The Mini-CEX (Clinical Evaluation Exercise): A Preliminary Investigation

John J. Norcini, PhD; Linda L. Blank, BA; Gerald K. Arnold, PhD; and Harry R. Kimball, MD

■ *Objective:* To gather preliminary data on the mini-CEX (clinical evaluation exercise), a device for assessing the clinical skills of residents.

■ *Design:* Evaluation of residents by faculty members using the mini-CEX.

■ *Setting:* 5 internal medicine training programs in Pennsylvania.

■ *Participants:* 388 mini-CEX encounters involving 88 residents and 97 evaluators.

■ *Measurements:* A mini-CEX encounter consists of a single faculty member observing a resident while that resident conducts a focused history and physical examination in any of several settings. After asking the resident for a diagnosis and treatment plan, the faculty member rates the resident and provides educational feedback. The encounters are intended to be short (about 20 minutes) and to occur as a routine part of training so that each resident can be evaluated on several occasions by different faculty members.

■ *Results:* The encounters occurred in both inpatient and ambulatory settings and were longer than anticipated (median duration, 25 minutes). Residents saw either new or follow-up patients who collectively presented with a broad range of clinical problems. The median evaluator assessed two residents and was generally satisfied with the mini-CEX format; residents were even more satisfied with the format. The reproducibility of the mini-CEX is higher than that of the traditional CEX, and its measurement characteristics are similar to those of other test formats, such as standardized patients and standardized oral examinations.

■ *Conclusions:* The mini-CEX assesses residents in a much broader range of clinical situations than the traditional CEX, has better reproducibility, and offers residents greater opportunity for observation and feedback by more than one faculty member and with more than one patient. On the other hand, the mini-CEX may be more difficult to administer because multiple encounters must be scheduled for each resident. Exclusive use of the mini-CEX also prevents residents from being observed while doing a complete history and physical examination. Given the promising results and measurement characteristics of the mini-CEX, however, the American Board of Internal Medicine encourages the use of this method in conjunction with or as an alternative to the traditional CEX.

Ann Intern Med. 1995;123:795-799.

From the American Board of Internal Medicine, Philadelphia, Pennsylvania. For current author addresses, see end of text.

In 1972, the American Board of Internal Medicine abandoned the oral examination of residents for logistic and psychometric reasons. It then delegated to program directors the task of evaluating the essential components of residents' clinical competence, including clinical skills. Over the years, the American Board of Internal Medicine has worked with program directors to develop efficient, effective local evaluation systems, and it has recommended that the clinical evaluation exercise (CEX) be part of the process (1). The traditional CEX is conducted by an experienced physician who observes a resident while that resident interviews a single patient (unfamiliar to the resident), does a complete physical examination, presents findings, and plans the patient's management. After the exercise, the evaluator gives the resident substantive feedback and documents the experience on a form provided by the Board. Later, the resident gives the evaluator a written record of the patient work-up for review. The traditional CEX takes about 2 hours. Approximately 82% of residents have one such evaluation during their first year of training, and a much smaller percentage (32%) have more than one (2).

As a measurement device, the traditional CEX is limited in three ways. First, the resident is observed by only one evaluator, and studies have shown that even experienced physicians differ from one another when observing exactly the same events (3). Second, the resident is observed with only one patient; patient problems vary considerably, and this is probably one reason why physician performance in one case does not predict performance in others (4). Third, because most physician–patient encounters are short and focused, the traditional CEX, with its emphasis on completeness, is a somewhat less relevant measure of clinical skill. The unreliability of the observer, the variation of resident performance from patient to patient, and the artificiality of the task mean that the traditional single-interaction CEX is not a dependable measure of a resident's clinical competence. As a result of these shortcomings, the reproducibility or generalizability of the CEX ratings has been found to be less than 0.30, and the CEX has been appropriately criticized as a measurement instrument (3, 5, 6). Our goal was to respond to those concerns by exploring the logistic and psychometric properties of a variation on the traditional CEX format, called the mini-CEX.

The mini-CEX is designed around both the skills that residents most often need in actual patient encounters and the educational interactions that attending physicians routinely have with residents during teaching rounds. A single faculty member observes and evaluates a resident while that resident conducts a focused history and physical examination (expected to take about 20 minutes). Because the interaction should be relatively short and because it occurs as a natural part of the training envi-

ronment, each resident can be evaluated on several occasions by different faculty members. In this study, we report initial data on the psychometric and logistic characteristics of the mini-CEX.

