Adv in Health Sci Educ (2010) 15:633–645
DOI 10.1007/s10459-010-9224-9

Internal structure of mini-CEX scores for internal medicine residents: factor analysis and generalizability

David A. Cook · Thomas J. Beckman · Jayawant N. Mandrekar · V. Shane Pankratz

Received: 2 October 2009 / Accepted: 1 February 2010 / Published online: 21 February 2010

Abstract The mini-CEX is widely used to rate directly observed resident-patient encounters. Although several studies have explored the reliability of mini-CEX scores, the dimensionality of mini-CEX scores is incompletely understood. *Objective*: Explore the dimensionality of mini-CEX scores through factor analysis and generalizability analysis. *Design*: Factor analytic and generalizability study using retrospective data. *Participants*: Eighty five physician preceptors and 264 internal medicine residents (postgraduate years 1–3). *Methods*: Preceptors used the six-item mini-CEX to rate directly observed resident-patient encounters in internal medicine resident continuity clinics. We analyzed mini-CEX scores accrued over 4 years using repeated measures analysis of variance to generate a correlation matrix adjusted for multiple observations on individual residents, and then performed factor analysis on this adjusted correlation matrix. We also performed generalizability analyses. *Results*: Eighty-five preceptors rated 264 residents in 1,414 resident-patient encounters. Common factor analysis of these scores after adjustment for repeated measures revealed a single-factor solution. Cronbach's alpha for this single factor (i.e. all six mini-CEX items) was $\geq$0.86. Sensitivity analyses using principal components and other method variations revealed a similar factor structure. Generalizability studies revealed a reproducibility coefficient of 0.23 (0.70 for 10 raters or encounters). *Conclusions*: The mini-CEX appears to measure a single global dimension of clinical competence. If educators desire to measure discrete clinical skills, alternative assessment methods may be required. Our approach to factor analysis overcomes the limitation of repeated observations on subjects without discarding data, and may be useful to other researchers attempting factor analysis of datasets in which individuals contribute multiple observations.

D. A. Cook (✉) · T. J. Beckman
Division of General Internal Medicine and Office of Education Research, Mayo Clinic College of Medicine, Baldwin 4-A, 200 First Street SW, Rochester, MN 55905, USA
e-mail: cook.david33@mayo.edu

J. N. Mandrekar · V. S. Pankratz
Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55905, USA

Keywords Medical education · Educational measurement · Clinical competence · Assessment · Reproducibility of results · Psychometrics · Internal consistency · Factor analysis

Introduction

The mini-clinical evaluation exercise (mini-CEX) is widely used to assess physician trainees during directly observed clinical encounters. In a typical mini-CEX, (Norcini et al. 1995) the supervising physician observes a trainee during a patient encounter and rates competence in six domains (interviewing, physical examination, humanistic qualities/professionalism, clinical judgment, counseling, and organization/efficiency) and overall clinical competence.

Research on the validity of mini-CEX scores has evaluated multiple sources of validity evidence, (Kogan et al. 2009) including the reliability and accuracy of ratings. Although interrater reliability is modest, overall reliability is acceptable if several encounters are averaged (Norcini et al. 2003; Kogan et al. 2003; Hatala et al. 2006). Research has also found that mini-CEX scores discriminate unsatisfactory, satisfactory, and superior performance in domains of interviewing, exam, and counseling (Holmboe et al. 2003; Cook and Beckman 2009).

However, validity evidence exploring the internal structure dimensionality of mini-CEX scores is limited. Dimensionality affects how an instrument's scores are interpreted and used. For example, it would be inappropriate to use mini-CEX scores to distinguish residents' "interviewing and examination skills" from their "humanistic qualities and clinical judgment," if the scores lack the ability to discriminate between these dimensions of performance. The six mini-CEX competence domains intercorrelate highly, (Norcini et al. 1995; Norcini et al. 2003; Kogan et al. 2003; Cook and Beckman 2009; Margolis et al. 2006) and one study reported a Cronbach's alpha of 0.9 across all domains (Durning et al. 2002). One study also found that error correlated across domains, implying that ratings in one domain influence those in another (Margolis et al. 2006). These findings suggest that mini-CEX scores may measure a unidimensional construct.

