Chapter 7: Evaluation of the Clerkship: Clinical Teachers and Program

Debra K. Litzelman, MD and Judy A. Shea, PhD, Lead Authors

Jennifer R. Kogan, MD and Paula S. Wales, EdD, Co-Authors

Introduction

This chapter focuses on evaluation of clinical teachers and programs. We will discuss four interrelated topics: (1) practical issues concerning evaluation; (2) sources of and methods for evaluating clinical teachers; (3) multiple viewpoints for constructing program evaluation data (within a clerkship, within an institution, and external to the institution) and processes for improving education programs; and (4) the importance of conducting research about evaluation. A comprehensive evaluation process is practical, and includes careful evaluations of faculty (and other clinical preceptors/teachers, such as residents), the individual events that occur within a clerkship (such as lectures and conferences), and the overall clerkship experience (such as organization and relevance). These data can be collected and used systematically to continuously improve the quality of a clerkship. Rigorous and/or innovative methods used concurrently will expand the knowledge base about evaluation.

Linking Evaluation and Curriculum Objectives

Evaluation of clinical teachers and the clerkship program needs to be tied logically to the stated clerkship objectives and goals (see Chapter 3, Creating a Clerkship Curriculum). The scope of learning, and thus teaching, should include both cognitive and non-cognitive skills, realizing details are certain to vary from site to site. National organizations provide leadership in defining the curriculum. These include the Association of American Medical Colleges (AAMC) and national organizations representing clerkship directors (i.e., Clerkship Directors of Internal Medicine). Medical schools such as Brown University (http://bms.brown.edu/omfa/handbk/md2000.html), Indiana University (http://meded.iusm.iu.edu/Programs/ComptCurriculum.htm), the University of Massachusetts (www.umassmed.edu/som/uploads/competencies.pdf), and the University of North Carolina (www.med.unc.edu/curriculum/Description/Where%20core%20competencies%20are%20addressed%2020100304.pdf), have addressed specific learning objectives by developing competency-based curricula.

The Accreditation Council for Graduate Medical Education (ACGME) mandated competency-based curricula for residency programs in 2001 (www.acqme.org/OutcomeD), paralleling the action at the undergraduate level. Defining and instituting objectives such as those found with competency-based curricula have obvious implications for student assessment (see Chapter 6, Evaluation and Grading of Students).

Program Evaluation Should Include Formal and Informal Curriculum

Program evaluation needs to address both the formal and the informal curriculum. The formal undergraduate medical education curriculum includes basic and clinical science knowledge; clinical skills; behavioral and social science knowledge; and knowledge, skills, and abilities
needed to practice the art and science of medicine. These elements are identified in the Medical School Objectives Program (MSOP) initiative and are typically defined by undergraduate and graduate medical educators as competencies. The chapter focuses mainly on evaluation of the formal curriculum.

However, a comprehensive clerkship evaluation also includes the informal or “hidden” curriculum, which refers to the experiences of medical students outside of the formal curriculum. These informal experiences have a profound and enduring impact on the development of a medical trainee's professional identity. Thus far, a few medical schools have used in-depth qualitative inquiry methods to assess their informal curriculum as a complement to their ongoing formal curricular reform efforts. Recently, rigorous psychometric methods were used to develop a tool to evaluate the multiple domains of patient-centeredness of the hidden curriculum in medical schools. It is reasonable to expect that aspects of the informal curriculum will become a routine part of clinical faculty and curricular evaluation in the next few years.

**Practical Issues Concerning Evaluation**

A clerkship director must carefully consider several practical issues concerning evaluation of clinical teachers and programmatic elements. Practical considerations include issues related to (1) the process of evaluation, (2) maintaining a secure evaluation system, and (3) reporting and reviewing the evaluation data.

**Process of Evaluation**

**Standardization**

The process of evaluation needs to be as standardized as possible. Ideally, all educational events that occur in the clerkship are evaluated, including lectures, conferences, seminars, and rounds. When students evaluate clinical faculty, the minimal duration of contact between an evaluator (i.e., student) and the one being evaluated (i.e., faculty member or resident) should be specified. For example, one week of contact between a student and faculty member may be considered enough for an evaluation to be submitted. The process of evaluation should be built into the curriculum and students should be taught from the outset that evaluation is a bi-directional process. Faculty and students should both receive and provide evaluation data that are helpful, timely, and professional.

**When and How to Collect Data**

A decision needs to be made about whether to collect data with a “just-in-time” perspective, making evaluations available as teaching events occur, or wait until the end of the clerkship and ask students to reflect back upon the entire experience. The method or tool used to collect evaluation data will likely help define the process. Paper evaluations are still used frequently, but Web-based options and programs for computers and personal digital assistants (PDAs) are available and becoming increasingly user-friendly. Regardless of the tools, the evaluation system may be designed so that evaluators cannot complete an evaluation until the end of a contact period (i.e., students cannot fill out evaluations of faculty before they have met them!). Accuracy is important. Therefore, any changes to syllabi including times, titles, and teachers of sessions need to be updated and reflected in the evaluation. If multiple students and/or teachers are being evaluated at one time, it is helpful to affix their pictures to the respective evaluation forms. Electronic evaluation systems can be programmed to merge digital images to evaluation instruments. It is also useful to provide evaluators with forms preprinted with the names of those
individuals being evaluated. Preprinting helps to prevent errors of omission, commission, and selection bias, which can happen when names are selected from lists or drop boxes.

Systems can be developed that ensure that evaluations are completed promptly. A high response rate in evaluation is important to make reproducible and valid assertions about the clerkship and its teachers. Evaluations completed at the time of a final examination yield high response rates, but ratings are generally lower than those made during a session separate from the examination. Having the teacher in the room when students are completing the clinical teaching evaluation forms can bias results positively. Therefore, it may be preferable to have a proctor or clerkship assistant, rather than the teacher, at the evaluation session to collect evaluation forms or to use electronic evaluation systems. Separate evaluation sessions are particularly challenging with multi-site clerkships.

Signed vs. Confidential vs. Anonymous Evaluations
One of the most important practical concerns is the decision about whether to have signed, confidential, or anonymous evaluations. The issue is as follows:

- If the form is signed, student identity is openly linked to the completed evaluations.
- If evaluations are confidential, student identity is known to a central evaluation unit but not to individual instructors or unit leaders
- If evaluations are anonymous, the student identity cannot be linked in any way to completed evaluations.

Pros and Cons
Students may craft more carefully worded, constructive feedback if they are signing evaluations. However, signed evaluations may not be ideal for students evaluating clinical teachers, given the mismatch in authority and possible conflicts related to grading. Many educators believe that students are unlikely to submit frank assessments of teachers or clerkship content if they believe they will be identified. Clerkship directors can create mechanisms to ensure student confidentiality and share with students how their identity is kept concealed. In courses with small numbers of students, anonymous evaluation systems may be virtually impossible to achieve. For clerkships with many students, it may be sufficient to wait to review the course evaluation data until after clerkship grades are determined. If students identify a serious or egregious problem with a clerkship site or preceptor, confidential evaluation systems allow a neutral person to return to the student for more extensive follow-up. In the rare instances where a student uses the evaluation system inappropriately, professionalism issues can be addressed with a confidential system.

Maintaining student confidentiality concerning their evaluations of faculty or resident teaching effectiveness is important. Clerkship directors can establish criteria for the minimum number of evaluations a faculty member must obtain before his/her teaching performance data are formally reviewed, ideally based on generalizability analyses (see section on Research about Evaluation-precision of scores/number of raters in this chapter) to help assure reproducible assessment. Achieving a minimum number can be particularly challenging for formative feedback, especially for the faculty member who only teaches once or twice per year and may only work with one or two students annually. For summative feedback reports, it is useful to compile students’ comments over an extended period and then transcribe them as aggregate data. Clerkship directors can help students maintain their confidentiality by advising them to avoid providing any potentially decodable information, such as the month or year of their rotation, in the open-ended comments section.
Maintaining a Secure Evaluation System

Evaluations of clinical teachers and programmatic elements should be maintained in a secure manner. Consult information technologists to guarantee that the evaluation system and data are safeguarded when electronic evaluation systems are used. Clerkship directors may want to define unambiguous criteria and written policies regarding who has access to clerkship data and reports and when these data will be made available. Using electronic evaluation systems, quantitative data and comments are easily compiled, aggregated, and can be made available on Web sites. Code can be written to limit "read-rights" and the timing of when data can be “opened for viewing” to protect student anonymity. Interested parties such as deans, program chairs, members of award committees, and promotion committees will quickly appreciate good quality data on a clerkship and its teachers. Paper and/or electronic data back-up systems also are needed.

