# The Mini-CEX: A Method for Assessing Clinical Skills

John J. Norcini, PhD; Linda L. Blank; F. Daniel Duffy, MD; and Gregory S. Fortna, MSEd

Objective: **To evaluate the mini–clinical evaluation exercise (mini-CEX), which assesses the clinical skills of residents.**

Design: **Observational study and psychometric assessment of the mini-CEX.**

Setting: **21 internal medicine training programs.**

Participants: **Data from 1228 mini-CEX encounters involving 421 residents and 316 evaluators.**

Intervention: **The encounters were assessed for the type of visit, sex and complexity of the patient, when the encounter occurred, length of the encounter, ratings provided, and the satisfaction of the examiners. Using this information, we determined the overall average ratings for residents in all categories, the reliability of the mini-CEX scores, and the effects of the characteristics of the patients and encounters.**

Measurements: **Interviewing skills, physical examination, professionalism, clinical judgment, counseling, organization and efficiency, and overall competence were evaluated.**

Results: **Residents were assessed in various clinical settings with a diverse set of patient problems. Residents received the lowest ratings in the physical examination and the highest ratings in professionalism. Comparisons over the first year of training showed statistically significant improvement in all aspects of competence, and the method generated reliable ratings.**

Conclusions: **The measurement characteristics of the mini-CEX are similar to those of other performance assessments, such as standardized patients. Unlike these assessments, the difficulty of the examination will vary with the patients that a resident encounters. This effect is mitigated to a degree by the examiners, who slightly overcompensate for patient difficulty, and by the fact that each resident interacts with several patients. Furthermore, the mini-CEX has higher fidelity than these formats, permits evaluation based on a much broader set of clinical settings and patient problems, and is administered on site.**

*Ann Intern Med.* 2003;138:476-481. www.annals.org

For author affiliations, see end of text.

When the American Board of Internal Medicine (ABIM) stopped administering the oral examination in 1972, it asked program directors to evaluate the clinical competence of candidates for certification. The ABIM has since recommended that program directors use a traditional clinical evaluation exercise, or CEX, as one form of assessment for residents, particularly first-year residents. In this exercise, which is based on the bedside oral examination, a faculty member evaluates the resident as he or she performs a complete history and physical examination on an inpatient and then reaches diagnostic and therapeutic conclusions (1). The CEX takes about 2 hours, and residents are assessed during their first year of training (2).

The CEX has been criticized as an evaluation instrument because the results are unlikely to be generalizable beyond the observed encounter (3–6). Physician performance is case specific, and the CEX assesses the performance of the resident with only one patient. Moreover, faculty members vary in stringency, and the CEX incorporates the ratings of only one examiner. In response to these problems, the ABIM proposed the mini-CEX. In this variation on the traditional CEX, one faculty member evaluates a resident with one patient in a 15- to 20-minute encounter; several of these assessments are conducted throughout the year. The encounters can occur in various settings (ambulatory, emergency department, and inpatient), so the patients present a broader range of challenges. This allows the residents to be evaluated by different faculty members as they interact with several patients who pose a wider range of problems.

In addition to offering better evaluation, the multiple-patient mini-CEX differs from the traditional CEX in what it can evaluate. The traditional CEX focuses on the resident's thoroughness in an environment that is uninfluenced by the time constraints of medical practice. In contrast, multiple mini-CEX encounters are more variable because the challenges they pose depend on a broader range of settings, patients, and tasks. Furthermore, mini-CEXs assess the resident's ability to focus and prioritize diagnosis and management within the context of real clinical practice.

Preliminary study of the mini-CEX verified that it 1) assessed residents in a broader range of clinical situations than the traditional CEX, 2) produced scores that were more reliable than those based on the traditional CEX, and 3) offered the residents more opportunity for observation and feedback (7, 8). However, these findings were based on only 88 residents from five programs in a limited geographic region. Moreover, the study did not address a series of important questions about changes in competence throughout the first year of training, the complexity of patient problems, the focus of the encounter, and the relative amount of time spent observing resident performance and offering feedback. The current study replicated and expanded on the previous work. Specifically, we collected the data for the study during a 1-year period from 21 training programs, representing a range of resident ability.

