Effects of Training in Direct Observation of Medical Residents' Clinical Competence

A Randomized Trial

Eric S. Holmboe, MD; Richard E. Hawkins, MD; and Stephen J. Huot, PhD, MD

Background: **Faculty observation of residents and students performing clinical skills is essential for reliable and valid evaluation of trainees.**

Objective: **To evaluate the efficacy of a new multifaceted method of faculty development called direct observation of competence training.**

Design: **Controlled trial of faculty from 16 internal medicine residency programs using a cluster randomization design.**

Setting: **Academic medical centers.**

Participants: **40 internal medicine teaching faculty members: 17 in the intervention group and 23 in the control group.**

Measurements: **Changes in faculty comfort performing direct observation, faculty satisfaction with workshop, and changes in faculty rating behaviors 8 months after completing the training.**

Intervention: **The direct observation of competence workshop combines didactic mini-lectures, interactive small group and videotape evaluation exercises, and evaluation skill practice with standardized residents and patients.**

Results: **37 faculty members (16 in the intervention group and 21 in the control group) completed the study. Most of the faculty in the intervention group (14 [88%]) reported that they felt significantly more comfortable performing direct observation compared with control group faculty (4 [19%]) ($P = 0.04$), and all intervention faculty rated the training as outstanding. For 9 videotaped clinical encounters, intervention group faculty were more stringent than controls in their evaluations of medical interviewing, physical examination, and counseling; differences in ratings for medical interviewing and physical examination remained statistically significant even after adjustment for baseline rating behavior.**

Limitations: **The study involved a limited number of residency programs, and faculty did not rate the performance of actual residents.**

Conclusion: **Direct observation of competence training, a new multifaceted approach to faculty development, leads to meaningful changes in rating behaviors and in faculty comfort with evaluation of clinical skills.**

Ann Intern Med. 2004;140:874-881. www.annals.org

For author affiliations, see end of text.

See related article on pp 902-909 and editorial comment on pp 927-928.

Medical educators have a major responsibility to evaluate the clinical competence of medical students and residents and to provide them with timely, useful feedback to ensure continued progress and correction of deficiencies. Despite tremendous advances in technology, the clinical skills of interviewing, physical examination, and counseling remain essential to the successful care of patients. The Association of American Medical Colleges (AAMC), Accreditation Council for Graduate Medical Education (ACGME), and American Board of Internal Medicine (ABIM) strongly endorse the evaluation of students and residents in these clinical skills through direct observation (1–3).

Numerous studies continue to document substantial deficiencies in the clinical skills of medical interviewing, physical examination, communication, and counseling among students, residents, and practicing physicians (4–24). For example, Mangione and Nieman (19) demonstrated that students and residents could successfully identify fewer than one third of 12 important cardiac sounds, and Braddock and colleagues (10) found that practicing physicians performed core elements of informed decision making in fewer than 10% of patient encounters. Direct observation of a student or resident performing these skills is mandatory for reliable and valid assessment and is essential for providing formative feedback to improve these skills. Residency is the last structured experience to ensure that young physicians have sufficient clinical skills, but evaluation of these skills by faculty is often neglected. The AAMC, for example, visited 97 medical schools between 1993 and 1998 and found that faculty rarely observed student interactions with patients and that most of a student's evaluation was based on faculty and resident impressions of student presentation skills and knowledge (3). Similar findings have been reported for residents (25–27).

Furthermore, when faculty observe the clinical skills of trainees, the quality of the evaluation is often poor. In a study of faculty ratings using the long form of the clinical evaluation exercise (CEX), faculty failed to detect 68% of errors committed by a resident when observing a videotape scripted to depict marginal performance (28). The use of checklists prompting faculty to look for specific skills increased error detection from 32% to 64% but did not improve overall accuracy; approximately two thirds of faculty still rated the overall performance of the marginal resident as satisfactory or superior. Other studies have also attested to the low reliability of faculty observations (28–33). Perhaps the most notable finding in these studies is that brief faculty training interventions (for example, a 30-

minute description of the CEX and its purpose) failed to improve the quality of faculty evaluation (28–31).

Given the growing concerns over patient safety and quality of care, both public and professional organizations are calling for a renewed emphasis on the teaching and evaluation of clinical skills. To address these concerns, better methods for training faculty in evaluation of clinical competence are urgently needed. The purpose of our study was to evaluate a novel approach to training faculty in direct observation skills—direct observation of competence training—in 3 domains: faculty satisfaction with the training, change in comfort in performing direct observation, and changes in faculty rating behavior.

Context

Reliable methods for evaluating clinical competence of medical trainees are needed.

Contribution

This cluster randomized trial involving 16 internal medicine programs evaluated a 4-day course that taught faculty direct observation methods for evaluating clinical competence. Eight months later, course participants reported greater comfort with direct observation to evaluate residents and rated videotaped clinical encounters between standardized residents and patients more stringently than did faculty not taking the course.

Cautions

Larger studies that use ratings of actual residents and that include resident feedback are needed to establish the transportability and efficacy of the program.