Reporting and Reviewing Evaluation Data

The data need to be organized meaningfully once clerkship evaluations of clinical teachers and program elements are submitted. This is usually accomplished by producing summary reports. Summary reports allow the clerkship director to view the clerkship as a whole, determine curricular strengths (which might promote maintaining the status quo) and weaknesses (which can become the impetus for change). Summary reports might contain data from a single clerkship or data from several clerkships.

Individual Clerkship Reports
Individual clerkship reports might include summaries of faculty teaching effectiveness, overall ratings of clerkship experiences, impressions of specific clerkship events (i.e., lectures, physical diagnosis rounds), and free text comments. This information can be interpreted much more easily when the report is organized thoughtfully. Reports can be organized by different times in the academic year or by comparing one year to prior academic years to determine trends in ratings. Data might be organized by clerkship site to determine strengths and weakness of different clinical assignments.

Multiple Clerkship or Multiple Year Reports
The medical school might consider creating reports that compare evaluations from different clerkships within a single year or across several years. The latter strategy can be helpful when trying to determine if changes in course evaluations from one year to the next reflect changes in the quality of the curriculum and its teachers or the rating characteristics of a cohort of students. A clerkship director who fails to look at the course evaluations across clerkships may make errors in data interpretation. For example, some medical school classes might be critical evaluators ("hawks"), while other student cohorts may rate events more highly ("doves").

Creating Reports that Provide Helpful Feedback to Faculty
The method used for collating, summarizing, and generating feedback reports on clinical teaching effectiveness can aid in interpreting student ratings. Student feedback on teaching effectiveness is useful if collated for clinical instructors as either visual or numerical summary reports. Norm-referenced standards can also guide teachers by helping them to interpret their ratings relative to other teachers of similar rank or level of teaching experience. This is particularly important, because teaching effectiveness questionnaire data that use rating items tend to show a skewed distribution. Placing the positive end of the response scale on the right side of the scale mitigates this “generosity” effect somewhat.
Summaries of recent ratings compared to prior ratings for each instructor help clinical teachers assess changes in students’ perceptions of their teaching. This type of breakdown may also help course directors identify performance patterns before nominating teachers for awards or recommending that certain teachers be removed from the teacher pool. Such data also may be useful for assessing the effectiveness of faculty development programs and for selecting the most critical focal areas. However, the number of ratings for individual faculty are often quite small and thus estimates fluctuate from year to year (see section on Research about Evaluation-precision of scores/number of raters in this chapter).

**Free Text Narrative Evaluation Comments**

Free text comments from students about their clinical teachers' effectiveness can be an invaluable source of data. Quantitative scores on rating scales and the qualitative assessment of comments about teaching effectiveness produce similar faculty rankings. However, the qualitative information, used alone or to complement student ratings, can provide detailed information about an individual faculty member's strengths and weaknesses. Data generated from open-ended comments can also be reviewed more systematically, and then classified using content analytic methods. For example, Ullian used content analysis of teaching effectiveness comments to define the role and characteristics of clinical teachers, as assessed by trainees at varying postgraduate levels. Differences were noted in the comments made by first- and third-year residents. Qualitative analysis software packages (e.g., NVivo, NUDIST, and Atlas.tiScholari) are available for conducting a more systematic approach to creating and sorting content themes from qualitative data. However, software packages require training, purchase and site license costs, and time, especially for novice users.

The clerkship director may want to think critically about the organization of the summary report and decide when the reports should be created and reviewed. Reviews that are too frequent may exaggerate the importance of a small number of ratings, whereas reviews that are too infrequent may fail to identify potentially important weaknesses in the clerkship. Independent of the number of reports generated in a given year, the reports should be reviewed when there are logical opportunities to institute change in the clerkship. For example, an ideal time to review a clerkship report might be several weeks before a new academic year, so there is enough time to make necessary modifications.

**Summary**

More specific methods of evaluating teachers or programmatic elements will be discussed in the following sections. Sample instruments and summary reports are provided as examples. Clerkship directors and coordinators can benefit from knowing about and using a variety of evaluation methods and instruments that have undergone rigorous psychometric testing. Each institution will most likely adapt certain evaluation methods and select evaluation tools that are most relevant to its unique setting and educational challenges.
### Recommendations Regarding Practical Issues Concerning Evaluation

<table>
<thead>
<tr>
<th>Standardize your process of evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Decide who/what to evaluate</td>
</tr>
<tr>
<td>• Decide how often to evaluate</td>
</tr>
<tr>
<td>• Create a paper or electronic system</td>
</tr>
<tr>
<td>• Decide whether evaluations will be signed/confidential/anonymous</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Maintain a secure evaluation system</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Decide who and when individuals/administrative units will have access to data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Standardize reporting and reviewing of evaluation data</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Decide what and when data will be reported</td>
</tr>
<tr>
<td>• Report comparative and trend data to help identify performance patterns</td>
</tr>
</tbody>
</table>

### Evaluating Clinical Teachers

Multiple methods can be used to evaluate clinical teaching effectiveness. We will discuss student ratings of inpatient clinical and ambulatory teachers, including open-ended comments, peer evaluation, self-assessment, standardized students, and learner outcomes. The methods of collection and data collection instruments impact the reliability and validity of data about clinical teaching effectiveness. We will briefly discuss instruments and central methodological issues for each method of evaluation.

### Student Ratings of Inpatient Clinical Teaching Effectiveness

Student ratings remain the most commonly used method for evaluating teaching performance. The method is logical, convenient, and economical.

**Validity and Reliability of Student Ratings**
(See Chapter 6, Evaluation and Grading of Students, for definitions; also section on Research on Evaluation in this chapter)

Although ratings are not without their problems, abundant evidence in the general education and medical education literature indicates that student ratings of teachers are reliable and valid. Student ratings are correlated significantly with ratings from recent graduates of training programs, ratings from graduates several years removed from contact with their clinical instructors, peer assessment measures, and self-assessment measures. Most importantly, student ratings of teaching effectiveness can predict certain learner outcomes.

Medical students spend more time with medical residents than with their attending physicians, especially in the hospital setting. Residents are instrumental in students' education and should be included in the evaluation process. Resident teachers appear to have teaching strengths that complement those of faculty teachers in certain areas. Observational studies show that residents play an important role in demonstrating patient evaluation skills at the bedside, giving brief lectures, asking questions, giving feedback, referring students to the literature, and demonstrating techniques and procedures.

Third-year medical students can discriminate the relative strengths of teachers at various levels. In one study, students rated attending physicians more highly on cognitive and experiential...
characteristics compared to residents, while they rated residents higher on interpersonal skills. In another study, students gave higher scores to residents for "control of session" and lower ratings on "promotion of self-directed learning," relative to attendings. This suggests that resident teachers are more effective at pacing and focusing teaching opportunities while on busy inpatient services, but less experienced at modeling their approaches to gaining new knowledge, compared to attending teachers.

**Instruments for Student Assessment of Clinical Teaching Effectiveness**

Multiple instruments are available that assess effectiveness of clinical teaching. Beckman et al. published a systematic review of the psychometric characteristics of published instruments for measuring clinical teaching effectiveness.

**Construct Validity of Teaching Assessment Instruments**

The review provides evidence that supports the construct validity of teaching assessment scales from studies (see Table 2 of the Beckman review, page 973). The authors identified 14 recurring domains of teaching. Seven of the 14 domains are categories from the well-designed and widely used Stanford Faculty Development Program (SFDP) for Clinical Teachers educational framework:

- establishing a positive learning climate
- control of the teaching session
- communication of goals
- enhancing understanding and retention
- evaluation
- feedback
- self-directed learning.

The two most common recurring domains were more general categories, clinical-teaching and interpersonal skills.

The remaining domains, listed in order of frequency, were:

- global teaching effectiveness
- availability
- punctuality
- ability to delegate
- ability to motivate.

**Convergent Validity**

Shorter instruments have been assessed for convergent validity compared to more comprehensive teaching evaluation instruments. For example, the brief University of Michigan (UM) Global Rating Scale, when compared to seven constructs on the SFDP showed strong convergent validity (see a copy of a modified UM Global Rating Scale in Appendix A and the SFDP in Appendix B). Brief instruments have the advantage of ease and relatively low cost of administration, but often only assess interpersonal and/or global clinical teaching domains. The results may adequately assess the proficiency of clinical teachers. However, they do not provide specific feedback to help individual teachers improve their teaching.
**Student Ratings of Outpatient Clinical Teaching Effectiveness**

Effective teaching skills differ depending on the teaching venues and levels of learners (see Table 1 of the Beckman review, page 972).\(^\text{20}\) Well-developed instruments for evaluating clinical teachers in the ambulatory setting differ from those for inpatient teaching, even when based on the same theoretical educational framework.

