Construct Validity of Three Clerkship Performance Assessments

Ming Lee, Ph.D., and Paul F. Wimmers, Ph.D.

David Geffen School of Medicine at UCLA

Paper presented at the 2010 American Educational Research Association Annual Meeting in

Denver, Colorado, April 30 –May 4, 2010

Abstract

This study examined construct validity of three commonly used clerkship performance assessments: preceptors' evaluations, OSCE-type clinical performance measures, and the NBME medicine subject examination. Six hundred and eighty-six students taking the inpatient medicine clerkship from 2003 to 2007 participated in the study. Exploratory and confirmatory factor analyses using the structural equation modeling procedure were adopted to examine the latent domains underlying various indicators assessed by these three measures and the pattern of indicator-domain relationships. Three separate, though correlated, constructs, labeled Clinical Performance, Interpersonal Skills, and Clinical Knowledge, were confirmed by factor analyses. The three domains tapped a common higher-order construct, Clinical Competence, in varied degrees of magnitude (.56, .69, .77, respectively). This study demonstrated that although the three commonly used tools for assessing clerkship performance contributed uniquely to the understanding of clinical performance, they also attested a shared domain of clinical competence in their assessment. The study thus confirmed the need for a multiple-trait-multiple-method approach to clerkship assessment. Findings also revealed that clerkship preceptors need to differentiate their judgment of students' performances.

Purpose

This study analyzed the construct validity of three commonly used measures in the assessment of medical students' clerkship performances in order to better understand their conceptual structures and utility in the explanation of clinical competence.

Theoretical Framework

Clinical competence is considered multifaceted and encompasses diverse aspects such as information acquisition ability, knowledge of basic sciences and clinical medicine, application of knowledge to clinical settings, and communication skills. Given this complexity, a multiple-method approach is usually recommended by clinical educators and researchers in the evaluation of clerkship performance (Hull et al., 1995; Wimmers & Fung, 2008).

The use of evaluation methods by clerkship programs has changed substantially over the past decade. A national survey of internal medicine clerkship directors found that the traditional method of using faculty-developed examinations has declined from 46% to 27%, and the use of a standardized patient examination has increased sharply from 2% to 27% (Hemmer, Szauter, Allbritton, & Elnicki, 2002). In addition, the use of the National Board of Medical Examiners (NBME) subject examination has increased from 66% to 83%.

Our School of Medicine has adopted all three of these commonly used evaluation methods. Preceptors' evaluation of students' clerkship performance uses a seven-point Likert scale (Medical Student Summative Evaluation, MSSE) which assesses students' performance in nine areas, including history taking, physical examination, oral case presentation, write-ups, fund of

knowledge, clinical judgment, physician-patient interaction, professional attitudes and behaviors, and overall performance.

A multiple-station OSCE-type Clinical Performance Examination (CPX) was implemented in 1996. Students rotate through seven or eight stations and spend fifteen minutes at each to conduct a focused work-up on a trained standardized patient (SP). Students' clinical skills are rated by SPs based on a checklist created by faculty members. The ratings are summarized as percentage of correct scores in four areas, including history taking, physical examination, information sharing, and physician-patient interaction.

In addition to these two performance-based evaluation measures, students also take relevant NBME subject examinations at the end of clerkships. The standardized exams provide additional information on the amount of knowledge in basic sciences and clinical medicine that students have obtained and their ability to apply the knowledge to clinical settings via written clinical vignettes.