## Methods

### Procedure

For each mini-CEX encounter, one examiner observed the resident conduct a focused interview or physical examination in an inpatient, outpatient, emergency department,

 

or other setting. After asking the resident for diagnostic or therapeutic decisions, the examiner completed the rating form (**Appendix Figure**, available at www.annals.org) and provided feedback. The rating form was a card that produced two copies, and the same form was used across all sites of the study.

For each encounter, the examiner recorded the date, the complexity of the patient's problem on a 3-point scale (low, moderate, and high), the sex of the patient, the type of visit (new or return), the setting (ambulatory, inpatient, emergency department, or other), the number of minutes spent observing the encounter, and the number of minutes spent giving feedback. The examiner also noted whether the focus of the encounter was data gathering, diagnosis, treatment, or counseling.

Using a 9-point scale (in which 1 to 3 were "unsatisfactory," 4 was "marginal," 5 and 6 were "satisfactory," and 7 to 9 were "superior"), the examiner rated the resident on interviewing, physical examination, professionalism, clinical judgment, counseling, organization and efficiency, and overall competence. The examiner also rated his or her own satisfaction with the method as a valid and efficient assessment device on a 9-point scale in which 1 was "dissatisfied" and 9 was "very satisfied." For all items, the examiner could select "not applicable" where appropriate and a small number of examiners assigned ratings in 0.5-point increments. In addition, a total score was calculated as the mean of the six component ratings.

### Participants

Twenty-one residency programs nonsystematically volunteered to assess residents by using this format. They were primarily, but not exclusively, from the northeastern United States. The programs represented a broad range of resident abilities, as judged by their program pass rates on the ABIM certifying examination.

A total of 1228 usable evaluation forms, representing data from 421 residents, were returned to the investigators. The residents were evaluated on 1 to 8 encounters, with a mean (±SD) of 2.9 ± 1.5 and a median of 3. The evaluations were conducted by 316 examiners who evaluated a mean of 3.9 ± 4.2 residents and a median of 2 residents.

### Statistical Analysis

Analysis was conducted at the level of the individual encounter, the resident, and the examiner. For the individual encounters, we describe the complexity, setting, and focus of the encounters, as well as the patients' sex and type of visit. Performance across the four quarters of the academic year are also described. For the residents, we average their ratings over all their encounters and analyze the relationships among the components of competence and the reliability of the total score. For the examiners, we also average their ratings and analyze their satisfaction with the method and its relationship with other aspects of the encounter. Nonparametric statistical tests, including the Mann–Whitney *U*, Kruskal–Wallis (with a chi-square approximation), and Spearman $\rho$, were applied to these data because some of the variables are categorical and others may not meet assumptions of normality. Analyses were conducted by using SAS software (SAS Institute, Inc., Cary, North Carolina) and SPSS software (SPSS, Inc., Chicago, Illinois).

*Table 1.* **Number of Examiners, Residents, and Encounters Included in the Study**

| Program | Examiners | Residents | Encounters |
|---|---|---|---|
| | ← *n* → | | |
| 1 | 1 | 2 | 2 |
| 2 | 11 | 9 | 50 |
| 3 | 5 | 14 | 34 |
| 4 | 8 | 14 | 19 |
| 5 | 4 | 10 | 12 |
| 6 | 25 | 42 | 139 |
| 7 | 14 | 12 | 47 |
| 8 | 18 | 22 | 118 |
| 9 | 4 | 16 | 17 |
| 10 | 9 | 11 | 20 |
| 11 | 43 | 34 | 135 |
| 12 | 7 | 20 | 30 |
| 13 | 2 | 21 | 29 |
| 14 | 14 | 26 | 81 |
| 15 | 38 | 31 | 110 |
| 16 | 10 | 25 | 46 |
| 17 | 7 | 11 | 40 |
| 18 | 19 | 30 | 67 |
| 19 | 19 | 28 | 79 |
| 20 | 4 | 9 | 16 |
| 21 | 54 | 34 | 137 |
| Total | 316 | 421 | 1228 |

All educational tests are fallible, and generalizability theory offers a family of analysis-of-variance–based indices to quantify how well their scores are generalizable beyond the specific behaviors they elicit (9–11). Details of the application of generalizability theory in this study can be found in the Appendix (available at www.annals.org), but briefly, the total scores were analyzed to estimate variance components for the residents and for measurement error. These reflect the variability in ratings that would occur if each resident were examined by a large number of evaluators while seeing a large number of patients.