Methods

Instrumentation

For each mini-CEX, a single faculty member observed and evaluated a resident while that resident conducted a focused history and physical examination in an inpatient, outpatient, or emergency department setting. After asking the resident for a diagnosis and treatment plan, the faculty member completed a short evaluation form and gave the resident feedback. For most residents, the mini-CEXs were collected within a 4-month period. All formal mini-CEX evaluation data were collected on a two-sided, 1-page form that was the same for all study sites. Information was gathered from the resident on level of training and satisfaction with the mini-CEX; the latter was rated on a 9-point scale anchored with the words "lowest" and "highest." Ratings were collected from the evaluator for the resident's overall clinical competence and for four components of competence: history-taking skill, physical examination skill, clinical judgment and synthesis, and humanistic qualities. The ratings were made on a 9-point scale, on which 1, 2, and 3 were unsatisfactory; 4 was marginal; 5 and 6 were satisfactory; and 7, 8, and 9 were superior. A box was also provided so that the evaluator could select "insufficient contact to judge." For the purposes of our study, we averaged the four component skill ratings, and the result is referred to as the calculated average score.

In addition to providing data on the resident, the evaluator recorded information on the site of the evaluation (inpatient service, clinic, or emergency department) and the patient's major medical problems and diagnoses. Unlike the traditional CEX, which focuses on an exhaustive assessment, the mini-CEX concentrates on the resident's ability to solve patient problems regardless of the nature of the encounter. Therefore, new and return visits with a particular resident were included and identified. Finally, the evaluator was asked to rate his or her level of satisfaction with the mini-CEX format.

Because each resident was evaluated on several occasions by different faculty members, the mini-CEX certainly differs from the traditional CEX in terms of quantity. However, it also differs qualitatively. The traditional CEX rewards the resident for being thorough and thoughtful in an environment unconstrained by the realities of clinical practice. In contrast, the exact nature of any mini-CEX encounter is more variable, because the challenge it poses is more dependent on the particular setting and patient. Consequently, this method of evaluation rewards the resident's ability to focus and prioritize diagnosis and management within the context of clinical practice.

Participants

Five residency programs in Pennsylvania volunteered to assess residents using the mini-CEX. In general, the mix of programs reflects the heterogeneity found in residency education.

Of the 397 useable evaluation forms collected, 9 were for residents who had received only a single mini-CEX. These 9 evaluations were excluded from the study to permit calculation of the psychometric characteristics of the format, so 388 evaluations formed the basis for analysis. The evaluations reflected the performance of 88 residents; each resident had between 2 and 10 evaluations (mean, 4.4 ± 2.2; median, 4). These evaluations were collected throughout the year of training; 78 of the residents (89%) were in their first year, 9 (10%) were in their second year, and 1 (1%) was in his or her third year. The process incorporated the efforts of 97 attending physicians, each of whom conducted between 1 and 26 evaluations (mean, 4.0; median, 2.0). Thirty-three attending physicians each conducted only 1 evaluation.

Analysis

Our results are reported in three parts. First, we describe the nature of the 388 encounters in terms of the setting of the mini-CEX, the types of patients seen, and the patients' diagnoses or problems. Second, we examine the mini-CEX from the perspective of the 97 evaluators, emphasizing their ratings of the residents and their satisfaction with the mini-CEX format. Finally, we analyze the performance of the 88 residents in terms of their competence ratings, the reproducibility of those ratings, and resident satisfaction with the mini-CEX format. Results are given ± SD.

Results

Nature of the Encounters

Of the 388 encounters, 54% occurred in an inpatient setting, 38% in an outpatient setting, and 14% in the emergency department (information on setting was missing for 4% of encounters). The overall competence ratings were similar in all settings ($P = 0.59$, data not shown), but the duration of the encounter differed significantly according to setting ($P = 0.003$). The average inpatient encounter lasted 34.9 minutes, the average outpatient encounter lasted 27.7 minutes, and the average emergency department encounter lasted 28.2 minutes.

Fifty-eight percent of the encounters represented the first visit of a patient to a particular resident, and 36% were return visits to that resident. There were no statistically significant differences in overall clinical competence scores or component scores for encounters with new and returning patients. The average first visit (35.3 minutes) was longer than the average follow-up visit (25.3 minutes; $P < 0.0001$).

As a group, the mini-CEXs took between 5 and 100 minutes (mean, 31.5 ± 19.3 minutes; median, 25 minutes). There were small but statistically significant Pearson product moment correlations of 0.10 to 0.16 between the duration of the evaluation and performance ratings ($P \leq 0.05$), indicating that longer encounters were generally associated with higher resident performance ratings. Duration also correlated with satisfaction with the mini-CEX format on the part of residents (0.19, $P = 0.002$) and faculty members (0.27, $P < 0.001$).