Although the three assessment measures have been used for many years, construct validity of these measures remains to be established. It is not clear (1) whether each of the three measures assesses the same or different latent domains; (2) what components of the three measures are associated with each latent domain; and (3) to what extent the latent domains are related to each other. Several studies have examined correlations among a number of student clinical performance measures (e.g., Ferguson & Kreiter, 2004; Hasnain, Connell, Downing, Olthoff, & Yudkowshky, 2004; Hull et al., 1995; Wilkerson & Lee, 2003; Wimmers & Fung, 2008); only

one simultaneously assessed the validity of three clinical performance assessments similar to the ones adopted by our school. This particular study (Hull et al, 1995) used the multitrait-multimethod (MTMM) matrix approach (Campbell & Fiske, 1959), a method that has limitations in objectively evaluating interactions between traits (e.g., clinical performance) and methods (e.g., MSSE) and separating trait, method, and random variance in observed indicators (e.g., history taking) despite its important contribution to our understanding of validation procedures through convergent and discriminant validity (Brown, 2006; Schmitt & Stults, 1986). To overcome this problem, confirmatory factor analysis (CFA) is recommended as it can "readily accommodate the elements of the MTMM matrix by offering more objective criteria for the evaluation of construct validity as well as more refined measurement of key constructs" through adjusting for measurement error (Andrasik, 2006, p. 27). Brown (2006) suggested that CFA be an "indispensable analytic tool for construct validation" (p. 2). The present study therefore used the CFA approach to answer the aforementioned three questions.

As the components assessed by the three measures show an overlap, we hypothesized that (1) there is a common domain assessed by all three measures; (2) the three measures also provide unique contributions to the understanding of clinical competence; and (3) the latent domains assessed by individual measures are correlated with each other at different levels.

Methods and Data Sources

Six hundred and eighty-six students who were in their third year of medical school from August 2003 to July 2007 participated in the study. Institutional Review Board (IRB) approval for using archival data of the students was obtained before the study began.

The three instruments used in the study were (1) MSSE: mean scores of two site directors' summative evaluation of the third-year students' performance during their inpatient medicine clerkship; (2) a multiple-station CPX administered at the end of the third-year, and (3) the NBME medicine subject examination.

During the eight-week inpatient medicine clerkship, all third-year students were randomly assigned to a pair of six training sites. A two-factor multivariate analysis of variance revealed no statistically significant difference in the ratings by year, rotation, or an interaction of year by rotation.

To explore constructs underlying the three clinical performance assessments, students' scores were first analyzed by exploratory factor analysis (EFA) using the principal axis factoring and promax rotation as it is hypothesized that the three measures are correlated. Hypothetical models were developed based on the EFA results and verified by CFA using the structural equation modeling (SEM) procedure. Maximum likelihood estimation was the method used in CFA. For goodness-of-fit tests, chi-square, comparative fit index (CFI), non-normed fit index (NNFI), and root mean square error of approximation (RMSEA) were used as the last three tests were found to be among the measures least affected by sample size (Fan, Thompson, & Wang, 1999), an important factor for consideration given the relatively large sample size of this study. The analyses were conducted using SPSS v. 17.0 and EQS v. 6.

Results

Factorial Validity

Using listwise deletion in the initial principal factor analysis, 498 (73%) of the 686 students had scores on all three clerkship performance measures. Three factors with an eigenvalue of 1 or higher were extracted. Table 1 shows the factor pattern and inter-factor correlations after Promax rotation.

All nine areas of MSSE clustered strongly under Factor 1. Two of the four components of the CPX, information sharing and patient-physician interaction, clustered under Factor 2, whereas the remaining two CPX components, history taking and physician exam, and the NBME medicine exam score clustered under Factor 3. The three factors were hence labeled Clinical Performance (Factor 1), Interpersonal Skills (Factor 2), and Clinical Knowledge (Factor 3). Cronbach's alpha coefficients for the three factor subscales varied, showing .96, .61, and .50, respectively. The inter-factor correlation coefficients showed modest to low degrees of correlation between Clinical Knowledge and Interpersonal Skills (.41), Clinical Knowledge and Clinical Performance (.38), and Clinical Performance and Interpersonal Skills (.30).

