UME-21 Local Evaluation Initiatives: Contributions and Challenges

Judy A. Shea, PhD; Patrick D. Bridge, PhD; Bruce E. Gould, MD; Ilene B. Harris, PhD

Background and Objectives: The 18 medical schools involved in the UME-21 initiative developed innovative curricula and evaluation strategies. While there was significant variation in how schools approached the evaluation process, there were common methodological issues and challenges affecting the reliability of scores and validity of interpretations regarding outcomes. This paper explores these issues and challenges, using experiences from selected UME-21 schools. Methods: Four evaluation issues and strategies are discussed: instrument development, study design, process evaluation using formative evaluation methods, and qualitative strategies. Within each discussion, examples from a UME-21 school are presented. Results: The four evaluation strategies offered the flexibility to match local evaluation needs with an effective approach to evaluations. Conclusions/Implications: The school-level evaluation requirements by the UME-21 initiative provided schools the flexibility to design individualized evaluation strategies, yet also encouraged collaboration among evaluators. While this strategy resulted in many successes at the school level, it also served to identify common methodological challenges that can be used as a guide for other schools in implementing and evaluating curricula.

Curricular development in US schools of medicine is an ongoing process. Often, however, evaluation of curricula is not a preplanned part of the change process, except when curricular change is part of a funded project, such as the Undergraduate Medical Education for the 21st Century (UME-21) initiative.

There are common and significant challenges to evaluation of educational programs, and these are outlined in Table 1. Each of these challenges affects the reliability and validity of evaluations. Nevertheless, reliability and validity can be enhanced by using, to the extent possible, recognized methodological guidelines for instrument development, study design, formative evaluation methods, and qualitative methods. The benefits and challenges of integrating standard methodological guidelines to evaluate local evaluation projects is described, using examples from the UME-21 curricular initiative.

Specifically, this paper’s purpose is to illustrate approaches to four important issues affecting the reliability of scores and validity of interpretations regarding outcomes: (1) instrument development, (2) study design, (3) formative evaluation, and (4) qualitative approaches. Our objective is to highlight how selected UME-21 projects addressed evaluation issues using high-quality, although imperfect, methods.

Instrument Development

Well-known and widely used assessment instruments (e.g., “shelf” exams, National Board of Medical Examiners Step scores) are constructed using sound psychometric principles. Unfortunately, many well-known instruments do not fit the unique needs of a local study, and thus, instruments must be developed. While there are multiple sources that give good advice on how to create instruments such as surveys,1 rating scales,2 and knowledge tests,3,4 three important components of the process deserve brief mention here: blueprinting, pilot testing, and scoring. Blueprinting is used to generate content-valid instruments and specify the objectives of an intervention, linking the objectives to the items on the instrument. This may be the most important step in the instrument development process,5 although many evaluation efforts do not follow the test blueprinting process.6 Once the evaluation instrument has been designed, there is a need for pilot testing, where items are tested and modified if necessary, prior to broad-scale implementation. Even brief pilot testing, followed by debriefing and “think aloud” exercises, can improve
Table 1

Common Challenges to Robust Evaluation Programs

- Lack of funds
  - Budgets for evaluation are disproportionately small compared to what
    is given for the innovation and its implementation.
- Lack of measurement expertise
  - By necessity, much of the leadership in curricular innovation is
    provided by clinicians who are rarely measurement experts. Moreover,
    at many schools, access to evaluation expertise is rare.
- Lacks importance
  - Unfortunately, evaluation is often considered a necessary evil and often
    receives the least amount of attention.
- Short timelines
  - Curricular innovations often have clearly defined and inflexible times.
    Few support long-term points of observation that would allow a
    conclusion that sustained change has occurred.
- Late planning
  - Innovations occur in a dynamic environment, giving insufficient time
    to put into practice the “best evaluation practices.” Thus, data available
    for evaluation are suboptimal to support the research.
- Moving targets
  - Educational interventions are carried out live in active laboratories
    where students are maturing, learning, and changing—all totally
    independent of any intervention.
- Lack of specificity
  - Often, evaluation plans are generic, for example broadly stating that
    there will be changes in learners’ knowledge, attitudes, and behaviors.
    Exactly how these domains will be operationalized is typically left
    unspecified.
- No gold standard
  - Often, results are reported as changes in means scores or significance
    of correlations; however, without a gold standard or robust outcomes
    measure, it is impossible to tell whether these scores are good or bad.
- Volunteer subjects
  - Often, research subjects are students or residents. The extent to which
    volunteerism, based on goodwill, impacts the results is unclear but
    should be considered.

an instrument. Scoring depends on the type of instrument developed; scoring might be objective, as in the case of multiple-choice exams, but it is often subjective, such as when it is dependent on the interpretation of a rater or of multiple raters with differing opinions. Controlling for subjectivity can be difficult, but there are methods that can be integrated into the evaluation process.

