PERSPECTIVES

The Structure of Program Evaluation: An Approach for Evaluating a Course, Clerkship, or Components of a Residency or Fellowship Training Program

Steven J. Durning
Paul Hemmer
Louis N. Pangaro
Department of Medicine
Uniformed Services University of the Health Sciences
Bethesda, Maryland, USA

Background: *Directors of courses, clerkships, residencies, and fellowships are responsible not only for determining whether individual trainees have met educational goals but also for ensuring the quality of the training program itself. The purpose of this article is to discuss a framework for program evaluation that has sufficient rigor to satisfy accreditation requirements yet is flexible and responsive to the uniqueness of individual educational programs.*
Summary: *We discuss key aspects of program evaluation to include cardinal definitions, measurements, needed resources, and analyses of qualitative and quantitative data. We propose a three-phase framework for data collection (Before, During, and After) that can be used across undergraduate, graduate, and continuing medical education.*
Conclusions: *This Before, During, and After model is a feasible and practical approach that is sufficiently rigorous to allow for conclusions that can lead to action. It can be readily implemented for new and existing medical education programs.*

Teaching and Learning in Medicine, *19*(3), 308–318

Directors of courses, clerkships, and residencies are responsible not only for determining whether individual trainees have met educational goals but also for ensuring the quality of the training program itself. This "program evaluation" is more than the aggregate of individual trainee data; it requires academic directors to employ a dynamic, longitudinal evaluation process that tracks multiple contributing factors and outcome measurements.

The desirability and necessity for academic directors to have a framework for program evaluation is evident in the Accreditation Council for Graduate Medical Education (ACGME) outcomes project, which, starting in 2006, expects residency and fellowship program directors to document the relationship between process measurements (what we do to our trainees) and product measurements (educational outcomes).[1] Likewise, the Liaison Committee on Medical Education (the accrediting body for U.S. medical schools) requires that clerkship directors specify the problems all students are expected to encounter during a clerkship (ED-2).[2]

The purpose of this article is to discuss a framework for program evaluation that has sufficient rigor to satisfy accreditation expectations and still be flexible and responsive to the uniqueness of individual educational programs. We recognize that there are existing models of program evaluation, such as models of quality management[3] and models such as CIPP (context, input, process, product[4,5]), and our framework draws from and builds on these models. Our intent is to demonstrate how practical our framework can be for conducting program evaluation in medical education

Disclaimer: The opinions expressed in this article are solely those of the authors and do not reflect the official policies of the Department of Defense, the United States Air Force, or other federal agencies. We thank Dr. Eric Holmboe for his assistance with suggested revisions for this article.

Correspondence may be sent to Steven J. Durning, USUHS–Department of Medicine (NEP), 4301 Jones Bridge Road, Bethesda, MD 20814, USA. E-mail: sdurning@usuhs.mil

training programs. This framework can be implemented quickly and without difficulty in a course, clerkship, residency, or fellowship training program; as such, our article is intended for course, clerkship, and residency or fellowship directors as well as those who oversee them, such as chairs of departments, deans, and perhaps chief operating officers of hospitals. We emphasize a uniform structure that can be applied to all, irrespective of the level in the medical education continuum.

The proposed framework emphasizes the role of baseline, process, and product (outcome) information, both quantitative *and* qualitative, for describing program "success." It should be clear that this emphasis includes outcome measurements, such as patient care parameters. Nevertheless, despite the recent emphasis of the ACGME on "outcomes," we emphasize the importance of process measurements (such as the number and kinds of patients seen during training and the level of proficiency obtained), as many available outcome measurements have uncertain reliability and validity and because, eventually, the outcomes should be used to refine the curricular process. For example, with the established reliability and validity of the mini-clinical evaluation exercise (mini-CEX)[6,7] the ABIM now requires programs to directly observation trainee's clinical skills in a mini-CEX or similar format. Finally, we advocate for the inclusion of baseline measurements, which may allow us to determine how much of an eventual outcome depends on the curriculum, as opposed to the prior characteristics of trainees.

Summary tables show how different types of information fit within the framework and practical, and specific examples illustrate how this framework can work for complex educational programs. Because implementation of a comprehensive system for program evaluation will often be gradual, we also provide "essential" and "desirable" recommendations for using this approach to accomplish this high-stakes task. While many of the measurements used in this article are derived from expectations of the LCME or Residency Review Committees of the ACGME, this article is not meant to serve as an exhaustive checklist of measurements.

Defining the Task of Program Evaluation

In this article, we define "success" for program evaluation (PEv) as achieving information that can relate "inputs" to "outputs" and therefore be used to help understand sources of success or failure. This relationship can be used in many ways—for instance, to help determine if a new component of a preclinical course improves performance on a standardized exam, if a multisite clerkship provides consistent evaluation across all training sites, or if a residency or fellowship training program compares favorably with ACGME standards. In other words, our goal is to *understand* how the program is working rather than simple classification of graduates (e.g., competent or not competent).