For 239 of the 388 encounters (62%), the patient's problems or diagnoses were specified by the evaluator; they covered a broad range and included a representative array of common problems such as hypertension, diabetes, congestive heart failure, coronary artery disease, arthritis, chronic obstructive pulmonary disease, abdominal pain, and peptic ulcer. In addition, several patients had problems that crossed from traditional internal medicine into other disciplines, including neurology (seizures), substance abuse (alcohol and drugs), psychiatry (depression and schizophrenia), dermatology (rashes), gynecology (ovarian cysts), prevention (obesity or smoking), and ophthalmology (cataracts). Many patients, particularly elderly persons and persons with the acquired immunodeficiency syndrome, had multiple problems. Some patients had acute problems, such as myocardial infarction or respiratory failure.

Evaluator Performance

The mean ratings given by the 97 evaluators were 6.6 or 6.7 for history-taking skill, physical examination skill,

clinical judgment and synthesis, and overall competence; the mean rating for humanistic qualities was 7.2. The correlations among the components of competence ranged from 0.65 to 0.81 ($P < 0.001$), and the correlations between the components and the overall competence rating ranged from 0.61 to 0.68 ($P < 0.001$).

The evaluators were generally satisfied with the mini-CEX format; their ratings ranged from 2 to 9 (median, 6.0; mean, 6.1). These ratings had a Pearson product moment correlation of 0.21 ($P < 0.05$) with overall competence, indicating that satisfaction with the format was related to resident performance.

Resident Performance

The mean ratings of the 88 residents were 6.5 ± 0.6 for history-taking skill, physical examination skill, and clinical judgment and synthesis; 7.0 ± 0.7 for humanistic qualities; 6.5 ± 0.6 for overall competence; and 6.6 ± 0.6 for the calculated average score. As with the evaluators, correlations among the components were high and statistically significant ($P < 0.001$); they ranged from 0.62 to 0.81. Similarly, the correlations between the components and the overall competence ratings were high; they ranged from 0.66 to 0.90 ($P < 0.001$).

When a test is administered, it is useful to know that the same person would receive the same score if tested again. This is called reproducibility, and it was estimated for the overall clinical competence ratings and the calculated average scores using generalizability theory (7, 8). First, the universe score variance component was computed for overall clinical competence (0.1586) and for the calculated average score (0.1510); this is the variation in ratings that would be seen if each resident could be assessed by a large number of evaluators as that resident interacted with a large number of patients. The error variance component was also computed for overall clinical competence (0.7394) and for the calculated average score (0.4991); this is the within-resident variation in ratings that would be seen over many encounters with different patients and evaluators. For an assessment composed of several mini-CEX encounters, the error variance is divided by the number of encounters.

These data were used to generate the reproducibility coefficients and standard errors of measurement (Table 1) for 1 to 14 encounters. The reproducibility coefficients are ratios of universe score variance to total variance; thus, they range from 0 (all error) to 1.0 (no error). The standard error of measurement is simply the square root of the error variance, so adding and subtracting two times (rounded from 1.96) the standard error of measurement from a resident's rating produces a 95% CI.

Between 12 and 14 encounters are required for the mini-CEX to reach a reproducibility of 0.80, and this number of encounters is similar to that needed for similar formats (such as standardized oral examinations and standardized patients) (9, 10). However, the CIs provide additional information that permits test length to be tailored to specific situations. For example, the average resident in our study had roughly four encounters that produced a calculated average score of 6.6. The reproducibility of the score is only 0.55, but the standard error of measurement is 0.35. The 95% CI is ± 0.70 (2 × 0.35), or 5.9 to 7.3 (6.6 − 0.7 = 5.9, and 6.6 + 0.7 = 7.3). Such precision is sufficient for many measurement purposes.

Table 1. The Reproducibility and Standard Error of Measurement of the Rating of Overall Competence and of the Calculated Average Score for 1 to 14 Encounters

Encounters, *n*	Rating of Overall Competence		Calculated Average Score	
	Reproducibility	Standard Error of Measurement	Reproducibility	Standard Error of Measurement
1	0.18	0.86	0.23	0.71
2	0.30	0.61	0.38	0.50
4	0.46	0.43	0.55	0.35
6	0.56	0.35	0.65	0.29
8	0.63	0.30	0.71	0.25
10	0.68	0.27	0.75	0.22
12	0.72	0.25	0.78	0.20
14	0.75	0.23	0.81	0.19

Because there were so few second-year ($n = 9$) and third-year ($n = 1$) residents, our study did not have considerable power to detect differences in ratings related to level of training. Nonetheless, the second-year residents did significantly better than the first-year residents for mean overall clinical competence (6.98 compared with 6.45; $P < 0.05$) and for the mean calculated average score (7.09 compared with 6.55; $P < 0.05$).

