis Blumberg, Ph.D., Professor of Psychology & Director Teaching and Learning Center University of the Sciences in Philadelphia
May 20, 2004

In this session a classical, iterative version of problem-based learning (PBL) is described, in which the case discussion stimulates learning. All material is discussed twice, first without prior preparation and then after researching the questions raised in the first session (called learning issues). Next seven learning outcome categories are outlined according to Fink's (2003) taxonomy of significant learning that guide our options for evaluating student learning in PBL. These categories are: learning how to learn, motivation/interest/values/respect for others, human dimension, integration/connection, application/problem solving/critical thinking, knowledge, and skills. Specific embedded assessments that are congruent with this taxonomy of learning that can be used at each step are identified. For example, the summaries of learning issues can be evaluated for: deep-learning (learning for understanding and meaning, and many connections are formed among concepts learned) , use of evidence-based decision making to evaluate information, synthesis of knowledge, evidence of self-directed learning, information literacy skills, and written communication. Many different types of categories of outcomes can be evaluated throughout all in-class PBL activities including: professional behaviors, leadership effective team behaviors, and management of complex projects. These evaluations are based on repeated observations of in-class interactions. Faculty, peers and the students can assess themselves on these dimensions. few examples of non-embedded, authentic evaluation tools that

are consistent with the PBL process, such as the triple jump are discussed.

An evaluation framework is proposed for selecting what to evaluate and how that considers the outcome category, the rationale for selection, the specific outcome to be evaluated, how the outcome should be measured and how to collect data to measure the outcome. Finally, the framework is applied to examples of how to evaluate deep learning and information processing. Deep learning falls in the categories of learning to learn, application and problem solving. Problem solving is hard to measure directly, but evidence of deep learning is a prerequisite for problem solving. Deep learning can be evaluated from the student discussions of cases, particularly on the second go around with the material. Students collectively can create concept maps of their understanding of the case and the underlying basic science that explains the disease process. Scoring rubrics can be used to evaluate students' concept maps. Usually a group grade is given and then individual students can earn more or less than the group grade for performance that was markedly above or below the standard performance. Peer feedback is helpful in determining the individual points. Information literacy standards for higher education have been established by the Association of College and Research Libraries including: determination of information needs, acquisition of information effectively and efficiently, critical evaluation of information and its sources, incorporation of selected information into one's knowledge base and use of information legally and ethically. The process of generating, researching and reporting on learning issues allows us to evaluate students on information literacy

Association of College and Research Libraries
www.ala.org/acrl/ilintr.html Fink, LD. (2003) Creating Significant Learning Experiences. San Francisco: Jossey-Bass

Computer-Based Assessment of Medical Knowledge and Skills

Sebastian Uijtdehaage, Ph.D.
Assistant Professor of Medicine
UCLA David Geffen School of Medicine
Co-Director, Health Education Assets Library (HEAL)
June 1, 2004

For centuries, medical educators have used traditional means for assessing medical knowledge and skills: paper-and-pencil tests, microscope-based exams, and clinical skills exams with simulated patients. Some of these trusted methods, however, have serious drawbacks. For instance, in a typical microscope-based exam students are given little time to examine a specimen and are not allowed to review their answers. Not uncommonly, specimens change or become damaged during the examination process.

Recent advances in web-based and robotic technology have remedied some of the disadvantages of traditional assessment methods. These new formats of assessment,

however, are expensive and introduce a new set of challenges. For instance, security concerns are raised because students need to be tested in shifts due to limited seating capacity in computer laboratories. Also, students could conceivably use the Internet inappropriately during the exam using "instant messaging" or surfing the World Wide Web to find answers. It has been UCLA's experience, however, that reminding students of the Honor Code is sufficient to avert widespread cheating.

In this seminar, emerging trends in the field of computer-based assessment were discussed. "Virtual patients" are computer-based simulations with which students can interact to sharpen their diagnostic reasoning and procedural skills without risks to patients. Virtual patients range from relatively simple web-based applications to very complex, high-fidelity computer-driven mannequins. These simulations can be used to assess clinical skills to the extent to which they can track and document students' clinical decisions and treatment choices.