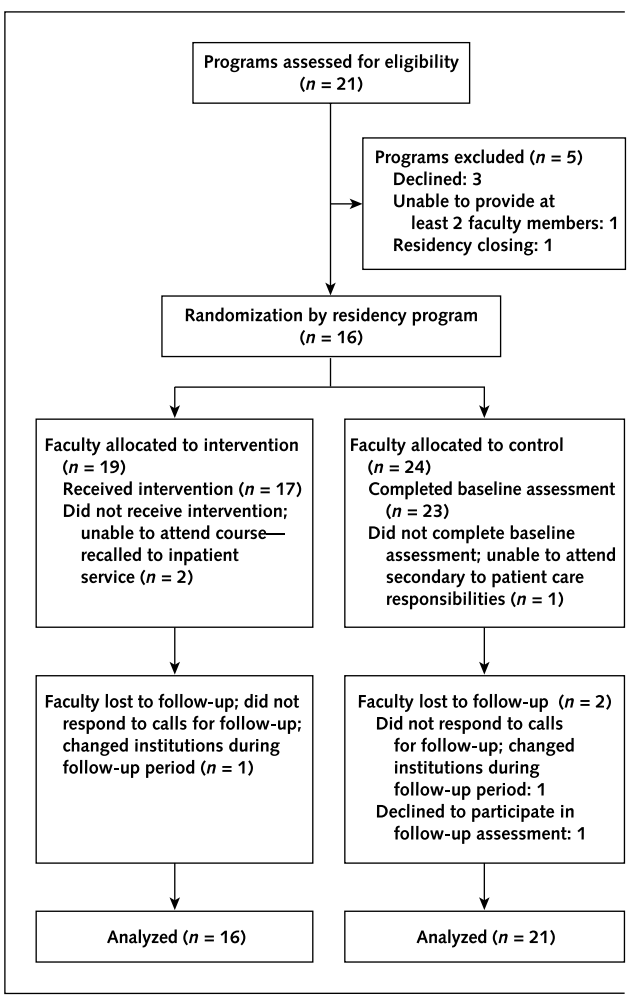
–The Editors

METHODS

Study Design

Our study was a cluster-designed randomized, controlled trial of an intensive 4-day faculty development course in evaluation of clinical competence (34). Twenty-one programs were approached, and 16 were enrolled (**Figure**). Invited programs were chosen from 2 regions—the Northeast (Connecticut, Massachusetts, and Rhode Island) and the Mid-Atlantic (Maryland, Virginia, and the District of Columbia)—and were balanced to represent both university and community-based programs. Of the 5 excluded programs, 3 declined to participate, 1 could not identify at least 2 faculty members to participate, and 1 was in the process of closing. The final cohort consisted of 5 university and 11 community-based programs. Randomization was stratified by program type and location and was blinded by using sealed envelopes. One smaller university program in the Mid-Atlantic group was randomly assigned along with the community-based programs because of the unbalanced number of university programs ($n = 3$) in this region.

Figure. **Flow of participants through the study.**



Participants

Forty faculty members participated in the study. Each program director was required to identify a minimum of 2 and a maximum of 4 faculty members before randomization, and program directors were encouraged to participate. Choice of study participants was left to the discretion of the program directors, but the study guidelines asked directors to choose faculty who were active in teaching and evaluation and who the program director believed would be willing to serve as agents to promote change in their local programs' evaluation practices. Program directors were not informed of other institutional allocations before participation. The Yale University Human Investigation Committee and the Uniformed Services University of the Health Sciences Institutional Review Board approved the study. Participants gave informed consent on the morning of the baseline assessment and before any research activities.

Evaluation Videotapes

Two sets of 9 videotapes were specifically produced for assessment of direct observation skills, 1 set each for the baseline and follow-up assessments. Scripts were written for standardized patients and standardized residents for each of 3 clinical skills: history taking, physical examination, and counseling. The baseline history skill videotapes depicted a male resident interviewing a 64-year-old woman presenting to the emergency department with acute shortness of breath and chest pain due to a pulmonary embolism. The physical examination skill videotapes depicted a male resident examining a 69-year-old man with progressive shortness of breath secondary to ischemic cardiomyopathy. The counseling skill videotapes portrayed a female resident counseling a 48-year-old man about treatment options for his recently diagnosed hypertension. The videotapes were scripted to demonstrate 3 levels of competence, and all contained some errors in clinical performance; none were designed to be a gold standard.

Level 1 videotapes were scripted to represent unequivocally poor performance and contained the most errors of omission and commission. Level 3 videotapes depicted the fewest number of errors. For example, on the baseline level 1 videotape for history taking, the resident fails to introduce himself and fails to ask about key thromboembolic risk factors and the patient's symptoms of leg swelling. In the physical examination videotapes, examples of errors include poor technique in pulmonary auscultation and failure to assess jugular venous distention. For each clinical skill, the average number of errors per videotape was 12 for level 1 videotapes, 6 for level 2 videotapes, and 2 for level 3 videotapes.