Finally, the residents were more satisfied than the evaluators with the new mini-CEX format ($P < 0.05$). Their ratings ranged from 1 to 9 (mean, 6.6; median, 7). Unlike that of the evaluators, the satisfaction of the residents was not correlated with the overall clinical competence ratings ($r = 0.09$; $P = 0.46$).

Discussion

Our purpose was to report preliminary logistic and psychometric data for the mini-CEX format, which clearly adapts itself to a broad range of clinical situations, has reasonable reproducibility, and is satisfactory according to both evaluators and residents. Three aspects of our study deserve particular note. First, the mini-CEX format was applied successfully in various settings that are directly relevant to the day-to-day activities of residents. Second, the range of patient problems faced by residents was broad and included many conditions that directly link traditional internal medicine with other disciplines within the purview of the generalist. Third, the format encouraged education as well as evaluation, because each resident had several interactions with attending role models and received feedback.

On the other hand, some of the program directors who participated in the study found the administration of the mini-CEX burdensome because they felt the need to schedule multiple encounters for each resident. This burden could be alleviated by encouraging more informal encounters that would reduce the need to schedule and that would allow a broader sampling of resident skills during routine teaching rounds. Internal medicine educators have been calling for such an increase in bedside teaching. A second limitation is that if the mini-CEX

were to be used exclusively, residents would not be observed while doing a complete history and physical examination. Finally, the mini-CEX encounters in our study took slightly longer than anticipated, particularly those in inpatient settings. This may be because of evaluator inexperience with the format or it may be because some encounters required additional time to be educationally meaningful; this area merits further study. Regardless, four mini-CEX encounters of about 30 minutes each would require the same faculty and resident resources as the traditional CEX.

Our study should be replicated for several reasons. First, the number of residents who participated in our study was relatively small, a convenience sample, and this cohort may not be completely representative of broader populations. Second, many evaluators were involved, but roughly one third ($n = 33$) each conducted only one mini-CEX, and their characteristics (whether they were practitioners or subspecialists, for example) were unknown. Third, for some variables, such as specific patient problems, large amounts of data were missing. Fourth, the participants were instructed to experiment with the mini-CEX format in different settings with different problems. This is appropriate in an exploratory study such as ours, but controlled investigations are now needed. Fifth, the study design did not permit estimation of whether the number of evaluators or the number of encounters would have a greater effect on reproducibility. However, previous work suggests that the number of encounters is more important than the number of evaluators, given that there are at least a few of the latter (9). Sixth, because residents were evaluated over time, some of the variability in their performance reflects growth in skills rather than random error; actual standard errors may be lower than those reported here. Finally, we collected little information on the validity of the mini-CEX. Although it is reasonable to conclude that the content validity is high and that first-year and second-year residents differ significantly, relations between mini-CEX scores and other markers of competence would provide important supportive evidence.

Despite these limitations, the reproducibility of the mini-CEX is higher than that of the traditional CEX, and its measurement characteristics are similar to those of other test formats, such as standardized patients and standardized oral examinations (9, 10). The evaluators differed in stringency, but the effects of this on resident ratings are reduced when multiple assessments are collected. The means and SDs of the ratings are similar to those obtained for national groups, and the high correlations among components of competence are also typical (11). The ratings for residents in different years of training also differed significantly.

Our findings on the reproducibility of the mini-CEX are also consistent with those in the broader literature: Ten or more encounters are needed to reach a reproducibility of 0.80 (9–12). However, the data on the CIs are more important because they make clear that, even with only few encounters, useful information can be gathered. For example, a resident who had a calculated average score of 7 after four encounters might see her score increase to 7.7 or decrease to 6.3 after 12 encounters, but greater changes are unlikely. If the difference between a score of 6.3 (satisfactory) and a score of 7.7 (superior) is of little practical significance, four encounters are sufficient for that resident. On the other hand, for the resident who has a score of 4 (marginal) after four encounters, it would be of practical significance whether he would have a score of 3.3 (unsatisfactory) or 4.7 (satisfactory) after 12 encounters. Thus, it is reasonable to seek additional encounters in such a case. It also makes educational sense to increase the number of encounters for a resident judged to be borderline, because the mini-CEX provides for feedback at the end of each interaction and thus provides the opportunity to learn and correct deficiencies. As an aside, it is worth noting that even an overall rating of 9 on the traditional CEX would not produce the same level of confidence that a resident is other than marginal (5).