Factor analytic studies can explore score dimensionality with greater rigor, but a comprehensive literature search identified only one factor analytic study of mini-CEX scores. This study (Hill et al. 2009) found a single dimension for medical student mini-CEX scores. However, the investigators averaged item scores from multiple observations before factor analysis, which could diminish item discrimination and artificially reduce score dimensionality. We found no factor analyses of mini-CEX scores in postgraduate training. Given our incomplete understanding of mini-CEX score dimensionality, additional research is needed.

The purpose of this study was to explore the internal structure (dimensionality and reliability) of mini-CEX scores obtained from ratings of internal medicine residents by supervising faculty in ambulatory clinic. To do this, we performed factor analyses and generalizability studies. Based on prior research we hypothesized two equally likely factor analytic solutions. First, research exploring the dimensions of clinical performance of medical students and residents has consistently demonstrated two main factors or themes, namely a cognitive (knowledge/technical) dimension and an affective (interpersonal/professionalism) dimension (Wilkinson and Frampton 2003; Hojat et al. 2007; Greenburg et al. 2007; Nasca et al. 2002; Silber et al. 2004; Thomas et al. 1999; Volkan et al. 2004). Alternatively, the research cited above suggests a single dimension for mini-CEX scores.

Method

Setting and sample

This study took place in the internal medicine resident continuity clinics in the Mayo School of Graduate Medical Education (Rochester, Minnesota). Supervising physicians had been using the mini-CEX to rate resident-patient encounters for more than a year prior to this study. In April and May 2006 we conducted additional faculty training in the use of the mini-CEX with the objective of improving observation skills and interrater agreement; however, we found little effect on agreement (Cook et al. 2009). This training was extraneous to the present study design, but since training might impact factor structure we accounted for it in our analyses as noted below.

This retrospective study was deemed exempt by our Institutional Review Board.

Instruments and outcomes

We obtained mini-CEX results from all observations of internal medicine residents in continuity clinic between October 2004 and December 2008, excluding the faculty training period (April–May 2006). Clinic preceptors completed a computer-based mini-CEX rating form after observing a live resident-patient encounter and providing feedback to the resident. The traditional mini-CEX uses a nine-point scale with anchors of 1–3 = Unsatisfactory, 4–6 = Satisfactory, and 7–9 = Superior. However, educators at our institution questioned the ability of raters to make distinctions of one point on a nine-point scale, and thus adopted a modified five-point scale with anchors of 1 = Needs Improvement, 2–4 = Average, and 5 = Top 10%. In a previous study we found little difference between these scales (Cook and Beckman 2009).

Analysis

We split the ratings into two sets—Set 1, prior to faculty training; and Set 2, after training—and analyzed these sets separately. All analyses used SAS 9.1.

The dataset included multiple observations on each resident, and occasionally multiple observations for the same preceptor-resident pair. Performing factor analysis on this dataset without adjustment would be incorrect because it would fail to account for within-person correlation. Thus, we used mixed linear models accounting for repeated measures on preceptors and residents to create an estimate of the correlation matrix that appropriately accounted for the repeated observations on the study subjects. We then performed a factor analysis on this adjusted correlation matrix. Details are provided in the Appendix. As a sensitivity analysis we repeated the factor analyses on a reduced dataset that excluded all but the first observation on each resident.

Since our purpose was to understand the latent variable structure rather than data reduction, (Floyd and Widaman 1995) our main analysis employed common factor (principal axes) analysis of the scores from the six subdomains. We estimated initial communalities using squared multiple correlations. Since our research question required us to extract the proper number of factors, we used multiple methods to determine the number of factors to retain. First, we planned to retain all factors that accounted for at least 5% of the variance, stopping when >85% of the variance had been explained (a level of variance explained that would err toward more factors being retained) (Gorsuch 1983). Second, we used the scree test. Third, we used parallel analysis (O'Connor 2000) to compare the actual eigenvalues with the mean eigenvalues obtained from 1,000 random re-orderings of the data. Finally, we repeated these analyses using principal component analysis, retaining all factors with eigenvalue $\geq$1. We retained items with loadings $\geq$0.3. We planned to use varimax rotation if >1 factor was found. As sensitivity analyses we repeated the common factor analysis using maximum absolute correlation to estimate communalities and again using maximum likelihood extraction, using the number of discrete preceptor-resident pairs as the effective sample size in the latter analysis.