PEv should begin with a definition of success for the specific program. It is not sufficient to say that "good patient care" or "patient safety" is the goal. The description of success must be construed with sufficiently precise words to allow, at a minimum, the determination of whether success has been met, in a dichotomous *yes*/*no* fashion. If possible, it is also desirable to describe degrees of success. In either case, these goals and expectations should be clear, specific, and tangible. For educators who are uncertain regarding how to define success, consider answering the following question: "I would be happy about my program if I knew that . . ." For instance, "Do all my residents/students pass their board exams?" However, we all recognize that demonstrating success is often defined externally, by accrediting bodies such as the LCME or ACGME. In fact, one can see that much of the data gathered is material that the LCME or ACGME explicitly define as necessary. In some cases, however, measurements required by regulating bodies are not sufficient to answer what the program director wishes to know. For instance, "Do I attract the best applicants to my own program?"

After listing goals, objectives should be constructed for determining success in achieving the goals. For example, a 3rd-year clerkship may define consistency across geographically separated teaching sites as a goal (site-to-site consistency). This is based on the assumption that any differences in student outcomes on a clerkship should be due to factors inherent to the students and not to curricular factors such as the geographic site. Objectives for achieving this goal might include no difference in examination scores or teacher evaluations of students based on site.

However, defining success is only the first step; understanding how to approach the question is more difficult. We have found that relatively little practical guidance exists for systematic evaluation of educational programs; the limited guidance that does exist is restricted to graduate medical education arena.[8–10] Indeed, the research pertaining to program evaluation in medical education is less developed than other educational fields and is largely descriptive.[11] Educational programs have been felt to be effective if its graduates appeared to have learned the stated objectives and successfully completed requirements of the program.[8] The literature that does address specifics primarily deals with evaluating individual domains within an educational program or accreditation, without providing a practical and rigorous approach for evaluating the entire program.[9] The literature also emphasizes the specific tools, rather than a framework in which to put the tools. In other words, it is assumed that having several

outcomes per domain ("triangulation") will be sufficient for the task. This is based on the assumption that the outcome measurement(s) reflect what is essential for the trainee to achieve during the program. Evaluation models have not existed, and there has been no overarching theory-driven structured approach to program evaluation that can be applied to programs across the educational continuum.[9]

PEv Framework Overview

We advocate a three-phase framework for program evaluation. This framework allows for establishing relationships among baseline, process, and product measurements—Before, During, and After. Baseline measurements address information about the trainee from *before* entry into the program. Process measurements[11] are those collected about the trainees, the faculty, and the program that are made *during* the curriculum. Outcome measurements occur *after* (or at least at the very end of) the course, clerkship, or residency training program. All of these measurements have often been based on what the graduate does under testing circumstances (such in vitro measurements would include licensure or certifying examinations), but we also want to include in vivo observations such as what trainees do in patient care and graduates do in their practice.

Before (baseline) measurements are necessary to determine "how learners change," and they are especially important to determine the effect of curriculum as opposed to selection of trainees. Typical examples of baseline measurements that a director of an academic program may wish to collect are illustrated in Table 1. Several studies suggest that most of the explained variance in trainee performance in a course or clerkship is due to baseline characteristics of the learners.[12,13] Thus, baseline measurements allow academic managers to put program outcomes into context. For example, assume a class of post-graduate year 2 residents in internal medicine takes the In-training Examination. Three of the 20 residents score below the 35th percentile. Determining if this is "success" (have these 3 residents improved?) or an indicator of a problem (is there a deficiency in the program?) requires knowing how these residents performed on prior standardized examinations.

During (process) measurements are those that monitor the activities of learners during the training program (Table 2). These measurements are often collected prospectively, that is, in real time. The number and types of process data that can be gathered could be overwhelming to a director of an academic program, and therefore a clearly articulated statement of program evaluation success facilitates the choice of process measurements. A director of an academic program also needs to consider unexpected results—collateral unexpected benefits or shortcomings. "During" measurements, therefore, need to be prioritized for program evaluation purposes so that response to critical, potentially unexpected, information is not delayed or potentially overlooked.

After (product) or outcome measurements, in the clinical research literature are analogous to primary and secondary end points. Primary end points indicate the overall success of management. These are analogous to the classifying of graduated trainees as competent or not competent. Secondary clinical end points indicate intermediate success or complications. As the majority of outcomes in medical education are as complex as in clinical studies, data gathering should optimally be done through multiple measurements (triangulation). Examples of product measurements are also illustrated in Table 1. A recent article lists some additional outcome measurements to be considered,[8] including satisfactory completion of minimal numbers of identified clinical procedures and percentage of residents involved in community service activities. A second recent article discusses the use of a structured portfolio[10] as an outcome measurement for residency education.