**Instruments for Student Assessment of Outpatient Clinical Teaching Effectiveness**

There are several options for evaluating teaching effectiveness for students who have limited contact with teachers in outpatient clinics.

**SFPD**
Three domains from the SFDP -- learning climate, communicating goals, and stimulating self-directed learning -- had high internal consistency and explained 67% of the variance in the clinical teaching evaluations.\(^\text{34}\)

**MedIQ**
The MedIQ instrument, developed for evaluating instructional quality in the ambulatory setting, includes four domains for assessing outpatient teaching:

- preceptor activities
- learning environment
- learning opportunities
- learner involvement.\(^\text{35, 36}\)

Kernan et al.\(^\text{37}\) developed a Teaching Encounter Card that has demonstrated accuracy in detecting nine outpatient teaching activities considered important for successful outpatient precepting. Jackson et al.\(^\text{38}\) developed the Teacher and Learner Interactive Assessment System (TeLIAS) to assess qualitative audiotapes of teacher-learner speech, providing a more detailed evaluation of outpatient teaching.

**Peer Assessment of Clinical Teaching Effectiveness**

Peer review is most commonly defined as the process of having a colleague observe a fellow teacher during classroom or clinical teaching.

**Validity and Reliability of Peer Assessment**

Data from classroom peer observations have produced reasonably reliable data for college teachers on ratings of their teaching effectiveness.\(^\text{39}\) The reliability and validity of peer assessment of clinical educators have not been studied extensively. However, evaluators in a general internal medicine hospital service showed significant agreement.\(^\text{40}\) Another study demonstrated that peer evaluations resulted in higher inter-rater and internal reliability ratings than residents' ratings for assessing bedside teaching.\(^\text{20}\)

**Types and Uses of Peer Assessment**

**Direct Observation**

Peer assessment is valuable to clerkship directors to document teaching improvement over time or the effectiveness of newly acquired teaching skills, in addition to summative purposes. Both the faculty member being reviewed and the faculty observer find peer review of inpatient
teaching valuable.20 Barriers to more regular use of peer assessment include faculty time constraints, consensus about the "qualifications" of peer reviewers, the appropriate frequency for conducting peer review, and mistrust about how the resulting reports are used. Despite these concerns, department chairs of medicine and chairs of promotion and tenure committees consider peer evaluation of teaching to be an important component of promotion and tenure dossiers.41, 42 Including peer review documentation in promotion packets can be valuable for faculty who emphasize the scholarship of teaching as an area of excellence. In one study, peer review documentation had a significantly positive impact on promotion of faculty who included peer evaluations compared to those who did not.39

Review of Videotapes
Direct observation is only one common method of peer observation clerkship directors should consider.43 Peer review and rating of videotapes of live teaching encounters may be less intrusive to learners and more comfortable for teachers. However, compared to direct observation, peer review of taped sessions has limitations, including poor quality of some tapes (e.g., missed nonverbal cues and inaudible communications) and small number of tapes reviewed for a given faculty member. The psychometric limitations of a small number of raters apply to peer as well as student assessment.

Instruments for Peer Assessment of Clinical Teaching Effectiveness
A literature review revealed only one study evaluating the reliability of a tool for peer assessment of clinical teaching.40 The Mayo Teaching Evaluation Form (MTEF) (see Table 2, page 133),40 based on the SFDP seven-category educational framework, has very good overall internal consistency for peer review of inpatient clinical teachers on general internal medicine services. The MTEF domains with the highest internal consistency were learning climate, communication of goals, evaluation, and self-directed learning. Forms for peer assessment of clinical teaching and classroom teaching based on Irby's clinical teaching model44 have been published, but have not yet undergone psychometric testing. At the Imperial College of London's Faculty of Medicine, a similar, internally developed clinical teaching observation record sheet (www.imperial.ac.uk/educationaldevelopment/opportunities/openlearning/peerobservation.htm) has been used for peer assessment.45

Self-Assessment of Clinical Teaching Effectiveness
Self-assessment of clinical teaching effectiveness involves reflection, interpretation, and appraisal of one's own skills and impact on learners. Self-assessment can be valuable for teachers and course administrators.

Validity and Reliability of Self-assessment
Numerous studies have compared teachers' self-assessment ratings with those of students. Correlations between student and teacher self-evaluation in areas relating to "preparation" and "meeting objectives" are particularly high, and students and teachers ranked strengths and weaknesses similarly.46, 47 Students' ratings of clinical teaching effectiveness did not correlate well with residents' self-assessment of their teaching skills.48 Similarly, residents' ratings of surgical educators compared to the surgeons' self-assessment of teaching differed significantly in the majority of categories evaluated. Faculty who chose not to complete self-assessment forms had the lowest ratings from residents.49
Uses of Self-assessment
Discrepancies between student and instructor self-ratings can provide teachers with important complementary information and motivate changes in teaching. Motivation to change teaching may be related to the direction and degree of the discrepancies, the teacher’s perceived importance of the teaching behavior, and the teacher’s perceived self-efficacy and self-concept. Instructors have been found to improve their teaching ratings in areas where student evaluations are lower than their self-evaluations. The complex factors motivating change in teaching when there are discrepancies between student ratings and self-assessments need further investigation in the clerkship setting.

Faculty Development and Self-assessment
Self-assessment may be impacted by the faculty development programs and methods of collecting self-assessment data. Clinical teachers' self-assessment of their teaching skills may be significantly different after taking faculty development courses on clinical teaching effectiveness than before. Compared to pre-course assessment of their skills, retrospective pre-course ratings tend to be lower, perhaps reflecting a better understanding of the dimensions being measured.

Instruments for Self-assessment of Clinical Teaching Effectiveness
Tools for self-assessment of clinical teaching skills have been published, based on well-accepted educational frameworks. These self-assessment tools are based on the SFDP seven-category educational framework and Irby's well-researched clinical teaching model, respectively. However, these self-assessment instruments have not yet been subjected to rigorous psychometric analyses.

Standardized Students' Evaluation of Clinical Teaching Effectiveness
Several educator-investigators have incorporated "standardized students" into faculty development programs in an effort to develop more objective evaluative measures.

Objective Structured Teaching Exercises (OSTE)
The concepts and principles used for objective structured clinical examinations (OSCEs) have been applied to assessment of clinical teaching. Trained "standardized students" have been included in OSTE as part of faculty development courses focused on improving and assessing clinical teaching skills of faculty and residents. The standardized students are trained to simulate clinical teaching challenges, to complete rating scales on the participants' use of effective and ineffective teaching strategies, and to provide feedback to the participant at the end of the simulated encounter with or without a facilitator present. The resources needed for training standardized students, and reimbursement for training and testing time are comparable to those for standardized patients.

Validity and Reliability of Instruments for Standardized Students' Assessment of Clinical Teaching Effectiveness
Morrison et al. documented the reliability and validity of scores on OSTE rating scales, adapted from the SFDP26 instrument, to assess improvement in residents' teaching skills. Similarly, a tool to evaluate clinical instructors' ability to address professional teaching issues during an OSTE has been shown to be reliable, especially for low inference behaviors. Zabar developed and tested an instrument to evaluate the competency of residents as teachers through OSTEs.
The four domains evaluated were:

- establishing rapport with the learner
- assessing the learner's needs
- demonstrating instructional skills
- fund of knowledge.

Together these domains explained 79% of the variance. Compared to a written test to evaluate teacher's feedback skills, an OSTE rating provided more sensitive measures of changes in faculty's teaching skills following a faculty development program.66

**Learner Outcomes as a Measure of Clinical Teaching Effectiveness**

The most critical measure of clinical teaching effectiveness is a teacher's impact on learner outcomes. Student ratings of teaching effectiveness are associated with what students believe they have learned and objective measures of achievement.

**Relationship between Ratings of Teaching Effectiveness and Learner Performance**

Research on education of health professionals revealed significant correlations between specific teaching behaviors and learner outcomes.67 Associations between gains in standardized test scores and specific teaching behaviors, such as "use of objectives," "provision of clear performance related feedback," and "asking good questions" have also been reported.68

Medical student ratings of clinical teaching during a neurology block of a clinical skills course were highly related to OSCE scores.69 Investigators compellingly documented the association between ratings of teaching effectiveness and the learner outcomes on internal medicine clerkships. At two different institutions, there were statistically significant associations between teaching effectiveness scores based on student ratings and students' performance on the end-of-medicine clerkship National Board of Medical Examiners (NBME) subject examination after adjusting for students' baseline standardized exam scores.21-23 High-quality teaching by surgical faculty has also been associated positively with NBME surgery subject exam scores and end-of-surgery clerkship OSCE scores.24

The ratings of resident teaching quality are associated with learner outcomes.22, 23 Griffith22 found that resident teachers had an impact on third-year students' NBME Subject Examination performance and that interns influenced students' performance on a clerkship practical examination that emphasized clinical skills and interpretation of common medical tests. The negative impact of the most poorly rated housestaff tended to have a greater influence than the positive impact of the highest rated residents.

