BMC Medical Education

Research article

How well do second-year students learn physical diagnosis? Observational study of an objective structured clinical examination (OSCE)

Claus Hamann*[1], Kevin Volkan[2], Mary B Fishman[3], Ronald C Silvestri[4], Steven R Simon[5] and Suzanne W Fletcher[5]

Address: [1]Geriatric Medicine Unit, Massachusetts General Hospital, 100 Charles River Plaza, Fifth Floor, Boston MA, USA, [2]Program in Psychology, California State University Channel Islands, Professional Building, University Drive, Camarillo, CA 93012, USA, [3]Division of General Internal Medicine, Georgetown University Medical Center, 3800 Reservoir Rd. NW, Washington DC 20007, USA, [4]Department of Medicine, Beth Israel Deaconess Medical Center, 330 Brookline Ave, Boston, MA 02215, USA and [5]Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, 133 Brookline Avenue, Sixth Floor, Boston, MA, 02215, USA

E-mail: Claus Hamann* - chamann@partners.org; Kevin Volkan - kevin.volkan@csuci.edu; Mary B Fishman - MPF3@gunet.georgetown.edu; Ronald C Silvestri - rsilvest@caregroup.harvard.edu; Steven R Simon - steven_simon@hms.harvard.edu; Suzanne W Fletcher - suzanne_fletcher@hms.harvard.edu

*Corresponding author

Published: 10 January 2002

BMC Medical Education 2002, **2**:1

Received: 11 October 2001
Accepted: 10 January 2002

This article is available from: http://www.biomedcentral.com/1472-6920/2/1

Abstract

Background: Little is known about using the Objective Structured Clinical Examination (OSCE) in physical diagnosis courses. The purpose of this study was to describe student performance on an OSCE in a physical diagnosis course.

Methods: Cross-sectional study at Harvard Medical School, 1997–1999, for 489 second-year students.

Results: Average total OSCE score was 57% (range 39–75%). Among clinical skills, students scored highest on patient interaction (72%), followed by examination technique (65%), abnormality identification (62%), history-taking (60%), patient presentation (60%), physical examination knowledge (47%), and differential diagnosis (40%) ($p < .0001$). Among 16 OSCE stations, scores ranged from 70% for arthritis to 29% for calf pain ($p < .0001$). Teaching sites accounted for larger adjusted differences in station scores, up to 28%, than in skill scores (9%) ($p < .0001$).

Conclusions: Students scored higher on interpersonal and technical skills than on interpretive or integrative skills. Station scores identified specific content that needs improved teaching.

Background

Learning the skills of physical diagnosis is a critical part of the medical school curriculum. While there is widespread agreement on what skills should be learned [1,2], there is little information on how well those skills are learned, especially among second-year students. Measuring skill acquisition objectively is the essential first step in improving clinical competence throughout undergraduate and postgraduate training [3,4].

During the past 25 years, the objective structured clinical evaluation or examination (OSCE) has become an impor-

tant method of assessing skills at all levels of medical training [5,6], complementing traditional evaluations of knowledge that use written multiple choice questions and essay questions. Compared with other levels of training [7], little is known about the use of the OSCE in physical diagnosis courses for second-year medical students.

Several studies have used the OSCE to assess the effect of educational interventions on specific skills at the second-year level, such as history-taking for smoking [8], or examination of low back pain [9] or the breast [10,11]. Others have examined the use of different examination personnel as examiners or patients [12–14], compared students' course feedback to their OSCE performance [15], examined costs [12,16] or reliability and generalizability [7], compared training locations [17] or provided general descriptions of their OSCE's [18–22]. We found no studies that have used the OSCE to report comprehensively on the spectrum of skills learned in a second-year physical diagnosis course. A comprehensive investigation is likely to help determine what aspects of the educational process should be improved.

We used the OSCE to examine how well second-year students learned clinical skills in the second-year physical diagnosis course at Harvard Medical School. We were particularly interested which skills students performed best and which were most difficult. We assessed what factors affected their performance on the overall OSCE, and on individual skills and stations. Finally, we examined whether student OSCE scores varied from year to year, medical students performed differently from dental students, learning at different teaching sites affected student performance, and preceptors and examination logistics affected student scores.

Methods

Setting

This study took place at Harvard Medical School as part of the required second-year physical diagnosis course, Patient-Doctor II [4]. The course is taught from September to May in the same general sequence at 9 clinical sites affiliated with the medical school. Each site is assigned 6–45 students for the entire 220-hour course, including a total of 30 second-year students from Harvard School of Dental Medicine. These dental students are preparing for careers in consultative dentistry and are required to learn the same clinical skills as Harvard medical students. The course involves a total of almost 700 faculty members. One or two faculty members at each site function as site director(s) and are intimately involved in teaching the students and organizing other faculty to teach in the course.

Teaching sessions are organized by organ system. Students first learn skills by practicing on each other and by taking histories and performing physical examinations on selected patients. Each year, approximately 130 medical students and 30 dental students participate in the course. Site directors meet monthly as a group to determine the curriculum, teaching techniques, and evaluation of the course.

Objective structured clinical examination (OSCE)

Development

We developed our OSCE primarily for educational purposes: to identify skills that each student has learned well and those that need improvement during the final portion of the course. Performance on the OSCE is not formally factored into a student's grade for the course, but individual student OSCE scores are reviewed by site directors.

