진료수행시험에서 동일한 사례를 연기한 다수의 표준화 환자 간 채점결과 신뢰도

경희대학교 의과대학 의학교육학교실

고진경 · 윤태영 · 박재현

= Abstract =

Inter-rater Reliability in a Clinical Performance Examination Using Multiple Standardized Patients for the Same Case

Jinkyung Ko, PhD, **Tai-Young Yoon,** MD, PhD, MHA, **Jaehyun Park,** MD, PhD

Department of Medical Education, School of Medicine, Kyung Hee University, Seoul, Korea

Purpose: The "standardization" of standardized patients (SP) is one of the most crucial factors for a successful clinical performance examination (CPX). This study aimed to examine the inter-rater reliability among SPs who portrayed the same case during a CPX.

Methods: The context was a CPX conducted under the supervision of CPX Seoul-Gyeonggi Consortium in K medical school in August 2007. K medical school ran 12 stations consisting of duplicated sets of 6 cases. In total, thirty SPs participated with 5 SPs acting each of the 6 cases. The SPs evaluated the student's performances in addition to portraying the cases. ANCOVA (analysis of covariance) was used to compare scores rated by the different SPs. The dependent variables were the case scores and the 4 subcomponent (history taking, physical examination, Clinical courtesy, and Patient-physician interaction) scores for each case; the independent variable was the SPs; and the covariate was the CPX total score.

Results: The Headache and Cough stations showed an acceptable level of reliability. Otherwise, Weight Loss and Facial Flushing failed to show consistent scores in all 4 subcomponents. Diarrhea and Lt. hemiparesis showed partial consistency. In terms of the subcomponents, the physical exam scores were most consistent and the patient-physician interaction scores were most inconsistent.

Conclusion: This study tested the level of "standardization" of one set of CPX cases with mixed results. The authors hope that our results will contribute to quality assurance of CPX.

Key Words: CPX, Standardized Patient, Inter-rater Reliability, Analysis of Covariance, Assessment

교신저자: 고진경, 경희대학교 의과대학 의학교육학교실, 서울시 동대문구 회기동 1번지
Tel: 02)961-9102, Fax: 02)969-0792, E-mail: michkay@khu.ac.kr

* 이 연구는 교육부 2단계 BK21 사업에 의해 지원되었음.

서 론

표준화 환자를 이용한 교육과 훈련은 의학교육에서 중요한 교육방법 중 하나가 되었다. 미국의 경우 대다수의 의과대학에서 교육과 평가에 표준화 환자를 활용하고 있으며, 우리나라에서도 2004년 서울-경기지역에서 8개 의과대학이 컨소시엄을 형성하여 표준화 환자를 이용한 진료수행시험을 실시한 이후 점차 확대되어 현재는 참여 학교가 18개 학교로 늘어났다. 또한 각 학교단위에서도 표준화 환자를 이용한 다양한 교육과 평가를 시도하고 있다 (CPX Seoul-Gyeonggi Consortium, 2007). 표준화 환자를 활용하면 임상현장의 복잡하고 맥락의존적인 활동을 현실과 가장 유사한 환경에서 안전하게 연습할 수 있고, 학생의 수행에 대해 직접적인 피드백을 줄 수 있어 의도한 교육성과를 크게 높일 수 있다 (Solomon & Ferenchick, 2004).

표준화 환자를 활용한 교육과 평가는 높은 비용과 시간, 복잡한 사전준비와 까다로운 실행과정, 그리고 다수의 인력과 특정한 공간 등을 필요로 한다. 그럼에도 불구하고 더 많은 임상전문가와 교육전문가들이 표준화 환자를 활용한 교육과 평가가 학생들의 진료능력 수행 향상에 필요한 중요한 교육방법이라는 점에 의견을 모으고 있다. 이러한 흐름 속에서 한국보건의료인 국가시험원에서는 2010년부터 의과대학 졸업생들의 임상의사 자격시험에 진료수행시험을 포함시키기로 결정하고 이를 위한 준비를 하고 있다.

사실 졸업시험이나 자격시험과 같이 결과의 중요성이 높은 평가에 표준화 환자를 이용하기 위해서는 반드시 눈여겨봐야 할 측정학적 논의가 있다. 표준화 환자를 활용한 평가는 주로 수행평가의 형식으로 이루어진다. 그런데 수행평가는 의학교육뿐 아니라 다른 여러 교육분야에서 중요한 평가방법 중 하나로 활용되고 있으면서도 평가의 객관성에 대한 논란이 끊이지 않고 있는 평가방법이다 (Sung, 1994).

