Research Article

KME Review

Korean Medical Education Review 2021; 23(3): 185-193
https://doi.org/10.17496/kmer.2021.23.3.185

pISSN 2092-5603 | eISSN 2093-6370

Messick의 타당도 틀을 활용한 임상실습 전 실기시험의 타당도 평가

이혜윤, 윤소정, 이상엽, 임선주

부산대학교 의과대학 의학교육학교실

Assessing the Validity of the Preclinical Objective Structured Clinical Examination Using Messick's Validity Framework

Hye-Yoon Lee, So-Jung Yune, Sang-Yeoup Lee, Sunju Im

Department of Medical Education, Pusan National University School of Medicine, Busan, Korea

Students must be familiar with clinical skills before starting clinical practice to ensure patients' safety and enable efficient learning. However, performance is mainly tested in the third or fourth years of medical school, and studies using the validity framework have not been reported in Korea. We analyzed the validity of a performance test conducted among second-year students classified into content, response process, internal structure, relationships with other variables, and consequences according to Messick's framework. As results of the analysis, content validity was secured by developing cases according to a pre-determined blueprint. The quality of the response process was controlled by training and calibrating raters. The internal structure showed that (1) reliability by generalizability theory was acceptable (coefficients of 0.724 and 0.786, respectively, for day 1 and day 2), and (2) the relevant domains had proper correlations, while the clinical performance examination (CPX) and objective structured clinical examination (OSCE) showed weaker relationships. OSCE/CPX scores were correlated with other variables, especially grade point average and oral structured exam scores. The consequences of this assessment were (1) making students learn clinical skills and study themselves, while causing too much stress for students due to lack of motivation; (2) reminding educators of the need to apply practical teaching methods and to give feedback on the test results; and (3) providing an opportunity for faculty to consider developing support programs. It is necessary to develop the blueprint more precisely according to students' level and to verify the validity of the response process with statistical methods.

Keywords: Educational measurement, Reliability and validity, Undergraduate medical education

Corresponding author
Sunju Im
Department of Medical Education, Pusan National University School of Medicine, 49 Busandaehak-ro, Mulgeum-eup, Yangsan 50612, Korea
Tel: +82-51-510-8021
Fax: +82-51-510-8125
E-mail: sunjuim11@hanmail.net
https://orcid.org/0000-0002-3038-3570

Received: February 8, 2021
1st revised: June 16, 2021
Accepted: July 8, 2021

서 론

의과대학생들은 임상실습 전에 기본 임상술기를 습득하는 것이 필요하다. 임상실습 전 시기(pre-clinical phase)에는 기초의학과 임상의학에 대하여 주로 강의실에서 교육이 이루어지는 반면, 임상실습(clinical phase)은 병원 현장에서 환자를 통해 학습이 일어나기 때문이다. 임상실습에 진입하는 순간 학생들은 환자와 직접 대면하여 임상술기를 시행해야 한다. 따라서 학생들은 효과적인 임상실습을 수행하기 위해서 뿐만 아니라 환자 안전 측면에서 실습 전에 기본 임상술기를 습득해야 하고, 학교는 학생들이 습득했는지 평가해야 한다.

표준화환자(standardized patient)를 활용한 실기시험은 3, 4학년에서 주로 시행되지만, 최근 임상실습 전 교육의 평가에 시행되기도 한다. 학교는 실습 전에 표준화환자를 활용한 교육과 평가를 시행하여 임상실습으로의 전환을 원활하게 유도하고, 임상실습입문 교육 후에 실기시험을 활용하여 학생의 실습 전 술기능력을 점검하기도 한다[1-3]. 그러나 임상실습 전에 시행되는 실기시험을 4학년 학생들과 동일한 형태로 시행한다면 과연 타당한지 의문이 있다.

타당도는 검사 또는 측정도구가 본래 측정하고자 하였던 것을 충실히 측정하고 있는지를 나타내는 지표로, 측정점수에 가치를 부여할 뿐만 아니라 비용 또는 실행 가능성 등의 실행 측면을 포함한다. 전통적으로 '적합성'을 나타내는 타당도를 '일관성'을 나타내는 신뢰도와 구별하고, 준거타당도(예측타당도, 공인타당도), 내용타당도,

구인타당도 등으로 구분하여 설명하여 왔으나, 최근 다양한 종류의 타당도를 통합하려는 시각이 제시되어 왔다[4-6].

Messick [6]은 타당도를 여러 가지 측면, 즉 내용(content), 응답과정(response process), 내적 구조(internal structures), 일반화 가능도(generalizability), 다른 변수와의 상관성(relations to other variables), 영향(consequences)의 6가지 근거를 수집하여 종합적으로 판단하고자 하였다. 그러나 미국교육연구학회, 미국심리학회, 미국교육측정협회는 신뢰도(reliability) 또는 일반화 가능도를 타당도에 함의를 주는 점수의 특성으로 간주하고 있고, 연구자들은 타당도 내에서 신뢰도 관련 증거를 제시하고 있다[7]. 예를 들어 의료인 교육의 연구에서는 신뢰도 또는 일반화 가능도를 '내적 구조'에서 제시하면서, 총 5가지의 타당도 증거를 'Messick의 타당도 틀(Messick's validity framework)'이라 부르고 있다[4,8-11]. 이 타당도 틀은 내용적인 측면에서 시험의 내용은 측정하고자 하는 것을 충실히 포함하는지, 시험의 응답과정은 점수로 나타나기까지 질적으로 통제되는지, 시험의 내적 구조는 신뢰성 있게 구성되어 있는지, 같은 구인을 측정하는 다른 시험과 연관성이 있는지 또는 다른 구인을 측정하는 시험과 구별이 되는지, 학생, 교육자 또는 교육기관에 긍정적 또는 부정적으로 미친 영향은 무엇인지를 평가한다(Table 1). Messick의 타당도 틀은 타당도를 각각 평가하기보다 전반적으로 시험의 적합성을 분석하는 평가 틀을 제공하는 장점이 있다.