We designed the OSCE stations in 1994, pilot-tested them at evaluation sessions held in 1995 and 1996, and reported on our results for 1996 [23]. Following established methods [7,24,25], the course director and a committee of site directors and 4th-year student representatives developed case scenarios, detailed instructions and checklists consisting of questions or tasks for 16 stations focused on specific clinical areas. From 1994–1996, we refined the content of the stations and the OSCE organization through frequent discussions with all site directors and through feedback from students and OSCE preceptors. We made no changes to the exam during 1997–1999. Site directors determined that all OSCE questions reflected essential skills to be mastered by second-year students. We did not weight OSCE questions, stations or skills according to degree of difficulty. Annual feedback from students and faculty endorsed the face validity of the OSCE. In 1999, 90% of students and 91% of faculty agreed that the OSCE represented an appropriate and fair evaluation method, and that enough time was given to complete the stations.

In the 16-station OSCE, nine different formats were used alone or in combination: question and answer, preceptor role play, standardized patients, actual patients, mechanical or structural models, 35-mm slides, audiotape, videotape, and CD-ROM (Table 1). OSCE committee members designated each question or task in the 16 stations as one of 7 clinical skills, defined as follows: asking appropriate questions for the history (history-taking); performing the physical examination correctly (physical examination technique); understanding the pathophysiology of physical findings (physical examination knowledge); identifying abnormalities on physical examination (identification of abnormalities); developing appropriate differential diagnoses for the clinical information obtained (differential diagnosis); utilizing appropriate patient-doctor interaction techniques (patient interaction); and orally presenting the history and differential diagnosis after taking a clinical history (patient presentation). The total number

Table 1: Number of questions devoted to each clinical diagnosis skill and OSCE station

OSCE STATIONS	Format	Content	CLINICAL DIAGNOSIS SKILLS							TOTAL[2]	No. of skills per station[3]
			Patient Interaction	Physical Examination Technic	Indentification of Abnormality	History-taking	Patient Presentation	Physical Examination Knowl	Differential Diagnosis		
Abdominal Pain	Role play	History, review of systems Pain and differential diagnosis of subacute abdominal pain	8			27			11	46	2
Alcohol / Abdominal Examination	Human model	Examining abdomen of normal Abdominal individual, stating potential Examination findings in an alcoholic patient		18				9		27	2
Arthritis[1]	Videotape	Describing and identifying hand and forearm findings in osteo- and rheumatoid arthritis			11				5	16	2
Breast	Silicone model	Examination technique, and detecting 5 lumps of varying sizes		9	5					14	2
Calf Pain	Question and answer	Describing findings relevant and answer to differential diagnosis of calf pain						9	30	39	2
Ear	Model, 35 mm slides	Otoscopy, and identifying 35 mm slides normal, bulging, and perforated drums		3	8			5	1	17	4
Headache	Role play	History, review of systems and differential diagnosis for environmental cause of headache	1			26			10	37	2
Heart	CD-ROM	Describing and identifying aortic stenosis and mitral regurgitation			13				2	15	2
Hemoptysis	Question and answer	History, review of systems, and answer differential diagnosis	5			13			12	30	2
Knee	Human model	Performing knee exam on normal individual and describing potential findings		12				14		26	2
Lung	Audiotape	Identifying wheezes and rhonchi, describing associated findings			2			12		14	2
Mental Status	Role play	Probing at least 5 domains of cognitive and affective function		10				10		20	2
Presentation	Question and answer	Orally presenting the history and answer and differential diagnosis from abdominal pain station					33			33	1
Rectal / Prostate	Plastic model	Verbally identifying possible /Prostate rectal conditions, and tactily identifying prostate findings			2			11		13	2
Skin	35 mm slides	Describing and identifying psoriasis, melanoma and basal cell epithelioma			20				3	23	2
Thyroid	Patient	Describing and performing the thyroid exam in a patient with findings		8	4					12	2
TOTAL No. % of questions			*14* 4%	*60* 16%	*65* 17%	*66* 17%	*33* 9%	*70* 18%	*74* 19%	382[4]	

[1]OSCE 1999 had one less question than OSCE 1997 and 1998. [2] Mean number of questions per skill: 55 (range 14–74) [3]Mean number of skills per station: 2.25 (range 1–4) [4]Mean number of questions per station: 24 (range 12–46)

of OSCE questions each year was 382, and the mean number of questions per skill was 55 (range 14–70), evenly distributed except for patient interaction and patient presentation.

Implementation

Each year, we held 10 sessions of the OSCE on 3 days (Monday, Wednesday and Friday afternoons) during a one-week period in April for all second-year students. Two consecutive, early and late afternoon sessions each consisted of the same 16 stations and lasted 2.5 hours. To accommodate all students, sessions were conducted simultaneously on 2 floors of the medical school's education center, for a total of 10 OSCE sessions. Other than by date and time, the sessions varied only in the assignment of preceptors. With the help of guides, timers and a strict schedule, students rotated through the 16 clinical stations, each precepted by a faculty member. All preceptors received standardized guidelines for checklists and feedback prior to each OSCE session, as did the standardized patients or actors for the abdominal pain, alcohol/abdominal exam, knee and thyroid stations. Fourteen stations were each 6 minutes in duration, and two – abdominal pain and headache – were 12 minutes in duration.