In a unique clinical problem-solving experiment, residents and chief residents influenced students more than experienced physicians.70 The rationale for this finding is thought to be that students' medical reasoning skills and ways of storing knowledge in memory are more congruent with residents, who use more "elaborated," or step-by-step, discourse than attendings. Attendings, who more often use a "compiled" structure, may fail to provide their novice learners with the intermediate, step-wise thought processes leading to medical decision-making.71 Thus, the resident teachers may demonstrate medical reasoning skills that are closer to those of the students than the attendings.72 Such studies can provide medical education administrators with valuable information on how to assign teachers to various teaching venues to optimize student learning.
Summary of Clinical Teacher Evaluations

A variety of methods for evaluating clinical instructors have been described in the literature. Student ratings with open-ended comments are the main source of data on clinical teaching effectiveness. Student ratings are an economical source of reliable and valid information. Peer and self-assessments and, more recently, ratings from standardized students during OSTE examinations, provide important complementary perspectives about teaching quality.

Significant progress has been made in the last 5 years in the development of reliable and valid tools for measuring clinical teaching effectiveness. However, most research has been conducted with trainees in internal medicine and only a small number of studies include family medicine, emergency medicine,

Several well-tested instruments are available for student evaluation of clinical teachers in various teaching venues, including OSTEs.

Our literature review revealed only one peer-review assessment tool that had undergone psychometric assessment and little or no psychometric information on self-assessment tools for the evaluation of clinician-educators. Clearly, psychometric evaluation of new or existing tools for peer and self-assessment is needed. Snell et al. emphasized the importance of moving toward a more comprehensive 360 degree evaluation of clinical teachers, including not only learners, peer, and self evaluations, but also perspectives from patients, institutional administrators, and payers in the health care systems.

<table>
<thead>
<tr>
<th>Recommendations for Evaluating Clinical Teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Employ a variety of methods for evaluating teachers</strong></td>
</tr>
<tr>
<td>• Student ratings</td>
</tr>
<tr>
<td>• Open-ended comments</td>
</tr>
<tr>
<td>• Peer and self assessment</td>
</tr>
<tr>
<td>• Standardized students</td>
</tr>
</tbody>
</table>

| Use reliable, validated instruments when available |

**Employ a tested framework to facilitate comprehensive evaluation**

Program Evaluation

There are several forms of programmatic evaluations of clerkships:

- program evaluation internal to the clerkship (i.e., data collected on multiple aspects of the clerkship, including didactic and clinical experiences)
- program evaluation internal to the organization (i.e., data across clerkships within an institution)
- program evaluation external to the organization, which can be useful for benchmarking purposes.

Each of these perspectives is considered in this section, followed by a discussion of a process for using program evaluation data to improve education programs.
Program Evaluation Internal to the Clerkship

Effectiveness of the clerkship's programmatic elements can be determined in multiple ways, including:

- collection of student ratings of their clerkship experience
- review of students' clinical encounters
- assessment of student competency (see Chapter 3).

Review of specific programmatic and global clerkship assessments should include clerkship ratings as a whole and aggregated information for each clinical site when multiple clinical sites are used. Inter-site consistency is an important method of program evaluation if the data are corrected for baseline attributes of students.78

Ratings of Specific Programmatic Components
Clerkship directors need to determine the content/wording of the questions to be asked on rating forms as well as deciding what to evaluate. Questions might ask about whether lecture objectives were achieved or the quality of teaching materials. The qualities of the clerkship events (e.g., lecture, bedside teaching) that are evaluated should remain constant across events to facilitate aggregation of clerkship data into useful reports. Space for free text comments provides useful information in addition to ratings. Balance the desire for thorough evaluation with the meaningfulness of the data and students’ compliance with completing the evaluation. Students are less likely to complete evaluations carefully if there are numerous specific questions.

Global End-of-Clerkship Assessment
Students’ evaluations of the entire clerkship experience at the end of the rotation can be valuable, in addition to asking them to rate specific programmatic components. In end-of-clerkship evaluations, students can rate the overall quality of the clinical rotation using a rating scale, open-ended comments, or both. See Appendix C for an example of item content and format from an end-of-course evaluation form.

End-of-clerkship evaluations also can include summative ratings of recurring clerkship events. For example, physical diagnosis rounds might happen weekly during a clerkship. While each session might be rated individually, clerkship directors might want to know about how students valued the experience as a whole. Examples of such summative questions are available in Appendix D, which also shows data by clinical site.

Review of Students' Clinical Encounters
Student Logs
Student logs can be very helpful in evaluation of the clerkship, especially when trying to assess the types of patients that students see and the patients' disease severity.79 Student logs also can be used to document procedures they observed or performed (See also Chapter 4, Technology in Clerkship Education).

Review of students' logs of clinical encounters with patients, specifically the diagnoses of patients seen is more important than in the past because of the Liaison Committee on Medical Education (LCME) standard, ED-2. ED-2 states, "the objectives for clinical education must include quantified criteria for the types of patients (real or simulated), the level of student responsibility, and the appropriate clinical settings needed for the objectives to be met. Each
course or clerkship that requires interaction with real or simulated patients should specify the numbers and kinds of patients that students must see to achieve the objectives of the learning experience. It is not sufficient simply to supply the number of patients students will work up in the inpatient and outpatient setting. The school should specify, for those courses and clerkships, the major disease states/conditions that students are all expected to encounter. They should also specify the extent of student interaction with patients and the venue(s) in which the interactions will occur. A corollary requirement of this standard is that courses and clerkships will monitor and verify, by appropriate means, the number and variety of patient encounters in which students participate, so that adjustments can be made to ensure that all students have the desired clinical experiences [80] (See also Chapter 15, The Clerkship Director and the Accreditation Process).

Some medical schools have developed systems to track students’ clinical experiences across clerkships. Products that can be customized to meet the needs of a medical school or clerkship include ClinicalWebLog software with a Web and PDA application from the Uniformed Services University [available gratis for modification and use] and the PDA Project used at the Southern Illinois University [available gratis for modification and use]. Both these systems provide cross-departmental tracking across all 4 years of medical school. They include fields for capturing LCME-required information while concurrently meeting HIPPA regulations.

Individual clerkship directors are responsible for identifying the numbers and kinds of patients that students must see and need to review patient logs to make certain that the pre-determined goals are met. Therefore, clerkship directors should make certain that patient encounter data can be aggregated and reviewed in a meaningful way to determine the breadth and depth of patient exposure and exposure to clinical procedures. For example, if a goal of a clerkship is to manage a patient with human immunodeficiency virus, and the majority of students do not care for, or observe care given to a HIV positive patient, the clerkship director will need to identify other experiences (either clinical or simulated) to give students an equivalent experience.

Patient logs become even more important in multi-site rotations where the type and severity of patients’ illnesses may vary substantially across sites [81]. Clerkship directors will need to evaluate whether patient exposure is markedly different across clinical sites and, if so, how this will be handled so that curricular goals can be achieved at all sites for all patients (See also Chapter 9, Directing a Clerkship Across Geographically Separated Sites). Finally, Clerkship directors must verify the reliability of log entries, because false-positive patient encounters and patient problems have been reported [82].

**Surveys**

A less-detailed view of clinical encounters might come from surveys. For example, clerkship directors might survey graduates of their school to determine the types of patients students should see and procedures they should be proficient in performing as a means of assessing and modifying curricular objectives [83]. Surveying graduates who have entered generalist fields may be particularly useful. The data allow clerkship directors to review the clerkship learning objectives to make certain that they are relevant to the future needs of practicing physicians.

**Assessment of Student Competency as a Means of Program Evaluation**

Assessment of students’ competencies (e.g., acquisition of knowledge and clinical skills) can be used to evaluate whether the clerkship curriculum results in students’ achieving curricular goals.
Knowledge Assessment
Students' acquisition of knowledge can be assessed through examinations such as the NBME Subject Examination or locally written "in-house" examinations. If performance on such examinations will be used to evaluate the clerkship curriculum, it is important to know if exam content parallels clerkship content. For example, the Clerkship Directors in Internal Medicine and Society of General Internal Medicine developed a core curriculum for the internal medicine core clerkship. A content analysis of the questions on the NBME Medicine Subject exam was performed to determine if questions were representative of the recommended core curriculum. Congruence between curricular and examination content is critical to ensure valid assessment for individual students and for program evaluation. Performance on national exams also can be used to measure student competency when alternative clerkship structures (i.e., multidisciplinary clerkships or variations in clerkship length) are used.