현재까지 국내에서는 임상실습 전 실기시험에 대한 보고가 드물고, 실기시험은 3, 4학년을 대상으로 주로 시행되고 연구되고 있으며, Messick의 타당도 틀을 활용하여 타당도를 연구한 사례가 없다. 해외에서도 타당도 틀을 사용한 분석이 늘어나고 있으나 임상실습 전 실기시험의 타당도 분석에서 타당도 틀을 사용한 사례는 드물다.

따라서 본 연구는 임상실습 전에 시행되는 실기시험의 타당도를 Messick의 틀을 사용하여 체계적으로 조사하고자 한다. 구체적인 연구문제는 다음과 같다. (1) 실기시험의 평가내용은 2학년 학생들의 수행능력을 평가하기에 합당한가? (2) 실기시험 과정은 질적으로 통제되었는가? (3) 실기시험의 내적 구조는 신뢰성을 확보하고 있는가? (4) 실기점수는 다른 측정점수와 관련이 있는가? (5) 실기시험은 학생, 교수 또는 교육에 어떤 영향을 미쳤는가?

연구대상 및 방법

1. 연구대상

기초임상통합교육을 마치고 임상실습에 진입하기 전인 부산대학교 의과대학의 2학년 학생 128명을 대상으로 하였다. 이 학생들은 각 통합과정에서 임상수기 내용을 교육 받았는데, 예를 들면 구토는 소화기학, 척추천자는 신경과학에서 학습하였다. 위 학생들을 대상으로 2학년 종료 시점에 종합시험(summative assessment)으로 기초종합평가, 임상종합평가, 구두시험, 실기시험 등 4가지 평가를 시행하였으며, 이 중 실기시험의 타당도를 분석하였다. 종합시험은 1주일 간격으로 시행되었고, 학생들은 통합과정 후에 개별 연습시간이 있었다.

실기시험은 진료문항(clinical performance examination, CPX)과 수기문항(objective structured clinical examination, OSCE)으로 구성되어 있다. 실기시험은 총 12개의 사례로 이루어져 있으며, 홀수 번호는 CPX, 짝수 번호는 OSCE로 구성되었고, 이틀간 시험을 시행하였다. 이틀간 시험에서 시험일정은 학생들을 무작위로 배정하였고, 사례는 학생들의 정보교환을 고려하여 일부 문항을 변경하여 사용하였다. 각 CPX 사례의 평가는 병력청취, 신체진찰, 환자 교육, 환자-의사 상호작용(patient-physician interaction, PPI) 등 4개의 세부영역(domain)으로 구분하였고, PPI는 표준화환자가, CPX의 나머지 영역과 OSCE는 교수 채점자가 평가하였다.

2. 연구방법

실기시험의 타당도는 Messick의 타당도 분석 틀을 이용하여 내

Table 1. Messick's five sources of validity evidence and their implementation in this study

Source of evidence	Definition	Evidence in this study
Content	- Identifying the characteristics of the content to be evaluated and reflecting them sufficiently	- Examination blueprint - Expertise and experience of case writers - Matching assessment content with objectives and teaching methods of the second year
Response process	- Quality control of the assessment	- Calibrating assessors for checklist ratings - Calibrating standardized patients for patient-physician interaction ratings
Internal structure	- Ensuring the reliability of scores in tasks that measure the same construct	- Case difficulty - Correlation analyses between scores of subcategories - Generalizability analysis
Relations to other variables	- Evaluating the correlations between test scores and other independent measures	- Correlations between CPX/OSCE scores and GPA, basic science written exam, clinical medicine written exam, and structured oral exam
Consequences of the test	- Defining the positive and negative impacts of the test on various aspects	- Effects on examinees - Effects on educators - Impacts on the curriculum or institution

CPX, clinical performance examination; OSCE, objective structured clinical examination; GPA, grade point average.

용, 응답과정, 내적 구조, 다른 변수와의 상관성 및 영향으로 구분하여 다음을 조사하였다(Table 1).

1) 내용타당도

내용타당도를 평가하기 위해 출제계획표(blueprint)대로 시험 사례를 선정하였는지 확인하였다. 또한 사례 개발자의 전문성과 경험을 확인하였다. 평가항목에 대한 교육이 적절하게 시행되었는지 파악하기 위해 항목별로 교육 여부와 교육방법을 조사하였다.

2) 응답과정

응답과정을 평가하기 위해 교수 채점자와 표준화환자 채점자에게 교육을 시행하여 오차를 줄이기 위한 노력을 하였는지 조사하였다.

3) 내적 구조

사례별 점수, CPX 영역별 점수, CPX 점수, OSCE 점수, 그리고 실기 총점을 각각 100점 만점으로 환산하여 보고하였다. CPX 영역별 점수, CPX 점수, OSCE 점수 및 실기 총점 간에 상관분석을 시행하여 구성이 타당한지 보고자 하였다. Pearson's correlation test를 이용하여 분석하였고, IBM SPSS ver. 22.0 (IBM Corp., Armonk, NY, USA)를 사용하였다.