The variance components were used to estimate what the 95% CIs would be for total scores based on 1 to 14 encounters. Analogous to its counterparts in biostatistics, the CI has the advantage of simultaneously addressing issues of psychometric and practical significance.

## Results

**Table 1** presents data on the numbers of programs, evaluators, residents, and encounters involved in the study. The degree of participation varied greatly. Unless otherwise noted, data reported are the mean (±SD).

### Encounters

The mean age of the patients was 55 ± 18 years; 45% were men and 49% were women (sex was not recorded for

*Table 2.* **Ratings and Complexity of Patient Problems for the Academic Quarters***

| Ratings | First Quarter | | Second Quarter | | Third Quarter | | Fourth Quarter | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Rating | Encounters | Rating | Encounters | Rating | Encounters | Rating | Encounters | Rating | Encounters |
| | | *n* | | *n* | | *n* | | *n* | | *n* |
| Interviewing† | 6.36 ± 1.16 | 201 | 6.46 ± 1.17 | 237 | 6.74 ± 1.09 | 270 | 6.91 ± 1.06 | 236 | 6.63 ± 1.14 | 944 |
| Physical examination† | 6.14 ± 1.23 | 199 | 6.28 ± 1.30 | 216 | 6.60 ± 1.28 | 278 | 6.79 ± 1.09 | 237 | 6.47 ± 1.25 | 930 |
| Professionalism† | 7.01 ± 1.01 | 245 | 7.01 ± 1.08 | 307 | 7.24 ± 0.99 | 354 | 7.37 ± 1.06 | 274 | 7.16 ± 1.05 | 1180 |
| Clinical judgment† | 6.28 ± 1.30 | 185 | 6.50 ± 1.07 | 241 | 6.77 ± 1.02 | 315 | 7.01 ± 1.15 | 258 | 6.68 ± 1.15 | 999 |
| Counseling† | 6.45 ± 1.25 | 121 | 6.73 ± 1.21 | 155 | 6.93 ± 1.13 | 225 | 7.02 ± 1.05 | 175 | 6.82 ± 1.17 | 676 |
| Organization and efficiency† | 6.22 ± 1.31 | 222 | 6.50 ± 1.21 | 253 | 6.84 ± 1.10 | 335 | 6.99 ± 1.11 | 253 | 6.67 ± 1.21 | 1063 |
| Total† | 6.40 ± 1.05 | 248 | 6.54 ± 1.03 | 319 | 6.81 ± 1.01 | 363 | 6.97 ± 0.99 | 280 | 6.69 ± 1.04 | 1210 |
| Complexity | 1.98 ± 0.66 | 226 | 1.95 ± 0.62 | 284 | 1.92 ± 0.66 | 330 | 2.04 ± 0.60 | 251 | 1.97 ± 0.64 | 1091 |

* Values with the plus/minus sign are the means ± SD.
† Kruskal–Wallis test is statistically significant at $P < 0.001$.

6% of the encounters). The complexity of patient problems was rated as low in 20% of the encounters, moderate in 54% of encounters, and high in 17% of encounters (9% were missing complexity ratings). Thirty-eight percent of the encounters took place in the inpatient setting, 52% were in the outpatient setting, 7% were in the emergency department setting, and 2% were in "other" settings (mostly the critical care unit) (2% were missing settings). Thirty-seven percent of the encounters were based on new patients, and 42% were follow-up visits (and for 22% of the encounters, this information was not specified). Each encounter called for residents to perform one or more tasks. Of the 1228 encounters, 56% required data gathering, 40% required diagnosis, 34% required therapy, and 26% required counseling.