수행평가는 평가자의 판단에 근거해 피험자의 수행에 점수가 부여된다. 그러나 평가결과의 신뢰성을 확보하기 위해서는 채점자의 주관적 편견을 배제하고 객관성을 유지해야 하는 모순적 특성을 가지고 있다. 수행평가의 신뢰도 확인을 위해서는 한 채점자가 다른 채점자와 얼마나 유사하게 점수를 주었는지 혹은 한 채점자가 여러 학생들에 대해서 얼마나 일관성 있게 점수를 주었는지를 조사한다. 이 때 전자를 채점자 간 신뢰도라 하며 후자는 채점자 내 신뢰도라 한다. 또한 채점자가 이해하기 쉽고 명확한 채점척도와, 충분한 채점자 훈련, 채점자의 피로가 신뢰도에 나쁜 영향을 주지 않도록 평가시간을 적절히 안배하고 충분한 휴식을 취하도록 하는 등 여러 요소들을 함께 고려하도록 권고한다 (Sung, 1994; Kim & Song, 2001). 이와 같이 신뢰도의 문제는 수행평가에서 되풀이되어 제기되고, 주의를 요하는 쟁점사항이다.

진료수행시험에서 수험자들은 매우 다양하고 흥미로운 행동들을 보인다. 그러므로 이러한 행동들을 모두 고려하여 타당한 평가틀을 만드는 것이 매우 어려운 것으로 알려져 있다 (Solomon *et al*., 2000). 또한 표준화 환자와 학생의사와의 대면이 평가과정의 대부분을 차지하는데 이들 간의 상호작용은 평가의 객관성을 저하시킬 수 있는 잠재적 요소들을 다수 포함하며, 더구나 상호작용의 한 주체인 표준화 환자가 채점을 하는 경우에는 평가 신뢰도에 다시 한 번 부정적인 영향을 줄 수 있다. 쉽게 말하면 진료수행시험은 일반적인 수행평가보다 평가의 신뢰도에 위협이 될 수 있는 요소들을 더 많이 포함하고 있는 것이다.

현재 컨소시엄에서 주관하는 진료수행시험은 다수의 학생들을 제한된 시간에 평가하기 위해서 사례를 복제해서 운영하기도 하고, 때로는 표준화 환자의 피로도를 낮추기 위하여 같은 사례를 여러 표준화 환자가 나누어 연기하고 학생들을 평가하기도 한다. 이때 표준화 환자들이 적절한 훈련을 받았고, 평가기준을 충분히 이해하여 공정하게 평가를 했다면 학생들의 능력차를 통제하고 점수를 비교했을 때 각 표준화 환자가 평가한 학생그룹의 점수 간에는 차이가 없어야 한다. 한 사람의 표준화 환자의 점수가 다른 표준화 환자의 것과 통계적으로 유의

미한 차이를 보였다면 이는 이 연기자가 사례 재현을 다른 사람과 다르게 했거나, 평가기준을 다르게 해석했거나 혹은 공정하지 않게 평가를 했을 가능성이 있는 것이다. 이러한 이론적 명제에 근거하여 현재 진료수행시험의 연기와 채점의 일관성을 검사해 볼 수 있다. 검사결과, 학생의 능력차가 아닌 표준화 환자의 연기나 평가에 의해 유발된 점수의 편차의 폭이 크다면 이는 학생들에게 불이익을 줄 수 있는 불공정한 평가로서 반드시 수정되어야 한다 (Boulet *et al*., 2003).

이 연구는 표준화 환자의 연기와 평가의 신뢰도 수준을 검증하기 위하여 일개 의과대학에서 실시한 진료수행시험의 점수를 분석하였다. 동일한 사례를 다수의 표준화 환자가 연기하고 학생을 평가한 상황에서 각 표준화 환자가 대면하고 채점한 학생들의 점수를 비교함으로써 같은 사례에 참여한 표준화 환자들이 산출한 평가점수 간에 편차가 있는지 조사하였다. 이론적으로, 측정된 점수는 학생의 능력 차이와 측정오차를 포함한다. 이 중에서 학생의 능력차를 조정하여 일정하게 해 주면 측정오차만이 남게 되는데 이 오차의 크기를 추리통계의 방법에 의해 추정하여 표준화 환자 간의 편차가 통계적으로 유의한지 여부를 판단할 수 있다. 학생의 능력차를 통제하기 위해서 학생의 진료수행능력을 반영하는 진료수행시험의 전체 점수를 관계변수로 사용하였다. 구체적인 연구문제는 다음과 같다.