실기점수의 신뢰도는 일반화 가능도 이론(generalizability theory)에 근거하여 평가하였다[12]. 일반화 가능도 이론은 실기점수와 같이 점수에 미치는 요인들이 다양할 때 신뢰도를 구하는 방법이다. 실기점수에 영향을 미치는 다양한 요인 중, 사례와 문항의 2개의 측면(2-facet)으로 분석을 진행하였다. 분석에 이용한 설계는, 사례(case, C)는 학생(person, P)과 cross되어 있으며, 문항(item, I)은 C 안에 nest되어 있으므로 P×(I:C) 설계를 이용하였다. 각 요인과 요인들의 상호작용, 즉 P, C, I:C (I nested within C), PC, PI:C가 평가점수의 총 변동에 미치는 영향을 파악하여, 이를 백분율로 표시하였다(각 변수로 인한 변동[variance]/총 변동[variance]). P는 학생의 실력 차이, C는 사례의 쉽고 어려운 정도, I:C는 사례 문항의 쉽고 어려운 정도, PC는 학생의 사례별 차이(사례 특이성, case specificity), PI:C는 나머지 알 수 없는 오차를 의미한다. 일반화 가능도 이론을 통해 도출된 변동을 이용하여 평가의 신뢰도를 추정하였고, 일반화 가능도 G계수(G-coefficient)와 파이계수(phi-coefficient)를 제시하였다. 분석에는 G-string IV 프로그램(Bloch & Norman, Hamilton, ON, Canada)을 이용하였다.

4) 다른 변수와의 상관성

실기점수와 평점평균 및 다른 종합시험 시험점수(기초종합평가, 임상종합평가, 구두시험)와의 상관관계를 분석하였다. 분석에는 Pearson's correlation test 방법을 사용하였다.

5) 영향

본 실기시험이 학생에게 미친 영향을 확인하기 위해, 해당 시험에 응시한 학생 128명을 대상으로 학년 말에 설문조사를 시행하여 그 결과를 분석하였다. 학년말 설문조사에서 실기시험 또는 종합시험 관련 문항은 20문항이었다. 연구자 2인이 객관식과 주관식 응답을 단독으로 검토하고 의미를 골라낸 후, 서로 토의하여 주요 사항을 협의하였다. 실기시험 결과를 토대로 교육자, 교육과정 및 학교가 개선을 모색하고 있는 사항을 조사하여 실기시험의 영향을 분석하고자 하였다.

결 과

1. 내용타당도

실기 사례는 출제계획표대로 선정하여 개발하였다. 1차 사례 작성은 실기시험 사례 개발의 경험이 있는 해당 분야의 전문가가 수행하였다. 이후 12명의 전문가 집단이 토의를 통해 사례를 수정하여 최종 개발을 완료하였다. 사례 개발이 완료된 후, 표준화환자를 대상으로 교육을 시행하고, 드레스 리허설을 통해 제대로 표현이 되었는지 점검하였다. 평가항목에 대한 교육이 시행된 교육방법을 조사한

Table 2. Matching assessment cases with teaching of second year

Case	Teaching during courses	Teaching methods
Day 1		
Vomiting	Yes	LGL
Chest X-ray presentation	Yes	LGL
Syncope	Yes	LGL
Visual acuity examination	Yes	SGP
Suicide	Yes	LGL
Papanicolaou smear	Yes	SGP
Knee pain	Yes	LGL
Suture	Yes	SGP
Polyuria	Yes	LGL
Venipuncture for blood culture	Yes	SGP
Hematuria	Yes	LGL
Wound dressing	Yes	SGP
Day 2		
Polyuria	Yes	LGL
Chest X-ray presentation	Yes	LGL
Skin rash	Yes	LGL
Anorectal examination	Yes	SGP
Oliguria	Yes	LGL
Wet smear of the vagina	Yes	SGP
Palpitation	Yes	LGL
Spinal tap	Yes	SGP
Smoking cessation counseling	Yes	LGL
Venipuncture for blood culture	Yes	SGP
Mood change	Yes	LGL
Burn dressing	Yes	SGP

LGL, large-group lecture; SGP, small-group practice.

결과 OSCE 항목은 실제 소그룹 실습으로 진행된 경우가 많았으나 CPX는 강의식으로 교육이 이루어졌다(Table 2).

2. 응답과정

PPI의 평가를 담당하는 표준화환자와 의학적 내용의 평가를 담당하는 교수 채점자를 대상으로 채점자 간 편차를 줄이기 위해 교육을 시행하였다. 표준화환자는 4시간의 사전교육 동안, CPX 영상을 보고 채점 후, 다른 채점자들과 의견을 교환하는 과정을 반복하여 PPI를 동일한 기준으로 평가할 수 있도록 하였다. 교수 채점자는 시험 당일 1시간의 교육을 시행하여 채점기준을 숙지하도록 하였다.