It is evident that we emphasize the importance of process measurements; we believe that we must not ignore systematic process measurements, or baseline measurements, at the expense of focusing on "outcomes" for several reasons: Our current outcomes in medical education are imprecise, process measures are required by regulating bodies (LCME with ED2, ACGME with outcomes project), and process measurements are indispensable to explaining the variance in the outcomes (products) of interest.

A variety of measurements can, and should, be used in this three-phase framework (see Table 2). Both qualitative (descriptive) as well as quantitative (numerical scores) assessments can, and in many cases, should be used for program evaluation[14] as the quantitative measurements alone may overlook important findings that are revealed through qualitative analysis. Particularly for new innovations (a new educational program and/or a new component of an existing educational program) and/or formative program evaluation, we advocate at least one qualitative measurement in addition to quantitative measurements because qualitative measurements complement the quantitative data and can be particularly useful for detecting unexpected outcomes.[10] Qualitative data can also serve as a rich database for hypothesis generation that can subsequently be confirmed with further qualitative and/or quantitative data. Indeed, some have argued that qualitative comments may be the most important source of data for program evaluation when dealing with something as complex as competence.[15]

The basic three-phase structure that we propose for program evaluation readily applies to both existing

Table 1. *Using a Framework of Before, During, and After Measurements to Assist in Program Evaluation and Determining "Success"*

Undergraduate medical education			Graduate medical education		
Baseline Measurements "Before"	**Process Measurements "During"**	**Product Measurements "After"**	**Baseline Measurements "Before"**	**Process Measurements "During"**	**Product Measurements "After"**
Student self-assessment	Patient logs	Surveys (4th year, graduates, program director)	Med school grade narratives	ACGME toolbox items	Graduate surveys
Exams (USMLE, pretest)	Procedure logs	Placing graduates (i.e., location, choice)	Clinical Performance Evaluations (e.g., OSCE, SP)	Procedure logs	Employer surveys
Clinical performance evaluation (OSCE)	Hours	Future exams (USMLE, ITE)	Dean's letter, other recommendations	Hours	Success placing graduates (academics, practice)
Compliance with requirements (e.g., immunizations)	End-of-clerkship critiques	Clinical performance evaluations	Interviews	Patient logs	Exams (e.g. board certification
GPA prior to clinical years	Student exit interviews	Narratives of trainee performance (e.g., 360 degree evaluation, peer assessment)	Exams (e.g., USMLE)	End of rotation teacher critiques	Professional society participation/advancement
Introduction to Clinical Medicine Grades	Clinical performance evaluation	Review of write-ups	GPA	Learner exit interviews	Academic productivity
MCAT scores	Portfolios	Attitudinal	Class rank	Examinations (ITE, local)	Research productivity
Critical Incident Reports	Reflective exercises	Student portfolios	AOA membership	Attitudinal	Disciplinary (e.g., NPDB, license)
Exams (USMLE, pretest)	Attendance rates	Professional development plans	4th-year clerkship grade	Patient care measurements	Patient care measurements
		End of clerkship exams (NBME, analytic, pattern)		Patient surveys	Patient surveys
		Critical incident reports			Career pathways
		Peer Evaluations			
		360 degree evaluations			

Note. This approach can also assist with prioritization of data to collect. This list is not intended to be all inclusive but serves as an illustration of how the framework can guide decisions about data to collect. USMLE = United States Medical Licensing Examination; ITE = in-training examination; GPA = grade point average; NBME = National Board of Medical Examiners; AOA = Alpha Omega Alpha Honor Medical Society; NPDB = National Practitioners Data Bank; OSCE = Objective Structured Clinical Examination/Exercise; SP = standardized patient.

programs and new programs, curricula, or interventions. Furthermore, it is important to realize that the data collected can be organized for different units of analysis—for instance, to compare groups of learners from year to year, geographic site to geographic site, or before and after a course or clerkship.