The median (mean [±SD]) time the examiner spent observing the resident interact with the patient was 15 (18 ± 12.1) minutes, and the time spent providing feedback to the resident was 5 (7.6 ± 5.3) minutes. The amount of time spent observing ($P < 0.001$) and giving feedback ($P < 0.001$) was greater for new visits than return visits. Likewise, time spent observing ($P < 0.001$) and giving feedback ($P = 0.007$) increased with the complexity of the patient's problem.

The examiner filled in the patient's problems or diagnoses for 1131 encounters (92%). The problems covered a broad range of presenting symptoms, including abdominal pain, chest pain, cough, dizziness, fever, headache, low back pain, shortness of breath, and weight gain. The content represented common internal medicine problems, such as arthritis, asthma, chronic obstructive pulmonary disease, congestive heart failure, coronary artery disease, diabetes, and hypertension. Routine physical examinations, pelvic examinations, and breast examinations were included, as were problems such as seizure, substance abuse, depression, dementia, and rash. Many of the patients had several problems (such as congestive heart failure, hypertension, and diabetes) or acute problems (such as sepsis and myocardial infarction).

Problem complexity varied significantly by the setting ($P < 0.001$) and the type of patient ($P = 0.001$) but not the patient's age ($\rho = 0.06$; $P = 0.08$) and sex ($P = 0.11$). The mean complexity rating was 1.78 ± 0.62 for ambulatory encounters, 2.18 ± 0.58 for inpatient encounters, 2.15 ± 0.57 for the emergency department, and 2.64 ± 0.63 for "other" settings (mainly intensive care unit). The mean complexity rating was 2.02 ± 0.66 for new visits and 1.88 ± 0.62 for follow-up visits. The mean complexity rating was 1.99 ± 0.64 for male patients and 1.93 ± 0.63 for female patients.

The total score calculated by examiners varied significantly by the complexity of the patient ($P = 0.002$), the setting of the encounter ($P < 0.001$), the type of patient ($P < 0.001$), and the patient's sex ($P = 0.002$). The mean score was 6.60 ± 1.02 for low-complexity problems, 6.66 ± 1.04 for moderate-complexity problems, and 6.94 ± 1.07 for high-complexity problems. The mean score was 6.56 ± 0.92 for ambulatory settings, 6.73 ± 1.16 for inpatient settings, 7.54 ± 0.90 for emergency department settings, and 6.60 ± 0.75 for "other" settings (mainly critical care unit). The mean score was 6.82 ± 1.08 for new visits and 6.58 ± 1.01 for follow-up visits.

**Table 2** presents the means and SDs for the complexity of patient problems and the ratings during the four academic quarters starting with the summer (1 July). During this time, the complexity of patient problems did not differ significantly ($P = 0.14$). However, differences for all of the ratings were statistically significant, indicating growth throughout the year. The largest gains were in clinical judgment and organization and efficiency, and the smallest gain was in professionalism.

### Residents

The mean ratings of the 421 residents were highest for professionalism (7.1 ± 0.9) and lowest for physical examination (6.4 ± 1.1); the other ratings were 6.6 ± 1.0 for interviewing, 6.6 ± 1.0 for clinical judgment, 6.8 ± 0.9 for counseling, 6.6 ± 1.0 for organization and efficiency,

and 6.7 ± 0.9 for overall competence. The mean total score was 6.6 ± 0.9.

The correlation coefficients among the components of competence were very high and statistically significant ($P < 0.001$), ranging from 0.61 to 0.78. Similarly, the correlations between the components and the rating of overall competence ranged from 0.73 to 0.86 ($P < 0.001$). There were also small but statistically significant positive relationships between the total score and problem complexity ($\rho = 0.16$; $P = 0.001$) and the amount of time the examiner spent observing ($\rho = 0.13$; $P = 0.009$) but not the time spent giving feedback ($\rho = 0.09$; $P = 0.065$).

**Table 3** presents the CIs for the mini-CEX's total score based on 1 to 14 encounters. As expected, the CIs decrease as the number of encounters increase because the residents' scores are based on interactions with more patients and examiners. **Table 3** also provides the CI for a resident whose total score is 5.0. This number was chosen because the status of a resident with this rating could be significantly affected by measurement errors. With only one or two encounters, the CI is relatively wide and a resident with a total score of 5.0 could, on retesting, be unsatisfactory (<4) or nearly superior (>6). Ten or more encounters produce relatively tight CIs, and an increase in the number of encounters beyond this produces only small gains in consistency.