Table 3. Scores of 12-case CPX/OSCE (n=128)

Case	Mean±SD	Low	High
Day 1 (n=63)			
Vomiting	63.4±11.7	31.5	86.1
Chest X-ray presentation	60.1±20.5	5.9	100.0
Syncope	49.2±9.8	30.8	69.8
Visual acuity examination	77.1±14.0	21.1	100.0
Suicide	53.8±12.5	26.3	87.0
Papanicolaou smear	85.1±10.6	50.0	100.0
Knee pain	55.9±10.3	32.4	81.9
Suture	69.4±19.6	20.8	100.0
Polyuria	52.6±10.9	31.6	78.3
Venipuncture for blood culture	65.3±18.9	25.0	95.0
Hematuria	65.6±12.1	39.3	92.5
Wound dressing	82.0±12.7	35.3	100.0
Domain			
History-taking	63.9±8.7	41.1	82.2
Physical examination	45.0±11.5	10.2	71.0
Patient education	35.5±18.2	0	77.8
Patient-physician interaction	52.8±7.5	34.4	70.4
CPX	56.8±7.0	37.0	76.0
OSCE	73.2±10.1	43.1	90.1
Total score %	62.2±6.4	47.5	80.4
Day 2 (n=65)			
Polyuria	57.5±12.6	32.9	81.9
Chest X-ray presentation	67.1±20.2	23.5	100.0
Skin rash	54.2±8.2	37.5	68.3
Anorectal examination	84.2±8.9	60.0	100.0
Oliguria	51.9±9.3	33.9	71.1
Wet smear of the vagina	84.4±11.3	46.7	100.0
Palpitation	50.0±11.4	27.5	78.3
Spinal tap	64.8±19.5	18.8	100.0
Smoking cessation counseling	63.0±12.5	31.7	85.0
Venipuncture for blood culture	77.3±16.8	30.0	100.0
Mood change	55.9±10.5	31.8	76.8
Burn dressing	79.7±12.1	50.0	100.0
Domain			
History-taking	57.0±8.5	37.2	77.3
Physical examination	46.7±12.9	16.6	79.8
Patient education	41.5±14.6	10.0	71.1
Patient-physician interaction	57.8±7.9	44.7	75.3
CPX	55.4±7.2	42.3	69.3
OSCE	76.2±8.1	52.0	94.1
Total score %	62.4±6.6	47.5	75.5

The perfect score for each measurement item was 100 points.
CPX, clinical performance examination; OSCE, objective structured clinical examination; SD, standard deviation.

3. 내적 구조

1) 기술통계

첫째 날의 12가지 사례 중 자궁경부 펴바름(Papanicolaou smear)의 평균 점수가 가장 높았고(85.1±10.6), 실신이 가장 낮았다(49.2±9.8). 둘째 날 평균 점수가 높은 사례는 질분비물 검사(84.4±11.3), 평균 점수가 낮은 사례는 두근거림이었다(50.0±11.4). 세부영역별 평가결과를 살펴보면, 신체진찰 점수와 환자교육 점수가 낮은 편이었다. OSCE는 CPX보다 높은 점수를 보였다(Table 3).

2) 영역별 상관분석

상관 정도의 해석은 Dancey와 Reidy [13]가 제시한 기준에 따라 | r | <0.1은 없음, 0.1≤ | r | <0.4는 약함, 0.4≤ | r | <0.7은 중등도, 0.7≤ | r | <1은 강함, | r | =1은 완전히 일치함으로 해석하였다. 세부영역별 점수 및 CPX와 OSCE 점수 간의 상관관계를 살펴본 결과, 병력청취-OSCE를 제외한 모든 변수 간에 통계적으로 유의한 상관관계를 보였다. CPX 점수는 병력청취 및 신체검진과 강한 상관관계(r=0.813 and r=0.712), 환자 교육 및 PPI와는 중등도의 상관관계(r=0.568 and 0.637)를 보였다. PPI는 신체진찰 및 환자교육과 중등도의 상관관계를 보였다(r=0.429 and r=0.456). CPX와 OSCE는 약한 상관관계를 보였다(r=0.347) (Table 4).

3) 신뢰도 분석

첫째 날의 경우 사례에 포함되어 있는 문항(I:C)의 변동이 54.73%로 가장 높았으며, 학생의 사례별 변동(PC, 사례 특이성)은 3.44%, 학생에 의한 변동(P, 학생 변별도)은 1.19%로 나타났다. 일반화 가능도 G계수는 0.724였다. 둘째 날의 경우, 사례에 포함된 문항(I:C)의 변동이 63.93%로 가장 높았으며, 학생의 사례별 변동(PC)은 2.56%, 학생에 의한 변동(P)은 1.32%로 나타났다. 일반화 가능도 G계수는 0.786이었다(Table 5).

Table 4. Correlation analysis between subcategory scores

	HT	PE	ED	PPI	CPX	OSCE	Total score %
HT	1						
PE	0.345**	1					
ED	0.273**	0.243**	1				
PPI	0.250**	0.429**	0.456**	1			
CPX	0.813**	0.712**	0.568**	0.637**	1		
OSCE	0.146	0.274**	0.275**	0.431**	0.347**	1	
Total score %	0.664**	0.651**	0.545**	0.670**	0.895**	0.728**	1

HT, history-taking; PE, physical examination; ED, patient education; PPI, patient-physician interaction; CPX, clinical performance examination; OSCE, objective structured clinical examination.
**p<0.01.

