Scores Gains on Performance Tests for Repeat Examinees: An Evaluation of Construct and Criterion-Related Evidence

Mark R. Raymond, Nilufer Kahraman, Kimberly A. Swygert,
and Kevin P. Balog

National Board of Medical Examiners

Paper Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April 2011

Correspondence: mraymond@nbme.org.

Abstract

Examinees who repeat performance tests experience large score gains. Limited evidence from multiple-choice testing programs suggests that the validity of scores from the second occasion is compromised. This study investigated the internal and external validity of scores for repeat examinees on a performance-based clinical skills test in medicine. Multi-group, confirmatory factor analysis indicated that the factor structure for repeat examinees on their first-attempt was markedly different from the structure for single-take examinees, but that by the second attempt the factor structure for repeat examinees differed only slightly. Scores on the second attempt were found to correlate more highly with three external measures taken at three points in time. Both sources of evidence – internal factor structure and external correlations – suggest that scores for repeat examinees based on the second administration exhibit improved validity.

Scores Gains on Performance Tests for Repeat Examinees: An Evaluation of Construct and Criterion-Related Evidence

The United States Medical Licensing Examination (USMLE) includes an assessment of clinical proficiency known as the Step 2 Clinical Skills (CS) Examination. Step 2 CS uses standardized patients (SPs) to assess an examinee's ability to acquire medical information through patient interviews, perform physical examinations, and summarize and communicate their findings. Determining the extent to which scores on the exam actually assess these skills is an important aspect of construct validity.[1] Correlations and factor analyses of Step 2 CS subscores provide evidence supporting the structural properties of scores on the clinical skills exam.[2,3] In addition, Step 2 CS scores make a unique contribution to the assessment of competence, as suggested by the low to moderate correlations between Step 2 CS and other exams comprising the USMLE sequence.[3]

An issue that has received limited attention is the validity of score interpretations for examinees who initially fail and then later repeat a performance assessment such as Step 2 CS. Although scores should increase for examinees who remediate deficiencies, score gains can also occur for reasons that compromise validity. Construct-irrelevant variance is a concern when the score increase can be attributed to an improvement in skills unrelated to purpose of the assessment, such as self confidence, appearance, or test sophistication.[1,4] Prior research has reported large score gains for repeat examinees on performance assessments such as oral exams, and it seems likely that construct-irrelevant variance explains some portion of the those increases.[5,6] Scores might also increase for repeat examinees who remember previously-seen test materials. If the increase is specific to the memorized test content and does not reflect an overall improvement in skill, then scores on the second attempt will overestimate true proficiency.[7]

Researchers have extensively studied retest effects in the context of multiple-choice tests. Results generally indicate that repeat examinees obtain significantly higher scores on their second attempt, and that this benefit is considerably more pronounced for examinees who see identical test items on their second occasion.[8] One particularly relevant study reported score gains exceeding a standard deviation (*SD*) on the reused portion of a medical school admissions test used in Belgium. Furthermore, the factor structure changed for repeat examinees from their first to second attempt, and scores on the second attempt were less predictive of performance in medical school.[9] The advantage of seeing the same form twice does not appear to occur on licensure and certification tests, at least for the few studies that have been conducted. Although

scores improve on the second attempt, the average increase appears to be the same whether examinees see the same test form or a different form on the second occasion.[10] Nor does the same-test advantage appear to hold for clinical skills exams. One study of Step 2 CS found gains averaging about 0.87 standard deviation (*SD*) units across all examinee groups and skill domains, but there was no additional advantage for examinees who saw one or two of the same cases on their second attempt.[7] Studies at medical schools have reported similar findings.[11-13]

The cumulative findings suggest that while repeat examinees experience large score gains on performance tests, the increase cannot be attributed to memorization of test content. However, the possibility remains that construct-irrelevant variance explains some portion of the score gain. The purpose of the present research was to further investigate the validity of scores for examinees who repeat the Step 2 CS exam by evaluating its internal structure as well as its relationship to external variables. The internal and external characteristics of a test are important aspects of validity,[1] but to our knowledge no studies have investigated both properties for a sample of repeat examinees. The present study used correlations and factor analysis to evaluate the internal structure of Step 2 CS scores separately for single-take and repeat examinees. Although a dissimilar correlational structure for single-take and repeat examinees would suggest differences in the constructs being assessed, such results would leave unanswered any questions regarding which scores were more or less valid. Therefore, we also examined the relationships between Step 2 CS scores and scores on three external measures of physician knowledge and skill for repeat examinees. Prior studies report external correlations that range from about .10 to .40;[3] similar correlations in the present study would support the external validity of scores for repeat examinees.