The CIs provide information that allows the number of encounters to be tailored to specific testing situations. For example, if the purpose of the evaluation is to determine which residents are satisfactory, fewer encounters are needed for residents who get very high or very low scores, while more are needed for residents who are close to passing or failing.

### Examiners

Not surprisingly, the ratings given by the 316 evaluators followed the same pattern as those for the residents. The mean ratings were highest for professionalism (7.3 ± 1.0) and lowest for physical examination (6.5 ± 1.1); the other ratings were 6.7 ± 1.0 for interviewing, 6.7 ± 1.0 for clinical judgment, 6.8 ± 1.0 for counseling, 6.7 ± 1.1 for organization and efficiency, and 6.8 ± 1.0 for overall competence. The mean total score was 6.8 ± 0.9. Higher ratings were associated with more complex patient problems ($\rho = 0.15$; $P = 0.008$) and satisfaction with the format ($\rho = 0.34$; $P < 0.001$).

As a group, the examiners were very satisfied with the mini-CEX. Their ratings ranged from 1 to 9 with a mean of 7.0 ± 1.3. More satisfied examiners tended to spend slightly more time observing the resident ($\rho = 0.12$; $P < 0.05$) but not giving feedback ($\rho = 0.09$; $P = 0.12$). In addition, there was a low positive correlation with the satisfaction ratings and the complexity of the patients' problems ($\rho = 0.21$; $P < 0.001$).

*Table 3.* **The 95% CI for the Total Score Based on 1 to 14 Encounters and for a Total Score of 5.0***

| Encounters, *n* | 95% CI | 95% CI for a Total Score of 5.0 |
|---|---|---|
| 1 | ±1.47 | 3.53–6.47 |
| 2 | ±1.04 | 3.96–6.04 |
| 4 | ±0.73 | 4.27–5.73 |
| 6 | ±0.60 | 4.40–5.60 |
| 8 | ±0.52 | 4.48–5.52 |
| 10 | ±0.46 | 4.54–5.46 |
| 12 | ±0.42 | 4.58–5.42 |
| 14 | ±0.39 | 4.61–5.39 |

* This table provides an example of how the variability for a resident with a total score of 5.0 changes with the number of mini–clinical evaluation exercise (mini-CEX) encounters completed. The components that influence the variability include the resident, program, patient, and examiner. Therefore, the CIs would probably be different for another sample of programs, patients, and examiners.

## Discussion

Our study replicated previous work on the mini-CEX with a larger and more representative group of training programs. Compared with the traditional CEX, the results again showed that the mini-CEX, with its multiple encounters, evaluated residents in a greater variety of clinical settings with a much more diverse set of patient problems. Furthermore, the multiple encounters with different examiners and patients produced reliable ratings, and the data presented in **Table 3** can help tailor it to particular measurement situations. In addition to its role in evaluation, the mini-CEX increased the opportunity for education because each resident had several interactions with attending role models and received feedback as part of each.

Our study also expanded on previous work concerning the mini-CEX. Most important, the ratings were compared over the four quarters of the first year of training, demonstrating growth in all aspects of competence. This supports the validity of the method, especially given the finding that there was no difference in the complexity of patient problems over time. Consistent with the type of learning that occurs during the first year of training, the biggest gains were in clinical judgment and organization and efficiency. Although professionalism showed the smallest improvement, it is reassuring that some growth occurred despite the rigors of the first year of residency training.

In the first study, some participants encountered logistic problems with the mini-CEX because they treated it in the same way as the traditional CEX and attempted to schedule each encounter. In this study, the programs experimented with administrative procedures, and some programs discovered more efficient methods. For instance, one participant asked his faculty to perform a mini-CEX with a resident's first visit of the day. This rarely disrupted the flow of the clinic but allowed for multiple observations over time.