The evaluators were generally satisfied with the new format. Their level of enthusiasm correlated with their ratings of resident performance, suggesting that a "halo effect" was operating or that satisfaction was influenced by how well the residents did. Residents were more satisfied than evaluators with the mini-CEX, but they were probably unfamiliar with their ratings when they made their judgments. Their satisfaction may be related to increased educational interactions with faculty during the exercise. The mini-CEX format may also produce less anxiety than the traditional CEX format, because the assessment is less formal and less dependent on a single, high-stakes encounter with one faculty member and one patient.

Finally, our results clearly show that, as a measurement device, the mini-CEX is superior to the traditional CEX. As an educational exercise, however, each method has advantages and disadvantages. The traditional CEX permits observation of and feedback on a complete history and physical examination, but it does so with only one evaluator and one patient at one point in time. The mini-CEX does not permit observation and feedback on a complete history and physical examination, but it does give the resident instruction from several faculty members on different patients over time. Consequently, the American Board of Internal Medicine plans to endorse the use of the mini-CEX in conjunction with, or as an alternative to, the traditional CEX.

Acknowledgments: The authors thank the following program directors for their participation: Dr. Robert L. Benz (Lankenau Hospital, Philadelphia, Pennsylvania); Drs. John J. Kelly and David G. Smith (Abington Memorial Hospital, Abington, Pennsylvania); Dr. Oksana M. Korzenowski (Medical College of Pennsylvania, Philadelphia, Pennsylvania); Dr. Frank L. Kroboth (University of Pittsburgh, Pittsburgh, Pennsylvania); and Dr. Lisa J. Wallenstein (Albert Einstein Medical Center, Philadelphia, Pennsylvania). The authors thank Nancy L. Grant and Jane M. Luistro for their help in coordinating the study.

Grant Support: This work was supported by the American Board of Internal Medicine but does not necessarily reflect the views or opinions of that Board.

Requests for Reprints: John J. Norcini, PhD, American Board of Internal Medicine, 3624 Market Street, Philadelphia, PA 19104-2675.

Current Author Addresses: Drs. Norcini, Arnold, and Kimball and Ms. Blank: The American Board of Internal Medicine, 3624 Market Street, 2nd Floor, Philadelphia, PA 19104-2675.

References

1. Guide to Evaluation of Residents in Internal Medicine—A Systems Approach. Philadelphia: American Board of Internal Medicine; 1994.
2. **Day SC, Grosso LG, Norcini JJ Jr, Blank LL, Swanson DB, Horne**

MH. Residents' perceptions of evaluation procedures used by their training program. J Gen Intern Med. 1990;5:421-6.
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.
4. **Elstein AS, Shulman LS, Sprafka SA.** Medical Problem-Solving: An Analysis of Clinical Reasoning. Cambridge, MA: 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.
6. **Woolliscroft JO, Stross JK, Silva J Jr.** Clinical competence certification: a critical appraisal. J Med Educ. 1984;59:799-805.
7. **Crocker L, Algina J.** Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart & Winston; 1986.
8. **Brennan RL.** Elements of Generalizability Theory. Iowa City, IA: American College Testing Publications; 1983.
9. **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.
10. **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.
11. **Norcini JJ, Webster GD, Grosso LJ, Blank LL, Benson JA Jr.** Ratings of residents' clinical competence and performance on certification examination. J Med Educ. 1987;62:457-62.
12. **Gao XH, Shavelson RJ, Baxter GP.** Generalizability of large-scale performance assessments in science: promises and problems. Applied Measurement in Education. 1994;7:323-42.

It is like that in the woods and even in the wide world generally—the rescue of men and women, alive or dead, comes first. Of course, some step on the gas and leave them lying on the pavement, where they landed and some sneak off; like Egyptian bas-relief with their profiles looking one way and their bodies going the other way. But most people think they can be of help, and some even seem born to rescue others, as poets think they are. The best of them goof, especially at first, because only a few have the opportunity to keep them in practice. Then as they catch on again they become beautiful in performance if one can step back for a moment to look. Almost as beautiful as when, having completed their job of deposing death, they fade into complete anonymity.

Norman Maclean
Young Men and Fire
Chicago: Univ Chicago Pr; 1992

Submitted by:
Paul M. Ford, MD
Stanford Medical Group
Palo Alto, CA 94303-2205

Submissions from readers are welcomed. If the quotation is published, the sender's name will be acknowledged. Please include a complete citation, as done for any reference.—*The Editors*