Method

Instrumentation

Step 2 CS is designed to measure the clinical skills in four domains: communication-interpersonal skills; spoken English proficiency; data gathering; and patient note documentation. Successful completion of Step 2 CS is required for entry into graduate medical education (residency) in the U.S.; therefore, students generally take this exam just prior to graduating from medical school and/or immediately prior to entering residency. The exam is administered five or six days a week, year-round, at five testing centers throughout the U.S.(Atlanta, Chicago, Houston, Los Angeles, and Philadelphia). Exam forms generated daily within and across test centers to ensure that examinees who test on one day do not see the same cases that were

administered on the previous or subsequent days. Exams are assembled according to a detailed blueprint to ensure that different forms are comparable in terms of case difficulty and content. Due to logistic constraints some examinees who repeat Step 2 CS may see the same case on two occasions, but this occurs for only 6% of all encounters.

Examinees encounter 12 cases during a testing session, with each case portrayed by a different SP. During each encounter, examinees have up to 15 minutes to interact with the SP. Examinees are informed of the reason for patient's visit prior to entering the SP's room, and are instructed to take a medical history and perform a physical examination. At the conclusion of the encounter, examinees have 10 minutes to document their findings in a structured patient note. The SPs use these 10 minutes to complete the checklist and rating scales that result in scores for data gathering, communication- interpersonal skills, and spoken English. Patient note ratings are assigned subsequent to the examination by trained physicians.

Approximately 34,000 examinees take Step 2 CS each year. To pass examinees must exceed cut scores in each of three areas: (a) communication-interpersonal skills; (b) spoken English proficiency; and (c) a composite consisting of data gathering and patient notes. An examinee who fails in one or more of these three areas and wishes to take the exam again must repeat the entire Step 2 CS. Fail rates average about 14% each year. Of those who fail, 59% fail communication-interpersonal skills, 50% fail the composite of data gathering and patient notes, and 22% fail spoken English proficiency. Some examinees fail more than one area; thus, the sum of these percentages is greater than 100. Although pass-fail decisions are based on the three areas just described, this study analyzes scores on all four domains (i.e., communication-interpersonal skills, spoken English, data gathering, and patient notes).

Scores were also available for the three written examinations: The Step 1 exam is a measure of the basic science (BS) knowledge; the Step 2 CK exam is a measure of clinical knowledge; and the Step 3 exam assesses one's ability to apply clinical knowledge to patient management (PM). These tests are designated as Step 1 BS, Step 2 CK, and Step 3 PM in this paper. Although the time interval from the first test to the last test varies, particularly for IMGs, the three tests are generally taken in that order.

Participants

The potential sample consisted of all examinees completing Step 2 CS between July 2007 and September 2009 under normal test administration conditions. Participants had given prior approval for their scores to be used for research purposes and were deemed by an NBME

research panel to be exempt from IRB approval. All personal identifying information had been removed from examinee records prior to analysis. The data of interest initially included 5,184 examinees who completed Step 2 CS on two occasions. As previously noted, 22% of repeat examineesfailed spoken English proficiency. Given that Step 2 CS performance is positively influenced by English proficiency in the general population of examinees,[3] we were concerned that the data contained a potential confounding factor: that any differences in correlations between single-take and repeat examines might be a function of differences in English language skills rather than retest status. Therefore, we matched single-take and repeat examinees on spoken English proficiency scores.