The scripts were developed by 1 of the authors and were then reviewed and edited independently by the other 2 authors. The scripts were revised until consensus was reached among the 3 authors. Scripts were then sent to an outside reviewer for additional comments before the final edits. The standardized resident and patient rehearsed the scenarios before the final videotapes were made.

For the follow-up assessment, the medical interview videotapes depicted a male resident interacting with a 69-year-old man who presented to an emergency department with chest pain secondary to coronary artery disease. The physical examination videotapes showed a different male resident examining a 55-year-old man with cough and shortness of breath secondary to right-middle-lobe pneumonia, and the counseling videotape showed a female resident counseling a 69-year-old woman about therapeutic options for hyperlipidemia. The same standardized residents from the baseline videotapes appeared on the medical interview and counseling follow-up videotapes, while all patients in the follow-up videotapes were different.

Baseline Assessment

In October 2001, all participants observed and rated the 9 baseline clinical encounter videotapes in random order. Participants rated the videotapes using a modified version of the ABIM's 9-point mini-CEX form, where 1 to 3 denotes unsatisfactory performance, 4 to 6 denotes satisfactory performance, and 7 to 9 denotes superior performance. Participants were instructed to rate any of the 7 dimensions of clinical competence listed on the mini-CEX form if they felt the videotape had given them sufficient information to do so.

After completion of the baseline assessment, all participants received a comprehensive 3-inch notebook "toolkit" that included reviews of several evaluation methods, the ABIM's mini-CEX evaluation booklet, a CD-ROM and paper copy of the ABIM's portfolio, and a floppy disk with multiple prefabricated evaluation tools and forms. No specific instruction on use of the comprehensive notebook was provided to either group. The control group faculty returned to their home institutions without further instruction while the intervention group faculty participated in a 4-day intensive course on evaluation. Direct observation of competence training occurred on day 2.

Direct Observation of Competence Training

Direct observation of competence training, developed by 2 of the authors, uniquely combines several previously validated rater training and educational methods (35–38). Direct observation of competence training was specifically designed to address deficiencies in faculty observation skills. The workshop began with a review on the state of clinical skills among physicians and trainees, followed by a summary of problems and deficiencies in faculty evaluation and observation skills. Videotapes were then used to facilitate an interactive exercise called performance dimension training, where faculty work in small groups to define the key components of competence for specific clinical skills and develop criteria for satisfactory performance (35, 38). The group then received instruction on how to prepare for and structure the direct observation of trainees.

To facilitate understanding and application of knowledge, skills, and concepts from the didactic and interactive exercises, the intervention faculty worked in groups of 3 to 4 for a 3-hour period with actual standardized patients and standardized residents applying a technique called frame-of-reference training (35, 38). Each participant performed at least 2 direct observation exercises for 2 of the 3 clinical skills for a minimum of 4 exercises per faculty member. Other members of the group watched on a monitor or in the rear of the room, out of view of the individual performing the evaluation. Following the observation, the faculty member provided his or her mini-CEX scores along with feedback to the resident with the other group members present. The encounter ended with a group discussion of how each member of the group rated the encounter and the reasons for their scores and evaluation. The facilitator then discussed what the scenario was scripted to depict and why some of the faculty members' ratings might differ. Performance dimension training and frame-of-reference

Table 1. **Components of Direct Observation of Competence Training**

Rater Training Method	Description	Example
Performance dimension training	Familiarize faculty with appropriate performance dimensions or standards to be used in their own evaluation system by reviewing the dimensions of a performance or competency. Faculty then work in groups to improve their understanding of these definitions with review of actual performance and videotape.	Faculty discuss the elements of an effective counseling session for a patient starting a new medication for a common medical condition, such as hypertension.
Frame-of-reference training	Faculty practice observation skills with standardized patients and residents performing at various levels of competence. One faculty member provides feedback to a standardized resident. Group discusses evaluation after each "encounter," focusing on reasons for the differences between faculty ratings.	Faculty watch a standardized resident counsel a standardized patient starting a new medication for hypertension, with the resident performing the counseling at a poor, satisfactory, and superior level in random order. The elements of informed decision making are used to calibrate faculty evaluations and feedback.*
Behavioral observation training	Focuses on use of observational aids, proper preparation, and correct positioning for observation.	Faculty learn principle of triangulation for proper positioning and how to prepare for the counseling observation by having the resident present the history and physical examination with focused questions before observing the resident–patient interaction.

* Based on Braddock et al. (10).

training are essential to help faculty develop criteria for effective evaluation of clinical skills at different levels of performance. **Table 1** lists the training components of direct observation of competence training.

Evaluation Course

The evaluation course included the evaluation and feedback modules from the Stanford Faculty Development Program (39) and comprised interactive lectures and workshops on rating scales, medical record audits, standardized patients, the mini-CEX, chart-stimulated recall, and the ACGME general competencies. In addition, separate sessions were held on the ACGME general competencies of practice-based learning and improvement and systems-based practice. On the final day of the course, the intervention group completed an individual satisfaction questionnaire and listed personal evaluation goals for the year.