We conducted generalizability analyses to evaluate score reproducibility (Shavelson and Webb 1991). Since each resident was observed by multiple preceptors, and each preceptor observed multiple residents, we used a residents (p) crossed with preceptors (j) crossed with items (i) design: $p \times j \times i$. We used variance components to estimate reliability for varying numbers of preceptor-resident observations, and also to calculate the standard error of measurement (SEM) (Brennan 2001). To facilitate interpretation, we adjusted the SEM to the traditional nine-point mini-CEX scale by multiplying by 9/5. We repeated this analysis using patient encounters, using an encounter (j) nested within residents (p) crossed with items (i) design: $(j{:}p) \times i$. Since learner maturation could lead to real changes in competence, we also repeated these analyses using subsets grouped by year.

Results

Set 1 contained scores from 307 encounters coding all six domains, while Set 2 contained scores from 1,107 encounters coding six domains. These sets comprised ratings by 85 supervisors of 264 residents in postgraduate years 1–3 over a 4-year period. Supervisors participating in the May 2006 training exercise contributed 803 (73%) of the observations in Set 2.

Factor analysis

Common factor analysis revealed a single-factor solution explaining 100% of the variance for both Set 1 and Set 2 (see Table 1). The scree plot was consistent with a single factor solution. All six mini-CEX domains loaded $\geq$0.38 (Table 2) on this factor, and thus all items were retained. Cronbach's alpha for this single factor was 0.88 for Set 1 and 0.86 for Set 2. The mean (SD) score for this factor (average of all six domains, using the five-point scale described above) was 3.8 (0.5) for Set 1 and 3.7 (0.5) for Set 2. Sensitivity analyses using a reduced dataset that included only the first observation on each trainee ($N = 129$ for Set 1, $N = 236$ for Set 2) revealed virtually identical results (not shown).

Common factor analysis requires that an initial *communality* (amount of variance in each variable explained by the factors) be estimated for each variable. Since different initial estimates can affect the analysis results, we conducted sensitivity analyses varying the method of communality estimation. The analysis using maximum absolute correlation to estimate communalities revealed a second factor explaining 14.5% of the variance for both Set 1 and Set 2, however, in each case the first factor alone explained >82% of the variance. Other than this, analyses using alternate communality estimates, maximum likelihood methods, and parallel analysis yielded similar one-factor solutions (not shown).

Principal components analysis revealed a less straightforward factor solution (Appendix Table 5). Considering the results for Set 2 (Set 1 revealed very similar results), the

Table 1 Common factor analysis of mini-CEX scores

Factor	Set 1 (307 observations)		Set 2 (1,107 observations)	
	Eigenvalue	% Variance explained[a]	Eigenvalue	% Variance explained[a]
1	2.06	116.9	1.82	124.0
2	0.22	12.5	0.18	12.0
3	−0.06	−3.7	−0.06	−4.0
4	−0.10	−5.6	−0.14	−9.6
5	−0.14	−7.7	−0.15	−10.5
6	−0.22	−12.4	−0.17	−11.8

[a] Variance explained indicates the proportion of total variance uniquely explained by this factor, expressed as a percentage. Variance explained exceeds 100% for Factor 1 due to unavoidable inaccuracy in estimating the initial communalities. Negative eigenvalues arise because the final total variance must equal zero

Table 2 Factor loading for mini-CEX domains

Domain	Set 1	Set 2
Interviewing	0.72	0.67
Examination	0.49	0.38
Counseling	0.65	0.59
Clinical judgment	0.66	0.65
Humanistic/professionalism	0.53	0.48
Organization/efficiency	0.41	0.48

Numbers reflect loading on Factor 1 (Table 1). Cronbach's alpha for this single factor was 0.88 for Set 1 and 0.86 for Set 2

eigenvalue for the second factor was 0.98. While strictly <1, this was sufficiently close to prompt further evaluation. Moreover, the scree plot showed some ambiguity regarding a possible second factor, and the first factor explained only 43% of the variance. However, varimax rotation (see Appendix Table 6) revealed ambiguous loading of two items (interviewing and clinical judgment). The scree plot did not suggest the need for additional factors, and even exploratory analyses with three or more factors did not resolve the problem of intercorrelated factors [suggesting that these were trivial factors (Gorsuch 1983)]. In the end, we concluded that these data most likely reflect a single factor.

Inspection of the adjusted and unadjusted (i.e., treating all observations as independent) correlation matrices revealed substantially higher domain intercorrelation in the unadjusted analysis (Appendix Table 7).