Matching proceeded in two steps. First, examinees who failed spoken English were excluded from the sample. Second, all single-take examinees and repeat examinees were matched on spoken English proficiency at every score level. It was apparent that the matching process would result in a ratio of single-take to repeat examinees of approximately three to one. Therefore, a random sample of scores for single-take examinees was drawn with the constraint that for every spoken English score (60, 61, 62, …, 82) there would be three times as many single-take examinees as repeat examinees. This matching produced almost identical score distributions on spoken English for single-take and repeat examinees.

The final sample included 12,090 single-take examinees and 4,030 repeat examinees. Table 1 summarizes the demographic characteristics of the two groups. The repeat group contains a larger proportion of males and smaller proportion of females than the single-take group. In addition, compared to single-take examinees the repeat group has a smaller proportion of graduates of U.S. medical schools and a larger proportion of U.S. citizens who graduated from an international medical school. The proportion of true IMGs (i.e., IMGs who are *not* U.S. citizens) is nearly identical in the two groups of examinees (64%, 63%).

Analyses

Examinee scores were assigned to three "groups" based on their repeat status: single-take examinees; repeat examinees on their first attempt (repeat-1); and repeat examinees on their second attempt (repeat-2). That is, all repeat examinees were measured twice on all four Step 2 CS domains. We completed three sets of analyses. We first obtained descriptive statistics to determine the magnitude of the score gains for repeat examinees, and to compare correlations among the four subscores internal to Step 2 CS. Second, we used multigroup confirmatory factor analysis to evaluate the similarity of the correlations for single-take and repeat examinees.

The primary purpose of the confirmatory factor analysis for this investigation was to formally assess the equivalence of the correlation matrices and factor structure of test scores for two groups of examinees.[14] At the first stage the model assumes that the factor structure is the same by constraining factor loadings to be equal across groups. At the second stage, the factor loadings are unconstrained, allowing each group to have its own factor structure. Model fit is evaluated at each stage using a conventional χ^2 goodness-of-fit test, which indicates the degree to which the observed correlation or covariance matrix is predicted by the factor model.[15] If the unconstrained model at the second stage provides a significantly better model fit than the constrained model at stage one, then it can be concluded that each group is best described by its own factor structure.[14] The χ^2 is useful for statistical testing, but does not lend itself to useful interpretation because large sample sizes tend to produce large and statistically significant results. Therefore, the comparative fit index (CFI) was also reported; it is an R^2 type of statistic ranging from 0 to 1, with values close to 1 indicating good fit. We conducted two confirmatory factor analyses. The first compared single-take examinees to repeat-1 examinees, while the second compared single-take examinees to repeat-2 examinees. Finally, we obtained correlations between Step 2 CS scores and scores on the three written external measures separately for each group. These correlations were based on scores from the first attempt for the three external measures. The SPSS software package[16] was used to compute descriptive statistics and correlations, while MPlus[14] was used for the factor analyses.

Results

Means and Correlations

Table 2 presents means, *SD*s, and correlations for the three sets of scores. The nearly identical means and *SD*s on spoken English proficiency for single-take and repeat-1 examinees is a consequence of the matching process for the present sample. Repeat examinees exhibited increases in mean scores for all four areas, but most notably for CIS. Eighty percent of repeat examinees passed all areas, a value which approaches the pass rate of 86% for first-time examinees. The score increases are slightly less than those reported by Swygert,[7] which might be due in part to matching on spoken English scores. The *SD*s for single-take and repeat examinees are 7.2, 7.3, and 6.8. This is noteworthy because the similarity in *SD*s contributes to the interpretability of subsequent analyses because too little variability for one or more groups would suppress the correlations.

Consistent with prior research, correlations for single-take examinees are positive and moderately strong, ranging from 0.24 to 0.56.[3] However, correlations for the repeat-1 examinees are mostly low, with three of the six being negative. Some of the differences in correlations between single-take and repeat-1 examinees are large, particularly those involving the two communication scales. For single-take examinees, the correlation between communication-interpersonal skills and data gathering is 0.47 and the correlation between communication-interpersonal skills and patient notes is 0.53. In contrast, those correlations for repeat-1 examinees are – 0.25 and – 0.15. The correlations of spoken English with data gathering and with patient notes are also unexpectedly low (– 0.17 and 0.15). In other words, for repeat-1 examinees, performance on the two communication measures appears to be independent of obtaining a history, performing a physical, and documenting findings. Meanwhile, for repeat-2 examinees, the negative and low correlations move in a positive direction. The largest shift in correlation is for, communication-interpersonal skills and data gathering, which changes from – 0.25 to 0.36. The results suggest that the constructs assessed for repeat-1 examinees are different from the constructs assessed for single-take examinees, but that the differences diminish by the time repeat examinees complete their second attempt. The confirmatory factor analysis provides a formal evaluation of this observation.