Collecting multiple measurements in each phase of our framework can require significant time and human resources. This is one of the reasons that we believe that data gathering for program evaluation can, and optimally should, be supported by numerous stakeholders in the academic program. All those responsible for the program's performance should have input into defining the questions to be asked and data to be gathered about the program. For example, medical school deans and hospital chief executive officers have a stake in deciding the length of clinical clerkships and the number of funded residency positions, respectively, and so their views are a critical element in determining measures of program evaluation. Indeed, program evaluation information should inform multiple decision makers in

Table 2. *Examples of Needed Resources in an Educational Program*

Time	Human resources	Funding
Reviewing evaluation forms (Global evals, mini-CEX forms)	Clerical support	Supplies
Feedback from and to instructors	Education or oversight committees	Informatics (personal computers, programs)
Exit interview and/or survey of trainees	Mock review/external evaluators	Research
Curriculum coordination	Patients	Office, clinic, and/or hospital overhead
Reviewing exam scores (NBME, ITE, boards)	Hospital and clinic personnel	
Reviewing call and other schedules		
Comparing and compiling data (e.g., student logbooks, resident procedure logs, student and resident critiques of faculty)		

Mini-CEX = mini-clinical evaluation exercise; NBME = National Board of Medical Examiners; ITE = in-training examination.

the program; if this is not the case, then the utility is too limited. Stakeholders include anyone with an interest in the course, clerkship, or residency training program such as medical school deans, department chairs, university presidents, hospital chief executive officers, and the public. The course, clerkship, residency, or fellowship director who oversees the program, however, should play the primary role in interpreting data for specific program evaluation.

In addition, our three-phase framework to program evaluation could foster collaboration across the medical education continuum, as the baseline measurements for a 3rd-year clerkship director may comprise the outcome measurements of a 2nd-year course director and the outcome measurements of a clerkship director can serve as baseline measurements for a residency training program director.

To illustrate this method of using multiple stakeholders for collecting data for program evaluation, we use an example of an educational intervention that might improve patient care.

Educational Intervention Example

A residency program director plans to begin a yearly procedure workshop in which residents rotate through stations, practicing such common inpatient procedures as inserting a central venous catheter or nasogastric tube, but the intervention uses mannequins and simulations rather than real patients. "Success" for this innovation might be defined, in part, as a reduction in complications from procedures and improved patient satisfaction with these procedures. The component of the residency training program that is being evaluated is resident procedural skill ("Patient Care" in the ACGME core competencies).

For this intervention, different stakeholders provide important viewpoints, based on their role in the system, which should be considered:

Individual *residents* can provide information on prior training in the skills and/or prior experience with the procedures (survey), feedback on the effectiveness of the workshop (i.e., survey and/or focus group), on their confidence in procedure performance before and after the intervention (i.e., survey and/or focus group), and on the change (if any) in procedure complications after the intervention (i.e., reviewing their procedure logbook).

Observing faculty might comment on change in resident skill with the workshop (i.e., checklist and/or global evaluation form), cost/benefit of the time needed for workshop (i.e., survey), and overall skills and deficiencies of these residents. They also might be the evaluators of procedural performance after the workshop and may refine their teaching of procedural skills on the wards based on observing housestaff during the workshop (focus groups after the intervention, critiques from trainees).

Patients can comment on their comfort and perhaps their perception of resident comfort and skill with the procedure (i.e., checklist and/or interview). They might also comment on the resident's explanation of risks/benefits and their satisfaction with the procedure performed (i.e., interview).

Data collection for program evaluation can be performed by all stakeholders in a program; indeed, well-designed program evaluation has high utility to most, if not all, participants in a program. The data analysis optimally is performed by the academic director for the program (with statistical consultation, if needed), as he or she understands how measurements from different sources and of different nature (qualitative vs. quantitative) can and should be triangulated and appraised to provide answers to program evaluation questions.

Collection of data for program evaluation may require obtaining permission from the Institutional Review Board if the goal of data collection is peer-reviewed publication or presentation. When data are

collected for quality control, with no consideration for presentation outside of institutional leadership, Institutional Review Board permission may not be needed.

Distinctions and Definitions

Next, we define the desirable attributes for the assessment tools that are placed within the model (the *micro* level) and afterward the desirable attributes of the overall framework (model) for the *macro* level. Optimally, a tool for assessing the success of a program, is *feasible, reliable,* and *valid*. For our purposes at the micro level, feasibility means the percentage of possible measurements that are actually obtained and the unit cost per measurement (i.e., cost of printing, mailing, and/or entering survey data); *reliability* means the internal consistency of specific assessment tools, and v*alidity* is the confidence that the inferences drawn from the data are true. Thus, in our model, validity will mean that process measurements have a significant, meaningful, predictive association or correlation with outcomes measurements and that outcomes measurements have a similar correspondence with patient outcomes

At macro level, academic directors optimally should be able to collect the same set of data the same way for each trainee, at each site, each year to help ensure reliable and valid observations collected for program evaluation purposes. The overall method for program evaluation must be feasible—measurements must be obtained *effectively* (method should allow at least 90% of possible observations about trainees, faculty, etc., to be captured each evaluation time), *consistently* (observations and ratings are recorded, transferred, and stored without degradation), *efficiently* (no more than 10% of a course, clerkship, program, or fellowship director's time must be consumed, and no more than 10% of an administrator's time is needed), *economically* (cost of program evaluation should be no more than 5% of the operating budget for the course, clerkship, or graduate medical education program), and *securel*y (trainees are protected from their data being shared, and any risk is minimized).