Follow-up Assessment

All participants were asked to return 8 months later for follow-up assessment, which included a survey about changes they had made in their personal evaluation practices during the year and a self-rating of their evaluation skills. The same modified mini-CEX form used at baseline was used to assess resident performance on the follow-up videotapes.

Statistical Analysis

All participants completed a preassessment questionnaire. Descriptive statistics were used to describe the demographic characteristics of the control and intervention groups. Intervention group faculty were asked to rate each component of the evaluation course, including direct observation of competence training, on a 5-point scale, where 1 represented poor and 5 outstanding. Intervention faculty were also asked to comment on whether they would recommend the entire evaluation course to another faculty member on a 7-point scale, where 1 represented definitely no and 7 definitely yes. A commitment-to-change strategy was used to assess personal goals with the intervention group (40). At the end of the 4-day course, each participant was asked to write down specific goals for the year. At follow-up 8 months later, these individual goals were given back to the intervention group. Participants were asked to state whether they had implemented each goal fully, partially, or not at all. Frequencies for partially and fully implemented goals regarding observation were tabulated.

To evaluate changes in comfort in performing direct observation, ratings were compared between the groups with the Wilcoxon rank-sum test. Because the errors on each videotape were scripted and thus not random (fixed), we could not use traditional methods to calculate reliability. Therefore, interrater agreement for the mini-CEX scores for each group was estimated using confidence intervals and range of scores.

For the rating scores on the videotapes, bivariate analysis with the faculty as the unit of analysis was performed for each clinical skill, comparing rating behavior between the 2 groups at baseline and follow-up. To examine differences in rating behavior between the baseline and follow-up periods, we used a 3-level hierarchical regression model that accounted for clustering by residency program and adjusted for participants' baseline rating behavior. A log transformation function was used to convert rating

Table 2. **Demographic Characteristics of the Study Sample**

Characteristic	Intervention Group ($n = 16$)	Control Group ($n = 21$)
Women, *n (%)*	7 (44)	11 (53)
Program directors, *n**	7	7
Mean age, *y*	41.2	40.3
Community-based programs, *n (%)*	9 (56)	17 (81)
University-based programs, *n (%)*	7 (44)	4 (19)
General internists, *n (%)*	14 (88)	18 (86)
Associate professor or higher rank, *n*	2	3
Mean time at current job, *y*	2.9	3.3

* Includes associate directors.

scores for the 3 clinical skills, with the control group serving as the reference. The analysis used the intention-to-treat principle, and a *P* value of 0.05 was considered statistically significant. Statistical analysis was performed by using SAS, version 8.0 (SAS Institute, Inc., Cary, North Carolina).

Role of the Funding Sources

The funding sources had no role in the collection, analysis, or interpretation of the data or the decision to submit the manuscript for publication. Statisticians from ABIM provided advice on methods to measure reliability only. However, all analysis was performed solely by the research team. The investigators had full access to data in the trial and had final responsibility for the decision to submit the results for publication.

RESULTS

Forty-three faculty members were originally randomly assigned. Forty were enrolled at baseline: 17 in the intervention group and 23 in the control group (**Figure**). Three faculty members were excluded because of unanticipated clinical responsibilities. Thirty-seven faculty members (93%) (16 in the intervention group and 21 in the control group) completed both the baseline and follow-up assessments. One participant in the intervention group missed the direct observation of competence training at baseline but completed all videotape reviews. One participant each from the intervention and control groups left their institutions during the academic year and were unavailable for follow-up. One participant from the control group declined to complete the follow-up assessment. The demographic characteristics of the 2 groups are shown in **Table 2**.

Table 3. **Ratings of the Clinical Skills Videotapes at Baseline***

Clinical Skill†	Mean Scores in the Control Group [95% CI]	Mean Scores in the Intervention Group [95% CI]	*P* Value
History taking			
Level 1	3.4 [3.0–3.8]	3.2 [2.7–3.7]	
Level 2	5.6 [4.9–6.3]	5.1 [4.5–5.7]	
Level 3	6.6 [6.1–7.2]	6.3 [5.6–6.9]	
Overall			>0.2
Physical examination			
Level 1	3.2 [2.6–3.8]	3.2 [2.5–3.9]	
Level 2	5.4 [4.7–6.2]	4.4 [3.8–5.0]	
Level 3	7.5 [7.0–8.0]	6.7 [5.7–7.6]	
Overall			0.10
Counseling			
Level 1	2.1 [1.6–2.5]	2.3 [1.8–2.8]	
Level 2	4.5 [3.9–5.2]	4.7 [4.1–5.3]	
Level 3	6.0 [5.2–6.7]	6.0 [5.3–6.7]	
Overall			>0.2

* Twenty-one faculty members participated in the control group, and 16 participated in the intervention group.
† Level 1 = less skill and many deficiencies depicted; level 2 = moderate skill and moderate number of deficiencies depicted; level 3 = excellent skill and few deficiencies depicted.