At each station, the student performed the indicated tasks for two-thirds of the time, while the faculty preceptor observed and checked off the tasks performed correctly, as defined by checklists, one for each student. All tasks performed or questions answered by each student were scored dichotomously as correct (1) or left blank (0) on the checklists. During the final one-third of time at each station, the preceptor provided feedback on the student's performance, as advocated by others [26]. Each year, approximately 150 preceptors participated in the OSCE, and 60% have had experience with this OSCE and the checklists from prior years.

Data collection and analysis

Correct answers to all OSCE questions were recorded on checklists by preceptors, double-entered by research staff into an ASCII file, and analyzed in SPSS [27]. Total OSCE, skill and station scores were calculated as follows. Each task or question counted one point, and the sum of tasks performed or questions answered correctly for each station was designated the station score. The sum of station scores produced a total OSCE score for each student. Means of students' scores ± one standard deviation for each of the 16 stations were computed. To compute the skills score, each task or question on the checklist for every station was designated as one of 7 skills. The sum of tasks performed or questions answered correctly for each skill produced a student's skill score. Means of students' scores for each of the 7 skills were computed. We combined the data from the 1997, 1998 and 1999 OSCE's.

Total OSCE score, scores for each clinical skill, and scores for each station were the primary outcome variables. In addition to the checklists completed by faculty preceptors at each station for each student, we collected data on student, preceptor and examination variables to examine factors that might predict students' OSCE scores. Student variables were type of student (medical or dental), and teaching site (Site A-I). The preceptor variables were the floor (first or third) and session group (early or late afternoon) assigned to each OSCE preceptor. Examination variables consisted of OSCE year (1997, 1998 or 1999), the day each student took the OSCE (first, second or third), and sequence of stations.

For all predictor variables, total OSCE, skill and station score means were compared with one-way ANOVA. Predictor variables significantly associated at $p < .05$ with students' total OSCE in univariate analysis were entered into a linear regression model, with the single dependent variable being a student's total OSCE score. The predictor variables were also entered into two multivariate analysis of variance (MANOVA) models, each of which included multiple dependent variables. As dependent variables, one model used clinical skill scores, and the second model used station scores. Separate models were used due to the high co-linearity between the skill and station scores, since both of these scores drew from the same item pool. P-values within each MANOVA model were adjusted for multiple comparisons. In addition, we set the threshold for judging statistical significance at $p <= .001$ to further reduce the influence of multiple comparisons on p values.

Because it was not logistically possible to obtain inter-rater reliability due to the large number of preceptors, we used generalizability theory analysis [28]. This analysis accounts statistically for rater error by parsing out the variance relevant to the instrument in question. By modeling the variances as separate characteristics, we isolated the variance due to student ability, which in classical test theory is equivalent to true score variance. Other variances related to the test are treated as error variances. In this framework, we treated error due to differences in raters as error variance.

We calculated the Kuder-Richardson-20 coefficient of reliability, KR-20, for the total OSCE score, clinical skill and station scores. The KR-20 [29] is used for binary items and is comparable to Cronbach's alpha. This measure of internal consistency is the best measure of reliability when there are many more than two raters. It is equivalent to the generalizability or G coefficient which examines total scale scores across raters in a D-study scenario (total scores