학생들의 능력 차이를 통제했을 때 동일한 사례를 연기하고 학생을 평가한 다수의 표준화 환자들은 일관된 평가결과를 산출했는가?

대상 및 방법

가. 연구의 맥락

이 연구는 2007년 8월 서울-경기 CPX 컨소시엄 주관 하에 K 의과대학에서 시행한 진료수행시험결과를 분석하였다. K대학은 진료수행시험을 위해 컨소시엄에서 개발한 사례 중 6개를 선정하고, 이를 복제하여 12개 스테이션을 운영하였다. 선정된 6개 사례는 각각 복통, 설사, 체중감소, 안면홍조, 좌반신마비, 기침을 주증상으로 개발되었다. 평가는 3일 동안 진행되었으며, 각 사례마다 5명의 표준화 환자가 동원되어 모두 30명의 표준화 환자가 평가에 참여하였다. 모든 표준화 환자들은 컨소시엄의 선발 및 훈련과정을 거쳐 검증된 자들이다. 평가대상은 의학과 4학년 학생 125명이었다.

학생들은 평가 시작 전 1분 동안 방 앞에서 준비된 상황소개 및 지침을 읽고, 입실하여 12분간 평가에 임하였다. 진료수행평가 후 퇴실하여 5분간 사이시험을 보고, 다음 방으로 이동하였다. 표준화 환자들은 학생들이 사이시험을 보는 5분 동안 학생평가지를 기록하였다. 한 스테이션에 소요되는 시간은 18분이며, 4개 스테이션 진행 후 표준화 환자의 휴식을 위해 10분의 시간이 주어졌고, 6개 스테이션을 모두 마친 후에는 20분 동안 학생들은 피드백 회합을 갖고, 운영위원들은 표준화 환자 교체 등 진행을 위한 준비를 하였다.

나. 연구도구

연구에 사용된 도구는 학생의 진료수행능력을 항목별로 측정하기 위해 개발된 사례별 학생평가표이다. 평가표의 기록은 학생의사를 대면한 표준화 환자가 맡았으며, 표준화 환자들은 사전에 사례 재현에 필요한 연기뿐 아니라 평가방법과 규칙에 대해서도 훈련 받았다.

학생평가표는 병력, 신체진찰, 임상예절, 환자-의사관계 등 4개 영역으로 이루어져 있는데 이중에서 병력과 신체진찰은 사례에 따라 각기 다른 문항들로 이루어져 있어, 문항수와 전체 평가표에서의 비중도 다르다. 임상예절과 환자-의사관계 영역은 모든 사례에서 동일한 문항으로 측정하도록 되어 있다. 사례별 문항수와 비중은 Table I에 제시하였다. 각 영역에 포함된 문항들은 평가방법이 달라 병력은 '예/아니오'로 수행여부를 기록하고, 신체진찰은 '제대로 했음/제대로 못했음/하지 않았음'으로 구분하여 수행의 수준에 따라 부분점수를 주도록 되어 있다. 임상예절 문항은 '예/아니오/해당 없음'으로 구분하여 문항이 지시하는 수행을 적용하

Table I. The Relative Contributions of Four Subcomponents in Each Case

	Total Score	HT*	PE†	CC‡	PPI§	GR‖	IPG¶
Headache	31	12 (39%)	6 (19%)	4 (13%)	7 (23%)	1 (3%)	1 (3%)
Diarrhea	32	10 (31%)	9 (28%)	4 (13%)	7 (22%)		
Weight Loss	31	16 (52%)	2 (6%)	4 (13%)	7 (23%)		
Facial Flushing	28	12 (43%)	3 (11%)	4 (14%)	7 (25%)		
Lt. Hemiparesis	30	6 (20%)	11 (37%)	4 (13%)	7 (23%)		
Cough	31	14 (45%)	4 (13%)	4 (13%)	7 (23%)		

* History taking, † Physical examination, ‡ Clinical courtesy, § Patient-physician interaction, ‖ Global rating, ¶ Initial patient greeting

기에 적합하지 않은 경우를 포함하도록 하였다. 마지막으로 환자-의사관계는 '최우수/아주 잘함/잘함/개선요망/최저수준/수준미달' 등 6점 척도로 학생들의 태도를 평가하도록 하였다.