Table 4. **Ratings of the Clinical Skills Videotapes at Follow-up***

Clinical Skills†	Mean Scores in the Control Group (Range) [95% CI]	Mean Scores in the Intervention Group (Range) [95% CI]	*P* Value
History taking			
Level 1	3.8 (2–7) [3.0–4.7]	2.7 (1–4) [2.1–3.3]	
Level 2	6.1 (4–8) [5.5–6.1]	5.1 (4–6) [4.7–5.6]	
Level 3	7.5 (6–9) [7.1–8.0]	6.0 (4–8) [5.4–6.6]	
Overall			<0.001
Physical examination			
Level 1	2.8 (2–5) [2.4–3.2]	2.1 (1–4) [1.7–2.6]	
Level 2	4.3 (3–7) [3.7–4.9]	3.3 (2–5) [2.9–3.7]	
Level 3	7.2 (4–9) [6.7–7.6]	6.0 (4–9) [5.4–6.6]	
Overall			0.06
Counseling			
Level 1	2.6 (1–6) [2.0–3.1]	2.3 (1–3) [1.8–2.7]	
Level 2	4.7 (3–7) [4.1–5.3]	3.5 (2–6) [2.9–4.1]	
Level 3	7.1 (3–9) [6.4–7.9]	6.7 (5–8) [6.0–7.4]	
Overall			0.13

* Twenty-one faculty members participated in the control group, and 16 participated in the intervention group.
† Level 1 = less skill and many deficiencies depicted; level 2 = moderate skill and moderate number of deficiencies depicted; level 3 = excellent skill and few deficiencies depicted.

There was a slightly higher percentage of women in the control group in the follow-up cohort, but there was no difference in the percentage of women enrolled at baseline (48% in both groups). Participants' mean age and duration of time in their current job was similar in both groups. Most faculty were general internists, and all were involved in both inpatient and outpatient resident education. The intervention group had more university-based faculty members because 2 university-based faculty in the control group did not complete the follow-up assessment. Overall, members of both groups were relatively young and early in their academic careers; the mean time spent in their current jobs was approximately 3 years.

Evaluation of the Course and Training

Participants gave high ratings to the quality of the direct observation of competence training. All 16 participants who completed the training (100%) rated the experience as a 5 on a 5-point scale, with 5 representing outstanding. This training was the most highly rated aspect of the 4-day faculty development course for the intervention group. In comparison, the 7 other workshops given during the evaluation course received mean ratings that ranged from 4 to 4.4. Finally, all faculty said they would definitely recommend the entire evaluation course to another colleague.

Follow-up Survey Results

All but 1 participant in the control group (97%) directly observed residents during the follow-up study period. Most participants (97%) used the mini-CEX, and 57% also used the traditional long-form CEX for direct observation. When participants were asked at follow-up to assess their change in comfort in performing direct obser-

vation, the intervention group reported being more comfortable than the control group. On a 5-point scale (1 representing less comfortable, 3 representing the same level of comfort, and 5 representing more comfortable), the median score was 3 for the control group and 4 for the intervention group. Most of the faculty in the intervention group (14 [88%]) reported that they felt significantly more comfortable performing direct observation compared with control group faculty (4 [19%]) ($P = 0.04$). Only 2 faculty members in the intervention group had the same level of comfort at follow-up, including the 1 faculty member who did not complete direct observation of competence training.

At the end of the initial 4-day course, the intervention faculty members were asked to write down personal goals for improvement in their evaluation skills as part of a commitment-to-change exercise. Thirteen of the 16 wrote a personal goal of performing more direct observations after the course, and 12 of these 13 (92%) reported they were at least partially successful in meeting this goal. The major barrier cited to full implementation was lack of time.

Changes in Rating Behavior

Table 3 shows that there were no statistically significant differences in the evaluation of the standardized residents between the control and intervention groups for any of the 9 videotapes during the baseline assessment. **Table 4** displays the results for the 2 groups at follow-up. The intervention group rated all 9 videotapes more stringently than did the control group, and the range was smaller in the intervention group for 6 of the 9 videotaped encounters. After we adjusted for baseline rating behavior and accounted for clustering by residency program, differences in the medical interviewing and physical examination ratings were statistically significant between the intervention and control groups. After adjustment, the regression model estimated that, on average, mini-CEX scores in the intervention group were 13% (95% CI, 4% to 24%) lower for medical interviewing ($P = 0.02$), 23% (CI, 8% to 35%) lower for physical examination ($P = 0.004$), and 8% (CI, −9% to 22%) lower for counseling ($P > 0.2$) compared with the control group. Of note, the follow-up ratings for the 3 medical interviewing tapes were all higher than at baseline in the control group and lower than at baseline in the intervention group.

DISCUSSION

Sound clinical skills remain one of the cornerstones of successful, safe, and effective medical practice. Studies continue to show significant deficiencies in clinical skills in students and practicing physicians (4–24). Direct observation of medical interviewing, physical examination, communication, and counseling by faculty is a fundamental requirement for the accurate evaluation of clinical skills.

