ORIGINAL ARTICLE

Experience of clinical skills assessment in the Busan-Gyeongnam Consortium

Beesung Kam[1], Young Rim Oh[2], Sang Hwa Lee[3], Hye Rin Roh[4], Jong Ryeal Hahm[5] and Sun Ju Im[1]

[1]Medical Education Unit, Pusan National University School of Medicine, [2]Department of Obstetrics & Gynecology, Kosin University College of Medicine, [3]Department of Microbiology, Dong-A University College of Medicine, [4]Department of Medical Education, Inje University College of Medicine, Busan, and [5]Department of Internal Medicine, Gyeongsang National University School of Medicine, Jinju, Korea

부산·경남 컨소시엄에서 임상수행평가의 시행 경험

[1]부산대학교 의학전문대학원 의학교육실, [2]고신대학교 의과대학 산부인과학교실, [3]동아대학교 의과대학 미생물학교실, [4]인제대학교 의과대학 의학교육학교실, [5]경상대학교 의학전문대학원 내과학교실

감비성[1], 오영림[2], 이상화[3], 노혜린[4], 함종렬[5], 임선주[1]

Purpose: The purpose of this study is to judge the quality of clinical skills assessment in Busan-Gyeongnam Consortium.

Methods: Fourth grade medical school students (n=350 in 2012 and n=419 in 2013) in the Busan-Gyeongnam Consortium were included in the study. The examination was consisted of 6 clinical performance examination (CPX) and 6 objective structured clinical examination (OSCE) stations. The students were divided into groups to take the exam in 4 sites during 3 days. The overall reliability was estimated by Cronbach α coefficient across the stations and the case reliability was by α across checklist items. Analysis of variance and between-group variation were used to evaluate the variation of examinee performance across different days and sites.

Results: The mean total CPX/OSCE score was 67.0 points. The overall α across-stations was 0.66 in 2012 and 0.61 in 2013. The α across-items within a station was 0.54 to 0.86 in CPX, 0.51 to 0.92 in OSCE. There was no significant increase in scores between the different days. The mean scores over sites were different in 30 out of 48 stations but between-group variances were under 30%, except 2 cases.

Conclusion: The overall reliability was below 0.70 and standardization of exam sites was unclear. To improve the quality of exam, case development, item design, training of standardized patients and assessors, and standardization of sites are necessary. Above of all, we need to develop the well-organized matrix to measure the quality of the exam.

***Key Words*:** Outcome and process assessment, Reliability, Psychometrics

Received: November 4, 2013 • Revised: November 18, 2013 • Accepted: November 21, 2013

Corresponding Author: Sun Ju Im
Medical Education Unit, Pusan National University School of Medicine, 20 Geumo-ro, Mulgeum-eup, Yangsan 626-700, Korea
Tel: +82.51.510.8021 Fax: +82.51.510.8125 email: sunjuim11@hanmail.net

Korean J Med Educ 2013 Dec; 25(4): 327-336.
http://dx.doi.org/10.3946/kjme.2013.25.4.327
pISSN: 2005-727X eISSN: 2005-7288

서론

2009년도부터 의사실기시험이 시행됨에 따라 교육 과정과 학생 평가 방법에 많은 변화가 있었다. 각 대학에서는 지식뿐만 아니라 임상술기와 태도 교육을 강화하고, 임상실습에서 실제 환자를 볼 수 있도록 기회를 제공하며, 3, 4학년 과정에서 표준화 환자 진료시험을 도입하는 등 긍정적인 변화를 보이고 있다. 또한 의과대학은 임상술기센터를 확보하여 실습 환경을 제공할 뿐만 아니라 사례 개발과 표준화 환자 훈련을 위해 지역별 컨소시엄에 가입하여 공동으로 대처하고 있다[1].

초기에는 4학년 중심으로 임상술기교육이 이루어지고 의사국가고시 준비에만 치중하는 등의 문제점이 있었으나[1], 점차 각 학년 단계별로 점진적이고 균형있게 실기 교육을 하는 대학들이 늘고 있다. 실기시험은 기본적으로 상대 평가가 아닌 절대적인 준거(criterion-based assessment)를 사용하는데, 실기시험을 통해 학생의 역량 수준과 진급 여부를 판단하는 종합시험의 특성을 갖기도 한다[2,3]. 이와 같이 성과중심교육(outcome-based education)의 도입으로 실기시험은 고비용 부담 시험임에도 불구하고 점차 확대되고 있다.

종합평가로서 실기시험은 학생들의 수행능력을 객관적으로 평가할 수 있어야 한다. 4학년 시험은 국가고시 실기시험을 앞두고 학생들의 수행 정도를 예측할 수 있어야 하며, 3학년 종합시험으로서 실기시험은 필수 임상실습을 마친 학생들의 진료 역량을 판단하는 기준을 설정함(standards setting)과 동시에 임상실습 교육과정의 성과(outcome)를 판단할 수 있어야 한다[4]. 따라서 실기시험의 신뢰도 등 질적 근거를 충분히 확보하는 것이 급선무이다.