학생평가표에는 4개의 하위영역 이외에도 전반적인 수행수준을 평가하는 (global rating) 문항과 자기소개 및 인사를 평가하는 문항이 있는데 이들의 결과도 총점 산출에 포함된다.

다. 분석방법

동일한 사례를 연기한 5명의 표준화 환자들이 산출한 평가점수 간 일치도를 검증하기 위하여 학생들의 진료수행시험 전체성적을 관계변수 (covariate)로 두어 공분산 분석하였다. 공분산 분석은 종속변수에 영향을 주는 외부요인을 통계적으로 통제하는 분석방법이다. 이 연구에서는 각기 다른 표준화 환자를 대면한 학생들의 점수차를 검증하되, 개인의 능력차를 통제하고 남은 점수의 분산만을 분석함으로써 진료수행시험에서 표준화 환자가 유발한 차이가 통계적으로 유의한지를 검증하였다. 분산분석 결과 통계적으로 유의한 차이가 나타나는 사례는 표준화 환자에 기인한 차이가 인정되므로 이들의 연기와 평가가 표준화되지 않은 것으로 판단할 수 있으며, 반대의 경우는 표준화된 것으로 인정할 수 있다.

이 연구의 종속변수는 진료수행평가의 사례점수 (case score)와 4개 영역점수 (subcomponent score)이고, 독립변수는 같은 사례를 연기한 각기 다른 표준화 환자이다. 학생들의 능력차를 통제하기 위한 관계변수는 6개 사례 점수의 합으로 산출된 진료수행평가의 총점 (CPX total score)이다. 통계분석은 SPSS 12.0를 활용하였고, 유의도 수준은 .01로 하였다.

결 과

가. 각 사례점수의 표준화 환자 간 일치도 분석결과

각 사례에서 5명의 표준화 환자가 평가한 사례점수의 평균과 표준편차는 Table II와 같다. 표준화 환자가 산출한 점수들을 일원분산 분석한 결과 6개 사례 모두에서 표준화 환자들의 점수가 일치하지 않았다 (두통 $F=5.10$, $p<.01$; 설사 $F=5.62$, $p<.01$; 체중감소 $F=12.68$, $p<.01$; 안면홍조 $F=14.43$, $p<.01$; 좌반신마비 $F=6.82$, $p<.01$; 기침 $F=4.66$, $p<.01$). 그러나 일원분산분석만으로는 학생들의 능력차가 유발한 점수의 차이를 배제할 수 없으므로 각 학생의 CPX 성적을 관계변수로 두고 공분산 분석을 실시하였다.

공분산 분석 결과에 따르면 6개 사례 모두 공분산의 효과가 확인되었다 (두통 $F=70.58$, $p<.01$; 설사 $F=61.49$, $p<.01$; 체중감소 $F=41.99$, $p<.01$; 안면홍조 $F=80.16$, $p<.01$; 좌반신마비 $F=76.45$, $p<.01$; 기침 $F=125.37$, $p<.01$). 채점자 간 평가결과 일치도

Table II. Descriptive Statistics and ANOVA Results for Case Scores

Variable	M (SE)					F
	1st SP	2nd SP	3rd SP	4th SP	5th SP	
Headache	14.14(2.54)	11.65(3.43)	12.80(3.16)	11.02(2.55)	13.26(2.82)	5.10*
Diarrhea	8.02(2.49)	9.56(2.61)	10.21(2.68)	9.88(1.88)	7.93(2.24)	5.62*
Weight loss	15.05(2.45)	11.57(2.16)	13.72(3.57)	13.88(1.89)	17.23(2.45)	12.68*
Facial flushing	14.74(3.12)	15.44(3.06)	16.66(2.94)	20.32(3.73)	15.36(2.78)	14.43*
Lt. Hemiparesis	16.64(3.06)	18.69(3.80)	13.47(2.34)	15.59(2.57)	16.78(3.77)	6.82*
Cough	18.18(2.56)	14.89(2.48)	15.59(2.81)	16.35(2.96)	15.70(2.51)	4.66*