Table 5. Generalizability analysis of the 12-case CPX/OSCE by test date

Effect	Coefficient	df	SS	MS	VC	% Variance
Day 1						
P	-	62	148.57	2.40	0.01	1.19
C	-	11	431.76	39.25	0.01	2.45
I:C	-	238	4,822.86	20.26	0.32	54.73
PC	-	682	434.17	0.64	0.02	3.44
PI:C	-	14,756	3,274.84	0.22	0.22	38.18
Phi	0.573					
G	0.724					
Day 2						
P	-	64	174.00	2.72	0.01	1.32
C	-	11	201.00	18.27	-0.01	-1.11
I:C	-	234	6,462.85	27.62	0.42	63.93
PC	-	704	396.75	0.56	0.02	2.56
PI:C	-	14,976	3,288.85	0.22	0.22	33.31
Phi	0.680					
G	0.786					

CPX, clinical performance examination; OSCE, objective structured clinical examination; df, degree of freedom; SS, sum of squares; MS, mean square; VC, variance component; P, person; C, case; I, item; I:C, item nested within case; Phi, phi-coefficient; G, G-coefficient.

Table 6. Correlation analysis between other independent measures

	GPA	Basic science written exam	Clinical medicine written exam	Structured oral exam
CPX/OSCE	0.605**	0.413**	0.476**	0.508**

GPA, grade point average; CPX, clinical performance examination; OSCE, objective structured clinical examination.
**p<0.01.

Table 7. Consequences

Subject	Impact
Students	Positive effects - Learning clinical skills - Learning a doctor's attitudes - Comprehensive understanding of learning content - Filling instructional gaps by studying broadly on their own Negative effects - Insufficient motivation due to poor understanding of the meaning of the assessment - Dissatisfaction with the exam due to unlearned content - Insufficient practice due to teacher-centered lectures, lack of models and materials, lack of practice time - Insufficient feedback - Excessive stress
Educators	Recognizing that - Small-group practice instead of classroom lectures is required - Feedback on the assessment is required
Curriculum, institution	Considerations - Reviewing both the curriculum related to OSCE and assessment blueprint - Developing a blueprint for the essential OSCE at the second-year level - Implementing a faculty development program for clinical skills instructors to improve their teaching and feedback skills - Implementing a student orientation program to introduce the meaning of the exam - Reducing students' excessive stress - Improving thepass/fail decision process

4. 다른 변수와의 상관성

실기시험의 점수와 평점평균, 기초종합평가, 임상종합평가, 구두시험의 점수와의 상관관계를 분석한 결과, 모두 중등도의 상관관계를 보였다. 그 중 평점평균과의 상관계수는 r=0.605, 구두시험과의 상관계수는 r=0.508였다(Table 6).

5. 영향

학생에게 미친 긍정적인 영향으로는 임상술기 및 의사의 태도에 대해 익힌 것, 학습내용의 포괄적인 이해, 스스로 학습하며 부족했던 부분을 보충할 수 있었던 점 등이 조사되었다. 부정적인 영향으로는 본 평가의 의미에 대한 이해 부족으로, 학습동기 부여가 불충분, 제대로 배우지 않은 내용이 출제된 점에 대한 불만, 교수자 중심의 강의형태의 수업과 실습모형과 도구 및 연습시간의 부족으로 인해 충분히 연습할 기회 부족, 평가결과에 대한 피드백이 불충분, 과도한 스트레스 등의 응답을 보였다(Table 7).

교육자에 미친 영향으로는 교실에서 이루어지는 강의형태의 수업보다 소규모의 실습 위주의 수업이 필요하다는 점, 평가결과에 대한 피드백이 필요하다는 점을 일깨워주는 효과가 있었다(Table 7).

교육과정 및 기관에 대해서는 실기 관련 교육과정 및 출제계획표에 대한 검토, 2학년 수준에 맞는 필수적인 실기내용에 관한 출제계획표 마련, 임상술기 교육자로 하여금 교육 및 피드백 능력을 향상시킬 수 있도록 하는 프로그램 개발, 본 평가의 의미에 대해 소개하는 학생 대상의 오리엔테이션 프로그램 제공, 큰 학점으로 인한 과도한 스트레스를 줄이는 방안, pass/fail 결정절차에 대한 논의 등에 대하여 고려하도록 한 효과가 있었다(Table 7).

고 찰

본 연구는 2학년에서 시행된 실기시험의 타당도를 분석하기 위해 Messick의 타당도 분석 틀을 활용하여 내용타당도, 응답과정, 내적 구조, 다른 변수와의 상관성 및 실기시험의 영향 등으로 세분하여 살펴보았다. 그 결과, 평가내용은 출제계획표에 근거하여 체계적으로 선정되었고, 시험과정은 교육 및 채점 훈련을 통해 질적으로 통제되었으며, 시험의 내적 구조는 신뢰성을 확보하고 있고, 다른 시험과 상관성을 보였다. 학생에게 술기를 익히도록 하고 스스로 학습하도록 하는 계기를 제공했으나, 동기부여 부족 및 지나친 스트

레스를 유발하였고, 교육자에게는 실기에 적합한 수업방법의 적용이 필요하며 평가결과에 대한 피드백이 필요하다는 점을 상기하였으며, 교육과정 및 기관으로 하여금 이를 지원하기 위한 프로그램 개발을 고려하도록 하는 효과가 있었다.