Computer adaptive testing (CAT) is being adopted increasingly in standardized testing but has not yet found widespread use in medical education. It was introduced in this presentation as a potential novel method to measure medical knowledge with great precision. Based on Item Response Theory, CAT selects a unique sequence of test items to estimate a student's proficiency. Difficulty level of the questions is based on the student's performance on previous questions. CAT, however, requires a large set of questions with established psychometric properties such as difficulty level. Therefore, this method may not be feasible for individual institutions unless medical colleges collaborate.

Fortunately, recent technological advances have facilitated collaborations among institutions. For instance, several XML metadata schemas have been developed to describe the content and characteristics of test items such as the IMS Question and Test Interoperability Specification (www.imsglobal.org/question/). As an increasing number of medical schools use electronic course management systems that are compatible with such metadata schemas, we may see more sharing, banking and re-deployment of test items in the near future.

In conclusion, computer-based testing resolves some problems associated with conventional assessment methods but at the same time introduces new challenges. Because computer-based assessment opens new ways to improve the validity and reliability of testing, it is worthwhile exploring how sharing of test items among medical schools can address the increased cost. Finally, but importantly, writing effective test items is and remains an art regardless of the sophistication of the assessment method.

Putting it Together: Planning an Effective Evaluation System

Rebecca Henry, Ph.D.

Professor, Office of Medical Education Research & Development

Michigan State University College of Human Medicine

June 17, 2004

This final session addressed how faculty might use many of the concepts presented in this evaluation series to create a broader system of evaluation. Initially the talk distinguished broad purposes of evaluation. This was accomplished by orienting the participants to Jacob's five-phase model for program evaluation that covers pre-curriculum evaluations (e.g. needs assessments and task analyses) accountability evaluations and program impact evaluations. Curriculum planning, implementation and evaluation were considered as integrated components of larger systems not independent activities.

Several tools were presented to participants to assist in designing program evaluation. First we discussed how the evaluation system could focus on: learners; courses; or the entire academic program and its related mission and outcomes. In determining what to evaluate, faculty can select the content, process, learners or outcomes; for each one can ask "what," "who," or "how" related questions.

Next, participants examined Kirkpatrick's hierarchy of levels of evaluation that has considerable overlap with Miller's hierarchy of competence. In evaluation, one can incorporate evaluation that emphasizes the reaction of learners (satisfaction), learner accomplishments (knowledge and skill acquisition), transfer of learning to new or real settings or ultimately, the impact of the program on important outcomes such as health care delivery or community.

The session then addressed how databases can be used as practical management tools in evaluation. One such tool used at the College of Human Medicine tracked all the performance-based assessments across the four-year curriculum. For each core area recognized by NBME (e.g. history taking) the database reflects where in the curriculum

the assessment occurs; classification of the assessment method (e.g. standardized patient); and if it is a primary or secondary source of evaluation data for the College. From this matrix we are able to: determine areas where we have evaluation gaps or redundancy; establish if we are using a desirable range of assessment strategies; and determine if our courses and rotations are incorporating the types of assessments valued by the College.

Finally, the "Evaluation System Checklist" was discussed that is designed to help faculty examine not just their own course evaluations but the entire program system and whether that system provides important information for decision-making. For example, a system for evaluation should have a broad mission statement that guides decisions about evaluation priorities and resources. Also, are there specific protections for student privacy and confidentiality?

The seminar finished with questions and observations about the challenges of creating practical evaluations that inform us on the progress of our academic programs and the learners served by them.

Series Summary

Several recurring themes can be seen in this series. One is that any method of evaluation of student learning must be carefully planned before the educational endeavor is undertaken. How they are to be used, and for what purpose they will be used must be determined beforehand. For example, this may be formative or summative evaluation of students, or program evaluation. This is especially true of interventions that will not be evaluated by typical objective exams. Second, there must be continuous and consistent feedback to the evaluators about the reliability and usefulness of the methods being used. Frequent refinement may be necessary. Third, methods must be consistent with the educational setting and methods by which students are learning. Finally, more than one method of evaluation should be in place (for example, direct observations by faculty, skills assessments, computer-based assessment, evaluation of student logs or student reports). The principles presented in the series will be extremely helpful to faculty and administrators assessing their own methods of evaluation of student learning.