Generalizability analysis

The dependability coefficient (phi) for a single rater and six items (i.e., all six mini-CEX domains) was 0.19 for Set 1 and 0.23 for Set 2 (see Table 3). The 95% confidence interval around the true score (Harvill 1991) for a single observation (for the traditional nine-point scale) was ±1.8 for Set 1 and ±1.6 for Set 2. In other words, using the estimate for Set 1, if the average score across the six domains was 6.5 for a single receptor-resident observation, the 95% confidence interval for that score would be approximately 4.7–8.3.

Table 3 Generalizability analysis results

Analysis	Set 1		Set 2	
	Coefficient	SEM[a]	Coefficient	SEM[a]
G-studies				
Dependability (phi), 1 rater	0.19	0.91	0.23	0.82
G-coefficient (Ep^2), 1 rater	0.26	0.75	0.33	0.65
D-studies (varying no. of raters)				
Dependability (phi), 2 raters	0.32	0.65	0.37	0.59
Dependability (phi), 4 raters	0.47	0.48	0.52	0.43
Dependability (phi), 6 raters	0.55	0.40	0.61	0.36
Dependability (phi), 8 raters	0.61	0.36	0.66	0.32
Dependability (phi), 10 raters	0.65	0.33	0.70	0.30
Dependability (phi), 12 raters	0.68	0.31	0.73	0.28
Dependability (phi), 14 raters	0.70	0.29	0.75	0.26

Values reflect analyses (residents × preceptors × items) assuming scores on a single occasion, including all six mini-CEX domains (six items rated for each resident)

[a] *SEM* standard error of measurement adjusted to reflect the traditional 9-point scale. For dependability coefficient this is upper case delta Δ; for G-coefficient this is lower case delta δ (see Brennan 2001). The 95% confidence interval around the true score = SEM × 1.96

Inspection of the variance components (Table 4) revealed that for both Set 1 and Set 2 the object of measurement (the resident) accounted for only about 12% of the variance, and this was less than the variance arising from either preceptor stringency or preceptor subjectivity (resident-preceptor interaction). D-studies exploring different numbers of preceptors observing each encounter are reported in Table 3. Only with 10–14 different preceptors does dependability approach levels commonly considered acceptable (reliability 0.70), although with four or more preceptors per encounter the 95% confidence interval would exclude uncertainty of ±1 point.

We note that multiple preceptors observing a single encounter would be infeasible in most settings, and educators may wish to know how many encounters must be observed (by different preceptors) to achieve reliable scores. Thus, we repeated the generalizability analyses looking at variability across *encounters*. Results (not shown) were virtually identical, and thus the results in Table 3 for preceptors can be reasonably construed as reflecting varying numbers of encounters.

Finally, while preceptors probably adjust ratings for resident maturity, it is nonetheless possible that competence could change over time, which might artificially increase between-encounter variance. Thus, we repeated the generalizability studies for Set 2 subsets grouped by academic year. While the variance components (and thus the dependability coefficient) fluctuated somewhat across subsets, as would be expected, the pattern (more variance from preceptors than from residents) remained the same in all analyses (data not shown).

Discussion

Part 1. Factor analysis of mini-CEX scores for internal medicine residents revealed a single factor accounting for score variance. Findings remained essentially unchanged for scores

Table 4 Generalizability analysis variance components

Variance	Set 1 Variance (% total)	Set 2 Variance (% total)
Resident	0.06 (10)	0.06 (12)
Preceptor stringency	0.07 (12)	0.07 (13)
Item	0.04 (7)	0.03 (6)
Resident x Preceptor (preceptor subjectivity)	0.13 (21)	0.08 (15)
Resident x Item	0.01 (2)	0.01 (2)
Preceptor x Item	0.01 (2)	0.02 (3)
Error	0.28 (46)	0.26 (49)
Total variance	0.60 (100)	0.54 (100)

Estimates obtained using SAS PROC VARCOMP, Type I method, with a fully crossed design (resident × preceptor × item)

obtained both before and after faculty training in resident observations, and when varying factor analytic methods. These results suggest that mini-CEX scores measure a single global dimension of resident physicians' clinical performance which might appropriately be labeled "overall clinical competence." This finding has important implications for trainee assessment since it suggests that mini-CEX domain scores may not discriminate unique aspects of residents' clinical abilities (i.e., history-taking, counseling, etc.). Our study corroborates a recent study suggesting a single dimension for medical student mini-CEX scores (Hill et al. 2009). By contrast, studies of other clinical performance measures typically suggest a two-factor model of performance (Wilkinson and Frampton 2003; Hojat et al. 2007; Greenburg et al. 2007; Nasca et al. 2002; Silber et al. 2004; Thomas et al. 1999; Volkan et al. 2004) and only rarely a one-factor model (Haber and Avins 1994).