Construct Validity

Based on the findings of exploratory factor analysis, three hypothetical models were proposed to test the hypothesis that the three clerkship performance assessments measured a unitary domain in clinical competence: (1) a one-factor model, (2) a three-factor model representing the original three measures; and (3) a three-factor model consisting of the three extracted factors.

First-order Factor Analysis: Findings of the comparison between the three proposed models showed that the one-factor model could not be retained given its lowest indices of goodness-of-fit (χ^2 = 471.75, CFI = .92, NNFI = .90, RMSEA = .10), suggesting that the components included in the three measures did not completely represent a single domain. Both of the three-factor models recognized a triple-domain structure in the three measures and showed improved fitness with the data. The model representing the three extracted factors, however, had slightly better results, with CFI = .96, NNFI = .95, and RMSEA = .077 indicating indices either better than or close to the conventional cutoffs for a good model fit (.95, .95, .05, respectively) (Garson, 2008). This multi-factorial structure of the three measures is graphically illustrated in Figure 1.

Second-order Factor Analysis: As the three factors correlated with each other, a second-order CFA was conducted to examine the degree to which the three factors tapped the same underlying dimension. This analysis tested the hypothesis that a latent construct, something that might be called Clinical Competence, could be found arching over the three domains measured by the three clinical performance assessments. The loadings of the three factors on the second-order factor were .56, .69, and .77 for Clinical Performance, Interpersonal Skills, and Clinical Knowledge, respectively.

Conclusions

A combination of EFA and CFA identified three latent constructs of the three clerkship performance assessments, which were labeled Clinical Performance, Interpersonal Skills, and Clinical Knowledge. These findings led to the construction and comparison of three structural models to determine which model described the data best. The CFA using the SEM procedure

ascertained that the three instruments assessed three separate, though correlated, domains of clerkship performance, and the three domains tapped the same latent dimension which was named Clinical Competence.

Although students' clerkship performance was assessed by their preceptors on nine different aspects, all of the nine ratings were clustered more fitly together than with any other components of the assessment measures. It is interesting to note that the preceptors' overall evaluation of a student's clerkship performance loaded most heavily on the underlying domain of clerkship performance, labeled Clinical Performance. This might be explained by the findings reported in the literature (Hasnain et al., 2004; Pulito, Donnelly, & Plymale, 2007) that faculty tended to form a general impression of students' performance and evaluate all aspects of their performance based on such an impression. This might also explain why the nine ratings clustered together under one domain which had extremely high internal consistency reliability, although the method effect could also contribute to the finding (Brown, 2006).

The finding that the NBME subject exam tapped the same domain as the history taking and physical exam components of CPX might be explained by the fact that students' performance on the two CPX components relied on their knowledge of disease, pathophysiology, diagnosis, and assessment. It was also understandable that this domain assessing mainly clinical knowledge was separate from the domain underlying the other two CPX components, information sharing and patient-physician interaction. The latter two focused more on assessing communication and interpersonal skills. The internal consistency reliability indices of both domains were only

modest, a finding consistent with that of other studies examining the psychometric properties of multiple-station OSCE-type assessments (Hull et al., 1995; Petrusa et al., 1987).

The evaluation of students' performance in pseudo-clinical encounters by standardized patients seemed to largely tap a different domain of clerkship performance than the one accounted for by preceptors' evaluation in real patient encounters in normal clinical settings. Although the two measures assessed many similar aspects of clerkship performance, the two assessment methods contributed uniquely to the explanation of clerkship performance. The separate domains might have resulted from either one or a combination of the following method effects or their interactions: (1) evaluator effect (SPs versus attending physicians), (2) setting effect (pseudo versus real clinical setting), and (3) scale effect (a mostly yes-no objective scale versus a seven-point subjective scale).

Although the three measures assessed three separate domains of clerkship performance, they all tapped a latent construct, Clinical Competence, in varied degrees of magnitude. The Clinical Knowledge domain seemed to play the most significant role in the relationship with this overarching construct, but the other two domains, Clinical Performance and Interpersonal Skills, also contributed substantially to it. The findings reflected the complexity of clinical competence and confirmed the recommendation made by many educators and researchers that a multiple-trait-multiple-method approach should be used in the evaluation of clerkship performance.