At the macro level, this means that each set of measurements (before, during, and after) adequately reflects the construct appropriate to the framework. In other words, the set of baseline measurements reflects the adequacy of the 2nd-year student's preparation for the clerkships or of a graduating student's preparation for internship. The set of process measurements accomplished during the curriculum adequately reflect the concept that the student did what was expected (i.e., saw the appropriate number of patients, wrote the appropriate number of histories and physicals).

For the program evaluation model to be valid, inferences that the process caused, or contributed to, the desired outcome must be reasonable. The usual standard for causality in clinical medicine is the prospective, randomized, blinded trial. This is difficult to achieve in the educational setting, except for subtotal modifications of the curriculum such as restructuring an individual clerkship or changing a 1-month rotation during residency. Therefore, validation of our program evaluation model would mean that proposed explanations for how variance in outcomes (skill of our individual graduates, in the aggregate, in taking care of their diabetic patients) was related to specific curricular elements (whether they actually performed the practice-based learning and improvement review of their own care of diabetics) would require two levels of evidence: initially statistical demonstration (e.g., through correlation or multiple regression models) and subsequently improvement in diabetic care given by residents graduating after further modification in their curriculum. The latter has not been published in the literature.

PEv Resources and Measurements

We think that it is helpful to define essential and desirable resources for program evaluation at the outset, as this can assist with requests for funding as well as facilitate best use of limited resources. It can also be helpful to divide needed resources into time, human resources, and funding to help generate this potential list of essential and desirable resources. We have included an example of potential needed resources for program evaluation in an internal medicine residency-training program in Table 2. With an understanding of what resources are available and what resources are needed, a director of an academic program can then decide which measurements to collect for program evaluation purposes.

PEv Practicalities

Timing—Identifying Problems Early

The appropriate timeline for data collection, analysis, and reporting can differ based on the academic program. When learners rotate through educational activities on a yearly basis, it is usual to collect, collate, analyze, and report program evaluation findings on an annual basis. However, in a medical school course, the time line for reporting summative findings might be necessitated by the curricular schedule (e.g., a trimester); in residency training in internal medicine, the optimal timeline for data collection might be duration of required training (e.g., 3 years).

We believe that a robust formative evaluation process, based on "during" or process measurements, is a critical component of successful program evaluation. For example, program information that might cause concern, such as having a sufficient number of patients for each intern, should be collected, analyzed,

Table 3. *Red and Yellow Flags: Examples of Benchmarks to Establish in Advance to Signal Problems Within Your Aacademic Program*

Undergraduate medical education		Graduate medical education	
Yellow flags	**Red flags**	**Yellow flags**	**Red flags**
Test scores > 1 *SD* from last year	Test scores > 2 SD from year prior	ITE scores > 1 *SD* from last year	ITE scores > 2 SD from year prior
Trend to higher/lower grades	Higher/lower clinical grades	Board scores > 1 *SD* from last year	Board scores > 2 SD from year prior
Change in students ranking a given residency	Few students choosing a given residency	Trend to lower evaluations	Lower global evaluations
Clerkship site receives poor ratings	Report of trainee abuse	Site receives lower ratings from trainees	Report of resident or student abuse
Little student autonomy	Trainee reports of "service" rather than "education" being a primary focus	"Light" patient rotation	Too few patient encounters
Change in key educational personnel	Deterioration of facilities (e.g., loss of library, sleeping rooms)	Narrow spectrum of disease	No spectrum of disease
"Light" patient load rotation	Too few patient encounters	Teachers marginalize residents on some rotations	Teachers ignore residents
Narrow spectrum of disease	No spectrum of disease (e.g., clerkship students seeing only oncology patients)	Change in key educational personnel	Deterioration of facilities (e.g., loss of library, sleeping rooms)
Some teachers marginalize students	Teachers ignore students	Lower rate of return of evaluations, critiques	Minimal rate of return of evaluations, critiques
Only 70% of students complete critiques of the program	Students fail to complete critiques of the program		
Only 50% of faculty complete evaluations of students	Faculty fail to complete evaluations of students		
Hours students participate in care are different across sites	Clerkship requirements are markedly dissimilar across sites (e.g., submission of required work, number of patients seen)		
Change in ownership of hospital/clerkship site			

Note. Red and yellow flags serve as *process measurements* for the model. Such measurements allow one to quickly identify and investigate unanticipated outcomes. Yellow flags signal a problem may have arisen; red flags require immediate attention. Therefore, mechanisms must be in place to monitor what academic leaders decide are important measures of quality, consistency, and effectiveness. These lists are examples and are not intended to be all inclusive. ITE = in-training examination.

and reported more frequently than on an annual basis. These concerns might be categorized as red and yellow "flags" (Table 3). Red flags represent program evaluation data that require immediate attention and action, whereas yellow flags represent data that require more frequent follow-up than the next routine planned evaluation. Potential examples of red and yellow flags in undergraduate and graduate medical education are shown in Table 3. As every program is unique in some respects, this categorization of red and yellow flags is largely dependent on the educational program.