Factor Analyses

The first multigroup confirmatory factor analysis compared correlation matrices for single-take examinees to repeat-1 examinees. Given the unusual correlations for repeat-1 examinees, reasonable model fit was achieved only by allowing certain error terms to correlate. Even so, the fit for the constrained model at stage one was quite poor (CFI = .78; χ^2 = 3,611), while the unconstrained model at stage two fit slightly better (CFI = .86; χ^2 = 2,349). The difference in model fit is statistically significant by the χ^2 difference test (χ^2_{diff} = 1,262, $P < .001$), indicating that the groups have different underlying factor structures. The second confirmatory factor analysis compared single take examinees to repeat-2 examinees. Fit indices for the constrained model at stage one were good (CFI = .96; χ^2 = 663), and improved only slightly for the unconstrained model at stage two (CFI = .96; χ^2 = 625). However, the unconstrained model did provide significantly better fit (χ^2_{diff} = 38, $P < .001$). Although statistically significant, the small value of χ^2_{diff} and the identical CFIs indicate that the differences between single take and repeat-2 examinees are small.

Correlations for the three groups were subjected to single-group factor analyses to further investigate differences. The results for the single factor solutions appear in Table 3. The factor loadings for single-take examinees are high and positive, ranging from 0.60 to 0.78. In addition, the model fit is good (CFI = .93). The factor loadings for the repeat-1 group are different from those for single-take examinees, ranging from – 0.39 to 0.75; furthermore, model fit is very poor (CFI = 0.51). The negative loadings indicate that the two communication scales are inversely related data gathering and patient notes for repeat-1 examinees. Meanwhile, the factor loadings for repeat-2 examinees range from 0.40 to 0.69; they are similar to the loadings for single-take examinees and different from the loading for repeat-2 examinees. The notable difference in factor loadings between single-take and repeat-2 examinees is for spoken English proficiency (0.64 vs. 0.40). That is, spoken English proficiency is less related to overall clinical proficiency for repeat-2 examinees than it is for single-take examinees.

Correlations with External Criteria

Correlations between the four Step 2 CS skill domains and the three external measures appear in Table 4. As noted in the footnote to Table 4, the number of examinees is different for the three external measures. Step 3 PM has the fewest examinees because many did not yet have scores available. The twelve correlations for single-take examinees range from 0.16 to 0.44, with a median of 0.33; these values are comparable to those reported in previous studies.[3] Correlations for repeat-1 examinees range from – 0.04 to 0.31, with a median of 0.15. The largest difference in correlations between repeat-1 examinees and single-take examinees is for communication-interpersonal skill and Step 3 PM (0.02 vs. 0.42). Meanwhile, correlations are slightly higher at repeat-2 than repeat-1, ranging from 0.0 to 0.37, with a median of 0.27. The correlations for repeat examinees generally approach the magnitude of the correlations for single-take examinees as repeaters move from their first to second attempt. The largest increases in correlation from repeat-1 to repeat-2 are for communication-interpersonal skill and Step 2 CK (– 0.01 to 0.24) and for communication-interpersonal skill and Step 3 PM (0.02 to 0.27). Taken as a whole, the correlations indicate that criterion-related validity of scores on Step-2 CS improves for repeat examinees on their second attempt.

Discussion and Conclusions

Two lines of evidence suggest that the construct underlying performance on Step 2 CS is markedly different for single-take examinees and repeat examinees on their first attempt. Not only were correlations and factor structure among Step 2 CS components weak and difficult to

interpret for repeat examinees on their initial attempt, but correlations of Step 2 CS scores with external measures of medical knowledge were lower than expected for this group . However, much of the difference between repeat and single-take examinees diminished with experience. By the time repeat examinees completed their second attempt, their factor structure was similar to that of single-take examinees and their correlations with external measures approached expectations. These outcomes extend previous research[7,11-13] by going beyond score gains to evaluate the impact of these gains on validity.