실기시험은 평가 내용과 평가 척도(scale) 등 문항을 개발하는 과정에서 복잡하고 일자별 문항조합에 따라 결과가 상이할 수 있다. 또, 실기시험 특성상 학생들을 그룹으로 나누어 여러 번 시험을 시행해야 하므로 장비 또는 기구 배치와 같은 시험실 환경, 표준화 환자 및 채점관에 따라 오차가 발생할 수 있다. 따라서 문항 수준과 설계, 환경 및 인력의 표준화에 많은 비용과 시간이 소요된다. 시험의 신뢰도를 확보하기 위해서는 많은 변수 중에서 문제점이 있는 부분을 파악하여 지속적으로 개선하는 노력이 필요하다[3].

부산·경남 실기시험 컨소시엄은 부산·경남 지역의 5개 의과대학·의학전문대학원으로 구성되어 있으며, 실기 문항과 표준화 환자를 공유하기 위해 2009년도에 결성되었다. 초기 2년 동안에는 각 학교 시험 일정에 맞추어 실기시험을 진행하다가 2012년과 2013년에는 4학년을 대상으로 공동 실기시험을 시행하였으며, 2014년부터는 3학년 학생을 대상으로 종합시험(summative assessment)을 시행하기로 협의하였다. 따라서 2년간의 공동시험에 대한 결과 분석과 질적 개선을 필요로 한다.

본 연구에서는 공동실기시험을 올바른 방향으로 개선하기 위해 현재의 질적 수준에 대해 판단하고자 하였다. 문항 수준과 설계, 표준화 환자 훈련 정도, 채점관, 시험장 및 일자별 변수를 고려하여 구체적인 연구 문제를 다음과 같이 설정하였다.

첫째, 공동실기시험의 전체 신뢰도(overall reliability)는 어떠한가?

둘째, 문항별 신뢰도는 어떠한가?

셋째, 시험 일자별로 다른 문항조합에서의 결과는 차이가 있는가?

넷째, 시험장 4곳에 따라 시험 결과가 차이가 있는가?

대상 및 방법

1. 연구 대상 및 방법

2012년, 2013년 부산·경남 실기 컨소시엄 4학년 학생들을 대상으로 공동 실기시험을 시행하였다. 응시자는 2012년도에 4개 대학 350명, 2013년도에 5개 대학 419명이었다. 문항은 장기별, 영역별로 고르게 평가할 수 있도록 설계를 시행한 후 매년 clinical performance examination (CPX) 12 문항, objective structured clinical examination (OSCE) 12 문항을 개발하였다. 1차 개발된 문항은 문항심의 작업을 통하여 수정·보완하였다. 컨소시엄에 소속된 표준화 환자 32명을 대상으로 환자-의사관계 채점을 포함한 기본 훈련과 시나리오 훈련을 12시간 시행하였다. 2012년에는 CPX와 OSCE 모두 교수 평가자가 채점하였으나, 2013년에는 CPX의 절반을 표

준화 환자가 채점하였으며, 이를 위하여 질적 근거가 확보된 표준화 환자를 선발하여 추가로 채점훈련을 시행하였다. 표준화 환자 간 환자-의사관계 채점 신뢰도는 연도별로 0.85과 0.88이었다.

시험은 6개의 CPX와 6개의 OSCE 문항으로 구성하여 3일 동안 시행하였는데, 둘째 날 시험은 첫날 시험의 CPX 3문항, OSCE 3문항을 새 문항으로 바꾸었고, 3일째 시험에서도 동일하게 변경하였다.

시험장은 4곳이었으며 각 학교의 임상술기센터를 이용하였다. 학생들은 가급적 다른 학교의 시험장에서 시험을 칠 수 있도록 배치하였고 채점자 교수도 시험장 4곳에 분산하였다. 시험장의 표준화를 위하여 사전에 시험실 배치, 장비 종류 및 세팅을 동일하게 맞추고, 시험 전날 교수 2인이 각 시험장을 순환하며 보완하였다. 표준화 환자는 문항당 4인으로 구성하고 훈련의 일치도를 확인하였다. 웹기반 채점시스템을 통하여 4곳의 시험장에서 동시에 접속하여 시험을 진행하였다. 동일한 시험안내방송 시스템을 통하여 같은 일정으로 시험을 시행하였다. 각 시험장에는 시험운영위원장이 교수 채점 안내 등 전반적인 사항을 운영하였다.

2. 연구 분석

연도별 평균점수는 독립표본 t검정을 시행하여 비교하였다.

공동 실기시험의 신뢰도를 알아보기 위하여 내적 일치도를 나타내는 Cronbach α값을 측정하였다. 그 중 첫 번째, 시험 일자별 전체 신뢰도는 12개 스테이션의 문항 점수에 따라 α값을 구하였고, 두 번째, 12개 문항의 체크리스트 응답에 따라 α값을 구하였으며, 세 번째, 문항 내에서 체크리스트 항목의 내적 일치도를 알아보기 위하여 문항별로 α값을 구하였으며 시험 일자별로 제시하였다. 첫 번째는 'across-stations', 두 번째는 'across-items over stations', 세 번째는 'across-items within a station'에 의한 α값을 의미한다.