* p<.01

Table III. Analysis of Covariance for Case Scores

Source	df	Headache		Diarrhea		Weight Loss	
		MS	F	MS	F	MS	F
CPX Total Score	1	356.00	70.58*	219.50	61.49*	223.92	41.99*
SP	4	6.72	1.33	33.21	9.30*	54.39	10.20*
Residual	119	5.04		3.57		5.33	
Total	124						
Source	df	Facial flushing		Lt. hemiparesis		Cough	
		MS	F	MS	F	MS	F
CPX Total Score	1	499.87	80.16*	502.17	76.45*	423.97	125.37*
SP	4	72.83	11.68*	12.22	1.86	10.40	3.08
Residual	119	6.24		6.57		3.38	
Total	124						

* p<.01

를 검증한 결과 두통 (F=1.33, p>.01) 좌반신마비 (F= 1.86, p>.01) 그리고 기침 (F=3.08, p>.01) 등 3개 사례에서 일치하였다 (Table III).

나. 4개 영역점수의 표준화 환자 간 일치도 분석결과

각 사례점수는 주로 4개의 영역점수에 의해 결정된다 (Table I). 각 사례점수를 영역별 점수로 해체하여 표준화 환자 간 일치도를 검증함으로써 채점자 간 신뢰도를 보다 세분하여 조사하였다. 각 사례에서 표준화 환자가 산출한 영역점수의 평균과 표준편차는 Table IV와 같다.

학생들의 CPX 성적을 관계변수로 두고 영역점수를 공분산 분석한 결과, 두통의 임상예절, 체중감소의 신체진찰과 임상예절, 그리고 안면홍조의 임상예절 영역에서만 공분산의 효과가 없었으며, 나머지 사례 및 영역에서는 공분산 분석이 타당하였다 (Table V).

Table IV. Descriptive Statistics and ANOVA Results for Scores of Subcomponents

Variable	Headache					F
	SP1	SP2	SP3	SP4	SP5	
HT†	5.10(1.63)	4.27(1.90)	4.70(2.29)	4.50(1.84)	4.47(1.88)	0.64
PE‡	1.59(1.08)	1.23(0.61)	1.33(0.75)	0.70(0.53)	1.24(0.81)	4.56*
CC§	2.48(0.69)	1.73(0.90)	2.65(0.88)	1.89(0.96)	2.71(0.76)	6.17*
PPI‖	4.37(0.72)	3.84(1.04)	3.69(1.03)	3.46(0.99)	4.22(0.74)	5.17*
Variable	Diarrhea					F
	SP6	SP7	SP8	SP9	SP10	
HT	3.27(1.38)	4.45(1.57)	4.94(1.30)	4.34(1.43)	3.14(1.38)	7.50*
PE	2.08(1.24)	2.36(0.71)	2.44(1.26)	2.43(1.11)	2.00(1.19)	0.86
CC	2.06(1.03)	2.18(0.87)	2.22(1.22)	2.60(0.55)	2.25(0.89)	1.64
PPI	4.07(0.86)	3.69(0.53)	4.17(0.48)	3.62(0.47)	3.86(0.70)	3.18
Variable	Weight loss					F
	SP11	SP12	SP13	SP14	SP15	
HT	6.93(1.92)	4.61(1.88)	6.36(2.00)	5.26(1.68)	8.35(1.97)	12.58*
PE	0.45(0.67)	0.00(0.00)	0.38(0.64)	0.43(0.59)	0.39(0.62)	1.96
CC	3.29(0.98)	3.06(1.06)	2.39(1.69)	3.65(0.71)	3.70(0.56)	6.24*
PPI	3.88(0.28)	3.40(0.79)	4.01(0.82)	3.96(0.36)	4.23(0.72)	4.56*
Variable	Facial flushing					F
	SP16	SP17	SP18	SP19	SP20	
HT	4.24(1.64)	4.69(1.99)	5.21(1.59)	7.12(2.04)	4.50(1.57)	12.40*
PE	2.71(1.14)	3.25(1.22)	3.26(1.25)	4.63(1.60)	2.79(1.23)	10.02
CC	2.83(1.07)	2.69(1.01)	2.88(1.01)	2.68(1.07)	2.67(1.15)	0.24*
PPI	4.35(0.64)	4.30(0.28)	4.71(0.48)	5.12(0.39)	4.72(0.96)	10.35*
Variable	Lt. hemiparesis					F
	SP21	SP22	SP23	SP24	SP25	
HT	3.06(1.16)	5.03(0.72)	2.55(1.04)	2.86(1.19)	3.30(1.23)	23.48*
PE	6.58(2.00)	6.79(2.39)	4.73(1.44)	5.83(1.82)	5.52(1.70)	3.39
CC	2.61(0.70)	2.53(1.05)	1.73(0.79)	2.09(0.51)	3.11(0.93)	8.31*
PPI	3.80(0.31)	3.82(0.56)	3.89(0.80)	4.23(0.53)	4.26(0.63)	4.34*
Variable	Cough					F
	SP26	SP27	SP28	SP29	SP30	
HT	7.94(2.11)	6.12(1.69)	6.60(1.10)	6.82(1.94)	7.15(1.71)	3.52
PE	1.50(0.59)	0.89(0.57)	1.20(0.91)	1.50(0.81)	1.15(0.79)	3.85
CC	2.71(0.59)	2.52(0.71)	2.77(0.94)	2.82(0.40)	2.32(0.77)	2.51
PPI	5.27(0.48)	4.72(0.65)	4.41(1.00)	4.60(0.36)	4.45(0.45)	1.90