Clinical Skills Assessment
It is tempting to simply look at students' performance on written examinations as proof of learning, but written examinations of knowledge often fail to predict students' ability to apply their knowledge in clinical settings. Therefore, students' performance on clinical skills assessments give a clerkship direction additional insight into the success of a clerkship curriculum. Aggregated student performance on Objective Structured Clinical Examinations (OSCE) and other standardized patient examinations can be used to assess whether students have acquired the clinical skills taught in the clerkship. Patient simulators also can be used to demonstrate acquisition of clinical skills.

Formative clinical skills assessments can also be aggregated to examine the clerkship curriculum. For example, aggregated performance on tools used to provide students with formative feedback about documentation skills, oral case presentation skills, or clinical encounters can demonstrate whether the students are achieving basic competency in such skills as communication, patient interviewing, and examination.

Program Evaluation Internal to the Organization (but External to the Clerkship)

Formative and Summative Uses of Program Evaluation
School of medicine administrators and clerkship directors often use formative program evaluation to improve program performance and help identify specific suggestions for improvement. However, program evaluation can also be used for summative purposes, including (1) judgments regarding success and efficacy of the program, (2) decisions regarding the allocation of resources, (3) influencing attitudes regarding the value of the curriculum, and (4) satisfying external requirements/mandates. When aggregated and reported for summative purposes, comparisons are typically made among multiple clerkships within an institution.

Using Existing Clerkship Data to Develop Comparative Multi-clerkship Reports
Program evaluation data across clerkships can be compared more easily when clerkship directors from different clerkship disciplines in an institution agree upon a core set of generic evaluation items for all of the core clerkships. An example is shown in Appendix E. The solid black line indicates the mean performance for all clerkships combined and the anchors reflect +/- one standard deviation from the mean. Appendix F contains a graphic grade distribution report for all clinical clerkships. Such aggregate reports allow a clerkship director to assess
whether there is grade inflation and/or inappropriately high failure rates on one clerkship compared to others.

Student performance on OSCEs can also be used for program evaluation. A sample OSCE score distribution is presented in Appendix G. One way to use this information for program evaluation is to compare students’ performance at different times in their curriculum, for instance during the third year and then again during the fourth year. Another is to compare performance across topic areas and/or across educational sites. Appendix H contains an example of OSCE performance data in the Self-Awareness, Self-Care, and Personal Growth competency and the Lifelong Learning competency by educational site.

**Developing an “Integration Ladder” to Assess a School’s Curriculum**

Organizational evaluation is not limited to a systematic, comparative, longitudinal analysis of the elements of traditional program evaluation data. Harden\(^9^0\) described an innovative approach using an 11-point "Integration Ladder" to evaluate a school’s curriculum. The ladder is a continuum, with the first anchor or rung labeled "isolation." Isolation occurs when departments or subjects are taught without attention to other courses or disciplines. The middle of the continuum, or the sixth rung in the ladder, is labeled "sharing" and is described as two disciplines agreeing to jointly plan and implement a complementary or overlapping program. The other anchor, or eleventh and highest rung in the ladder, is labeled "trans-disciplinary" and is described as the curriculum transcending the individual disciplines. One important measure of program evaluation is the degree to which the curriculum as a whole is integrated with other disciplines.

**Program Evaluation External to the Organization**

*Comprehensive Self-assessment and External Review*

Hagenfeldt and Lowry\(^9^1\) described a four-step evaluation model, involving external reviewers, which can be used in medical schools.

The Steps:

- **Step 1** is to collect baseline data, such as admissions criteria, curricular content, data analyses, and institutional resources.

- **Step 2** involves institutional self-assessment when learners, teachers, and administrators critically review all aspects of the school, particularly assessing the curriculum based on the stated educational objectives of the institution. This internal self-assessment should result ultimately in an action plan for the school.

- **Step 3** involves external reviewers examining data and providing objective analyses, resulting in a report back to the institution.

- **Step 4** is discussion of the results by all interested parties and creation of an action plan for the future.

*Less Comprehensive External Review*

Des Marchais and Bordage\(^9^2\) describe an external formative program evaluation whereby three external reviews were held over a 6-year period in an effort to continuously improve their new medical education curricular program in a formative way in preparation for their summative
LCME accreditation review (See also Chapter 15, Clerkship Director and the Accreditation Process).

Additional Sources of External Review Data

Association of American Medical Colleges’ (AAMC) Graduation Questionnaire

Reports or summaries published by external organizations are another source of external data. For example, AAMC conducts an annual survey, the AAMC Graduation Questionnaire (GQ). The GQ asks graduates about their medical school experience. One form of external program evaluation is to compare the performance of an individual medical school to the national cohort. Appendix I contains graphic representations of answers to questions on the AAMC GQ regarding clerkships. The AAMC GQ is a comprehensive questionnaire and a wealth of information is provided for each discipline through quantitative questions. Additionally, open-ended comments identifying the strengths and weaknesses of each school provide a wealth of program evaluation data. The AAMC sends the annual report to the dean of each medical school.

National Board of Medical Examiners (NBME) Subject Test Results

Reports from the NBME provide clerkship directors with information about student performance on NBME subject examinations across the United States, as well as mean performance by quarter.

Processes for Improving Education Programs

Medical schools often collect and analyze multiple data elements to address varying evaluation needs and requests. The evaluation process is optimized when it is theory-driven and data are collected, analyzed, and synthesized for students, individual clerkships, across clerkships, and across schools or institutions. There are numerous theories or frameworks that can be used to direct and organize the evaluation efforts in a comprehensive and meaningful way.

Three Models for Ongoing Evaluation

Dolmans, Wolfhagen, and Scherbier93 claim that medical schools seeking quality improvement must engage in an evaluation process that is systematic, structural (cyclical), and integrated into the culture of the school. The integration piece, which focuses on understanding and acceptance by members of the medical school culture to seek continuous improvement, is necessary to achieve improvement in the quality of medical education.

Berwick94 defined the central law of improvement as "every system is designed perfectly to achieve the results it achieves." He states that while variation in a system is inevitable, better or worse performance averaged over time cannot be achieved on demand. Berwick’s definition suggests that to improve performance and successfully adapt to the changes in educational environments, evaluation should be done on a systemic level.

One model for change, developed by Langley, Nolan, and Nolan95 for the business world, has been applied successfully in health care settings. It is applicable to medical education/clerkship evaluation and is based on continuous quality improvement (CQI) principles. The CQI model asks three basic questions: (1) what are we trying to accomplish? (2) how will we know if a change leads to improvement? and (3) what changes could we make that we think will result in improvement? This model, coupled with the Plan-Do-Study-Act (PDSA) cycle has been applied
in health professions educational settings to analyze educational systems and to improve their quality.96-100

**Continuous Quality Improvement (CQI) and Plan-Do-Study-Act (PDSA)**

The CQI and PDSA processes involve continuous comprehensive loops of collecting data and evaluating progress. Essentially, the purpose of this process is to continuously evaluate systems using the Shewhart Cycle for Learning and Improvement.96

Step 1 in the PDSA process is to **Plan** a change or a test aimed at improving quality.

Step 2 is to **Do** or carry out the change/test (preferably on a small scale)

Step 3 is to **Study** the results or the effects of the change by collecting and analyzing relevant data.

Step 4 is to **Act** on what has been learned by adopting or abandoning a change/test or deciding to run through the cycle again.

![Figure 1: The Plan-Do-Study-Act Cycle](image)

During the planning stage of the cycle, one might collect data in a variety of ways, including brainstorming, cause and effect diagrams, rating forms, check sheets, tree diagrams, pareto charts, flow charts, scatter diagrams, run charts, control charts, and histograms.101,102 A change is carried out and then studied/evaluated. Just as teachers evaluate students by a variety of methods and tools, clerkship directors evaluate the clerkship by a variety of methods and tools. Data can be gathered from clerkship directors, clinical teachers, discussion group facilitators, office personnel at sites where students rotate, students, and other stakeholders in the clerkship. The goal is to see if there has been improvement in accordance with the plan. Finally, a decision is made regarding how best to proceed or act. Additional data are collected and the iterative PDSA cycle begins again.