Program Evaluators

Internal evaluators are intra- or interdepartmental experts who evaluate your educational program. This could include an education committee, a medical education office, or other members of the medical school faculty and/or administration. *External evaluators* are often academic managers from different institutions who evaluate an educational program. As external evaluators are not involved with the local curriculum or program politics, external evaluators can potentially provide greater objectivity and a clearer perspective of programmatic strengths and needs. The program being evaluated benefits from ideas for improvement and high visibility and the external evaluators' programs might benefit from new ideas for conducting program evaluation at their own institutions. The effective use of voluntary external evaluators has been documented in graduate medical education.[16]

Visualization of Goals and Objectives

The business, quality, and clinical models all begin with the premise that there is a desired outcome: financial profit, a successfully performing product and patients' health. As stated previously, program evaluation for clinical training begins with defining one's goals in a way that can be visualized. It is not sufficient to say that "good patient care" or "patient safety" is the goal. The description of success must be construed with sufficiently precise words or quantifiable measurements to allow, at least, the determination of whether success has been met, in a dichotomous *yes*/*no* fashion.

Application of PEv Framework

Undergraduate Medical Education Example. To illustrate this framework, consider a complex undergraduate program—an individual clerkship site in a multisite clerkship. Because intersite consistency (i.e., there are no site to site differences and the individual site does not contribute to clerkship outcomes) is an explicit expectation in accrediting medical schools, we use the *site* as the unit of analysis, and we wish to detect any site-to-site (intersite) differences that do exist. As a principle, we choose the unit of analysis—the site—that would be most likely to reveal the problem (i.e., "inconsistency"). The goal could be stated as, "I would be happy to know that my students' educational experiences across my teaching sites are similar/consistent." A partial list of data that could be collected for this intersite consistency determination is shown in Table 1.

Graduate Medical Education Example. A program director of a large pediatrics residency training program wishes to determine if there has been a change in resident performance since implementation of the 80-hr workweek. In this example, the unit of analysis would be the year. Examples of baseline, process, and product measurements are shown in Table 1. Alternatively, the program director may wish to determine resident performance using the six ACGME competencies. Baseline, process, and product measurements that could be used for this form of assessment are shown in Table 4.

Data Analysis for Framework Examples. A data analysis example of how to use our pre-during-post model is shown in Table 5, returning to our intersite consistency example on a multisite clinical clerkship. Baseline measurements include an exam at the start of the clerkship (pretest) and preclinical grade point average. During or process measurements include the points given by teachers during one of the clerkship rotations, site and the total teacher points

Table 4. *Baseline, Process, and Product Measurements Using Accreditation Council for Graduate Medical Education (ACGME) Competencies*

	Baselines "Before"	Process "During"	Product "After"
Patient care	4th-year grade/narrative 3rd-year grade/narrative OSCE	ACGME Toolbox[a]	Employer survey Patient sat survey CEX Referral to NPDB State-board disciplinary action
Medical knowledge	USMLE Step 1 End of medical school GPA AOA status	ACGME Toolbox ITE and in-house exams	Board cert exam Graduate survey Employer survey Maintenance of certification
Communication	OSCE USMLE Step 2, Clinical Skills Exam Grade narratives	ACGME Toolbox	360 evaluations Employer survey Employer survey
Professionalism	Dean's letter Clerkship narratives	ACGME Toolbox	Portfolio[b] Referral to NPDB
Systems-based practice	Portfolio 360 degree evaluation	ACGME Toolbox	State-board disciplinary action 360 evaluations
Practice-based learning and improvement	Portfolio	ACGME Toolbox	Graduate, employer survey Portfolio

Note. OSCE = Objective Structured Clinical Examination/Exercise; CEX = clinical evaluation exercise; NPDB = National Practitioners Data Bank; USMLE = United States Medical Licensing Examination; ITE = in-training examination; GPA = grade point average; AOA = Alpha Omega Alpha Honor Medical Society.

[a] See ACGME Toolbox of Assessment Methods. Available at: http://www.acgme.org/outcome/assess/toolbox.asp. [b] Also consider structured portfolio, see Holmboe ES, Rodak W, Mills G, McFarlane MJ, Schultz HJ. Outcomes-based evaluation in resident education: Creating systems and structured portfolios. *American Journal of Medicine* 2006;708–14.