내용타당도를 확보하기 위하여 출제 전 출제계획표를 활용하여 이에 부합하도록 출제하였고, 전문가 집단의 검토와 표준화환자 교육 및 드레스 리허설을 시행하는 등, 사례 개발의 적절성과 표현의 정확성을 위한 절차가 충분히 반영되어 있었다. 그러나 사용된 출제계획표가 2학년의 수준에 적합하게 작성되었는지 검증이 필요하겠다. 2학년 수준에서 임상실습 진입을 위한 수준을 고려하여 출제계획표를 재설계하고, 면담 등 어려운 문항은 출제하지 않는 것이 바람직하겠다. 또한 본 시험에서는 출제계획표를 공식적으로 학생에게 공개하지 않았는데, 기존 연구에서 출제계획표를 공개하는 것은 교육자와 학생 모두 중요한 내용에 집중할 수 있도록 하며, 학생들로 하여금 시험의 타당도에 대해 긍정적인 인식을 갖게 한다고 알려져 있으므로[13], 출제계획표를 공개하여 소통하는 것을 고려할 필요가 있다. 한편, 각 평가항목에 대한 교육방법이 적합하지 않았던 경우가 확인되었다. 술기에 대한 수업은 강의형태보다 학생들이 실제 연습이 가능하도록 이루어져야 하며, CPX는 특히 실제 수업에서 표준화환자를 활용한 교육이 확대되어야 할 것이다.

응답과정에 대해서는 사전교육과 반복적인 모의채점 과정을 통해 채점자 간 편차를 최소화하고, 한 채점자가 반복되는 채점에서 동일한 기준으로 채점할 수 있도록 하였다. 다만 평가자 간 신뢰도 및 평가자 내 신뢰도를 분석하지는 못했다. Inter-rater agreement는 한 학생이 서로 다른 평가자로부터 받는 점수의 안정성을 보여주고, inter-rater reliability는 서로 다른 평가자들로부터 여러 학생들이 받은 점수의 일관성을 보여주는데, 이 두 가지가 항상 함께 충족되지는 않는 것으로 알려져 있으므로 inter-rater agreement와 inter-rater reliability가 각각 확보되었는지 객관적인 검증이 요구된다[14].

내적 구조의 타당도를 살펴보기 위해서는 기술통계 결과와, 영역별 점수의 상관분석, 일반화 가능도 이론을 이용한 신뢰도 분석을 이용하였다. 먼저 기술통계 결과에서, CPX보다 OSCE에서 높은 평균 점수를 보인 점과, 영역별 점수에서 환자 교육과 신체진찰이 낮은 점수를 보인 점을 통해 단순 술기보다 환자를 대면하여 진찰하고 설명하는 것에 학생들이 어려움을 느끼고 있는 것을 알 수 있고, 이 내용에 대한 충분한 교육이 이루어져야 할 것으로 생각된다.

영역별 점수를 살펴보았을 때, CPX와 OSCE는 약한 상관관계를 보여, 이 두 가지는 서로 다른 것을 평가하는 것이고, 교육도 달리 이루어져야 한다는 점을 시사한다. 병력청취 및 신체진찰 점수가 CPX 점수와 '강한' 상관관계를 보인 것은 병력청취와 신체진찰의 점수 비중이 CPX 내에서 차지하는 비율이 크기 때문인 것으로 생각된다. CPX 점수는 환자 교육 및 PPI와도 중등도의 상관관계를 보여, OSCE에서는 평가되지 않는 환자를 대면함으로써 발생하는 영역들을 CPX를 통해 타당하게 평가하고 있는 것으로 생각된다. PPI는 공감능력, 감정 인식 및 표현능력이 뛰어난 학생의 점수가 높다고 알려져 있고, 이러한 능력이 뛰어난 경우 임상상황을 더욱 정확히 인식하고, 환자의 입장을 고려한 문제해결방안을 모색할 수 있으며[15], 환자들은 자신에게 집중하는 의사로부터 더 많은 이득을 얻게 된다는 보고가 있다[16]. 본 연구에서 PPI는 신체진찰 및 환자 교육과 중등도의 관련을 보이는 것으로 나타나, PPI의 평가에는 환자의 신체를 접촉하게 될 때 충분히 설명하고 동의를 구하는 등 환자를 배려하는 내용과 환자에게 알기 쉽게 교육하는 것이 반영되었을 것으로 보인다.

일반화 가능도 이론에 의한 신뢰도는 첫째 날과 둘째 날 각각 0.724와 0.786으로 나타났다. 일반적으로 일반화 가능도 이론의 신뢰도는 0.7-0.8 정도면 이상적인 것으로 알려져 있어[17], 본 실기시험의 결과는 상당히 신뢰할 수 있는 것으로 판단된다. 측면별로 변동에 미친 영향 정도를 살펴보았을 때, 학생의 사례별 변동(PC)에 의한 변동은 첫째 날과 둘째 날 각각 3.44%와 2.56%로 낮게 나타나, 일반적으로 사례 특이성이 높게 보고되는 기존 연구들의 결과와는 일치하지 않았다[18,19]. 또한 학생으로 인한 변동은 첫째 날과 둘째 날 각각 1.19%와 1.32%로 나타나, 학생 변별도가 높지 않은 것으로 생각할 수 있다. 사례에 포함되어 있는 문항(I:C)으로 인한 변동이 가장 크게 나타나(54.73%, 63.93%), 본 실기시험에서는 '문항'으로 인한 변동이 가장 컸다는 것을 알 수 있다. 병력청취, 신체진찰, PPI 등에는 여러 사례에 공통적으로 있는 '일반적인(generalizable)' 문항과 사례 특이적인(case specific) 문항이 포함되어 있는데, 일반적인 문항들이 설명 가능한 변동에서 차지하는 비율이 큰 것으로 알려져 있다[19]. 본 시험에서는 일반적인 문항을 잘 대비한 학생과 그렇지 못한 학생이 있을 것으로 생각되고, 이것은 본 시험이 해당 학생들이 처음으로 경험해 본 실기시험이라는 점이 영향을 미쳤을 것으로 보인다. 각 사례에 특이적인 문항은 전체 학생의 수준에 비해 어려웠기 때문에 학생 변별도와 사례 특이성은 낮게 나온 것으로 유추할 수 있다. 일반적인 문항과 사례 특이적인 지식은 모두 임상에서 필수적인 요소이므로, 두 가지를 적절히 평가하는 것이 필요하며[19], 향후 본 시험의 의미와 대비방법 등에 대한 학생들의 준비가 향상된 상태에서 사례 특이적인 문항과 사례 특이성이 어떻게 변화되는지 분석할 필요가 있겠다.