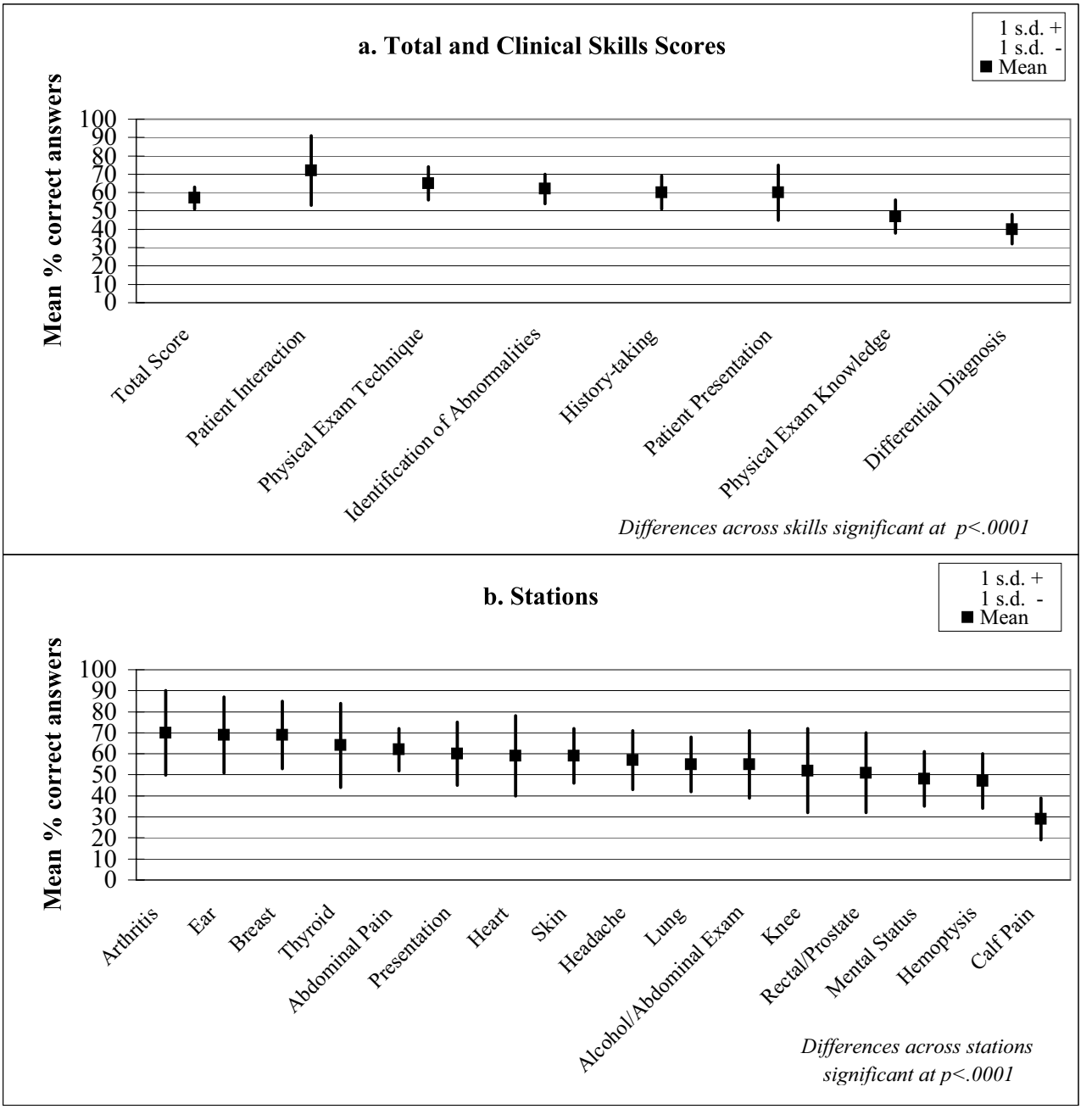

Figure 1
Students' mean scores on OSCE, 1997–1999

are normally distributed), when the main effect variance due to raters is assumed to be zero [30–32]. In our study, we assumed zero main-effect variance to be the average across the large pool of student raters, because student assignment to a preceptor for any given station was essentially random.

Results

Over three years, 489 second-year students (402 medical and 87 dental) and 445 faculty participated in the OSCE for second-year physical diagnosis course. Students answered slightly more than half of all the OSCE items correctly, 57% ± 6% (Figure 1a), with almost no change over

Table 2: Variables and groups showing the largest differences in OSCE scores

OSCE Score	Variable	Reference		Highest Scoring		Skill / Station	Largest Difference
		Group	Score (%)	Group	Score (%)		(95% C.I.)[2] %
Total	Student	Dental	51	Medical	57	n.a.[3]	6 (4, 7)
						Skill	
Clinical skills	Student	Dental	45	Medical	54	Patient Presentation	9(6, 12)
	Preceptor Group	Group 4	45	Group 3	53	Patient Presentation	8(4, 12)
	Teaching Site	Site I	40	SiteB	47	Physical Exam Knowledge	7 (3, 10)
	OSCE Day	Day 3	45	Day 2	51	Patient Presentation	6 (3, 9)
	OSCE Year	1999	62	1997	58	History-taking	-4 (-6, -2)
						Station	
Stations	Teaching Site	Site I	55	SiteH	83	Thyroid	28 (20–37)
	Student	Dental	46	Student	61	Rectal/Prostate	15 (11–19)
	Preceptor Group	Group 4	32	Group 2	46	Knee	14 (8–20)
	OSCE Day	Day 3	55	Day 1	66	Thyroid	11 (7, 15)
	OSCE Year	1999	60	1998	71	Skin	11 (8, 13)

We used MANOVA for comparisons among the predictor variables and multiple dependent variables. [1]Adjusted score difference denotes adjusted score difference between the highest scoring group and the reference group means. All differences were significant at p <= .001. [2]C.I. denotes confidence interval. [3]n.a. denotes not applicable.

3 years (p = .28). Individual student scores on the entire OSCE ranged from 39% to 75%.

For clinical skills, students scored highest on patient interaction (72%), followed by physical examination technique (65%), identification of abnormalities (62%), history-taking (60%), patient presentation (60%), physical examination knowledge (47%), and differential diagnosis (40%) (p < .0001, Figure 1a). Clinical skill scores remained stable from 1997–1999, with only slight improvement in history-taking (3% from 1997 to 1999, p = .0003).

For adjusted total OSCE scores, medical students scored 6% higher than dental students, 57% vs. 51% (p < .0001, Table 2). No other variable was found to predict total OSCE scores. For adjusted clinical skill scores, the largest score differences were associated with the student variable – medical vs. dental. Medical students' scores were 9% higher than dental students' scores for patient presentation (and were slightly but significantly higher for all other clinical skills except history-taking, not shown). Table 2 shows other significant differences among several tested variables and groups, but the absolute score differences for these variables were relatively small, 8% or less.

For OSCE stations, students scored highest on the arthritis (70%), ear (69%), breast (69%) and thyroid stations (64%), and lowest on the rectal/prostate (51%), mental status (48%), hemoptysis (47%) and calf pain stations (29%) (p < .0001, Figure 1b). We found statistically significant year-to-year differences among the means for 10 of 16 stations. However, absolute differences were small; the largest differences were 10% for the skin and mental status stations (data not shown).