시험 일자별, 시험장별 요소에 의한 차이를 알아보기 위해 일원배치분산분석(analysis of variance, ANOVA)과 그룹 간 변량(between-group variation)을 계산하였다. 데이터 분석은 Windows용 SPSS version 21.0 프로그램(SPSS Inc., Chicago, USA)을 이용하였다.

결과

1. 평균점수

연도별로 시험의 평균은 66.92와 67.12점이었으며, CPX는 62.91과 64.86점, OSCE는 37.47과 36.27점이었다. CPX/OSCE 총점은 연도별로 유의수준 0.05에서 차이가 없었으나 ($p=0.21$), CPX 영역별 점수와 OSCE 점수는 차이를 보였는데, 신체진찰과 환자교육 영역 점수는 증가하고 환자-의사관계 영역 점수는 오히려 감소하였다($p<0.01$) (Table 1).

Table 1. Clinical Performance Examination/Objective Structured Clinical Examination Scores

	2012 (n=350)		2013 (n=419)		p-value
	Mean	SD	Mean	SD	
CPX	62.91	6.27	64.86	5.42	<0.001
History taking	69.08	8.58	70.58	6.66	0.01
Physical examination	50.76	14.33	54.33	10.85	<0.001
Patient education	51.96	15.30	57.14	17.31	<0.001
Patient-physician interaction	64.14	7.51	59.80	7.48	<0.001
OSCE	37.47	4.54	36.27	4.19	<0.001
Total	66.92	5.83	67.42	5.24	0.21

OSCE items score was based on 50, the other score was based on 100.
SD: Standard deviation, CPX: Clinical performance examination, OSCE: Objective structured clinical examination.

2. 신뢰도 분석

12개의 스테이션 점수로 측정한 across-stations α값은 2012년도에 평균 0.66, 2013년도에 0.61이었다. CPX 6문항의 신뢰도는 연도별로 0.57과 0.51이었으며, OSCE 6문항의 신뢰도는 0.47과 0.40이었다(Table 2).

Table 2. Overall Reliability Across-Stations

	2012				2013			
	Day 1 (n=140)	Day 2 (n=140)	Day 3 (n=70)	Mean (n=350)	Day 1 (n=140)	Day 2 (n=140)	Day 3 (n=139)	Mean (n=419)
CPX (n=6)	0.45	0.63	0.62	0.57	0.47	0.49	0.58	0.51
OSCE (n=6)	0.49	0.44	0.47	0.47	0.40	0.42	0.39	0.40
Overall (n=12)	0.61	0.67	0.71	0.66	0.58	0.58	0.67	0.61

CPX: Clinical performance examination, OSCE: Objective structured clinical examination.

Table 3. Station Reliability according to Days

Stations	2012			Stations	2013		
	Day 1 (n=140)	Day 2 (n=140)	Day 3 (n=70)		Day 1 (n=140)	Day 2 (n=140)	Day 3 (n=139)
CPX				CPX			
Seizure	0.72			Insomnia	0.58		
Dyspnea	0.67			Dysuria	0.75		
Acute abdominal pain	0.64			Developmental delay	0.70		
Amenorrhea	0.66	0.56		Joint pain	0.71	0.46	
Polyuria	0.67	0.72		Chest pain	0.75	0.65	
Back pain	0.54	0.56		Hematemesis	0.77	0.66	0.68
Mood change		0.64	0.76	Bleeding tendency		0.70	0.70
Palpitation		0.75	0.67	Antepartum care		0.70	0.70
Constipation		0.68	0.69	Alcohol problems		0.58	
Vaccination			0.86	Weakness of limbs			0.85
Breat lump/pain			0.86	Hypertension			0.81
Delivery bad news			0.79	Dizziness			0.68
OSCE				OSCE			
Foreign body airway obstruction	0.92			Motor/sensory/reflex exam of limbs	0.60		
Local anesthesia	0.65			Report of patient	0.69		
K-MMSE	0.51			Lower extremities prehospital splinting	0.63		
Lumbar puncture	0.75	0.75		Venipuncture for blood culture	0.60	0.55	
Wet smear of the vagina and cervix	0.61	0.57		Otoscopic exam	0.61	0.49	
Chest/lung exam	0.67	0.66		Chest X-ray presentation	0.62	0.56	0.64
Basic life support		0.58	0.61	Intubation		0.72	
Exam of delivery		0.51	0.41	Burn dressing		0.61	0.40
Cranical nerve exam		0.72	0.68	Informed consent		0.54	0.59
Blood transfusion			0.76	Lumbar puncture			0.66
Electrocardiography			0.70	Neck exam			0.59
Anorectal exam			0.72	Defibrillation			0.74

CPX: Clinical performance examination, OSCE, Objective structured clinical examination, K-MMSE: Korean mini-mental state examination.

문항의 α값(across-items within a station)은, CPX 0.54~0.86, OSCE 0.51~0.92의 범위에 있었으며, 시험에 다시 출제된 경우 22문항 중 11문항의 α값이 감소하였고, 6문항의 α값은 증가하였으며, 3문항에서 변화가 없었다(Table 3).