**Examples of CQI Used for Education Program and Clinical Instructor Improvement**

CQI and the PDSA cycle have been used successfully to improve medical education programs. For example, the Indiana University School of Medicine, with nine geographically separated campuses, was working to ensure educational equivalence and faculty involvement from all educational sites using audio-visual conferencing (polycom) technology as one mechanism for
facilitating improved communication. Faculty from the nine sites acknowledged benefits to using the polycom, but also noticed some barriers to fully participating in critical educational conversations. A group of faculty and students engaged in a CQI process and used the PDSA cycle to reduce some of the technological barriers to full participation in key educational decisions. Interventions were implemented, such as changing seating arrangements in the polycom rooms to facilitate seeing the committee chairman’s face at all distant sites and stopping frequently to allow all members at each site to respond to issues. These simple changes dramatically improved the perceived inclusiveness in discussions and decision making, identification of resources that could be shared, and consensus building regarding the adoption of educational policies.

The CQI model has also been used to improve teaching effectiveness. When determining the uses and distribution of personal teaching effectiveness summary reports, a CQI project uncovered an unanticipated complex interaction between teachers' baseline performance and the impact of feedback from student ratings on future teaching performance. Perhaps explained by theories of self-efficacy, teachers with higher baseline teaching ratings benefited the most from augmented feedback, while teachers with low baseline teaching ratings may actually experience a decrement in ratings of their performance. The results of this CQI project suggest that clerkship directors and other program administrators interested in positively impacting teaching performance may need to provide more personalized, supportive feedback to teachers with low teaching effectiveness ratings. More specifically, a "faculty-centered" approach incorporating teacher-learner collaboration in small-group consultation sessions may stimulate a cycle of teaching successes along with heightened self-efficacy.

**Summary of Program Evaluation**

Comprehensive program evaluation includes assessment that is internal to the clerkship, internal to the organization (across clerkships in the same school), and external to the organization (input from reviewers outside the institution). The authors believe there is benefit to organizing and reviewing all program evaluation data in a systematic fashion that promotes continuous quality improvement. Due to the iterative characteristic of the PDSA cycle, the process is inherently formative. Since medical educators are constantly striving to improve educational programs, the authors view this formative process as appropriate for program evaluation.

<table>
<thead>
<tr>
<th>Recommendations for Evaluating Programmatic Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Evaluate both the formal and informal curriculum</strong></td>
</tr>
<tr>
<td><strong>Employ a variety of methods for program evaluation</strong></td>
</tr>
<tr>
<td>• Student ratings of specific elements and global scores</td>
</tr>
<tr>
<td>• Review students’ logs of clinical encounters</td>
</tr>
<tr>
<td>• Assess students’ performance/competence</td>
</tr>
<tr>
<td><strong>Employ a comprehensive approach</strong></td>
</tr>
<tr>
<td>• Collect and review data internal to the clerkship</td>
</tr>
<tr>
<td>• Review cross-clerkship data within your institution</td>
</tr>
<tr>
<td>• Review cross-clerkship data between institutions (national data)</td>
</tr>
<tr>
<td>• Systematically organize and review data that promotes continuous quality improvement</td>
</tr>
</tbody>
</table>
Research about Evaluation

For most clerkship directors, selecting and implementing evaluation processes and tools provides important data about how students, residents, faculty, and programs perform. Beyond the primary focus on evaluation, there can be a secondary agenda: research about evaluation. (See also Chapter 14, Educational Scholarship). In general, studies that focus on evaluation aim to examine or support the robustness of the evaluation processes by asking questions concerning feasibility, and/or measurement properties such as reliability and validity.23, 24, 105 Importantly, the data for such studies are often already being collected because of the clerkship director's focus on evaluating programs and faculty; however, frequently it is not turned into research projects and manuscripts. Doing so would serve at least three purposes: first, enhance the scholarship productivity of clerkship directors and their collaborators; second, add to the literature about robust evaluation methods; and third, improve clerkship education.

One framework for conceptualizing the types of research that can be conducted about evaluation is to organize the questions or topics into the commonly used subheadings in the methods section of a research manuscript. Research projects can focus on the evaluators or those being evaluated (sample), the evaluation tools (instruments); the evaluation process (design); the data collection details (procedure) or interpretation of evaluation data (analysis). Examples of research projects that fit each question are provided below. Consult texts on educational research methods and design for more detail,106-108 as well as recent papers that address areas where more research about evaluation is needed. 109 Finally, it is important to consider thoughtfully the requirements of Institutional Review Boards and know when program and faculty evaluation efforts are legitimate research projects that require a priori approval.

Types of Research Framed in Five Domains of a Manuscript Methods Section

Sample
Creating studies that examine questions of sample selection in evaluation can add to the body of knowledge about sample recruitment and associated biases, strengths, and weaknesses. Within the context of a clerkship, individuals who could both be evaluated and provide evaluations include students, residents, and peer faculty. When 360° evaluations are used, allied health professionals and patients also may provide assessments.110 Within these broad categories of individuals, there are often subcategories (e.g., inpatient versus outpatient faculty; community-based versus hospital-based faculty; interns versus upper-level residents; volunteer samples of students versus all students). How the people who provide the evaluation information are selected could impact the evaluation data and its interpretation. For example, one study attempted to determine the impact of the time of year and level of learner on evaluations of faculty teaching effectiveness.111 Similarly, Steiner et al.112 looked at evaluations provided by different levels of residents, while Horowitz et al.113 compared faculty peer evaluations to student evaluations.

Instruments
Research about evaluation “tools” occurs on two levels. The more global level involves specification (and perhaps comparison) of various types of tools. For example, indicators of “program effectiveness” may include surveys of past students assessing the value of the clerkship, students’ NBME Subject Test scores, and faculty evaluations of the curriculum. The second level centers on deciding the exact wording of the question[s] that will be asked and the format in which they will be presented. These choices must be tied to the purpose of the
The content and structure of an evaluation instrument can become the topic of a research study. Questions might involve the format of the evaluation tool, the impact of negatively worded questions, or the content of the item in terms of general versus specific. The item answers can be a topic of study, asking questions such as: are evaluators asked to make an endorsement statement (yes/no), a frequency assessment (never, some of the time, all of the time), a comparative judgment (below average, average, above average, one of the best), a quality judgment (poor, fair, good, excellent), or some other type of assessment? Powers has studied the differences in responses for ratings scales and text responses. Even questions relating to the structure of the evaluation rating scale itself can be studied. For example, should there be 5 or 7 choices on the rating scale, should there be a middle/neutral category, should the highest number connote the best or the worst performance, and should the better performance be on the left or the right of the rating scale? The latter questions may be excellent areas for collaboration between the clerkship director and evaluation psychometricians.

**Design**

When beginning a study related to evaluation, one must decide whether to use a qualitative or more traditional experimental or quasi-experimental design. Historically, randomized trials have been the “gold standard” in research design. However, they are usually impractical in the educational setting, since students are not always amenable to being “randomized.” Moreover, the target/focus of evaluation is often the program (i.e., clerkship) rather than individual learners. Quasi-experimental options that include a control group (historical or concurrent), or pre-test/post-test assessment can be acceptable alternatives. Alternatively, qualitative methods can be used, such as focus groups, observational studies, or interviews. Often, a combination of designs is used in sequence or as complementary approaches to a single research question.

Research studies can also focus on the process of evaluation. This can take the form of comparing alternative evaluation designs and assessing how outcomes and conclusions vary accordingly. For example, research questions may ask about how many groups of people/subjects will provide the data, or the feasibility of a new evaluation process.

**Procedure**

After decisions regarding who will provide evaluation data and the type of data that will be provided, one might consider the finer details of data collection. Comparative research questions in this area might focus on the method of data collection (paper and pencil versus web-based, the timing of data collection (after each lecture, end-of-clerkship versus end-of-year), and other features of the evaluation tool (anonymous versus signed), and process (required versus voluntary). None of these issues are directly linked to central issues of designing and running a clerkship. However, they do have implications for the evaluation data that are produced about a program and are potential areas for research.

**Analysis**

A large part of running a clerkship involves collecting, aggregating, and finding sensible ways to report evaluative data. A number of analytic questions that focus on the evaluation data itself underlie this general task. Often these questions are framed in terms of reliability and validity. Both constructs come into play when one is describing the development of an evaluation
A brief consideration of these concepts follows; for a more detailed review, the reader is referred to other resources.\textsuperscript{116, 130}

**Reliability**
Reliability has to do with the consistency or repeatability of assessments. There are numerous types of reliability statistics and coefficients, many of which are applicable to program and faculty evaluation. For example, one would like to see scores agree when provided by multiple raters, such as multiple clerkship students rating preceptors (inter-rater reliability) or the same rater making assessments on different occasions, assuming no change in the target of investigation (intra-rater reliability). Assessments by the same individual should remain stable, at least over a short time (test-retest reliability), for example, as students complete overall clerkship evaluation forms. Two applications of reliability theory are especially relevant to clerkship evaluation: estimating the internal consistency among items on an evaluation form and determining the number of responses needed to achieve precise scores (evaluation ratings).