When we examined the mean total, clinical skill and station scores according to student, preceptor and examination variables, we found many statistically significant associations in the univariate analyses. Multivariable analyses yielded fewer but still similarly significant results. Table 2 presents the highest scoring groups of predictor variables and the largest adjusted differences between the highest scoring and the reference groups.

Adjusted station scores demonstrated the largest differences, notably for teaching sites (Table 2). For the thyroid station, the scores of students at site H were 28% higher than scores for students at reference site I. Other predictor variables accounted for smaller differences. Medical students' adjusted scores on the rectal/prostate station were 15% higher than dental students' scores. They were also significantly – but less than 10% – higher for 8 other stations, and no different for 7 stations (not shown). Other variables – preceptor groups, OSCE day and OSCE year – also demonstrated some variation, with the largest differences being 14% among preceptor groups for the knee station.

Table 3: Mean OSCE station scores[1] and reliability[2] for teaching sites, 1997–1999

			Total OSCE[3]	Arthritis	Ear	Breast	Thyroid	Abdominal Pain	Presentation	Heart	Skin	Headache	Lung	Alcohol/ Abdominal Exam	Knee	Rectal/ Prostate	Mental Status	Hemoptysis	Calf Pain
		Number of Students	Mean scores (%)																
Teaching Site[3]	A	69	51	49	77	64	75*	55	50	45	55	60	**44***	35	32	42	58	44	27
	B	35	52	**41***	63	60	81*	53	43	54	60	66	55	35	**58***	48	55	43	24
	C	36	50	48	64	70	70*	54	45	45	56	62	49	43	39	50	**45***	40	30
	D	35	51	52	64	70	80*	55	42	53	**50***	57	50	37	46*	39	52	41	28
	E	35	50	57	**47***	67	68*	54	40	54	53	61	46	47	42	41	57	41	28
	F	135	52	51	71	67	73*	59	47	**56***	54	59	50	47	39	46	51	44	27
	G	42	**49**	57	73	62	61	57	43	43	58	65	49	36	27	45	54	43	27
	H	18	52	58	65	61	**83***	53	39	59	51	62	**44***	53	34	41	60	44	32
Largest difierence[4]			-2	-17	-23		28			11	-10		-11		25		-10		
Reference constant[5]			51	58	70	63	55	57	45	45	60	63	55	43	33	46	55	41	29
Reliability[2]			*0.86*	0.73	0.73	0.53	0.64	0.67	0.75	0.53	0.48	0.73	0.4	0.76	0.83	0.51	0.53	0.65	0.68

[1]Mean station scores were adjusted for type of student, teaching site, preceptor group, OSCE day, and OSCE year. Only the teaching site variable is shown here.

Linear regression models were calculated for total OSCE scores, and MANOVA models were calculated for the station scores. [2] Internal consistency was measured by the Kuder-Richardson-20 coefficient. [3] Reference Site I, N=84 students [4]Denotes largest difference of adjusted site mean (in bold font) from reference constant. [5]The reference constant comprises reference values for all predictor variables. * Means with asterisks are significant at p <= 001.

Because teaching sites demonstrated the greatest differences in OSCE station scores, even after adjustment for other variables, we examined detailed inter-site differences (Table 3). Eight adjusted station scores showed substantial and significant differences in student scores among teaching sites: thyroid (28%), knee (26%), ear (23%), arthritis (17%), heart (13%), mental status (11%), lung (11%) and skin (10%) ($p <= .001$). There were no significant inter-site differences for the breast, abdominal pain, presentation, headache, alcohol/abdominal exam, rectal/prostate, hemoptysis and calf pain stations. At every teaching site, adjusted scores for 1 or 2 stations were higher than at reference site I, while scores for 1 to 3 other stations were lower than those for the reference site.

The overall reliability coefficient for the OSCE of 382 items was .86 (Table 3), indicating good reliability of the OSCE total score [25,31,32]. The reliabilities of the clinical skill scores ranged from .57 to .77 (not shown). All but one of these scores – identification of abnormalities, .57 – had a reliability coefficient of .65 or higher. Reliabilities for clinical skill scores were generally higher than for station scores which ranged from .40 to .83 (Table 3).

Discussion

In an OSCE for a second-year physical diagnosis course, we found a similar pattern of clinical skill acquisition for three successive classes of students. Students performed better on interpersonal and technical skills – patient interaction, history-taking, physical examination technique, identification of abnormality, and patient presentation – than on interpretative or integrative skills – knowledge of the pathophysiology of physical examination findings, and differential diagnosis. Teaching sites differed widely from one another in performance on individual OSCE stations, only modestly on clinical skills, and not at all on total OSCE scores. Medical students scored somewhat better than dental students on the overall OSCE, all clinical skills except history-taking, and almost half of the stations.

To our knowledge, this study is the first to examine comprehensively student performance for general clinical skills and specific OSCE stations at the second-year student level. Other studies of OSCE's for second-year students have focused on specific skills or content [8–11], or logistics and psychometrics [7,12,16]. None of the other studies employed multivariable analysis in examining factors associated with OSCE performance. By including such analysis, we were able to hold student and examination variables constant in order to determine what parts of the curriculum students mastered best and which sites best taught specific physical diagnosis content.