3. 시험 일자별 분석

출제되었던 문항이 다음날 반복되었을 때 유의수준 0.05에서 CPX 11문항 중 8문항(무월경 p<0.001, 허리통증 p=0.01, 기분변화 p<0.001, 두근거림 p<0.001, 변비 p=0.03, 흉통 p= 0.06, 토혈, 쉽게멍이듦, 산전진찰 p<0.001), OSCE 11문항 중 5문항(척추천자 p=0.02, 폐진찰 p=0.01, 기본심폐소생술 p<0.001, 뇌신경기능평가 p=0.00, 환자보고하기 p=0.02)에서 점수가 상승한 것으로 분석되었다. 결과를 파악해야 했던 OSCE 문항 중 분만진행단계진찰은 일자별 변동이 없었으나(p=0.31), 흉부X선프리젠테이션(p=0.00)과 이경검사(p=0.04)의 점수는 일자별 차이가 있었지만 일관되게 상승하지는 않았다. 일자별 그룹간 변량은 모두 30% 미만이었다(Table 4).

2012년도 총점 평균점수는 66.92로 일자별 차이가 없었으며(p=0.42), 2013년도 총점 평균은 67.42로 유의미한 일자별

Table 4. Effect of Different Days

	Mean				p-value	Between-group variation (%)
	Day 1	Day 2	Day 3	Total		
2012						
Amenorrhea	60.67	65.43		63.05	<0.001	5.97
Polyuria	57.81	58.11		57.96	0.82	0.02
Back pain	63.27	66.21		64.74	0.01	2.47
Mood change		61.77	67.73	63.75	<0.001	7.48
Palpitation		56.34	64.79	59.16	<0.001	13.25
Constipation		70.06	73.46	71.20	0.03	2.15
Lumbar puncture	33.49	36.20		34.84	0.02	1.89
Wet smear of the vagina and cervix	43.29	41.40		42.35	0.02	1.95
Chest/lung exam	42.58	44.50		43.54	0.01	2.46
Basic life support		41.37	44.92	42.55	<0.001	7.64
Exam of delivery		29.70	30.77	30.05	0.31	0.50
Cranical nerve exam		32.58	37.24	34.13	0.00	4.91
CPX	61.13	62.99	66.31	62.91	<0.001	9.12
OSCE	38.55	37.62	34.98	37.47	<0.001	8.35
Total	66.46	67.07	67.53	66.92	0.42	0.50
2013						
Joint pain	67.17	66.98		67.07	0.85	0.01
Chest pain	63.63	66.03		64.83	0.06	1.31
Hematemesis	68.41	71.65	74.24	71.43	<0.001	7.07
Bleeding tendency		68.83	74.21	71.51	<0.001	6.70
Antepartum care		51.38	56.28	53.82	<0.001	5.71
Venipuncture for blood culture	38.46	38.75		38.60	0.72	0.05
Otoscopic exam	32.01	29.92		30.96	0.04	1.50
Chest X-ray presentation	31.45	34.29	31.70	32.48	0.00	2.65
Burn dressing		41.54	41.82	41.68	0.69	0.06
Informed consent		38.82	40.64	39.73	0.02	1.95
CPX	65.21	65.20	64.18	64.86	0.19	0.79
OSCE	33.90	36.49	38.44	36.27	<0.001	19.74
Total	66.07	67.79	68.41	67.42	0.00	3.58

OSCE items score was based on 50, the other score was based on 100.
CPX: Clinical performance examination, OSCE, Objective structured clinical examination.

차이가 있었으나(p=0.00), 그 차이는 크지 않았다.

4. 시험장별 분석

시험장별 평균점수는 유의수준 0.05에서 CPX 24문항 중 16문항(경련, 호흡곤란, 다뇨증, 허리통증, 기분변화, 변비, 예방접종, 유방종괴/유방통, 배뇨곤란, 관절통증, 흉통, 토혈, 음주문제상담, 팔다리근육힘약화, 고혈압, 어지러움), OSCE 24문항 중 14문항(국소마취, 척추천자, 질분비물검사, 폐진찰, 기본심폐소생술, 분만진행단계진찰, 심전도검사, 환자보고하기, 하지부목고정, 혈액배양을위한채혈, 기관내삽관, 화상드레싱, 동의서받기, 목진찰)에서 차이를 보였다(p<0.05). 시험장별 그룹 간 변량이 30% 이상이었던 항목은 유방통과 고혈압이었고, 다른 문항에서는 30% 미만이었다(Table 5).