Although internal consistency (Cronbach's alpha) was high, we found relatively low reproducibility across preceptors (and across encounters). In fact, variance arising from preceptors was larger than the variance from residents. Previous studies have demonstrated similar results (Margolis et al. 2006; Hill et al. 2009; Weller et al. 2009).

Limitations and strengths

This study has limitations. We used a modification of the mini-CEX with a five-point rather than a nine-point response scale. Although we doubt this would affect the score factor structure, particularly in light of our finding of similar reliability between the two scales, (Cook and Beckman 2009) it may be useful to replicate our study using the traditional nine-point scale. We conducted a training session midway through the study period. However, this was not a deliberate part of the present study's purpose, and moreover this provided an opportunity to analyze the factor structure before and after training. Preceptors were confounded with encounters, and the latter were nested within residents, but we performed separate analyses focused on each of these facets and found nearly identical results. Finally, the adjustment might not completely have captured all of the structure in the data.

Strengths of this study include a large sample size, results robust to variations in factor analysis technique and generalizability study design, and demonstration of similar factor structures before and after a 4-hour faculty training session.

Integration with other literature

These results add to a growing body of evidence informing the interpretation and use of mini-CEX scores. Although overall ratings have high internal consistency, (Hatala et al. 2006; Durning et al. 2002; Weller et al. 2009) the present study corroborates prior research demonstrating suboptimal interrater reliability (Cook and Beckman 2009; Margolis et al. 2006; Hill et al. 2009; Weller et al. 2009). Our previous work suggests that improving interrater reliability is difficult even with moderately intensive interventions (Cook et al. 2009) [although one study using already-experienced raters achieved high reliability (Sidhu et al. 2009)]. Thus, multiple observations and multiple raters are required for reliable scores. Studies exploring relationships between mini-CEX scores and ratings of standardized patient interactions have shown variable results (Hatala et al. 2006; Ney et al. 2009).

Implications

These limitations in score reliability and validity justify caution when using the mini-CEX in its present form for summative purposes. Mini-CEX domain scores may be combined to provide an overall rating of clinical competence, but it may be inappropriate to use scores from individual domains or combinations of domains for moderate- or high-stakes summative assessment.

However, the mini-CEX may have an important role in formative assessment. In formative applications, the actual numeric rating may not matter as much as the attention given to behaviors in each domain, which in turn facilitates meaningful discussions (i.e., feedback) between clinical teachers and learners. Despite its value in professional development, feedback is often insubstantial and infrequently given (Kroboth et al. 1996; Herbers et al. 1989; Srinivasan et al. 2007; Fernando et al. 2008). Directly observed trainee-patient encounters provide an ideal opportunity to provide feedback. Yet despite attempts to enrich feedback using the mini-CEX traditional form, (Fernando et al. 2008) variations of the mini-CEX, (Donato et al. 2008) and other instruments (Holmboe et al. 2001) there remains much room for improvement and further research in formative assessment and feedback in clinical teaching (Kogan et al. 2009).

A comment on methods

Datasets obtained in naturalistic settings often contain multiple observations on individuals. Treating each of these observations as independent in factor analysis is incorrect because it fails to account for within-subject correlation. However, averaging scores prior to analysis may diminish item discrimination and artificially reduce score dimensionality. Discarding repeated observations (i.e., including only the first observation for a given individual) is better, but again disregards data that might inform the analysis. Our two-step approach, first generating an adjusted correlation matrix and then performing factor analysis on this adjusted matrix, offers a potential solution that includes all available data while properly accounting for within-subject correlation, and can be performed without specialized programs. This adjusted correlation matrix shows lower domain intercorrelation than the unadjusted correlation matrix, as would be expected since the unadjusted matrix blurs within-subject and between-domain correlation. Although this did not affect the final factor structure in this study, it could potentially do so in other datasets. The effective sample size (weight) for the adjusted correlation matrix is debatable, but sample size only affects hypothesis tests (e.g. in maximum likelihood analysis) and thus does not affect the other common factor and principal components analyses.