Higher scores on technical and patient interaction skills, compared to integrative skills, are not surprising. Students at Harvard and in many medical schools begin to practice some interviewing, history-taking and patient interaction during the first year curriculum, and they spend the entire second-year physical diagnosis course learning the techniques of physical examination. Investigators have reported similar results in other settings. OSCE scores among clinical clerks were higher on history-taking/physical examination skills (mean score ± s.d., 61 ± 4%) and interviewing skills (69 ± 11%), and lower on problem solving (50 ± 6%) skills [33]. In a non-OS CE examination using patient management problems, second-year students scored 70 ± 9% on history, 66 ± 10% on physical examination, and 40 ± 15% on diagnosis [34]. However, in an OSCE for a second-year neurology skills course, this pattern did not hold: interpretative skill scores (76 ± 16%) were higher than technical performance scores (67 ± 17%), but no significance testing was reported [15].

Differential diagnosis has traditionally been considered a secondary goal of our physical diagnosis course, so performance might be expected to be lower. However, pathophysiology of disease is a major focus of the second-year curriculum. Lower performance in knowledge of the pathophysiology related to physical diagnosis, compared with technical performance of the physical examination, suggests that improvements integrating pathophysiology into the teaching of the history and physical examination are needed.

Our other key finding was the variable performance by students from different teaching sites on half the OSCE stations, despite similar performance by sites on the overall OSCE. Every site scored highest or next-to-highest on at least one station, and every site also scored lowest or next-to-lowest among sites on at least one station. Because of the large numbers of students in this study, even differences of 2% were statistically significant, but we consider differences greater than 10% to be educationally significant and worthy of targeted improvement efforts.

We found the largest differences for the thyroid, knee, ear, arthritis, heart, lung, mental status, and skin stations. While students may have acquired overall physical diagnosis skills similarly from site to site, our results suggest they did not learn equally at every site the skills required for adequate examination or understanding of these specific organ systems. Inter-site differences in content-specific station scores represent opportunities for teaching sites to learn from one another, using strategies such as structured clinical instruction modules [9,35] or reinforced practice [11] and developing more uniform learning objectives and curriculum.

Raw score results at one medical school must be interpreted with caution, since OSCE's at other schools may differ

in degree of difficulty. The mean total OSCE score of 57% ± 6% in our study compares favorably with results from one report on second-year students (52 ± 6%) [36], a report on clinical clerks (57 ± 4%) [33], and a study of third-year medicine students (58%) [3], but less favorably with another report on second-year students, 70% [12]. None of these studies adjusted their student scores.

Consistent with a prior study from the U.K. [37], we found that dental students scored lower than medical students, but not at a level which raises serious concerns about their participation in the physical diagnosis course. While dental students scored lower on the majority of stations, they performed as well as medical students on some stations with content that is not related to their ultimate professional focus, such as breast, mental status and abdominal pain.

This study has several limitations. We have not directly assessed inter-rater reliability because of logistical and cost constraints. To address this methodological concern, we used generalizibility theory (GT) to produce a measure of reliability similar in quality to inter-rater reliability [32].

There are a number of examples of the use of GT to account statistically for rater error [32,38–40]. Using GT can also overcome some problems inherent in inter-rater reliability, such as overestimating reliability [41]. Due to the large number of preceptors involved in our OSCE, we made the statistically reasonable assumption that any error due to rater differences is randomly distributed. Since randomly distributed error has a mean of zero, the error variance due to differences among all preceptors is zero. In our OSCE, the variation of individual raters around the mean station score of all raters is very close to 0 (e.g., .04 for the presentation station, data not shown), and the standard deviations of student scores are comparatively large (e.g., 15 for the presentation station). Finally, our GT-based assumption is especially appropriate when the test scores used in the analysis are created by summing many items across each scale. Summing in this fashion has the effect of further randomizing the error variance. The reliability, or internal consistency, of the overall OSCE was good at .86. The reliability of 6 of 7 skill scores, and 9 of 16 station scores, were acceptable at > .60.

Another benefit of the GT approach is that the reliability coefficient derived from the GT analysis is equivalent to Cronbach's alpha coefficient which, for binary items, is equivalent to the KR-20 reliability coefficient. The alpha coefficient is especially useful during test development because it gives a measure of how each item is contributing to the scale to which it has been assigned. This measure makes it easy to target items for editing or deletion if they are not performing well. Since we are ultimately interested in using the scale scores for our research study, the GT measure of reliability is appropriate for OSCE's involving many preceptors.

The validity of our OSCE is only partially established. While several features support its face and content validity, construct and criterion validity remain to be tested. Multiple refinements of stations over the two developmental years of the OSCE prior to this study yielded broad agreement among the teaching site directors that all OSCE questions reflected essential skills that should be taught to and mastered by second-year students. Five successive years of post-OSCE student and faculty evaluations have endorsed the OSCE as a highly appropriate and acceptable method of education and evaluation. Finally, a more recent investigation supports predictive validity of our OSCE. Physical diagnosis skills examined in the present study correlated with scores on the USMLE Step 1 exam, and the skills that foreshadow the clinical clerkships – identification of abnormality and development of differential diagnoses – best predicted USMLE scores [42].

Variation in skill scores may be due to different OSCE station content. Three of the skills drew their questions from a smaller number of stations: patient interaction, 3 stations; history-taking, 3 stations; presentation, 1 station. However, patient interaction and history-taking drew their questions from the same stations. More importantly, the remaining 4 skills each drew their questions from 6–8 stations. For these 4 skills (physical examination technique, physical examination knowledge, identification of abnormalities, and differential diagnosis), the range of case content is considerable and counters the concern that variation might be caused by case content rather than by student performance.