Local UME-21 Examples

At the University of Pittsburgh, investigators developed an end-of-clerkship evaluation for their ambulatory medicine clerkship. Two early decisions guided development of the evaluation instrument. The first is to let the nine domain areas from the UME-21 project serve as the blueprint. The second is to use the “critical incident technique” to guide item formation. The latter technique involves creating questions for each content area that are constructed to focus students’ attention on their clinical experiences. Questions for each UME-21 subject area are created in pairs that probe for positive and negative experiences. For example, for the health care finance content area, the following questions were asked: “Did you observe an incident where a preceptor was able to apply their understanding of health care finance/economics to bring about a positive outcome or occurrence for a patient?” and “Did you observe an incident where a preceptor’s lack of understanding of health care finance/economics led to a negative outcome or occurrence for a patient?” A key strength of this technique is its ability to elicit specific, focused comments from which themes can be derived through qualitative analysis.

Questions were written by faculty from the clerkship evaluation committee. A draft of the questions was critiqued by members of the curriculum committee and then questions were revised in response to critiques. In July 2000, the instrument was pilot tested with 40 students at the end of their clerkship. During pilot testing, students were able to describe memorable positive and negative experiences, but the investigators learned from students’ feedback that the instrument was too long, leading to a low completion rate and short answers. Subsequently, the instrument and delivery strategy was changed. A unique set of three item pairs is now given to students completing each 12-week block. Since the curriculum is repeated during each block, enough data are collected over time to evaluate each content area of the curriculum. While the instrument development process can be tedious, there are long-term gains for following strict measurement practices.

The University of Pennsylvania used a somewhat different approach. Faculty there dealt with scoring an evaluation instrument for the financial history component of their medical interviewing curriculum. In spring 1999, a class of 150 students participated in didactic sessions focusing on interviewing skills related to insurance and payment issues. In fall 1999, the students prepared six weekly write-ups of a patient seen in a medical setting. The consistency in scoring these write-ups illustrates an inter-rater reliability issue.

To assess reliability, trained research coordinators at the University of Pennsylvania developed a detailed checklist to abstract student write-ups of the financial interview. To increase the inter-rater reliability of the abstractions, they used an abstraction form, randomly selecting six student write-ups. After a review of the write-ups and verbal agreement on the scoring, each rater then abstracted 10 more write-ups and met together to assess the inter-rater reliability and to resolve differences. This technique generated consistency in the scoring procedure that increased reliability of scores.

Study Design

Many evaluators have been trained in disciplines that place value on randomized controlled trials. With many educational interventions, randomized trials are not feasible. Quasi-experimental designs that lack random-
ization are often used in education, yet the robustness of the design can be maintained by using a control group and pretest/posttest measures.

It is important to have a control group—ideally similar or equivalent to the intervention group. In educational research, the control group is often historical (ie, the preceding year of students) because one needs to give the intervention to all students enrolled in a class. Additionally, stronger designs have pretest and posttest outcomes assessment, the latter preferably administered at some interval distant from the intervention. Although best practices are not always feasible, investigators must understand the weakness of a chosen design and be able to effectively interpret the results. Classic designs are outlined in sources such as Campbell and Stanley.

Local UME-21 Examples

The UME-21 initiative at the University of Connecticut focused on developing and implementing a quality improvement curriculum. Their project used a classic pre-post test design to assess outcomes following several interventions. In one intervention, 77 second-year students, working in groups of two to four, abstracted charts at their clinics to assess the quality of treatment provided to patients with diabetes mellitus. A random sample of 513 charts was abstracted at baseline, and 380 were abstracted at a 6-month follow-up. Training in chart abstraction, as well as items and tools for abstraction, was provided by personnel from Connecticut’s state peer-review organization. Students observed that following the interventions, the rate of documentation of performance of foot and eye exams increased significantly from baseline to the 6-month follow-up, and the mean value for glycohemoglobin levels dropped significantly.

The investigators concluded that a project with an aim of improving quality of care was feasible. However, as acknowledged by the investigators, because there was no control group, it is not certain that observed changes were due to their curriculum rather than to other forces that may have heightened physician awareness of diabetes management issues.

In the University of Pennsylvania financial history-medical interviewing project mentioned earlier, the investigators used a historical control—students in the prior class who had not had the intervention. Outcomes from the intervention were assessed by comparing the two classes of students’ performance on write-ups and on a history-taking examination with a standardized patient. Outcomes showed that 34% of students in the intervention group noted a financial interview in their write-ups, and 75% asked about finances as part of the standardized patient encounter. Outcomes far exceeded the 3% of students in the control group who noted a financial interview in their write-ups. The investigators concluded that the curriculum had at least a modest effect, though there was room for improvement.

Formative Evaluation Methods

While there are many theoretical approaches to evaluating a program, two of the most basic methods are formative and summative evaluation. Formative evaluation improves the process of curriculum construction, teaching, and learning and is used for new or existing programs as a tool to document continual progress or as a systematic method to detect unanticipated events and strengthen program processes. Summative evaluation, on the other hand, measures the success of curricula in achieving learner objectives. While both methods are vital, formative assessment is often overlooked, possibly affecting the validity of research outcomes.