In performing generalizability studies it is important to clearly identify the object of measurement (in this case, residents) and the other relevant facets. In performance assessment, the occasion facet may be as or more relevant than the rater facet. It is also important to consider true changes over time when analyzing longitudinal datasets.

Finally, the finding that the psychometric properties of mini-CEX scores vary across applications reminds us that reliability and validity are properties of instrument scores, not the instruments themselves. The same instrument, applied in different settings, can yield results for which the confidence of our interpretations varies widely. Thus, instruments should be evaluated in contexts as similar as possible to the intended application.

Conclusions

We found that scores from the traditional six-item mini-CEX rating form reflect a single global dimension of clinical competence, and reproducibility is relatively low. Alternative assessment methods may be required if educators wish to measure discrete clinical skills.

Acknowledgments Financial support No external funding.

Authorship All authors were involved in the planning and execution of this study and in the drafting and revising of this manuscript.

Ethical approval Judged exempt by our Institutional Review Board.

Conflicts of interest statement The authors have no affiliation with an organization with a financial interest in the subject matter, and are not aware of any conflicts of interest.

Appendix

Part 1. Factor analysis accounting for multiple observations on each resident

Code is for SAS 9.1. Original dataset for the post-workshop data is “cex_post” with data columns:

- fac_id (unique preceptor identification code),
- res_id (unique resident identification code),
- rep_id (unique encounter identifier for preceptor-resident pairs with more than one observation; if preceptor A observed resident B three times, rep_id values would be 1 for the first encounter, 2 for the second, and 3 for the third)
- ratings for each mini-CEX domain: couns, ex, human, hx, judg, org.

Step 1. Create an adjusted correlation matrix using mixed linear models with repeated measures on preceptors and residents.

(a) First reformat the dataset for proc mixed:

```
data cex_mixed; set cex_post;
rating=couns; item="couns"; output;
rating=ex; item="ex"; output;
rating=human; item="human"; output;
rating=hx; item="hx"; output;
rating=judg; item="judg"; output;
rating=org; item="org"; output;
```

(b) Then create the adjusted correlation matrix. Note that you must be careful to determine the order of the variables in the resultant matrix (the item labels are not part of the matrix output).

```
data temp; set cex_mixed; *we do not use the temp dataset again;
proc sort; by item res_id fac_id rep_id;
proc mixed;
      class res_id fac_id item rep_id;
      model rating=item/solution;
      random fac_id res_id;
      repeated / subject=res_id*fac_id*rep_id type=un rcorr;
```

(c) The results of this analysis will appear as output on the screen. Again, it is essential to correctly identify which variable matches with each column in this matrix (column order is the same as the order of the variables in the parent dataset, in this case "cex_mixed").

```
Row    Col1   Col2   Col3   Col4   Col5   Col6
1      1.0000 0.1738 0.3839 0.4291 0.3843 0.2359
2      0.1738 1.0000 0.1389 0.2449 0.3088 0.2440
3      0.3839 0.1389 1.0000 0.3588 0.3033 0.1384
4      0.4291 0.2449 0.3588 1.0000 0.4560 0.3608
5      0.3843 0.3088 0.3033 0.4560 1.0000 0.3582
6      0.2359 0.2440 0.1384 0.3608 0.3582 1.0000
```

Step 2. Create a correlation data set for subsequent analysis

The values derived above can be manually used to create a data set of type "CORR" for subsequent analysis as shown below. Means and standard deviations can be determined using proc means. We estimated N (effective sample size) from the number of discrete preceptor-resident pairs (which can be found by counting the number of observations with rep_id = 1).

```
data rcorr_post(type=corr);
      input _TYPE_ $ _NAME_ $ couns ex human hx judg org;
    datalines;
MEAN . 3.767 3.533 4.014 3.815 3.716 3.579
STD . 0.692 0.677 0.646 0.684 0.636 0.711
N . 873 873 873 873 873 873
CORR couns 1.0000 0.1738 0.3839 0.4291 0.3843 0.2359
CORR ex 0.1738 1.0000 0.1389 0.2449 0.3088 0.2440
CORR human 0.3839 0.1389 1.0000 0.3588 0.3033 0.1384
CORR hx 0.4291 0.2449 0.3588 1.0000 0.4560 0.3608
CORR judg 0.3843 0.3088 0.3033 0.4560 1.0000 0.3582
CORR org 0.2359 0.2440 0.1384 0.3608 0.3582 1.0000
;
```