One puzzling outcome concerns the correlation between spoken English and other scores. Although the groups were matched, the relationship between spoken English proficiency and other scores still varied based on repeater status. Additional analyses confirmed that lower spoken English scores were associated with *higher* scores on data gathering for both native and nonnative speakers of English, suggesting that impaired performance on the first attempt is not a simple function of language differences. It may be, for example, that examinees who struggle with English speak enough in the encounter to ensure that they cover all or most data gathering checklist questions, but this exposes more fully the limitations of their English skills

These findings raise questions regarding the source and validity of the score gains. Although some of the gain can be attributed to examinees improving their clinical skills, other factors may contribute. Part of the increase can be attributed to random measurement error. Even in the absence of any improvement in proficiency, low scores on performance tests tend to regress toward the mean on retesting by nontrivial amounts.[17,6] The shift in correlations also implicates construct-irrelevant variance as a possible source. If construct-irrelevant variance is introduced after the first attempt, then the validity of scores for the second attempt will be compromised. This could occur when repeat examinees learn certain test-taking tactics between their first and second attempts, and these tactics are not available to most examinees on their first attempt. That is, scores may increase because examinees have become skilled test takers rather than skilled clinicians. In contrast, if construct-irrelevant variance is reduced after the first attempt, then score validity on the second attempt should improve. This occurs when performance on the first attempt is suppressed by factors irrelevant to the construct being measured – such as anxiety or unfamiliarity with an assessment format – but which become neutralized by the second attempt. The consequence is that the latter scores will more accurately reflect an examinee's true proficiency.

The existence of a large practice effect implies that some examinees are not well-prepared for the innovative SP format on their initial attempt. Given that the vast majority of U.S. medical schools now use SP-based clinical skills exams for student evaluation, it is likely that most US graduates now have considerable preparation with this format.[18] The situation for IMGs is not as clear, but prior experience with the SP format is certain to be less consistent, and it is more common for IMGs to lack formal training with the SP format. That much said the unusual pattern of correlations observed for repeat-1 examinees did not appear to be attributable solely to IMG status. Although IMGs are more likely to repeat Step 2 CS,[7] the sampling method employed here resulted in approximately equal percentages of IMGs in each group (single take = 64% IMG; repeat = 63% IMG). Regardless of country of medical education, medical schools likely differ in the extent to which their SP exams are similar to Step 2 CS, and examinees from schools with less similar assessment formats may feel less certain and more challenged on their first attempt.

Further research is needed to better understand the relationships between IMG status, English-language fluency, and test performance for repeat examinees. We have plans to evaluate scores gains and patterns of correlations for each of these groups; the practical problem is that sample sizes become exceedingly small as repeat examinees are partitioned into groups based on IMG status and English as a first language. It also would be informative to investigate practice effects *within* a testing session. Previous research detected a sequence effect by which examinees perform better after their initial few SP encounters;[19] future studies should evaluate the magnitude of the within-session effect for repeat or other low scoring examinees. Additional studies might seek to verify our assumption that lack of experience with the SP format is a source of construct-irrelevant variance for some examinees (e.g., by surveying repeat examinees). Such results could identify interventions that would help minimize the effect of test format. Plans are also underway to determine the portion of score increases that can be attributed to random measurement error, because such gains may have implications for the manner in which passing scores are established.[6,17]

In summary, we interpret the more consistent correlations for repeat examinees on their second attempt as a sign that construct-irrelevant variance was reduced and that inferences based on scores from the second assessment will be more valid than inferences based on the first attempt. It appears as if the first attempt served as a practice test for many repeat examinees, and once they learned the format, they were better equipped to demonstrate their skill.[4] However,

alternative explanations are certainly possible. For example, the improved correlation between Step 2 CS and Step 1 BS might be attributable to a type of test-taking skill that is common to most assessment formats.