다른 변수와의 상관성을 살펴본 결과, 평점평균과의 상관성이 높게 나타난 것은 학생의 일반적인 학습능력이 본 실기시험의 결과와도 상관성이 있는 것으로 생각할 수 있다. 실기시험과 구두시험과의 중등도의 상관성은, 기존 연구의 결과와 일치한다[20]. 또한 실기시험은 기초종합평가와 임상종합평가와도 중등도의 상관성을 보였는데, 아주 강한 상관을 보이지 않는다는 점에서 실기평가는 지필고사에서 평가하는 내용과는 다르며, 지필시험은 실기시험을 대체할

수 없다는 연구의 맥락과 일치한다[21]. 한편, Dadgar 등[22]에 의한 설문조사 연구에서 학생들은 OSCE/CPX, 구두시험, 지필시험 중 실기시험이 학습에 가장 도움이 되었다고 답했고, 구두시험이 두 번째, 지필시험이 그 다음으로 도움이 된 것으로 응답한 점과 OSCE/CPX는 최종 성적과 강한 상관관계를 보였고, 구두시험은 최종 성적과 중등도의 상관관계를 보인 점 등을 고려하면, 실기시험이 학습에도 도움이 되며 능력 측정에도 타당한 방법으로 생각된다.

본 실기시험은 학생 스스로 학습하며 부족한 부분을 보충할 수 있도록 하고, 임상술기 및 의사의 태도에 대해서 익힐 기회가 되었다는 것이 긍정적인 영향으로 조사되었으나, 예상치 못한 부정적 영향도 확인되었다. 부정적인 영향들은 대체로 시험의 난이도가 너무 높고 중요도가 크며, 수업방법이 평가방법에 적합하지 않았던 점, 시험의 의의에 대한 것이나 시험결과에 대한 의사소통이 부족했던 점에 의해 비롯된 것으로 보여, 향후 이에 대한 보완이 필요하다. 교육자와 교육과정 및 기관에 대해서는 위 내용에 대해 인지하게 하고 이를 극복하는 다양한 방안을 논의하도록 하였다는 효과가 있었다. 평가가 학생에게 미치는 영향은 시험 전, 시험 중, 시험 후로 나누어지는데[23,24], 본 평가는 시험 전의 부정적 영향으로 학생들에게 과도한 스트레스를 준 점과, 동기부여가 부족하고, 적합한 준비방법을 찾지 못했다는 것이 조사되었다. 기존 연구에서 시험에 대한 인식이 충분할수록 시험에 대해 더 잘 준비를 하도록 하여 평가 전 학습활동을 높인다고 알려져 있다[24]. 본 실기시험에서는 해당 학생들이 처음으로 접하는 형태의 실기시험이라는 점과, 2학년에서 종합시험이 시행된 첫 해라는 점 등으로 인해 시험에 대한 학생들의 인식이 낮았고, 시험 전 학습활동을 이끌어내는 데 충분하지 못했기에, 향후에는 이에 대한 보완이 필요하다. 또한 시험 후 학습의 효과를 극대화하기 위해서는 피드백이 필요하다. 특히 임상실습 진입 후에 환자를 직접 대면하게 되므로 지식, 술기뿐만 아니라 의사소통 부분에서 개별적인 피드백을 시행할 필요가 있다. 기존 연구에서 비디오 피드백은 실질적인 실력 향상의 효과는 미미하더라도 학생들의 만족도는 높고 큰 도움을 받은 것으로 여긴다는 것이 알려져 있고, 평가점수에 대한 통지 자체가 학생에게 학습동기 유발을 하게 되고, 이로 인해 실력이 향상되는 측면이 컸다고 밝히고 있어, 사례별, 항목별, 영역별로 학생의 성취도를 통지하는 것도 향후 학습에 도움이 될 것으로 생각된다[25].

본 연구에는 몇 가지 제한점이 있었다. 첫 번째, 일개 대학에서 시행한 시험에 대한 분석이므로 향후 보다 광범위한 연구가 필요하다. 두 번째, 응답과정의 타당도 분석에서 평가자 간 일치도, 평가자 내 일치도 등을 객관적으로 검증해보지 못했으며, 세 번째, 내적 구조에서는 모든 학생이 같은 내용의 시험을 보지는 않고 2일에 거쳐 시험을 보아 2일간 시험의 내용을 직접 비교할 수 없었으며, 일반화 가능도에 의한 신뢰도 분석에서 평가자 요인을 분석하지 못한 한계가 있었다. 마지막으로 이 시험은 해당 학생들이 처음 경험해 본 실기시험이었기 때문에, 시험에 대한 학생들의 충분한 인식이 없었다.