Our study has several limitations. Although the sample of training programs is relatively large, it has an East Coast bias and is not completely representative of the population.

Likewise, the number of evaluators involved was sizeable, but 32% conducted only one evaluation and their background and training in the method are unknown. Furthermore, there was a small but statistically significant correlation between the complexity of the encounter and the ratings given by examiners. This suggests that examiners may have been slightly overcompensating for more challenging patient problems.

For many of the variables, the amount of missing data was substantial. In most instances, the questions were simply not applicable to the encounter, but in other instances, the reasons for missing data are unclear. This problem was exacerbated by the fact that residents and evaluators were nested within training programs, they were crossed in some instances, many had only one observation, and the encounters were not evenly spaced throughout the year. This made it difficult to conduct multivariate analyses without significant data loss. It also means that the simple analyses reported here are subject to various biases and should be repeated in controlled settings.

Despite these limitations, the multiple-encounter mini-CEX is superior to the traditional CEX as an evaluation device, and its measurement characteristics are similar to those of other performance assessments, such as standardized patients and the standardized oral examination (12, 13). Unlike these assessments, however, the difficulty of the examination will vary with the patients that a resident encounters. The magnitude of this effect is mitigated to some degree by the examiners, who slightly overcompensate for patient difficulty, and by the fact that each resident interacts with several patients. Furthermore, compared with other formats, such as standardized patients, the mini-CEX has higher fidelity, permits evaluation based on a much broader set of clinical settings and patient problems, is administered on site, and is less expensive.

The increase in ratings throughout the year is consistent with growth expected as an outcome of training. Therefore, it would not be appropriate to compare a resident with four encounters in July to a resident with four encounters the following June. However, average annual growth is not very great in absolute terms (about 0.5 point on the 9-point scale), so ratings should be roughly comparable so long as all encounters do not occur at the beginning or end of the year.

Consistent with previous work, the examiners were satisfied with the new format. Their satisfaction was correlated with their evaluations, their ratings of problem complexity, and their time spent observing the resident. This may result from a halo effect or the fact that examiners found the experience more rewarding when the resident was good and faced with a complex patient problem. It may also reflect difficulty among the examiners in giving negative feedback.

Future research should focus on at least two issues. First, recognizing that there is no gold standard against which to judge performance, additional work is needed to establish the validity of the mini-CEX through its relationships with other markers of competence. For instance, data from this study will be compared with ABIM scores and program directors' ratings, as they become available. Second, additional work needs to be done on the reliability of the mini-CEX. The error estimates provided in this paper include the effects of examiner differences but cannot isolate them. Consequently, it is not known whether extensive efforts to train the examiners would significantly reduce the magnitude of the CIs reported here.

Finally, in addition to being superior to the traditional CEX as an evaluation device, the mini-CEX has advantages as an educational strategy. It does not permit observation of and feedback for the complete history and physical examination. However, it does ensure that different faculty members observe a reasonable sample of the resident's clinical skills over time. Moreover, the observation and feedback occur with a broad range of patient problems in various settings.

From the Foundation for Advancement of International Medical Education and Research and the American Board of Internal Medicine, Philadelphia, Pennsylvania.

**Disclaimer:** This research was supported by the American Board of Internal Medicine but does not necessarily reflect its opinions.

**Acknowledgments:** The authors thank the following participants: Edward Bollard, MD, and Richard J. Simons Jr., MD, Penn State College of Medicine and Milton S. Hershey Medical Center; R. Michael Buckley, MD, Pennsylvania Hospital; Rand David, MD, Elmhurst Hospital Center and Mt. Sinai School of Medicine; William Farrer, MD, Seton Hall University; Susan D. Grossman, MD, and Cynthia Wong, MD, St. Vincent's Catholic Medical Center of New York, Staten Island Region; Sheik N. Hassan, MD, Howard University Hospital; Eric Holmboe, MD, National Naval Medical Center; Brenda Horwitz, MD, Temple University Hospital; Stephen J. Huot, MD, PhD, Yale Primary Care Internal Medicine Residency; Gregory Kane, MD, Jefferson Medical College; David G. Kemp, MD, Easton Hospital; Nayan Kothari, MD, Robert Wood Johnson Medical School; Frank Kroboth, MD, University Health Center of Pittsburgh, Montefiore University Hospital; Jeanne Macrae, MD, State University of New York Health Center at Brooklyn; Dragica Mrkoci, MD, The George Washington University Medical Center; Richard S. Rees, MD, New York Harbor Veterans Affairs Health Care System Medical Service; Steven Reichert, MD, Englewood Hospital and Medical Center; David G. Smith, MD, Abington Hospital; Sara L. Wallach, MD, Monmouth Medical Center; Frederick K. Williams, MD, Washington Hospital Center; Jack Boulet, PhD; William Burdick, MD; Danette McKinley; and Gerald P. Whelan, MD, Educational Commission for Foreign Medical Graduates.