Step 3. Perform factor analysis on this adjusted correlation matrix:

```
proc factor data=rcorr_post priors=SMC scree;
```

Part 2. Additional tables: principal components factor analysis and adjusted/unadjusted correlation matrices

Table 5 Principal components factor analysis of mini-CEX scores (Set 2)

Factor	Eigenvalue	Variance explained (%)	Cumulative variance explained (%)
1	2.55	43	43
2	0.98	16	59
3	0.77	13	72
4	0.61	10	82
5	0.57	9	91
6	0.53	9	100

Table 6 Factor loading for mini-CEX domains following principal components analysis (Set 2)

Domain	Factor 1	Factor 2
Interviewing	**0.63**	**0.44**
Examination	0.01	**0.75**
Counseling	**0.76**	0.17
Clinical judgment	**0.49**	**0.57**
Humanistic/professionalism	**0.81**	−0.04
Organization/efficiency	0.15	**0.73**

Bolded items load substantially (>0.3) on specified factor

Table 7 Comparison of correlation matrices for adjusted and unadjusted factor analysis

	Interviewing	Examination	Counseling	Clinical judgment	Humanistic	Organization
Adjusted analysis						
Interviewing	1.00	0.24	0.43	0.46	0.36	0.36
Examination	0.24	1.00	0.17	0.31	0.14	0.24
Counseling	0.43	0.17	1.00	0.38	0.38	0.24
Clinical judgment	0.46	0.31	0.38	1.00	0.30	0.36
Humanistic	0.36	0.14	0.38	0.30	1.00	0.14
Organization	0.36	0.24	0.24	0.36	0.14	1.00
Unadjusted analysis						
Interviewing	1.00	0.46	0.61	0.63	0.57	0.53
Examination	0.46	1.00	0.42	0.51	0.40	0.43
Counseling	0.61	0.42	1.00	0.59	0.60	0.45
Clinical judgment	0.63	0.51	0.59	1.00	0.54	0.53
Humanistic	0.57	0.40	0.60	0.54	1.00	0.38
Organization	0.53	0.43	0.45	0.53	0.38	1.00

Data are for Set 2 ($N = 1,107$ observations). Note that all correlation coefficients in the adjusted matrix are substantially lower than those in the unadjusted matrix. Analysis of Set 1 revealed similar findings

References

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Cook, D. A., & Beckman, T. J. (2009). Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX. *Advances in Health Sciences Education: Theory and Practice, 14*, 655–664.

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. S. (2009). Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial. *Journal of General Internal Medicine, 24*, 74–79.

Donato, A. A., Pangaro, L., Smith, C., Rencic, J., Diaz, Y., Mensinger, J., et al. (2008). Evaluation of a novel assessment form for observing medical residents: A randomised, controlled trial. *Medical Education, 42*, 1234–1242.

Durning, S. J., Cation, L. J., Markert, R. J., & Pangaro, L. N. (2002). Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Academic Medicine, 77*, 900–904.

Fernando, N., Cleland, J., McKenzie, H., & Cassar, K. (2008). Identifying the factors that determine feedback given to undergraduate medical students following formative mini-CEX assessments. *Medical Education, 42*, 89–95.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*, 286–299.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Greenburg, D. L., Durning, S. J., Cohen, D. L., Cruess, D., & Jackson, J. L. (2007). Identifying medical students likely to exhibit poor professionalism and knowledge during internship. *Journal of General Internal Medicine, 22*, 1711–1717.

Haber, R. J., & Avins, A. L. (1994). Do ratings on the American Board of internal medicine resident evaluation form detect differences in clinical competence? *Journal of General Internal Medicine, 9*, 140–145.

Harvill, L. M. (1991). NCME instructional module: Standard error of measurement. *Educational Measurement: Issues and Practice, 10*(2), 33–41.

Hatala, R., Ainslie, M., Kassen, B. O., Mackie, I., & Roberts, J. M. (2006). Assessing the mini-clinical evaluation exercise in comparison to a national specialty examination. *Medical Education, 40*, 950–956.