Scholarly Significance of the Study

This study demonstrated that although the three commonly used tools for assessing clerkship performance contributed uniquely to the understanding of clinical performance, they also attested a shared domain of clinical competence in their assessment. Findings also revealed that clerkship preceptors need to differentiate their judgment of students' performances. Future investigation is needed to further examine the likelihood of a multifaceted underlying structure of CPX.

References

Andrasik, F. (Ed.). (2006). *Comprehensive handbook of personality and psychopathology, Vol. 2: Adult psychopathology*. Hoboken, NJ: Wiley.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation method, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*, 56-83.

Ferguson, K. J., & Kreiter, C. D. (2004). Using a longitudinal database to assess the validity of preceptors' ratings of clerkship performance. *Advances in Health Sciences Education, 9*, 39-46.

Garson, G. D. *Strucutral equation modeling*. Retrieved July 31, 2008, from http://www2.chass.ncsu.edu/garson/pa765/structur.htm.

Hasnain,. M., Connell, K. J., Downing, S. M., Olthoff, A., & Yudkowsky, R. (2004). Toward meaningful evaluation of clinical competence: The role of direct observation in clerkship ratings. *Academic Medicine, 79*, S21-S24.

Hemmer, P. A., Szauter, K., Allbritton, T. A., & Elnicki, D. M. (2002). Internal medicine clerkship directors' use of and opinions about clerkship examinations. *Teaching and Learning in Medicine, 14*, 229-235.

Hull, A. L., Hodder, S., Berger, B., Ginsberg, D., Lindheim, N., Quan, J., & Kleinhenz, M. E. (1995). Validity of three clinical performance assessments of internal medicine clerks. *Academic Medicine 70*, 517-522.

Petrusa, E. R., Blackwell, T.A., Rogers, L.P., Saydjari, C., Parcel, S., & Guckian, J. C. (1987). An objective measure of clinical performance. *The American Journal of Medicine, 83*, 34-42.

Pulito, A. R., Donnelly, M. B., & Plymale, M. (2007). Factors in faculty evaluation of medical students' performance. *Medical Education, 41*, 667-675.

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*, 1-22.

Wilkerson, L., & Lee, M. (2003). Assessing physical examination skills of senior medical students: Knowing how versus knowing when. *Academic Medicine, 78*, S30-S32.

Wimmers, P. F., & Fung, C. C. (2008). The impact of case specificity and generalisable skills on clinical performance: A correlated traits-correlated methods approach. *Medical Education, 42*, 580-588.

Table 1. Exploratory factor analysis* of the three clerkship performance assessments: Factor pattern and inter-factor correlation after rotation.

Measure	Variable	Factor 1	Factor 2	Factor 3
NBME Medicine Exam		.16	-.05	.46
CPX	History taking	-.19	.26	.48
	Physical exam	.04	.00	.47
	Info sharing	-.01	.58	.07
	Patient/physician interaction	.11	.74	-.02
CPA	History taking	.85	-.04	.09
	Physical exam	.85	-.03	.04
	Oral case presentation	.84	.01	.08
	Write-ups	.86	.02	-.03
	Fund of knowledge	.84	-.05	.12
	Clinical judgment	.88	-.01	.03
	Physician-patient interaction	.80	.16	-.19
	Professional attitudes	.84	.05	-.13
	Overall performance	.92	-.05	.02
Inter-factor correlations	Factor 1	1.00	.30	.38
	Factor 2		1.00	.41
	Factor 3			1.00

* Extraction method: Principal axis factoring; Rotation method: Promax. Pattern matrix is shown.