고찰

이 연구의 목적은 실기시험의 질적 개선을 위해 실기시험의 신뢰도와 문항의 일자별, 시험장별 차이를 알아보고자 했다. 그 결과는 첫째, α값으로 측정한 실기시험의 신뢰도(overall reliability, α across-stations)는 2012년도에 0.66, 2013년도에 0.61이었으며, 둘째, 문항 신뢰도(across-items within a station)는 CPX 0.6~0.8, OSCE 0.5~0.7의 범위가 많았으며, 문항 정보가 노출된 경우 약간 감소하였다. 셋째, 문항 정보가 알려진 경우 시험 일자에 따라 문항의 평균점수가 상승하였으나, 새로운 문항이 출제됨에 따라 일자별 평균은 동일하게 유지되었으며, 일자별 그룹간 변량은 30% 이하

Table 5. Effect of Different Examination Sites

	Mean					p-value	Between-group variation (%)
	A	B	C	D	Total		
2012							
Seizure	60.25	55.55	55.42	48.32	54.80	<0.001	21.68
Dyspnea	60.07	60.25	67.64	58.54	61.61	<0.001	16.03
Acute abdominal pain	69.31	67.23	68.43	69.59	68.64	0.68	1.09
Amenorrhea	62.55	65.58	62.27	61.81	63.05	0.09	2.32
Polyuria	53.49	52.01	60.10	66.05	57.96	<0.001	24.74
Back pain	68.29	60.75	65.73	64.25	64.74	<0.001	8.45
Mood change	70.45	60.36	57.98	62.93	63.75	<0.001	23.15
Palpitation	61.33	58.42	58.21	57.24	59.16	0.22	2.11
Constipation	76.30	70.02	63.45	71.09	71.20	<0.001	16.02
Vaccination	72.69	54.08			63.38	<0.001	22.64
Breat lump/pain	58.46	77.48			67.97	<0.001	51.59
Delivery bad news	58.88	62.19			60.54	0.21	2.27
Foreign body airway obstruction	41.47	38.67	32.86	35.65	37.12	0.09	4.57
Local anesthesia	35.11	40.89	40.63	38.02	38.68	0.01	8.87
K-MMSE	35.76	38.12	35.26	35.53	36.17	0.14	3.92
Lumbar puncture	40.92	32.02	30.55	35.96	34.84	<0.001	16.48
Wet smear of the vagina and cervix	43.22	39.69	43.47	43.01	42.35	0.00	5.19
Chest/lung exam	44.71	42.65	45.25	41.59	43.54	0.00	6.01
Basic life support	40.00	43.49	42.86	45.46	42.55	<0.001	10.61
Exam of delivery	31.59	28.30	31.43	29.12	30.05	0.02	4.47
Cranical nerve exam	32.91	35.61	32.76	35.00	34.13	0.32	1.69
Blood transfusion	21.22	24.35			22.79	0.15	3.01
Electrocardiography	41.59	35.00			38.29	<0.001	20.34
Anorectal exam	34.50	37.29			35.89	0.11	3.73
CPX	64.72	61.73	62.28	62.64	62.91	0.00	3.78
OSCE	37.55	36.55	37.86	38.32	37.47	0.06	2.12
Total	68.18	65.51	66.76	67.31	66.92	0.01	3.24

(Continued to the next page)

Table 5. Continued

	Mean					p-value	Between-group variation (%)
	A	B	C	D	Total		
2013							
Insomnia	66.64	63.89	68.17	70.01	67.18	0.07	5.10
Dysuria	63.63	55.96	56.61	53.78	57.49	<0.001	13.10
Developmental delay	66.02	67.44	66.75	69.20	67.35	0.55	1.50
Joint pain	65.30	63.98	72.98	66.03	67.07	<0.001	17.90
Chest pain	62.87	68.25	69.44	58.77	64.83	<0.001	16.70
Hematemesis	72.86	67.20	74.07	71.59	71.43	<0.001	8.40
Bleeding tendency	73.69	69.83	71.92	70.59	71.51	0.13	2.00
Antepartum care	52.33	53.09	55.32	54.56	53.82	0.30	1.30
Alcohol problems	70.49	67.48	61.07	66.18	66.30	<0.001	14.90
Weakness of limbs	68.61	62.35	52.82	55.59	59.87	<0.001	26.10
Hypertension	79.23	58.17	65.47	58.15	65.31	<0.001	39.90
Dizziness	50.44	54.46	56.92	58.90	55.15	0.00	9.70
Motor/sensory/reflex exam of limbs	40.77	39.11	37.56	38.45	38.97	0.31	2.59
Report of patient	28.38	32.00	25.81	25.43	27.90	0.01	8.53
Lower extremities prehospital splinting	38.81	32.54	36.51	30.48	34.58	<0.001	23.55
Venipuncture for blood culture	35.98	38.30	38.35	41.79	38.60	<0.001	9.53
Otoscopic exam	31.43	29.51	31.21	31.70	30.96	0.42	1.02
Chest X-ray presentation	33.03	31.23	33.87	31.79	32.48	0.07	1.71
Intubation	32.86	31.52	41.34	36.70	35.60	<0.001	17.06
Burn dressing	44.55	39.91	40.89	41.35	41.68	<0.001	8.91
Informed consent	43.70	34.75	40.84	39.64	39.73	<0.001	24.72
Lumbar puncture	41.18	38.66	39.92	36.76	39.15	0.14	4.01
Neck exam	46.95	40.38	42.29	41.76	42.85	<0.001	14.20
Defibrillation	34.82	34.82	36.61	32.72	34.76	0.29	2.72
CPX	66.22	63.42	66.07	63.71	64.86	<0.001	5.72
OSCE	37.45	34.87	36.83	35.91	36.27	<0.001	5.45
Total	69.12	65.53	68.60	66.42	67.42	<0.001	8.12

OSCE items score was based on 50, the other score was based on 100.
K-MMSE: Korean mini-mental state examination, CPX: Clinical performance examination, OSCE: Objective structured clinical examination.