Our randomized, controlled trial of a new faculty development method in direct observation, direct observation of competence training, found that the method was highly valued by the participants and led to statistically significant differences in self-reported comfort with direct observation as well as self-reported changes in frequency of observation. Equally important was the finding that 1 day of training led to changes in actual rating behavior that persisted over 8 months.

Given the call from the public and from professional organizations for increased emphasis on clinical skills as one of the initiatives to improve quality of care and patient safety, effective evaluation of clinical skills is critical. Such evaluation is needed to provide effective feedback and to ensure that trainees, especially residents, have attained sufficient skill for independent practice. Although simulations and standardized patients are well-tested methods for teaching and measuring competence in clinical skills, they cannot replace the direct observation of trainees' performance with actual patients during medical training (41–48). Teaching faculty must continue to accept the primary responsibility for observation of trainees' interactions with patients. The faculty development course we presented here describes a new method for improving faculty comfort and skill in direct observation.

Because of the competing demands on faculty, any new faculty training method needs to be effective, interactive, and given over a reasonable period. Direct observation of competence training meets these criteria. It is a multifaceted approach with interactive exercises and hands-on practice, it is based on methods previously developed and studied in the performance appraisal field, and it is consistent with adult learning principles (35–38, 49). This training can be accomplished in 1 day and appears to have effects that last at least 8 months.

Our study had several limitations. The sample size was limited by the number of programs able to participate, and thus the power of the study was reduced. A second limitation is that participants rated residents on videotapes, not actual residents. Therefore, although our study demonstrated that faculty training can change rating behavior, we do not yet know whether such change will be seen with actual residents. Third, a higher proportion of the intervention faculty were from university-based programs because 2 university-based faculty members in the control group did not participate in the follow-up assessment. Finally, the analysis cannot tell us whether the intervention group's ratings were more accurate than those of the control group. More stringent ratings are not necessarily more accurate. In addition, are more stringent ratings educationally important?

From the perspective of formative assessment, more stringent ratings can lead to higher standards of performance and more meaningful feedback to help residents attain better clinical skills. From the perspective of competency determination, identification of residents with unsatisfactory skills is an educational and professional obligation of residency programs. Our study suggests that interven-

tion group faculty more accurately recognized unsatisfactory performance. None of the intervention group faculty rated the 3 clearly unsatisfactory level 1 videotapes as satisfactory or superior, but several of the control group faculty did so. It is crucial that unsatisfactory resident performance be identified early to allow programs to intervene and ensure that graduating residents have attained a minimum acceptable level of competence in clinical skills.

In conclusion, direct observation of competence training appears to be a promising new method to train faculty in the observation of clinical skills. Direct observation of competence training can be accomplished in a single day, is well regarded by participants, and produces changes in faculty evaluation behaviors that last at least 8 months. Future work in faculty evaluation training for observation should focus on the accuracy of observation and on faculty rating behavior with actual residents after participation in direct observation of competence training.

From Yale University, New Haven, Connecticut, and the Uniformed Services University of the Health Sciences, Bethesda, Maryland.

Note: Dr. Hawkins was Associate Professor of Medicine at the Uniformed Services University of the Health Sciences during the duration of this study.

Acknowledgments: The authors thank Yun Wang, MS, for his statistical expertise, Leslie Galaty for her help in data entry and organization, Dr. Jeffrey Chung for his invaluable assistance in the research, Dr. Louis Pangaro for his assistance in reviewing videotapes, Dr. John Norcini and Rebecca Lipner for their statistical advice, and Drs. Patrick O'Connor and Daniel Duffy for their review of the research.

Grant Support: By the Robert Wood Johnson Foundation and the American Board of Internal Medicine Foundation.

Potential Financial Conflicts of Interest: *Honoraria:* E.S. Holmboe (American Board of Internal Medicine Foundation); *Grants received:* E.S. Holmboe (Robert Wood Johnson Foundation, American Board of Internal Medicine, Bureau of Primary Healthcare, Agency for Healthcare Research and Quality, Mitchell Center for Prisoner of War Studies, Reynolds Foundation, American College of Physicians Foundation).

Corresponding Author: Eric S. Holmboe, MD, Department of Medicine, Waterbury Hospital, Pomeroy 6, 64 Robbins Street, Waterbury, CT 06721.

Current author addresses and author contributions are available at www.annals.org.

References

1. Accreditation Council for Graduate Medical Education. ACGMA Outcome Project: The General Competencies. Accessed at www.acgme.org on 12 February 2003.

2. **American Board of Internal Medicine.** Portfolio for Internal Medicine Residency Programs. Philadelphia: American Board of Internal Medicine; 2001.

3. **Szenas P.** The role of faculty observation in assessing students' clinical skills. Contemporary Issues in Medical Education. 1997;1:1-2.