Local UME-21 Example

Faculty at Wayne State University developed a 6-month, 1 half day per week continuity clinic clerkship for year-3 students, which was integrated into the family medicine, internal medicine, and pediatrics clerkships. Because of the large class size (256 per class), and the placement of students in both urban and rural communities, evaluating program processes was a major challenge. Therefore, two types of formative evaluation strategies were developed: preceptor evaluations and student evaluations.

The preceptor evaluation focused on process questions related to their understanding of clerkship objectives and learning assignments, their ability to teach to the objectives, and their role and responsibility in the clerkship. Initially, some preceptors struggled with aspects of the clerkship. For example, the first formative evaluation in 1998–1999 showed that 10% of preceptors did not understand their role in the clerkship, and 14% did not understand the clerkship objectives. Consequently, a faculty development intervention was introduced, and improvement of the preceptors was evident. After the second formative evaluation in 1999, only 5% of preceptors reported that they did not understand their role in the clerkship; in 2001, fewer than 1% responded negatively to these questions.

Student evaluations revealed similar issues. In 1998–1999, 36% of students felt the expectations of the clerkship were not made clear, and 53% did not understand the relevance of learning exercises. But, with continual process and curricular refinements, later evaluations showed that only 5% of students thought clerkship expectations were not clear, and 15% did not understand the relevance of the learning exercises. Overall, integrating a formative evaluation into the UME-21 program was beneficial in identifying key process issues that would have affected the validity of program outcomes.
Qualitative Approaches
Increasingly, qualitative approaches are used to help evaluate educational interventions. Four qualitative methods are particularly applicable for assessing educational interventions. First, participants’ responses to open-ended questions on written surveys can be coded and analyzed using the methods of qualitative analysis to identify and validate the presence of themes. Second, focus groups bring a group of participants together to express views about a topic, allowing investigators to learn more about reactions of its members than might be learned with individual interviews. Third, individual interviews, ranging from very structured to open-ended questions may be appropriate for eliciting perspectives when a focus group cannot be arranged or when participants represent different stakeholders and may not be comfortable talking in a group setting. Fourth, observational studies provide for viewing the naturalistic process and outcomes of a particular intervention. Overall, qualitative methods provide a rich repertoire of approaches for assessing the process and outcomes of interventions. They usually provide different, yet complementary information than quantitative data, leading to more valid study outcomes.

Local UME-21 Example
At the University of Minnesota, one of the UME-21 initiatives was to assess students’ attitudes about managed care. During the last week of each 8-week required primary care clerkship rotation, students in groups of 35 participated in a 2-hour colloquium discussion with two managed care executives. The colloquium was designed for students to gain a better understanding of the functioning, strengths, weaknesses, and evidence-based decision making process of managed care organizations. Students were encouraged to ask challenging questions. All third-year medical students (class of 2002) taking the required clerkship participated in the intervention. A pre and post survey was administered to elicit attitudes about the value of managed care delivery systems for various aspects of health care. Students also responded to open-ended questions about the effect of the colloquium on their knowledge and attitudes and the role of various aspects of the colloquium and its impact.

The investigators found moderate but statistically significant changes in attitudes from pre-post colloquium in favor of managed care health care delivery systems. The open-ended questions were analyzed using standard procedures, helping evaluators understand the reasons for changes in attitudes and thus enhance validity. For example, students were impressed by the credibility and knowledge base of the executives and the evidence-based methods used to make decisions about reimbursement. This design did not incorporate a control; however, the changes in attitudes over a brief time period, combined with the analysis of students’ comments, contributed to a persuasive assessment of the value of this intervention, at least for the short-term.

Other Evaluation Challenges
Conducting local evaluations of educational interventions presents a host of challenges. One common theme is that evaluation requires planning, and rigorous planning takes time and money. It is eye-opening that even within UME-21, which provided more than the usual amounts of planning time and funding, challenges were numerous. In the spirit of encouraging investigators to consider what is needed to conduct a careful evaluation, we return to the more common challenges presented in Table 1, recognizing that only rare projects can successfully meet all of the challenges. The length of this list may give the impression that conducting a robust local evaluation project is a hopeless process. Careful planning and following sound measurement theory can often overcome many of the challenges and enhance both the reliability of instruments and validity of student outcomes.

Evaluation is at worst an afterthought. At best, it is a planned process that leads to useful data and credible conclusions. The UME-21 project afforded the opportunity for many local evaluators to think ahead, work together, make plans, and design good studies. Because of this strategy, common methods and evaluation challenges were identified. Recognizing and preparing for such challenges is a key factor toward successfully evaluating education programs.

Corresponding Author: Address correspondence to Dr Shea, University of Pennsylvania, School of Medicine, 1223 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021. 215-573-5111. Fax: 215-573-8778. sheaja@mail.med.upenn.edu.

REFERENCES