였다. 넷째, 시험장에 따라 문항의 평균점수는 차이를 보였으며, 시험장에 따른 그룹 간 변량은 2문항을 제외하고 30% 이하였다.

첫째, CPX 6문항, OSCE 6문항으로 구성된 실기시험의 신뢰도는 일반적으로 표준화된 시험을 의미하는 0.7보다 낮았으며, Brannick et al. [5]이 다수의 문헌 연구를 통하여 보고한 실기시험의 평균 신뢰도인 0.66과 비슷하거나 낮았다. 서울·경기컨소시엄에서 CPX 6문항의 신뢰도를 0.66으로 보고하였는데[6], 이것은 일반화가능도 이론을 이용한 신뢰도로서 본 연구에서의 α값과 직접 비교하기는 어렵다. 그러나 일반화가능도 계수가 α값보다 작게 나오는 경향이 있고[5], 문항이 6문항임을 감안할 때 본 시험의 신뢰도가 낮다고 판단된다.

α값이 낮은 원인에 대하여 먼저 스테이션 수와 시험 형태를 들 수 있다. 스테이션 수가 많을수록 신뢰도는 증가하는데, 스테이션 수가 10개 이하였을 때의 신뢰도는 0.56이었고, 10개보다 많았을 때 평균 신뢰도는 0.74였다고 보고하였다[5]. 본 시험 설계에서 총 12개의 스테이션이지만 CPX와 OSCE가 혼합되어 있어 명확하게 판단하기가 어렵다. 내적 일치도를 의미하는 α값은, 12개의 CPX 문항만으로 구성한 시험과 시험의 성격이 다른 CPX와 OSCE 문항을 혼합했을 때는 차이가 있을 것이다.

다음으로 문항 특이성을 들 수 있다. Table 6에서 2013년 첫날 시행된 흉통 문항의 경우 내적 일치도 α값은 0.75로 높

Table 6. Different Cronbach α

	Across-items within a station[a,b]		Across-items over stations[b]		Across-stations[b]	
	α	No. of items	α	No. of items	α	No. of stations
CPX	0.75	26	0.83	150	0.47	6
Clinical scale	0.68	20	0.71	116	0.41	6
History taking	0.46	14	0.57	80	0.28	6
Physical examination	0.75	5	0.68	23	0.36	6
Patient education	-	1	0.16	13	0.11	6
PPI scale	0.82	6	0.85	34	0.38	6
OSCE			0.75	88	0.40	6
Overall			0.86	238	0.58	12

Clinical scale was checklist scale, PPI scale was Likert scale.
CPX: Clinical performance examination, PPI: Patient-physician interaction, OSCE: Objective structured clinical examination.
[a]Chest pain case, [b]Day 1 on 2013.

았다(across-items within a station). 이 중 병력청취, 신체진찰 및 환자교육을 포함하는 임상 척도의 α값은 0.68, 환자의사관계 척도의 α값은 0.82로 각 항목 간의 높은 내적 일치도를 보여주었다. 병력청취 항목 간의 일치도는 0.46, 신체진찰 항목 간의 일치도는 0.75로 높았다. 당일 시험일에 시행된 CPX 문항의 신뢰도는 수면이상을 제외하고 0.7 이상이었으며 OSCE 신뢰도 또한 0.6 이상이었다.

그러나 2013년 첫날의 전체 신뢰도(across-stations)는 0.58, CPX 6문항의 신뢰도는 0.47, OSCE 6문항의 신뢰도는 0.40로 낮았다(Table 6). CPX의 각 영역별 신뢰도는 매우 낮았다. 문항 내의 일치도는 높지만 문항 간에 일치도가 낮게 나온 원인은 스테이션 수가 적고 문항이 서로 다른 내용을 측정하고 있다고 판단할 수 있다. 즉, 앞서 언급한 CPX와 OSCE 시험 형태에 의한 차이와 함께 CPX 항목 간의 문항 특이성이 α값을 낮게 할 수 있다.

한편, 동일한 시험 일자의 신뢰도를 문항의 체크리스트를 모두 합한 것을 변수로 하여 측정하였을 때(across-items over stations), 전체 신뢰도는 0.86, CPX 6문항의 신뢰도는 0.83, OSCE 6문항의 신뢰도는 0.75로 높았다. CPX 영역별로 병력청취 0.57, 신체진찰 0.68의 내적 일치도를 보였고, 환자교육은 0.16이었는데 문항 특이적이고 체크리스트 수가 적어 낮게 나온 것이라 판단된다. 이렇게 측정한 α값은 높지만 비슷한 항목이 중복되어 있는 경우가 많으므로 조심스런 판단이 요구된다.