4. **Platt FW, McMath JC.** Clinical hypocompetence: the interview. Ann Intern Med. 1979;91:898-902. [PMID: 517891]

5. **Meuleman JR, Caranasos GJ.** Evaluating the interview performance of internal medicine interns. Acad Med. 1989;64:277-9. [PMID: 2713013]

6. **Beaumier A, Bordage G, Saucier D, Turgeon J.** Nature of the clinical difficulties of first-year family medicine residents under direct observation. CMAJ. 1992;146:489-97. [PMID: 1737313]

7. **Sachdeva AK, Loiacono LA, Amiel GE, Blair PG, Friedman M, Roslyn JJ.** Variability in the clinical skills of residents entering training programs in surgery. Surgery. 1995;118:300-8; discussion 308-9. [PMID: 7638747]

8. **Pfeiffer C, Madray H, Ardolino A, Willms J.** The rise and fall of students' skill in obtaining a medical history. Med Educ. 1998;32:283-8. [PMID: 9743783]

9. **Ramsey PG, Curtis JR, Paauw DS, Carline JD, Wenrich MD.** History-taking and preventive medicine skills among primary care physicians: an assessment using standardized patients. Am J Med. 1998;104:152-8. [PMID: 9528734]

10. **Braddock CH 3rd, Edwards KA, Hasenberg NM, Laidley TL, Levinson W.** Informed decision making in outpatient practice: time to get back to basics. JAMA. 1999;282:2313-20. [PMID: 10612318]

11. **Stewart MA, McWhinney IR, Buck CW.** The doctor/patient relationship and its effect upon outcome. J R Coll Gen Pract. 1979;29:77-81. [PMID: 480298]

12. **Richards T.** Chasms in communication [Editorial]. BMJ. 1990;301:1407-8. [PMID: 2279152]

13. **Katz J.** The Silent World of Doctor and Patient. New York: Free Pr; 1984.

14. **Simpson M, Buckman R, Stewart M, Maguire P, Lipkin M, Novack D, et al.** Doctor-patient communication: the Toronto consensus statement. BMJ. 1991;303:1385-7. [PMID: 1760608]

15. **Wiener S, Nathanson M.** Physical examination. Frequently observed errors. JAMA. 1976;236:852-5. [PMID: 947266]

16. **Wray NP, Friedland JA.** Detection and correction of house staff error in physical diagnosis. JAMA. 1983;249:1035-7. [PMID: 6823058]

17. **Butterworth JS, Reppert EH.** Auscultatory acumen in the general medical population. JAMA. 1960;174:32-4.

18. **Raftery EB, Holland WW.** Examination of the heart: an investigation into variation. Am J Epidemiol. 1967;85:438-44. [PMID: 4225873]

19. **Mangione S, Nieman LZ.** Cardiac auscultatory skills of internal medicine and family practice trainees. A comparison of diagnostic proficiency. JAMA. 1997;278:717-22. [PMID: 9286830]

20. **Mangione S, Burdick WP, Peitzman SJ.** Physical diagnosis skills of physicians in training: a focused assessment. Acad Emerg Med. 1995;2:622-9. [PMID: 8521209]

21. **Li JT.** Assessment of basic physical examination skills of internal medicine residents. Acad Med. 1994;69:296-9. [PMID: 8155238]

22. **Johnson JE, Carpenter JL.** Medical house staff performance in physical examination. Arch Intern Med. 1986;146:937-41. [PMID: 3963985]

23. **Fox RA, Ingham Clark CL, Scotland AD, Dacre JE.** A study of pre-registration house officers' clinical skills. Med Educ. 2000;34:1007-12. [PMID: 11123564]

24. **Todd IK.** A thorough pulmonary exam and other myths. Acad Med. 2000; 75:50-1. [PMID: 10667875]

25. **Thompson WG, Lipkin M Jr, Gilbert DA, Guzzo RA, Roberson L.** Evaluating evaluation: assessment of the American Board of Internal Medicine Resident Evaluation Form. J Gen Intern Med. 1990;5:214-7. [PMID: 2341920]

26. **Haber RJ, Avins AL.** Do ratings on the American Board of Internal Medicine Resident Evaluation Form detect differences in clinical competence? J Gen Intern Med. 1994;9:140-5. [PMID: 8195912]

27. **Holmboe ES, Fiebach NF, Galaty L, Huot S.** The effectiveness of a focused educational intervention on resident evaluations from faculty: a randomized controlled trial. J Gen Intern Med. 2001;16:427-34. [PMID: 11520379]

28. **Noel GL, Herbers JE Jr, Caplow MP, Cooper GS, Pangaro LN, Harvey J.** How well do internal medicine faculty members evaluate the clinical skills of residents? Ann Intern Med. 1992;117:757-65. [PMID: 1343207]

29. **Herbers JE Jr, Noel GL, Cooper GS, Harvey J, Pangaro LN, Weaver MJ.** How accurate are faculty evaluations of clinical competence? J Gen Intern Med. 1989;4:202-8. [PMID: 2723833]