제한점을 고려하여 본 연구의 의의는 2학년에서 시행한 실기시험이 타당하며 신뢰성이 높다는 것을 확인하였고, 향후 교육과정과 평가방법에서 개선이 필요한 부분을 파악하는 데에 도움을 준 것으로 판단된다. 임상실습에 진입하기 직전 단계에서 학습내용을 총괄하는 데에 이러한 형태의 실기시험을 포함한 종합평가가 도움이 되는 것으로 보이고, 학생들이 처음으로 접하는 실기시험이라는 점을 고려하여 적정 난이도를 설정하고 최적의 평가방법과 교육방법을 모색하는 것이 필요할 것이다.

감사의 글

이 과제는 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.

저자 기여

이혜윤: 연구설계, 자료 분석과 원고 작성; 윤소정: 자료 분석과 원고 검토; 이상엽: 자료 분석과 원고 검토; 임선주: 연구설계, 자료 분석, 원고 작성과 검토

REFERENCES

1. Kim BS, Lee YM, Ahn DS, Park JY. Evaluation of introduction to clinical medicine by objective structured clinical examination. Korean J Med Educ. 2001;13(2):289-98.
2. Godefrooij MB, Diemers AD, Scherpbier AJ. Students' perceptions about the transition to the clinical phase of a medical curriculum with preclinical patient contacts; a focus group study. BMC Med Educ. 2010;10:28.
3. Graham R, Zubiaurre Bitzer LA, Anderson OR. Reliability and predictive validity of a comprehensive preclinical OSCE in dental education. J Dent Educ. 2013;77(2):161-7.
4. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 2015;49(6):560-75.
5. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. Adv Simul (Lond). 2016;1:31.
6. Messick S. Standards of validity and the validity of standards in performance assessment. Educ Meas Issues Pract. 1995;14(4):5-8.
7. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. Standards for educational and psychological testing. Washington (DC): American Educational Research Association; 2014.
8. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-7.
9. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts

as validity evidence?: examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ Theory Pract. 2014;19(2):233-50.
10. Borgersen NJ, Naur TM, Sorensen SM, Bjerrum F, Konge L, Subhi Y, et al. Gathering validity evidence for surgical simulation: a systematic review. Ann Surg. 2018;267(6):1063-8.
11. Pugh D, Hamstra SJ, Wood TJ, Humphrey-Murto S, Touchie C, Yudkowsky R, et al. A procedural skills OSCE: assessing technical and non-technical skills of internal medicine residents. Adv Health Sci Educ Theory Pract. 2015;20(1):85-100.
12. Brennan RL. Generalizability theory. New York (NY): Springer-Verlag; 2001.
13. Dancey CP, Reidy J. Statistics without maths for psychology. 4th ed. Harlow: Pearson Prentice Hall; 2004.
14. Liao SC, Hunt EA, Chen W. Comparison between inter-rater reliability and inter-rater agreement in performance assessment. Ann Acad Med Singap. 2010;39(8):613-8.
15. Kim SH, Ko JK, Park JH. Effect of emotional intelligence on patient-physician interaction scores of clinical performance examination. Korean J Med Educ. 2011;23(3):159-65.
16. Kiyohara LY, Kayano LK, Kobayashi ML, Alessi MS, Yamamoto MU, Yunes-Filho PR, et al. The patient-physician interactions as seen by undergraduate medical students. Sao Paulo Med J. 2001;119(3):97-100.
17. Downing SM, Yudkowsky R. Assessment in health professions education. New York (NY): Routledge; 2009.
18. Iramaneerat C, Yudkowsky R, Myford CM, Downing SM. Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. Adv Health Sci Educ Theory Pract. 2008;13(4):479-93.
19. Wimmers PF, Fung CC. The impact of case specificity and generalisable skills on clinical performance: a correlated traits-correlated methods approach. Med Educ. 2008;42(6):580-8.
20. Bakhsh TM, Sibiany AM, Al-Mashat FM, Meccawy AA, Al-Thubaity FK. Comparison of students' performance in the traditional oral clinical examination and the objective structured clinical examination. Saudi Med J. 2009;30(4):555-7.
21. Remmen R, Scherpbier A, Denekens J, Derese A, Hermann I, Hoogenboom R, et al. Correlation of a written test of skills and a performance based test: a study in two traditional medical schools. Med Teach. 2001;23(1):29-32.
22. Dadgar SR, Saleh A, Bahador H, Baradaran HR. OSCE as a tool for evaluation of practical semiology in comparison to MCQ & oral examination. J Pak Med Assoc. 2008;58(9):506-7.
23. Cilliers FJ, Schuwirth LW, Herman N, Adendorff HJ, van der Vleuten CP. A model of the pre-assessment learning effects of summative assessment in medical education. Adv Health Sci Educ Theory Pract. 2012;17(1):39-53.
24. Yune SJ, Lee SY, Im S. How do medical students prepare for examinations: pre-assessment cognitive and meta-cognitive activities. Korean Med Educ Rev. 2019;21(1):51-8.
25. Kim JH. The effect of remedial precepted video review on clinical performance examination scores. Korean Med Educ Rev. 2012;14(1):51-6.