Even if the practice effect explanation is correct, characteristics of the study design may limit the extent to which results generalize to other settings, including medical schools. The SP exam studied here was completed by a heterogeneous and highly motivated group of examinees. A more homogeneous group of less motivated examinees may change their behaviors less between their first and second attempts. Another potential limitation is that examinees in the present study had a low probability of seeing the same case or SP on two occasions. Score gains on exams where examinees see identical content may produce different outcomes than we observed,[9] although previous studies suggest not.[12,13] A related issue is that Step 2 CS is highly standardized, which means that many of its features remain the same across time. Even if the content changes, repeat examinees know what to expect. It is possible that results would be different for less standardized SP exams. In short, although the present results may not generalize to assessment contexts too dissimilar from the one studied here, they do serve as a reminder to medical educators in general that some caution is required when drawing inferences from test scores of examinees who have limited experience with a novel assessment format.[4]

References

1 Messick S. Standards of validity and the validity of standards in performance assessment. Ed Meas: Issues & Practice. 1995;14(4):5-8.

2 De Champlain AF, Swygert KA, Swanson DB, Boulet JR. Assessing the underlying structure of the USMLE Step 2 test of clinical skills using confirmatory factor analysis. Acad Med. 2006;81(10 suppl):S17-S20.

3 Harik P, Clauser BE, Grabovsky I, Margolis MJ, Dillon GF, Boulet JR. Relationships among subcomponents of the USMLE Step 2 clinical skills examination, the step 1, and the Step 2 clinical knowledge examinations. Acad Med. 2006;81(10 suppl):S21-S24.

4 Anastasi A. Coaching, test sophistication, and developed abilities. Am Psychol. 1981;36:1086-1093.

5 Rowland-Morin PA, Burchard KW, Garb JL, Coe NPW. Influence of effective communication by surgery students on their oral examination scores. Acad Med. 1991;66:169-171.

6 Raymond MR, Luciw-Dubas, UA. The second time around: Accounting for retest effects on oral examinations. Eval & the Health Prof. 2010;33:386-403.

7 Swygert KA, Balog KP, Jobe A. The impact of repeat information on examinee performance for a large-scale standardized-patient examination. Acad Med. 2010;1506-1510.

8 Hausknecht JP, Halpert JA, Di Paolo NT, Gerrard MO. Retesting in selection: A meta-analysis of practice effects for tests of cognitive ability. Jnl of Appl Psychol. 2007;92:373-385.

9 Lievens F, Reeve CL, Heggestad ED. An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. Jnl of Appl Psychol. 2007;92:1672-1682.

10 Raymond MR, Neustel S, Anderson D. Same-form retest effects on credentialing examinations. Educ Meas: Issues & Practice. 2009;28:19-27.

11 Boulet JR, McKinley DW, Whelan GP, Hambleton RK. The effect of task exposure on repeat candidate scores in high-stakes standardized patient assessment. Teach and Learn Med. 2003;15:227-232.

12 Niehaus AH, DaRosa DA, Markwell SJ, Folse R. Is test security a concern when OSCE stations are repeated across clerkship rotations? Acad Med. 1996;71:287-289.

13 Swartz, M.H., Colliver, J.A., Cohen, D.S., & Barrows, H.S. (1993). The effect of deliberate, excessive violations of test security on performance on a standardized patient examination. Acad Med. 1993;68(10 suppl): s76-s78.

14 Muthen LK, Muthen BO. Mplus user's guide. 3rd ed. Los Angeles: Muthen & Muthen; 2001.

15 Bentler PM. Comparative fit indexes in structural models. Psychol Bull. 1990;107:238-246.

16 SPSS, Inc. SPSS base 16.0 user's guide. Chicago: SPSS Inc.; 2008

17 Clauser BE, Nungester RJ. Classification accuracy for tests that allow retakes. Acad Med. 2001;(10 suppl):s108-s110.

18 Barzansky B, Etzel, SI. Educational programs in the U.S. medical schools, 2003-2004. JAMA 2004;292:1025-1031.

19 Ramineni C, Harik P, Margolis MJ, Clauser BE, Swanson DB, Dillon GF. Sequence effects in the United States Medical Licensing Examination (USMLE) Step 2 clinical skills (CS) examination. Acad Med. 2007;82(10 suppl):S101-S104.