Variation in skill scores may be also due to inherent differences in the degree of difficulty of exam questions. In our exam, we did not weight OSCE questions according to degree of difficulty. We were not trying to create an exam in which all items were of equal difficulty. Instead, we created an OSCE in which the course directors considered all test items essential to be mastered by the students. The results showed variation in the degree to which the students mastered different clinical skills. Remarkable stability of overall scores over the three years of this study with three different cohorts of students provides evidence that there has been no significant "teaching to the OSCE." This finding is consistent with a prior study of fourth-year students [43].

The successful implementation of the OSCE at our medical school is relevant to all medical schools that face the logistical challenges posed by multiple sites and preceptors for student training in physical diagnosis. Further-

more, the results from the second-year OSCE reported here and our pre-fourth year OSCE [44] have been useful in helping to identify areas of weakness that could benefit from remediation prior to the start of clinical clerkships. This benefit is especially true for students with the lowest performance on individual stations and skills. For site directors and faculty, the OSCE has also helped identify those parts of the curriculum students had difficulty mastering. Holding a second-year OSCE prior to the end of a physical diagnosis course helps medical school faculty identify opportunities for remediation, focus the remaining sessions of the course, and improve future physical diagnosis teaching.

Conclusions

Objective identification of skills acquired in a physical diagnosis course is a necessary first step in improving the quality of both the teaching and the learning of those skills. In our OSCE for a second-year physical diagnosis course, students scored higher on interpersonal and technical skills than on interpretive or integrative skills. Station scores identified specific content needing improvements in students' integrative and organ system-specific skills of physical diagnosis, and in the teaching of these skills.

Competing Interests

None declared

Acknowledgements

We are grateful for the participation of Patricia McArdle EdD and the Patient-Doctor II teaching site directors: Frederick Basilico MD, Hallowell Churchill MD, Gary Epler MD, Diane Fingold MD, Jerry Greene MD, David Hirsh MD, Linda Nelson DMD, Barbara Ogur MD, Joseph Rothchild MD, Barbara Scolnick MD, Ellen Spar MD, Valerie Stelluto-Pronio MD, Katherine Treadway, and John Whyman MD. We thank Harley Baker EdD for statistical consultation and Cary M. Williams for assistance in the preparation of the manuscript.