다음으로, 시험장의 환경, 표준화 환자 차이 및 채점자에 의한 원인이 있을 수 있다. 2012년도 3일째 시험의 α값이 가장 높았는데, 다른 시험일에는 4개의 시험장을 운영하였으나 이 때에는 2개의 시험장만을 운영하였고, 여러 그룹으로 나뉘어서 시험을 시행하는 데서 기인하는 오차가 적었을 것이다.

Pell et al. [3]은 α값이 낮은 원인으로 문항이 서로 다른 내용을 측정하고 있거나, 문항 설계를 잘못했거나, 학생들이 시험과는 다른 방법으로 배웠거나 또는 평가자가 제시된 기준에 따라 평가하고 있지 않는가를 살펴보아야 한다고 하였다. 특정 문항을 제거하였을 때 전체 신뢰도의 변화 양상을 살펴보아, 문항 제거 시 신뢰도가 올라가면 그 문항을 특히 개선할 필요가 있다고 하였다.

둘째, 날짜별로 시험에 출제된 문항이 다시 나왔을 때 문항 신뢰도는 하락하는 양상을 보였고, 특히 관절통증, 화상드레싱 및 이경검사는 2일째 시험에서 신뢰도가 대폭 감소하였다. 그러나 다뇨증, 기분저하와 같은 일부 문항은 2일째에 오히려 신뢰도가 올라갔고, 토혈과 흉부X선프리젠테이션 문항은 2일째에 신뢰도가 하락하였다가 3일째 다시 올라가는 양상을 보였다. 이경검사와 흉부X선프리젠테이션의 경우 일자별로 다른 자료의 결과를 해석하도록 하였는데, 이경검사는 첫 날, 흉부X선프리젠테이션은 3일째 자료가 신뢰도가 더 높았다. Schoonheim-Klein et al. [7]은 일주일간 지속되는 치과의사 실기시험에서 평균점수와 신뢰도의 일자별 변화가 없었다고 보고하였는데 흉부X선 사진, 치아 모형을 약간씩 변경하

면서 시험을 운영하였다고 하였다. 실기시험에서는 각 스테이션의 문항의 특이성에 따라 신뢰도의 변화가 있을 것이라 예상되므로 반복 출제 시 문항의 최소 신뢰도를 확보할 필요가 있겠다.

셋째, 시험에 사용했던 문항을 다시 출제 했을 때 22문항 중 13문항에서 평균점수가 상승하였지만, 그룹간 변량은 모두 30% 미만으로 일자별 효과는 미미한 것으로 판단할 수 있다. 또, 반복적으로 출제된 문항 이외에 새로운 문항이 추가됨에 따라 전체 평균점수는 차이가 없거나 근소한 차이를 보였다. 동일한 문항을 사용하는 경우 시험 일자별 점수 변화에 대하여 서울·경기컨소시엄의 6개 대학을 대상으로 시행한 시험에서는 문항이 반복 출제 되더라도 점수 변화가 없었다고 하였으나[8], 일개 대학의 연구에서는 본 연구와 동일하게 문항 정보가 노출되면 점수가 상승하였다고 보고하였다[9]. 학교 또는 컨소시엄에서 시험을 시행할 때, 일자별로 난이도가 유지될 수 있도록 문항 난이도를 고려하여 문항을 조합해야 하며, 일자별로 평균성적의 변화가 큰 경우 표준점수로 환산하여 비교할 수 있다[9].

넷째, 단기간에 시험을 시행하기 위해 각 학교의 시험장 4곳을 이용하였고 시험장의 표준화를 판단하는 것이 중요하였는데, 일원분산분석과 시험장별 그룹 간 변량을 사용하였다. 시험장 4곳의 평균점수는 절반 이상에서 차이가 있었고, 표준화 환자가 채점한 환자-의사관계, 교수 채점자가 평가한 CPX 영역과 OSCE에서도 차이를 보였다. 그룹간 변량은 30% 이하여야 하며, 40%가 넘으면 학생들의 수행의 차이가 아니라 평가자의 채점이 일관되지 않거나 특정 그룹의 문제로 판단될 수 있는데, 2문항을 제외하고 30% 이하였다[3,10]. 유방통과 고혈압 문항은 각 그룹 간 변량이 각각 51.59와 39.90으로 평가자에 기인한 차이로 판단되었고 채점관 훈련을 강화하여 문제점을 개선할 필요가 있다.

부산·경남 공동 실기시험의 특징은 시험장을 복제하여 짧은 기간 동안 시험을 진행한 것이다. 일자별 난이도는 새로운 문항을 추가함으로써 비교적 일정하게 유지하였으나 신뢰도를 확보하기 위해 다음과 같은 점을 개선해야 할 것으로 생각된다.