**Internal Consistency**
Clerkship directors who are choosing/developing evaluation forms with sets of items that will be summed or averaged should consider assessing how well the items appear to represent a single construct. Measures of internal consistency are based on a single administration of a test or survey instrument.\textsuperscript{116} They capture the homogeneity of items measuring a specific objective or construct. Theoretically, the items actually used in the instrument were selected from the universe of items that could have been asked and thus the selected items should represent the whole, i.e., be homogeneous. Practically speaking, homogeneity coefficients depend on the quality and quantity of items measuring each objective or construct.

Educational evaluations routinely report a Cronbach’s alpha, the most widely used assessment of homogeneity. The coefficient applies to a particular use of a scale with a particular sample of subjects. Consequently, reproducibility is not a stable characteristic of an instrument and it should be recalculated each time the instrument is used, particularly if the sample is different. Moreover, evaluators need to consider when it is appropriate to compute an internal consistency statistic. Sometimes, the intent is not to ask items representing a single domain. In fact, the intention is to represent several distinct domains, perhaps using tools such as factor analysis to help define and refine the domains. When educational evaluators develop new tools, they routinely address the consistency of items scores within multiple domains.\textsuperscript{30, 76, 131}

**Precision of Scores/Number of Raters**
Probably the single most important reliability concept relating to preceptor/faculty evaluation has to do with gathering enough data to provide a reasonably precise estimate of a person’s teaching/precepting ability. In the 1980s, educators started using a framework called generalizability theory as a tool to estimate multiple reliability statistics.\textsuperscript{132, 133}

**Generalizability Theory**
Based on an analysis of variance framework, the theory works in practice by looking at the total variability among a set of scores/ratings and partitioning it into multiple sources. For example, consider the case of preceptor ratings. There might be variability in scores because of the students (different students have different opinions and thus provide different answers), the items (ratings on items about clarity are not exactly the same as those about feedback), and the time of year (students taking the clerkship early in the year are less critical than student taking the clerkship later in the year). Generalizability theory has two important advantages over traditional approaches. First, it allows one to look at multiple sources of variability
simultaneously (often called the G-study). Second, one can manipulate the estimates of variability to address “what if” questions in D-studies. This means asking question such as “what if” one had 5, 10, or 15 ratings per preceptor rather than 2 or 3? What is the impact on the precision of the scores? (See Shavelson and Webb\textsuperscript{132} and Brennan\textsuperscript{133} for a more complete description of the theory and application).

**Precision: How Many Cases/Observations/Students Do I Need?**

Questions about the precision of evaluations of students were prominent in the 1980s, when many schools began assessing students with standardized patients and other case-based examinations. A common question was “how many cases does a learner need to see before a “stable” estimate of performance can be made?” A repeated finding was that performance on one case was not generalizable to performance on other cases.\textsuperscript{134,135} Since then, the general issue of ‘case specificity’ has been extended to other types of evaluation. For example, when judges or raters assess learners’ performances, using multiple judges produces better assessment – thus evening out any potential “hawk and dove” effects.\textsuperscript{87,136-138} Similarly, when abstracting charts to draw inferences about processes of clinical care, large samples of charts are required to reliably estimate a physician's performance.\textsuperscript{139-141} Very few of the studies are about evaluations of faculty/preceptors, but the same principles would apply. Multiple ratings are needed to draw a dependable conclusion regarding a clinical teacher’s abilities.\textsuperscript{8,142}

**Validity**

Validity has to do with the interpretation and use of the scores attached to evaluation instruments. Questions of validity are concerned with asking “have we measured (i.e., evaluated) what we intended to measure? Do scores behave the way we expect them to?” The traditional schema for thinking about validity involves content, criterion, and construct validity. In general, content validity has to do with assessments/descriptions of the methods with which the items/topics within an evaluation instrument were generated. For example, Wallace and colleagues\textsuperscript{143} discussed content validation of family practice resident recruitment questionnaires. Kernan\textsuperscript{37} described the development of an instrument for assessing ambulatory teaching, based on “research on students’ opinions and authoritative guidelines on teaching in the ambulatory setting.”

Criterion and construct validity are more concerned with scores on evaluation instruments and how such scores are related to theories and/or hypotheses of interest. It is quite common in medical education to use correlational analyses and ask how scores on one tool (e.g., self-assessment) compare to those on another (e.g., student ratings). Though valuable for the evaluation enterprise, correlational studies are often limited by difficulty in interpretation - how much is enough? More robust analytic methods are those that involve prediction, for example, how student ratings of faculty performance predict who gets awards for outstanding teaching, or examining differences in ratings before and after faculty participation in teaching enhancement workshops.\textsuperscript{144} Other types of validity studies that focus on evaluation processes include those that compare responses and examine hypotheses related to such things as teacher attributes, like sex or Board certification status,\textsuperscript{131} the site where they precept,\textsuperscript{145} or whether the preceptor was an inpatient attending, an outpatient attending, or a resident.\textsuperscript{146} A persuasive group of studies showed that teacher/preceptor ratings are related to student outcomes.\textsuperscript{21-24} These are among the most convincing data for faculty evaluations.

There are multiple opportunities and rationales for conducting research about evaluation within the context of directing a clerkship. Many of the evaluation research questions that have traditionally focused on the learner (i.e., student) can be extended to the evaluator and/or
moved from focusing on the individual to focusing on the clerkship or educational program. Many research questions concerning evaluation are important to consider, even when they are not turned into full-fledged research studies. However, when developed into public scholarship, they have the potential to assist and influence evaluation activities.

The Role of the Institutional Review Board

Students, residents, and faculty are considered “human subjects” (akin to patients in a clinical trial), and evaluations are considered “data.” In fact, students are often considered a “vulnerable” subject population, similar to children and pregnant women. A researcher would submit a clinical trial research protocol to the university's/hospital's office of regulatory affairs (i.e., Institutional Review Board [IRB]). Medical education research protocols must also be submitted to the IRB if the intent is to publish or disseminate (i.e., presentations at national meetings) the findings. Submissions to an IRB often require a description of the background to the study, the study question and hypotheses, the study design, and the proposed analyses. In certain circumstances, consent forms must be included. A study might be considered exempt or require expedited or full review by the IRB, depending on the nature of the research agenda. Clerkship directors must remember that research about evaluation may need to be approved by their university’s office of regulatory affairs. Policies vary from institution to institution and currently there is a lot of interest in defining how educational research fits into the more typical clinical and basic science purview of IRBs. Policies are in a state of flux, but are moving towards requiring more review of educational research projects. Clerkship directors planning to engage in educational research must familiarize themselves with the IRB requirements of their institution and participate in any training required for investigators conducting research involving human subject.

<table>
<thead>
<tr>
<th>Recommendations for Research about Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Turn evaluation into scholarship</strong></td>
</tr>
<tr>
<td><strong>Focus research on discrete aspects of the clerkship</strong></td>
</tr>
<tr>
<td>• Learners/teachers</td>
</tr>
<tr>
<td>• Development and testing of evaluation instruments</td>
</tr>
<tr>
<td>• Process of evaluation</td>
</tr>
<tr>
<td>• Data collection procedures</td>
</tr>
<tr>
<td>• Analysis of evaluation data</td>
</tr>
<tr>
<td><strong>Become familiar with Institutional Review Board requirements</strong></td>
</tr>
</tbody>
</table>

Conclusion

This chapter provides both practical and theoretical approaches for a thoughtful and comprehensive evaluation of clinical teachers and programmatic elements of a clinical clerkship. A clear vision of the educational purpose and goals of the clerkship are not only the most important starting point, but also the key checking point prior to continuing, eliminating, or modifying programmatic elements. With a growing national interest in creating competency-based curriculum for both undergraduate and graduate medical education, the importance of broadening the definition of formal curricular elements is paramount. All aspects of the formal and the informal curriculum deserve careful program evaluation. Applying a broad variety of methods and tools for program evaluation also helps assure that multi-faceted dimensions are considered. Gathering data from the many internal and external sources gives a clearer vision of the dimensions to be assessed. Using tested frameworks for the evaluation process promotes a systematic approach to the collection, analysis, and interpretation of the
assessment data. Ideally, program evaluation is not an episodic process only for accreditation purposes, but rather an on-going plan for continuous improvement and scholarship.

Appendix A

Modified University of Michigan Global Rating Scale

Faculty Teaching Global Rating Scale

Please rate each faculty member with whom you have worked this month according to the scales below.

Focus this evaluation on your experience with each faculty as a teacher. Please leave the value column line blank if you have had no contact with a faculty.