References

1. Bickley LS, Hoekelman RA: **Bates' guide to physical examination and history-taking.** *Philadelphia, PA: Lipincott* 1999
2. Novack DH, Volk G, Drossman DA, Lipkin M: **Medical interviewing and interpersonal skills teaching in U.S. medical schools: progress, problems and promise.** *J Am Med Assoc* 1993, **269:**2101-2105
3. Petrusa ER, Blackwell TA, Rogers LP, Rogers LP, Saydjari C, Parcel S, Guckian JC: **An objective measure of clinical performance.** *Am J Med* 1987, **83:**34-42
4. Makoul G, Curry RH, Novack DH: **The future of medical school courses in professional skills and perspectives.** *Acad Med* 1998, **73:**48-51
5. Harden RM, Stevenson M, Downie W, Wilson GM: **Assessment of clinical competence using objective structured examination.** *Br Med J* 1975, **1:**447-451
6. Stillman PL, Wang Y, Ouyang Q, Zhang S, Yang Y, Sawyer WD: **Teaching and assessing clinical skills: a competency-based programme in China.** *Med Educ* 1997, **31:**33-40
7. Carpenter JL, McIntire D, Battles J, Wagner JM: **Administration of a parallel, simultaneous objective structured clinical examination to accommodate a large class of students.** *Teach Learn Med* 1993, **5:**79-85
8. Boehlecke B, Sperber AD, Kowlowitz V, Becker M, Contreras A, McGaghie WC: **Smoking history-taking skills: a simple guide to teach medical students.** *Med Educ* 1996, **30:**283-289
9. Allen SS, Bland CJ, Harris IB, Anderson D, Poland G, Satran L, Miller W: **Structured clinical teaching strategy.** *Med Teacher* 1991, **13:**177-184
10. Lee KC, Dunlop D, Dolan NC: **Do clinical breast examination skills improve during medical school?** *Acad Med* 1998, **73:**1013-1019
11. Pilgrim C, Lannon C, Harris RP, Cogburn W, Fletcher SW: **Improving clinical breast examination training in a medical school: a randomized controlled trial.** *J Gen Int Med* 1993, **8:**685-688
12. Hasle JL, Anderson DS, Szerlip HM: **Analysis of the costs and benefits of using standardized patients to help teach physical diagnosis.** *Acad Med* 1994, **69:**567-570
13. McGaghie WC, Kowlowitz V, Renner BR, Sauter SV, Hoole AJ, Schuch CP, Misch MS: **A randomized trial of physicians and physical therapists as instructors of the musculoskeletal examination.** *J Rheumatol* 1993, **20:**1027-1032
14. Feickert JA, Harris IB, Anderson DC, Bland CJ, Allen S, Poland GA, Satran L, Miller WJ: **Senior medical students as simulated patients in an objective structured clinical examination: motivation and benefits.** *Med Teacher* 1992, **14:**167-177
15. Anderson DC, Harris IB, Allen S, Satran L, Bland CJ, Davis-Feickert JA, Poland GA, Miller WJ: **Comparing students' feedback about clinical instruction with their performances.** *Acad Med* 1991, **66:**29-34
16. Frye AW, Richards BF, Philp EB, Philp JR: **Is it worth it? A look at the costs and benefits of an OSCE for second-year medical students.** *Med Teacher* 1989, **11:**291-293
17. Barclay DM 3rd, McKinley D, Peitzman SJ, Burdick B, Curtis M, Whelan GP: **Effect of training location on students' clinical skills.** *Acad Med* 2001, **76:**384
18. Kowlowitz V, Hoole AJ, Sloane PD: **Implementing the objective structured clinical examination in a traditional medical school.** *Acad Med* 1991, **66:**345-347
19. Heard JK, Allen R, Tank PW, Cason GJ, Cantrell M, Wheeler RP: **Assessing clinical skills of medical students.** *J Ark Med Soc* 1996, **93:**175-179
20. Fields SA, Toffler WL, Elliott D, Chappelle K: **Principles of clinical medicine: Oregon Health Sciences University School of Medicine.** *Acad Med* 1998, **73:**25-31
21. Wilkes MS, Usatine R, Slavin S, Hoffman JR: **Doctoring: University of California, Los Angeles.** *Acad Med* 1998, **73:**32-40
22. Steele DJ, Susman JL: **Integrated clinical experience: University of Nebraska Medical Center.** *Acad Med* 1998, **73:**41-47
23. Hamann C, Hannigan R, Fishman MB, McArdle PJ, Silvestri RC, Fletcher SW: **Which skills do second-year medical students learn best in physical diagnosis?** *J Gen Int Med* 1997, **12(S1):**91
24. van der Vleuten CPM, Swanson DB: **Assessment of clinical skills with standardized patients: state of the art.** *Teach Learn Med* 1990, **2:**58-76
25. Wilkinson TJ, Newble DI, Wilson PD, Carter JM, Helms RM: **Development of a three-centre simultaneous objective structured clinical examination.** *Med Educ* 2000, **34:**798-807
26. Hodder RV, Rivington RN, Calcutt LE, Hart IR: **The effectiveness of immediate feedback during the objective structured clinical examination.** *Med Educ* 1989, **23:**184-188
27. **Statistical Package for the Social Sciences. Version 9.0, Base User's Guide.** *Chicago, IL* 1999
28. Brennan R: **Performance assessment from the perspective of generalizability theory.** *Applied Psychological Measurement* 2000, **24:**339-353
29. Carmines EG, Zeller RA: **Reliability and validity assessment.** *Thousand Oaks, CA: Sage Publications* 1979
30. Suen HK: **Principles of test theories.** *Hillsdale, NJ,: Lawrence Erlbaum* 199061
31. Pedhazur EJ, Schmelkin LP: **Measurement, design and analysis: an integrated approach.** *Hillsdale, NJ: Lawrence Erlbaum Associates* 1991
32. Braun HI: **Understanding score reliability: Experience calibrating essay readers.** *J Educ Statistics* 1988, **13:**1-18
33. Cohen R, Rothman AJ, Ross J, Poldre P: **Validating an objective structured clinical examination (OSCE) as a method for selecting foreign medical graduates for a pre-internship program.** *Acad Med* 1991, **9:**S67-S69

34. Palchik NS, Wolf FM, Cassidy JT, Ike RW, Davis WK: **Comparing information-gathering strategies of medical students and physicians in diagnosing simulated medical cases.** *Acad Med* 1990, **65:**107-113
35. Blue AV, Stratton TD, Plymale M, DeGnore LT, Schwartz RW, Sloan DA: **The effectiveness of the structured clinical instruction module.** *Am J Surg* 1998, **176:**67-70
36. Ross JR, Hutcheon MA, Cohen R: **Second-year students' score improvement during an objective structured clinical examination.** *J Med Educ* 1987, **62:**857-859
37. Ferguson DB, Rutishauser S: **A problem-based preclinical course for dental students.** *Br Dent J* 1997, **182:**387-392
38. Englehard G: **Examining rater errors in the assessment of written composition with a many-faceted Rasch model.** *Journal of Educational Measurement* 1994, **31:**93-112
39. Longford NT: **Models for uncertainty in educational testing.** *New York: Springer-Verlag* 1995
40. Clauser BE: **Recurrent issues and recent advances in scoring performance assessments.** *Applied Psychological Measurement* 2000, **24:**310-324
41. Fan X, Chen M: **Published studies of inter-rater reliability often overestimate reliability: Computing the correct coefficient.** *Educational & Psychological Measurement* 2000, **60:**532-542
42. Simon SR, Volkan K, Hamann C, Duffey C, Fletcher SW: **How well do second-year medical students' OSCE scores predict USMLE scores? (Abstract)** *J Gen Intern Med* 2001, **16(Suppl 1):**108
43. Rutala PJ, Witzke DB, Leko EO, Fulginiti JV, Taylor PJ: **Sharing of information by students in an objective structured clinical examination.** *Arch Int Med* 1991, **151:**541-544
44. Morag E, Lieberman G, Volkan K, Shaffer K, Novelline R, Lang EV: **Clinical competence assessment in radiology: introduction of an objective structured clinical examination (OSCE) in the medical school curriculum.** *Acad Radiol* 2001, **8:**74-81