**Potential Financial Conflicts of Interest:** None disclosed.

**Requests for Single Reprints:** John J. Norcini, PhD, Foundation for Advancement of International Medical Education and Research, 3624 Market Street, 4th Floor, Philadelphia, PA 19104; e-mail, jnorcini@faimer.org.

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

## References


1. Guide to Evaluation of Residents in Internal Medicine—A Systems Approach. Philadelphia: American Board of Internal Medicine; 1994.

2. **Day SC, Grosso LJ, Norcini JJ Jr, Blank LL, Swanson DB, Horne MH.** Residents' perception of evaluation procedures used by their training program. J Gen Intern Med. 1990;5:421-6. [PMID: 2231039]

3. **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]

4. **Elstein AS, Shulman LS, Sprafka SA.** Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge: Harvard Univ Pr; 1978.

5. **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]

6. **Woolliscroft JO, Stross JK, Silva J Jr.** Clinical competence certification: a critical appraisal. J Med Educ. 1984;59:799-805. [PMID: 6481776]

7. **Norcini JJ, Blank LL, Arnold GK, Kimball HR.** The mini-CEX (clinical evaluation exercise): a preliminary investigation. Ann Intern Med. 1995;123: 795-9. [PMID: 7574198]

8. **Norcini JJ, Blank LL, Arnold GK, Kimball HR.** Examiner differences in the mini-CEX. Advances in Health Sciences Education: Theory and Practice. 1997; 1:27-33.

9. **Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N.** The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley; 1972.

10. **Brennan RL.** Elements of Generalizability Theory. Iowa City, IA: American College Testing Publications; 1983.

11. **Crocker L, Algina J.** Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart and Winston; 1986.

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

13. **Maatsch JL, Huang R, Downing SM, Munger BS.** Studies of the reliability and validity of examiner assessments of clinical performance: what do they tell us about clinical competence? In: Hart IR, Harden RM, Walton HJ, eds. Newer Developments in Assessing Clinical Competence. Montreal: Heal Publications; 1986.


**Personae Prize**

***Annals of Internal Medicine* is offering a $500 prize for the best photograph submitted to *Annals* in 2003. In an effort to bring people to the pages and cover of *Annals*, the editors began publishing a section called Personae in 1999. Personae are black and white photographs of people that appeared in the body of the journal from 1999 to 2000 and on the cover since 2000. We invite the submission of photographs that catch people in the context of their lives and that capture personality. The images should speak for themselves, so we do not publish accompanying titles or captions.**

***Annals* will publish photographs in black and white, and black-and-white submissions are preferred. We will also accept color submissions, but the decision to publish a photograph will be made after the image is converted to black and white. Slides, prints, and digital photographs are acceptable. Print sizes should be standard (3″ × 5″, 4″ × 6″, 5″ × 7″, 8″ × 10″). Photographers should send two copies of each photograph. We cannot return photographs, regardless of publication. We must receive written permission to publish the photograph from the subject (or subjects) of the photograph or the subject's guardian if he or she is a child. The editors may make occasional exceptions to this requirement for photographs taken in public places where the identity of the subject is unknown. A cover letter assuring no prior publication of the photograph and providing permission from the photographer for *Annals* to publish the image must accompany all submissions. The letter must also contain the photographer's name, academic degrees, institutional affiliation, mailing address, and telephone and fax numbers.**

**Please submit photographs or questions to Christine Laine, Senior Deputy Editor, *Annals of Internal Medicine*, 190 N. Independence Mall West, Philadelphia, PA 19106-1572, claine@acponline.org. We look forward to receiving your photographs.**

**Current Author Addresses:** Dr. Norcini: Foundation for Advancement of International Medical Education and Research, 3624 Market Street, 4th Floor, Philadelphia, PA 19104.
Ms. Blank, Dr. Duffy, and Mr. Fortna: American Board of Internal Medicine, 510 Walnut Street, Suite 1700, Philadelphia, PA 19106-3699.