Table 5. *Analysis of Clerkship Intersite Differences for Clerkship Sites 1–5*

	Univariate Analysis of Internal Medicine Clerkship Intersite Differences (ANOVA)						
	1	**2**	**3**	**4**	**5**	***M* ± *SD***	**p**
Before clerkship							
Pretest (max = 100 points)	63.7	60.8	58.4	63.9	57.5	60.0 ± 7.3	.217
End of 2nd-year GPA	3.45	3.41	3.40	3.39	3.38	3.40 ± .5	.320
During clerkship							
Total teacher points, both sites (max = 72)	52.9	51.9	53.0	62.0	46.6	52.9 ± 22.2	.676
Site 2 teacher points (max = 42)	22.2	24.0	25.5	26.7	20.3	23.9 ± 11.7	.751
NBME Subject Exam	73.4	70.4	70.9	74.4	69.0	71.6 ± 7.7	.328
After clerkship							
USMLE Step 2	80.0	82.0	81.0	84.0	78.0	80 ± 6	.428
4th-year mini-CEX (overall performance = 0–9 scale)	7.1	7.2	7.1	7.3	7.0	7.2 ± .3	.412

Outcome Variance Explained by Variable (Linear Regression)		
Variables	**Adjusted R^2 (for Goodness of Fit)**	p
Clerkship site (1–5)	0.000 = none of variance explained	.532, .608
Block (time of year)	0.006 = 0.6% of variance	.310
Exam points + GPA	0.391 = 39% of variance	.000
End of second year GPA	0.272 = 27% of variance	.000

Site 4 = Highest before clerkship points (pretest, grade point average [GPA]) and during clerkship points (total teacher points, site teacher points) and after clerkship points (medicine subject exam). Site 5 = lowest before clerkship and during clerkship and after clerkship points. Without baseline variables, one might conclude there might be important differences in student performance at Sites 4 and 5. Clerkship sites = A, B, C, D, E; each student rotates at two different sites for the internal medicine clerkship. Outcome = total clinical points; NBME = National Board of Medical Examiners; mini-CEX = mini-clinical evaluation exercise; ANOVA = analysis of variance.

from the entire clerkship, and the as well as National Board of Medical Examiners subject examination scores. After or product measurements included results on USMLE Steps 2 and a 4th-year mini-CEX. Process and product measurements should be interpreted in light of the context of the baseline data, looking for measures of statistical or functional significance. A result can be statistically significant without being functionally or clinically significant. Measurements of functional significance include the effect size and analysis of variance (ANOVA). Generally speaking, an effect size of .5 or greater is considered to be functionally significant or meaningful.[17]Regression analysis is the amount of variance in the outcome that is explained by the variable of interest.

After entering the data into statistical package spreadsheet such as SPSS, an ANOVA calculation can be performed to determine if intersite differences exist. Table 5 illustrates ANOVA results for respective clerkship sites a to e. The overall mean, standard deviation, and *p* values are shown. No statistically significant differences are present, although students at Site 4 had higher baseline and subsequent measures and students at Site 5 had lower preclinical and subsequent measures. This finding reinforces the need to collect baseline characteristics, for if not included, one might conclude there was a difference across the sites.

The lower portion of Table 5 is a regression model. Note that site accounts for none of the variance and preclerkship grade point average accounts for the majority of the variance. From this analysis, the clerkship director could conclude that there is intersite consistency; that differences in student performance are due to what a student "brings to the clerkship" and not due to the experience at the site.[12] Nevertheless, much of the variance in student performance remained unexplained; qualitative measurements (e.g., students' career desires, concomitant personal or health concerns, opinions about the program requirements) could provide further insight into factors affecting performance, ones not easily measured by standard quantitative methods. Likewise, qualitative measurements can play an essential role in suggesting why variance was explained by a measurement(s) and next steps for analysis.

Limitations of the Approach

In the three-phase approach, learners serve as their own control, which is not necessarily an optimal design (i.e., randomized approach) to study the benefits of a curricular innovation. Complete data collection may not be feasible for directors of educational programs with limited resources. Success with using this model requires close cooperation from others (registrars, other course and clerkship directors). Data collection may be constrained by local Institutional Review Boards if you wish to use findings for more than

quality control in your own program. Further, the model illustrates correlation and not causation. As all factors cannot be controlled, the opportunity exists for multiple confounders with data analysis. Also, many measurements will have not been sufficiently studied to demonstrate reliability and validity in each institution. Despite these limitations, such a framework can be useful to guide efforts to evaluate the program.