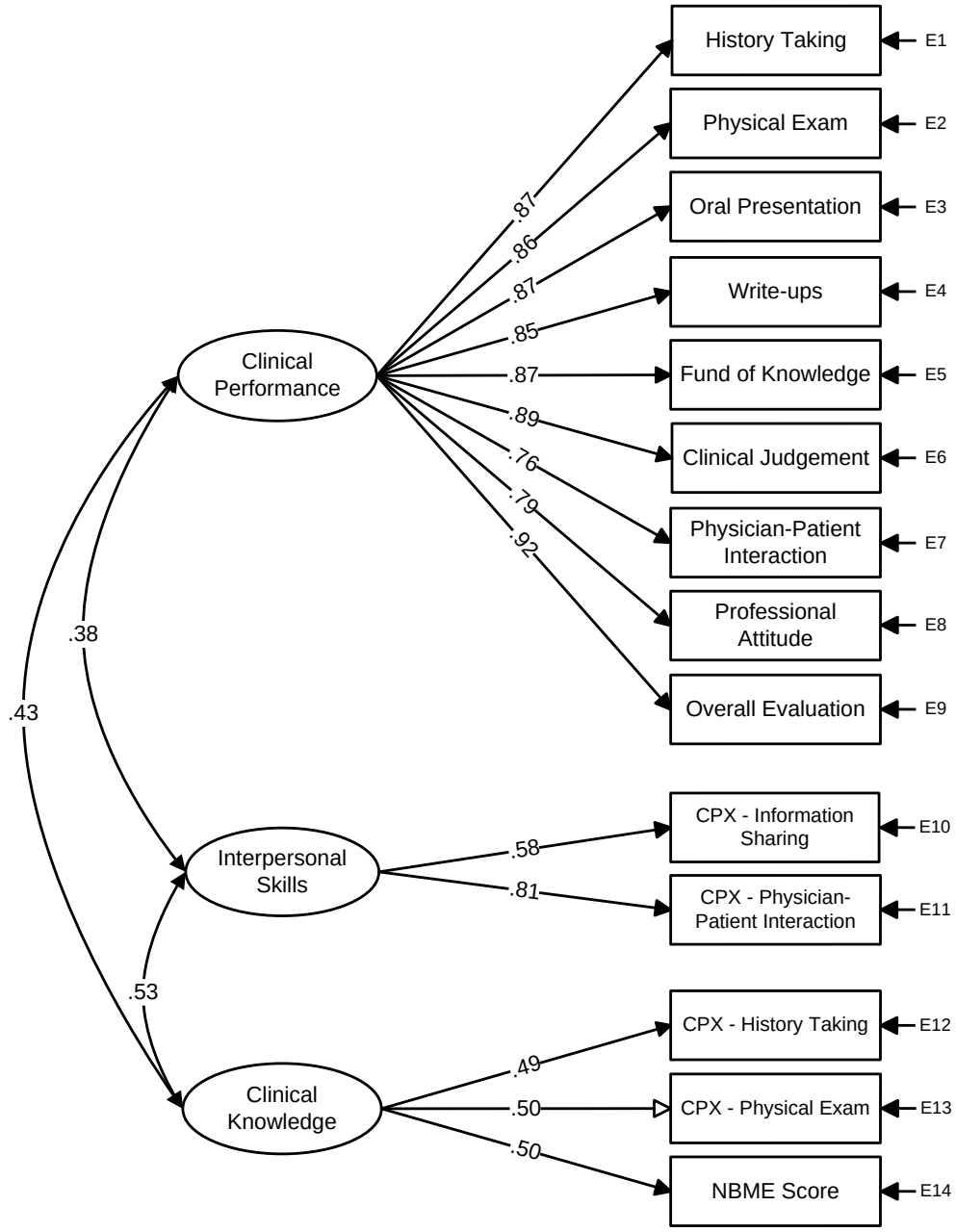

Figure 1. Underlying factor structure of the three clerkship performance assessments