Herbers, J. E., Jr., Noel, G. L., Cooper, G. S., Harvey, J., Pangaro, L. N., & Weaver, M. J. (1989). How accurate are faculty evaluations of clinical competence? *Journal of General Internal Medicine, 4*, 202–208.

Hill, F., Kendall, K., Galbraith, K., & Crossley, J. (2009). Implementing the undergraduate mini-CEX: A tailored approach at Southampton University. *Medical Education, 43*, 326–334.

Hojat, M., Paskin, D. L., Callahan, C. A., Nasca, T. J., Louis, D. Z., Veloski, J., et al. (2007). Components of postgraduate competence: analyses of thirty years of longitudinal data. *Medical Education, 41*, 982–989.

Holmboe, E., Fiebach, N., Galaty, L., & Huot, S. (2001). Effectiveness of a focused educational intervention on resident evaluations from faculty. *Journal of General Internal Medicine, 16*, 427–434.

Holmboe, E. S., Huot, S., Chung, J., Norcini, J., & Hawkins, R. E. (2003). Construct validity of the miniclinical evaluation exercise (miniCEX). *Academic Medicine, 78*, 826–830.

Kogan, J. R., Bellini, L. M., & Shea, J. A. (2003). Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. *Academic Medicine, 78*(10 Suppl), S33–S35.

Kogan, J. R., Holmboe, E. S., & Hauer, K. E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *JAMA, 302*, 1316–1326.

Kroboth, F. J., Hanusa, B. H., & Parker, S. C. (1996). Didactic value of the clinical evaluation exercise. Missed opportunities. *Journal of General Internal Medicine, 11*, 551–553.

Margolis, M. J., Clauser, B. E., Cuddy, M. M., Ciccone, A., Mee, J., Harik, P., et al. (2006). Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: A validity study. *Academic Medicine, 81*(10 Suppl), S56–S60.

Nasca, T. J., Gonnella, J. S., Hojat, M., Veloski, J., Erdmann, J. B., Robeson, M., et al. (2002). Conceptualization and measurement of clinical competence of residents: A brief rating form and its psychometric properties. *Medical Teacher, 24*, 299–303.

Ney, E. M., Shea, J. A., & Kogan, J. R. (2009). Predictive validity of the mini-clinical evaluation exercise (mCEX): Do medical students' mCEX ratings correlate with future clinical exam performance? *Academic Medicine, 84*(10 suppl), S21–S24.

Norcini, J. J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1995). The mini-CEX (clinical evaluation exercise): A preliminary investigation. *Annals of Internal Medicine, 123*, 795–799.
Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine, 138*, 476–481.
O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, & Computers, 32*, 396–402.
Shavelson, R. L., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park: Sage Publications.
Sidhu, R. S., Hatala, R., Barron, S., Broudo, M., Pachev, G., & Page, G. (2009). Reliability and acceptance of the mini-clinical evaluation exercise as a performance assessment of practicing physicians. *Academic Medicine, 84*(10 suppl), S113–S115.
Silber, C. G., Nasca, T. J., Paskin, D. L., Eiger, G., Robeson, M., & Veloski, J. J. (2004). Do global rating forms enable program directors to assess the ACGME competencies? *Academic Medicine, 79*, 549–556.
Srinivasan, M., Hauer, K. E., Der-Martirosian, C., Wilkes, M., & Gesundheit, N. (2007). Does feedback matter? Practice-based learning for medical students after a multi-institutional clinical performance examination. *Medical Education, 41*, 857–865.
Thomas, P. A., Gebo, K. A., & Hellmann, D. B. (1999). A pilot study of peer review in residency training. *Journal of General Internal Medicine, 14*, 551–554.
Volkan, K., Simon, S. R., Baker, H., & Todres, I. D. (2004). Psychometric structure of a comprehensive objective structured clinical examination: A factor analytic approach. *Advances in Health Sciences Education: Theory and Practice, 9*, 83–92.
Weller, J. M., Jolly, B., Misur, M. P., Merry, A. F., Jones, A., Crossley, J. G. M., et al. (2009). Mini-clinical evaluation exercise in anaesthesia training. *British Journal of Anaesthesia, 102*, 633–641.
Wilkinson, T. J., & Frampton, C. M. (2003). Assessing performance in final year medical students. Can a postgraduate measure be used in an undergraduate setting? *Medical Education, 37*, 233–240.