We do not assert that our model is superior to other program evaluation frameworks or that our model is more useful for changing educational programs to make than more effective than other models. However, the use of a similar conceptual model to what we propose in the quality assurance literature does enhance the validity of our approach.[3–5] We do believe that our model can be effective at monitoring educational programs to make them more effective. In our program, we monitored and subsequently minimized intersite inconsistencies. Although this did not necessarily lead to any specific changes, successful program evaluation informs the stakeholders and guides their decision making, whether or not the decisions lead to change. We also propose using red and yellow flags allowing a course, clerkship, residency, and/or fellowship director to identify and remediate potentially harmful curricular and/or teacher anomalies (a definition of quality).

PEv Recommendations

1. Begin with defining the goal: "I would be happy about my program if I knew that . . . "
2. It is essential to list outcomes measurements for key parameters of success. Use triangulation—collect at least two measurements for each domain. We advocate collecting at least three measurements for each phase.
3. Then list process measurements that you think will lead to successful outcomes.
4. It is desirable to specify baseline measurements that would attribute success to the learner rather than the program.
5. It is desirable to include qualitative information along with quantitative data measurements.
6. Define the needed resources—time, human resources, and money. This will assist with feasibility of program evaluation efforts.
7. Include red and yellow flags to prioritize if unexpected/undesirable process measurements or outcomes are observed
8. Define your unit of analysis. As a principle, we recommend using the unit of analysis that would be most likely to reveal problems.
9. The analysis of data should include measurements of both statistical and functional significance. Decide the functional significance that would constitute success. Decide the statistical significance that would constitute failure.

Summary

Academic directors are continuously challenged by the need to monitor and evaluate the quality of their education programs with relatively little guidance from the literature on how to conduct this essential task. We have discussed key aspects of program evaluation and have proposed a three-phase framework that can be used throughout the spectrum of medical education. This Before, During, and After model is a feasible, practical, approach that is sufficiently rigorous to allow conclusions that can lead to action, and can be implemented across the spectrum of medical education for new and existing programs.

References

1. Accreditation Council on Graduate Medical Education, Common Program Requirements, Section VII. C. 2. Available at: http://www.acgme.org/DutyHours/dutyHoursCommonPR.pdf. Accessed August 23, 2006.
2. Liaison Committee on Medical Education. *Functions and structure of a medical school: Standards for accreditation of medical education programs leading to the M.D. degree*. Accessed June 2006. Available at: http://www.lcme.org/functions2006june.pdf
3. Donabedian A. The quality of care: How can it be assessed? *Journal of the American Medical Association* 1988;260:1743–8.
4. Stufflebeam DL. *CIPP evaluation model checklist.* The Evaluation Center, Western Michigan University, Kalamazoo, MI, 2002. Available at: http://www.wmich.edu/evalctr/checklists/cippchecklist.htm#bibliography
5. Stufflebeam DL. The CIPP model for program evaluation. In GF Madaus, M Scriven, DL Stufflebeam (Eds.), *Evaluation models* (pp. 117–41). Boston: Kluwer-Nijhoff, 1983.
6. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine* 2003;138:476–81.
7. Durning S, Cation L, Markert R, Pangaro L. Assessing the reliability and validity of the mini-CEX in residency training. *Academic Medicine* 2002;77:50–4.
8. Musick DW. A conceptual model for program evaluation in graduate medical education. *Academic Medicine* 2006;81:759–65.
9. Kassebaum DG. The measurement of outcomes in the assessment of educational program effectiveness. *Academic Medicine* 1990;65:293–6.
10. Holmboe ES, Rodak W, Mills G, McFarlane MJ, Schultz HJ. Outcomes-based evaluation in resident education: Creating systems and structured portfolios. *American Journal of Medicine* 2006;708–14.
11. Gage NL. *Hard gains in the soft sciences.* Bloomington, IN: CEDR, Phi Delta Kappa, 1985.
12. Durning S, Pangaro L, Denton GD, et al. Inter-site consistency as a measurement of programmatic evaluation. *Academic Medicine* 2003:78:S36–8.
13. Roop SA, Pangaro L. Effect of clinical teaching on student performance during a medicine clerkship. *American Journal of Medicine* 2001;110:205–10.

14. Cox S, DaRosa DA, Margo K, Morgenstern BZ, Pangaro LN, Sierles FA. Chapter 7: Evaluation of the clerkship: Clinical teachers and program. In RME Fincher (Ed.), *Guidebook for clerkship directors* (3rd ed.). Omaha, NE: Alliance for Clinical Education, 2005. pp. 251–291.
15. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: From methods to programmes. *Medical Education* 2005;39:309–17.
16. Hejduk G, Kahn N, Ostergaard D. Twenty years of consulting excellence: The residency assistance program. *Family Medicine* 1997; 29:696–700.
17. Colliver JA. Call for greater emphasis on effect-size measures in published articles in *Teaching and Learning in Medicine*. *Teaching and Learning in Medicine* 2002;14:206–10.

Final revision received on January 22, 2007