Table 1. Type of Medical Education and Gender of Single-Take and Repeat Examinees Included in the Sample.

	Single-Take		Repeat	
Examinee Group	**N**	**%**	**N**	**%**
Medical Education				
U.S. medical graduates	3350	27.7	791	19.6
U.S. international medical graduates	986	8.2	702	17.4
International medical graduates	7754	64.1	2537	63.0
Total	12090	100.0	4030	100.0
Gender				
Male	6703	55.4	2898	71.9
Female	5387	44.6	1132	28.1
Total	12090	100.0	4030	100.0

Table 2. Mean Scores, Standard Deviations (*SD*s), Reliabilities, and Correlations on Step 2 CS for Single-Take Examinees and Repeat Examinees on Their First Attempt and Second Attempt.

				Internal Correlations*			
Group	**Step 2 CS Domain**	**Mean**	**SD**	**Comm-Interpers**	**Spoken English**	**Data Gathering**	**Patient Note**
Single Take (n = 12090)	Communication-Interpersonal Skill	70.6	6.6	*.81*	0.56	0.47	0.53
	Spoken English Proficiency	72.1	7.2		*.94*	0.24	0.47
	Data Gathering	67.2	9.7			*.69*	0.53
	Patient Note	69.6	9.7				*.73*
Repeat-1 (n = 4030)	Communication-Interpersonal Skill	62.7	6.3	*.70*	0.37	- 0.25	- 0.15
	Spoken English Proficiency	72.0	7.3		*.94*	- 0.17	0.15
	Data Gathering	56.2	10.4			*.70*	0.40
	Patient Note	60.2	8.8				*.67*
Repeat-2 (n = 4030)	Communication-Interpersonal Skill	70.3	5.8	*.72*	0.43	0.36	0.41
	Spoken English Proficiency	73.1	6.8		*.91*	0.07	0.24
	Data Gathering	64.4	9.5			*.65*	0.48
	Patient Note	67.1	8.8				*.67*

* Italicized values on the diagonal are reliability (phi) coefficients.

Table 3. Standardized Loadings from Exploratory Single-Group Factor Analyses of Step 2 CS for Single-Take Examinees and Repeat Examinees on their First Attempt and Second Attempt.

Step 2 CS Domain		**Single Take**	**Repeat-1**	**Repeat-2**
Communication-Interpersonal Skill		.78	.75	.65
Spoken English Proficiency		.64	.46	.40
Data Gathering		.60	– .39	.59
Patient Note		.74	– .20	.69
Model Fit	χ^2	1400	963	561
	CFI	.93	.51	.81

Table 4. Correlations of Step 2 CS Domains with External Measures for Single-Take Examinees and Repeat Examinees on Their First Attempt and Second Attempt.

		External Written Measures*		
Group	**Step 2 CS Domain**	**Step1 BS**	**Step 2 CK**	**Step 3 PM**
Single Take	Communication-Interpersonal Skills	0.27	0.31	0.42
	Spoken English Proficiency	0.16	0.17	0.36
	Data Gathering	0.30	0.34	0.32
	Patient Note	0.38	0.41	0.44
Repeat-1	Communication-Interpersonal Skills	-0.04	- 0.01	0.02
	Spoken English Proficiency	-0.01	- 0.01	0.15
	Data Gathering	0.21	0.22	0.16
	Patient Note	0.29	0.31	0.26
Repeat-2	Communication-Interpersonal Skills	0.19	0.24	0.27
	Spoken English Proficiency	0.01	0.00	0.17
	Data Gathering	0.27	0.30	0.26
	Patient Note	0.34	0.37	0.34

* External written measures: BS = basic science; CK = clinical knowledge; PM = patient. management. Sample sizes for single take examinees were 11,463, 10617, and 3,522 for Step 1 BS, Step 2 CK, and Step 3 PM; sample sizes for repeat examinees were 3,937, 3,789, and 1,328 for Step 1 BS, Step 2 CK, and Step 3 PM.