## Appendix: Generalizability Theory Analyses

Generalizability theory systematically identifies and quantifies errors of measurement in educational tests (9–11). For this study, a random-effects, encounter-within-resident design was the basis for the calculation of the variance components. In a well-designed and controlled study, it would be possible to include in the analysis the effects of various factors, such as the evaluators, training programs, or occasions, to determine how much each contributes to measurement error. In this naturalistic study, however, residents and evaluators were nested within training programs, occasionally they were crossed, many had only one observation, and the encounters were not well spaced throughout the year. Thus, a very simple analysis was performed, but it is subject to various potential biases and should be repeated in a controlled setting with more sophisticated analyses.

The total scores were submitted to the VARCOMP procedure (SAS Institute, Inc., Cary, North Carolina), and the MIVQUE0 method was used for estimation. The variance component for residents was 0.520 (48% of the total variance). This is the variability in ratings that would occur if each resident were examined by a large number of evaluators while seeing a large number of patients. The error variance component was 0.559 (52% of the total variance). This is the within-resident variation in ratings that would occur over a very large number of encounters with different patients and evaluators.

These data were used to generate the CIs reported in Table 3. To obtain the confidence intervals for a resident's total score, the error variance was divided by the number of encounters (from 1 to 14 encounters), the square root was taken and multiplied by 1.96. Adding or subtracting this from a resident's total score produced the range within which the resident is expected to fall within 95 times, if independent retesting occurred 100 times.

*Appendix Figure.* **The mini–clinical evaluation exercise (*mini*-CEX) form.**

Evaluator: _____ Date: _____

Fellow: _____ ○ R-1 ○ R-2 ○ R-3

Patient Problem/Dx:

Setting: ○ Ambulatory ○ In-patient ○ ED ○ Other

Patient: Age: _____ Sex: _____ ○ New ○ Follow-up

Complexity: ○ Low ○ Moderate ○ High

Focus: ○ Data gathering ○ Diagnosis ○ Therapy ○ Counseling

1. Medical interviewing skills (○ Not observed)

1 2 3 | 4 5 6 | 7 8 9

Unsatisfactory | Satisfactory | Superior

2. Physical examination skills (○ Not observed)

1 2 3 | 4 5 6 | 7 8 9

Unsatisfactory | Satisfactory | Superior

3. Humanistic qualities/professionalism

1 2 3 | 4 5 6 | 7 8 9

Unsatisfactory | Satisfactory | Superior

4. Clinical judgment (○ Not observed)

1 2 3 | 4 5 6 | 7 8 9

Unsatisfactory | Satisfactory | Superior

5. Counseling skills (○ Not observed)

1 2 3 | 4 5 6 | 7 8 9

Unsatisfactory | Satisfactory | Superior

6. Organization/efficiency (○ Not observed)

1 2 3 | 4 5 6 | 7 8 9

Unsatisfactory | Satisfactory | Superior

Overall clinical competence (○ Not observed)

1 2 3 | 4 5 6 | 7 8 9

Unsatisfactory | Satisfactory | Superior

Mini-CEX time: Observing: _____ Min Providing feedback: _____ Min

Evaluator satisfaction with mini-CEX

Low 1 2 3 4 5 6 7 8 9 High

Resident satisfaction with mini-CEX

Low 1 2 3 4 5 6 7 8 9 High

Comments:

_____ _____

Resident signature Evaluator signature

Dx = diagnosis; ED = emergency department; min = minutes; R-1 = first-year resident; R-2 = second-year resident; R-3 = third-year resident.