우선, 시험 문항의 질적 향상이 필요하다. 전체 신뢰도에 부정적인 영향을 미치는 문항을 중심으로 다른 시험 문항과 측정하는 내용이 다르지 않은지, 시험 설계가 잘못되지 않았는지 확인하여야 한다[3,11]. 평가 척도 또한 신뢰도에 영향을 줄 수 있으며 체크리스트(checklists)와 전반적 평가(global rating)의 장단점을 파악하여 신뢰도를 향상시킬 수 있는 척도를 개발하여야 한다[4,5,12,13]. 아울러 시험을 통하여 무엇을 평가해야 하는지 평가 목표를 분명히 정하는 것은 시험의 타당도를 높일 수 있을 뿐만 아니라, 문항 특이성을 줄이고 시험을 일관된 방향으로 끌고 갈 수 있기 때문에 신뢰도를 향상시킬 것으로 기대한다[4].

다음으로 시험 문항 조합의 개선이 필요하다. 성격이 다른 CPX와 OSCE 문항을 조합하는 것은 신뢰도를 떨어뜨리는 것으로 판단된다. 복부진찰과 같은 신체진찰 OSCE는 CPX와 결합하면 시험 형태의 차이를 줄일 수 있고 CPX 스테이션 수를 늘리는 효과가 있어 전체 α값이 증가할 것으로 예상되며, CPX 영역 간 신뢰도 또한 향상될 것이다.

또, 시험장의 표준화를 위해 노력해야 한다. 다양한 시험장은 짧은 기간 내에 시험을 시행할 수 있는 장점이 있지만 시험의 질적 확보가 힘들다. 시험장 환경 뿐 아니라, 표준화 환자 훈련을 강화해야 하며, 짧은 시간에 효과적으로 교수 채점자 훈련을 할 수 있는 방법을 모색해야 한다.

마지막으로 실기시험에 대한 질적 향상을 위해 다양한 근거들을 제시할 수 있어야 한다. α값 이외에 신뢰도를 나타내는 일반화가능도 계수(generalisability coefficients), 체크리스트의 타당도를 점검할 수 있는 체크리스트와 전반적 평가(global rating)의 상관성(R2) 또는 등급 간 구별(inter-grade discrimination) 등 질적 근거를 확보하는 방안들을 모색해야 한다[14,15,16].

본 연구의 의의는 2년간의 부산·경남 의사실기 컨소시엄의 시험 운영 결과를 점검함으로써 개선책을 마련하고자 하였다. 현재의 시험 운영은 신뢰도와 표준화 관점에서 미흡한 점이 관찰되며, 문항 수준, 시험 형태, 평가 척도(scale), 평가 목표 설정 등 전반적인 부분의 지속적인 개선을 필요로 한다.

Acknowledgements: None.

Funding: This work was supported by Pusan National University Research Grant, 2010.

Conflicts of interest: None.

REFERENCES

1. Park HK. The impact of introducing the Korean Medical Licensing Examination clinical skills assessment on medical education. J Korean Med Assoc 2012; 55: 116-123.
2. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach 2010; 32: 676-682.
3. Pell G, Fuller R, Homer M, Roberts T; International Association for Medical Education. How to measure the quality of the OSCE: a review of metrics - AMEE guide no. 49. Med Teach 2010; 32: 802-811.
4. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ 2004; 38: 199-203.
5. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ 2011; 45: 1181-1189.
6. Yim MK, Lee GM. The school effect on the reliability of clinical performance examination in medical schools. Korean J Med Educ 2010; 22: 215-223.
7. Schoonheim-Klein M, Muijtjens A, Habets L, Manogue M, Van der Vleuten C, Hoogstraten J, Van der Velden U. On the reliability of a dental OSCE, using SEM: effect of different days. Eur J Dent Educ 2008; 12: 131-137.
8. Han JJ, Park H, Kwon I, Ryu KH, Eo E, Kim N, Jung J, Kim KH, Lee SN. The comparison of clinical performance examination scores according to the different testing time: six medical schools in Seoul·Gyeonggi CPX Consortium 2005. Korean J Med Educ 2007; 19: 31-38.
9. Park HK, Kwon OJ. Sharing of information among students and its effect on the scores of clinical performance examination (CPX). Korean J Med Educ 2005; 17: 185-196.
10. Fuller R, Homer M, Pell G. Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability. Med Teach 2013; 35: 515-517.
11. Auewarakul C, Downing SM, Praditsuwan R, Jaturatamrong U. Item analysis to improve reliability for an internal medicine undergraduate OSCE. Adv Health Sci Educ Theory Pract 2005; 10: 105-113.
12. Wilkinson TJ, Newble DI, Frampton CM. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. Med Educ 2001; 35: 1043-1049.
13. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ 2003; 37: 1012-1016.
14. Hodges B, Turnbull J, Cohen R, Bienenstock A, Norman G. Evaluating communication skills in the OSCE format: reliability and generalizability. Med Educ 1996; 30: 38-43.
15. Reznick RK, Blackmore D, Dauphinée WD, Rothman AI, Smee S. Large-scale high-stakes testing with an OSCE: report from the Medical Council of Canada. Acad Med 1996; 71(1 Suppl): S19-S21.
16. Dauphinee WD, Blackmore DE, Smee S, Rothman AI, Reznick R. Using the judgments of physician examiners in setting the standards for a national multi-center high stakes OSCE. Adv Health Sci Educ Theory Pract 1997; 2: 201-211.