<table>
<thead>
<tr>
<th>Teaching Contact:</th>
<th>5 = Frequent (&gt;10 sessions)</th>
<th>4 = Moderate (7-9 sessions)</th>
<th>3 = Some (4-6 sessions)</th>
<th>2 = Minimal (1-3 sessions)</th>
<th>1 = None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education Value:</td>
<td>5 = Exceptional Teacher (in top 10-20% of faculty)</td>
<td>4 = Above Average Teacher (includes 15-25% of faculty)</td>
<td>3 = Average Teacher (includes 25-35% of faculty)</td>
<td>2 = Below Average Teacher (includes 15-25% of faculty)</td>
<td>1 = Poor Teacher (in bottom 10-20% of faculty)</td>
</tr>
</tbody>
</table>

Session = ½ day precepting in clinic or presiding at a conference.

<table>
<thead>
<tr>
<th>Faculty</th>
<th>Contact</th>
<th>Value</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty 5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix B
Clinical Teaching Effectiveness Evaluation
(SFDP 26)

Academic Year:
Rotation/Semester:
Course:
Instructor:

You must complete all items below in order to submit the evaluation.

Legend:
SD=Strongly Disagree
D= Disagree
N=Neither Agree nor Disagree
A= Agree
SA=Strongly Agree

Please indicate your agreement with the following statements. During this rotation, my instructor generally:

<table>
<thead>
<tr>
<th>Statement</th>
<th>SD</th>
<th>D</th>
<th>N</th>
<th>A</th>
<th>SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listened to learners</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encouraged learners to participate actively in the discussion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expressed respect for learners</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encouraged learners to bring up problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Called attention to time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avoided digressions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discouraged external interruptions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stated goals clearly and concisely</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stated relevance of goals to learners</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prioritized goals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repeated goals periodically</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presented well organized material</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explained relationships in materials</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Used blackboard or other visual aids</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluated learners’ knowledge of factual medical information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluated learners’ ability to analyze or synthesize medical knowledge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluated learners’ ability to apply medical knowledge to specific patients</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluated learners’ medical skills as they apply to specific patients</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gave negative (corrective) feedback to learners</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td>Rating</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explained to learners why he/she was correct or incorrect</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Offered learners suggestions for improvement</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gave feedback frequently</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explicitly encouraged further learning</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motivated learners to learn on their own</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Encouraged learners to do outside reading</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For the following question, use the scale 1=Poor to 5=Excellent.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall teaching effectiveness</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please type any comments about the instructor:

---

*Your individual responses will NOT be seen by anyone determining or contributing to your final grades. These data will only be seen in aggregate form AFTER your final grade has been submitted.*

Submit Evaluation

Copyright 1998 The Board of Trustees of the Leland Stanford Junior University. All Rights Reserved. Stanford Faculty Development Program, Stanford University School of Medicine.
Appendix C

Undergraduate Medical Education Standard Clerkship Evaluation

<table>
<thead>
<tr>
<th>PLEASE RATE EACH OF THE FOLLOWING. CIRCLE ONE RESPONSE FOR EACH ROW.</th>
<th>POOR</th>
<th>FAIR</th>
<th>GOOD</th>
<th>VERY GOOD</th>
<th>EXCELLENT</th>
<th>NOT APPLICABLE; unable to assess</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Clarity of course goals, objective and expectations.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>2. Overall course organization and coherency.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>3. Commitment of course director(s).</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>4. Educational value/amount learned.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>5. How well the course achieved stated goals.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>6. Professionalism of faculty involved with the course.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>7. Professionalism of residents involved with the course.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>8. Understanding of how you would be evaluated.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>9. Usefulness of feedback</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>10. How well the workload challenged you / level of material appropriate.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td>11. Overall rating / quality of course</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>X</td>
</tr>
</tbody>
</table>

COMMENTS: (Please comment on any strengths, weaknesses or suggestions for future improvements.)
Appendix D

Example of an End-of-Clerkship Course Evaluation Summarizing Repeating Clerkship Events

<table>
<thead>
<tr>
<th>Event</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resident Teaching Conference</td>
<td>128</td>
<td>4.02</td>
<td>0.90</td>
</tr>
<tr>
<td>Radiology Conference</td>
<td>131</td>
<td>4.58</td>
<td>0.64</td>
</tr>
<tr>
<td>EKG Conference</td>
<td>130</td>
<td>3.47</td>
<td>1.09</td>
</tr>
<tr>
<td>Physical Diagnosis Rounds</td>
<td>132</td>
<td>3.28</td>
<td>1.15</td>
</tr>
<tr>
<td>Clinical Problem Solving Session</td>
<td>132</td>
<td>4.00</td>
<td>0.86</td>
</tr>
<tr>
<td>Effectiveness of Attending Rounds</td>
<td>131</td>
<td>4.21</td>
<td>0.81</td>
</tr>
<tr>
<td>Effectiveness of Resident Work Rounds</td>
<td>131</td>
<td>3.89</td>
<td>0.90</td>
</tr>
<tr>
<td>Commitment of teachers (ward attendings/residents) to the course</td>
<td>132</td>
<td>4.25</td>
<td>0.86</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Site</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Site</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Site</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site A</td>
<td></td>
<td></td>
<td></td>
<td>Site B</td>
<td></td>
<td></td>
<td></td>
<td>Site C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effectiveness of Attending Rounds</td>
<td>103</td>
<td>4.29</td>
<td>0.91</td>
<td>48</td>
<td>3.42</td>
<td>1.44</td>
<td>58</td>
<td>4.24</td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effectiveness of Resident Work Rounds</td>
<td>103</td>
<td>1.05</td>
<td>0.96</td>
<td>48</td>
<td>3.52</td>
<td>1.24</td>
<td>58</td>
<td>4.16</td>
<td>1.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effectiveness of Physical Diagnosis Rounds</td>
<td>104</td>
<td>3.60</td>
<td>1.13</td>
<td>49</td>
<td>2.82</td>
<td>1.41</td>
<td>59</td>
<td>3.17</td>
<td>1.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commitment of Teachers (ward attendings/residents)</td>
<td>104</td>
<td>4.32</td>
<td>0.86</td>
<td>49</td>
<td>3.59</td>
<td>1.26</td>
<td>58</td>
<td>4.38</td>
<td>0.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Educational Value of Overnight Call</td>
<td>102</td>
<td>2.78</td>
<td>1.12</td>
<td>47</td>
<td>3.38</td>
<td>1.26</td>
<td>57</td>
<td>3.53</td>
<td>1.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Educational Value of ‘Short’ Call</td>
<td>100</td>
<td>3.64</td>
<td>1.12</td>
<td>47</td>
<td>3.38</td>
<td>1.26</td>
<td>57</td>
<td>3.53</td>
<td>1.34</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix E

Comparison of Overall Clerkship Evaluation Data

Data represent the rating of the "overall quality of the course" question from the clerkship evaluation.
Appendix F

Clerkship Grade Distribution Report

2002/2003 THIRD YEAR CLERKSHIP GRADE DISTRIBUTION

<table>
<thead>
<tr>
<th>DISCIPLINE/COURSE</th>
<th>ANES</th>
<th>FAMP</th>
<th>MED</th>
<th>NEURO</th>
<th>OB/GYN</th>
<th>PEDS</th>
<th>PSYC</th>
<th>SURG</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>44</td>
<td>129</td>
<td>50</td>
<td>156</td>
<td>25</td>
<td>57</td>
<td>39</td>
<td>50</td>
</tr>
<tr>
<td>HP</td>
<td>117</td>
<td>85</td>
<td>104</td>
<td>88</td>
<td>127</td>
<td>104</td>
<td>154</td>
<td>125</td>
</tr>
<tr>
<td>P</td>
<td>95</td>
<td>44</td>
<td>99</td>
<td>13</td>
<td>100</td>
<td>95</td>
<td>63</td>
<td>51</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

NUMBER OF STUDENTS
Appendix G

OSCE Scores

Raw Score Distribution 3rd yr 03

Raw Score Distribution 4th yr 03
Appendix H

OSCE Self Awareness Performance by Site

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Deviation</td>
<td>2.26</td>
<td>2.55</td>
<td>1.32</td>
<td>2.87</td>
<td>2.95</td>
<td>3.02</td>
<td>3.40</td>
<td>3.15</td>
<td>1.97</td>
<td>2.79</td>
</tr>
</tbody>
</table>
Appendix I
OSCE Self Awareness Performance by Site

2001-2003 AAMC Graduation Questionnaire Results
References


82. Lee JS, Sineff SS, Sumner W. Validation of electronic student encounter logs in an emergency medicine clerkship. Proceedings from the AMIA Annual Symposium, 2002:425-0.


111. Wood J, Collins J, Burnside ES, Albanese MA, Propeck PA, Kelcz F, Spilde JM, Schmaltz LM. Patient, faculty, and self-assessment of radiology resident performance: a 360-