30. **Kalet A, Earp JA, Kowlowitz V.** How well do faculty evaluate the interviewing skills of medical students? J Gen Intern Med. 1992;7:499-505. [PMID:

1403205]

31. **Kroboth FJ, Hanusa BH, Parker S, Coulehan JL, Kapoor WN, Brown FH, et al.** The inter-rater reliability and internal consistency of a clinical evaluation exercise. J Gen Intern Med. 1992;7:174-9. [PMID: 1487766]

32. **Elliot DL, Hickam DH.** Evaluation of physical examination skills. Reliability of faculty observers and patient instructors. JAMA. 1987;258:3405-8. [PMID: 3682139]

33. **Kassebaum DG, Eaglen RH.** Shortcomings in the evaluation of students' clinical skills and behaviors in medical school. Acad Med. 1999;74:842-9. [PMID: 10429595]

34. **Regehr G.** The experimental tradition. In: Norman GR, van der Vleuten CP, Newble DI, eds. International Handbook of Research in Medical Education. Dordrecht, the Netherlands: Kluwer Academic Publishers; 2002:5-44.

35. **Hauenstein NM.** Training raters to increase accuracy of appraisals and the usefulness of feedback. In: Smither JW, ed. Performance Appraisal: State of the Art in Practice. San Francisco: Jossey–Bass; 1998:404-44.

36. **Woehr DJ, Huffcutt AI.** Rater training for performance appraisal: a quantitative review. Journal of Occupational Organizational Psychology. 1994;67:189-205.

37. **Stamoulis DT, Hauenstein NMA.** Rater training and rating accuracy: training for dimensional accuracy versus training for rate differentiation. J Appl Psychol. 1993;78:994-1003.

38. **Murphy KR, Cleveland JN.** Performance Appraisal: An Organizational Perspective. Boston: Allyn and Bacon; 1991.

39. **Skeff KM, Stratos GA, Berman J, Bergen MR.** Improving clinical teaching. Evaluation of a national dissemination program. Arch Intern Med. 1992;152: 1156-61. [PMID: 1599342]

40. **Lockyer JM, Fidler H, Ward R, Basson RJ, Elliott S, Toews J.** Commitment to change statements: a way of understanding how participants use information and skills taught in an educational session. J Contin Educ Health Prof. 2001;21:82-9. [PMID: 11420869]

41. **Barrows HS.** An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC. Acad Med. 1993;68:443-51; discussion 451-3. [PMID: 8507309]

42. **van der Vleuten CPM, Swanson DB.** Assessment of clinical skills with standardized patients: state of the art. Teaching and Learning in Medicine. 1990; 2:58-76.

43. **Dupras DM, Li JT.** Use of an objective structured clinical examination to determine clinical competence. Acad Med. 1995;70:1029-34. [PMID: 7575931]

44. **Anderson MB, Stillman PL, Wang Y.** Growing use of standardized patients in teaching and evaluation in medical education. Teaching and Learning in Medicine. 1994;6:15-22.

45. **Scoles PV.** An evaluation of clinical skills in the United States medical licensing examination: A report from the National Board of Medical Examiners. Journal of Medical Licensure and Discipline. 2002;88:66-9.

46. **Ram P, van der Vleuten C, Rethans JJ, Grol R, Aretz K.** Assessment of practicing family physicians: comparison of observation in a multiple-station examination using standardized patients with observation of consultations in daily practice. Acad Med. 1999;74:62-9. [PMID: 9934298]

47. **Kopelow ML, Schnabl GK, Hassard TH, Tamblyn RM, Klass DJ, Beazley G, et al.** Assessing practicing physicians in two settings using standardized patients. Acad Med. 1992;67:S19-21. [PMID: 1388543]

48. **Rethans JJ, Sturmans F, Drop R, van der Vleuten C, Hobus P.** Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. BMJ. 1991;303:1377-80. [PMID: 1760606]

49. **Knowles M.** The Adult Learner: A Neglected Species. Houston, TX: Gulf Publishing; 1984.

Current Author Addresses: Dr. Holmboe: Department of Medicine, Waterbury Hospital, Pomeroy 6, 64 Robbins Street, Waterbury, CT 06721.
Dr. Hawkins: National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104-3102.
Dr. Huot: Yale University School of Medicine, 33 Cedar Street, 1074 LMP, New Haven, CT 06520.

Author Contributions: Conception and design: E.S. Holmboe, R.E. Hawkins, S.J. Huot.
Analysis and interpretation of the data: E.S. Holmboe, R.E. Hawkins.
Drafting of the article: E.S. Holmboe, S.J. Huot.
Critical revision of the article for important intellectual content: E.S. Holmboe, R.E. Hawkins.
Final approval of the article: E.S. Holmboe, R.E. Hawkins, S.J. Huot.
Provision of study materials or patients: E.S. Holmboe, R.E. Hawkins.
Obtaining of funding: E.S. Holmboe.
Administrative, technical, or logistic support: E.S. Holmboe, R.E. Hawkins.
Collection and assembly of data: E.S. Holmboe, R.E. Hawkins.