* $p<.01$, †History taking, ‡Physical examination, §Clinical courtesy, ‖Patient-physician interaction

Table V. Analysis of Covariance for Scores of Subcomponents

		Headache_HT†		Headache_PE‡		Headache_CC§		Headache_PPI‖	
Source	df	MS	F	MS	F	MS	F	MS	F
CPX Total	1	93.04	32.55*	6.19	10.28*	0.38	0.56	24.18	41.63*
SP	4	3.42	1.20	1.64	2.72	3.47	5.04*	0.64	1.10
Residual	119	2.86		0.60		0.69		0.58	
Total	124								
		Diarrhea_HT		Diarrhea_PE		Diarrhea_ CC		Diarrhea_PPI	
Source	df	MS	F	MS	F	MS	F	MS	F
CPX Total	1	33.73	19.89*	22.71	19.39*	10.73	14.59*	12.82	39.90*
SP	4	15.06	8.88*	1.84	1.57	1.01	1.37	2.48	7.71*
Residual	119	1.70		1.17		0.74		0.32	
Total	124								
		Weight loss_HT		Weight loss_PE		Weight loss_CC		Weight loss_PPI	
Source	df	MS	F	MS	F	MS	F	MS	F
CPX Total	1	89.61	30.84*	0.02	0.05	1.64	1.28	13.74	47.29*
SP	4	33.08	11.38*	0.67	1.93	6.41	4.98*	1.72	5.92*
Residual	119	2.91		0.35		1.29		0.29	
Total	124								
		Facial flushing_HT		Facial flushing_PE		Facial flushing_CC		Facial flushing_PPI	
Source	df	MS	F	MS	F	MS	F	MS	F
CPX Total	1	106.43	45.82*	63.36	50.98*	0.55	0.49	7.94	34.63*
SP	4	20.13	8.67*	7.70	6.20*	0.36	0.32	2.00	8.72*
Residual	119	2.32		1.24		1.12		0.23	
Total	124								
		Lt. Hemiparesis_HT		Lt. Hemiparesis_PE		Lt. Hemiparesis_CC		Lt. Hemiparesis_PPI	
Source	df	MS	F	MS	F	MS	F	MS	F
CPX Total	1	26.37	27.97*	107.98	35.99*	8.62	13.92*	9.82	41.51*
SP	4	13.21	14.01*	6.88	2.29	4.38	7.07*	2.40	10.16*
Residual	119	0.94		3.00		0.62		0.24	
Total	124								
		Cough_HT		Cough_PE		Cough_CC		Cough_PPI	
Source	df	MS	F	MS	F	MS	F	MS	F
CPX Total	1	101.64	52.61*	11.39	24.39*	5.51	10.47*	15.79	49.32*
SP	4	1.89	0.98	0.54	1.17	1.34	2.55	2.22	6.95*
Residual	119	1.93		0.47		0.53		0.32	
Total	124								

* $p<.01$, †History taking, ‡Physical examination, § Clinical courtesy, ‖Patient-physician interaction

Table VI. Inter-Rater Consistency of Subcomponent Scores and Case Scores of 6 cases

	HT*	PE†	CC‡	PPI§	Case Score
Headache	○	○	NA‖	○	○
Diarrhea	X	○	○	X	X
Weight loss	X	NA	NA	X	X
Facial flushing	X	X	NA	X	X
Lt. Hemiparesis	X	○	X	X	○
Cough	○	○	○	X	○

* History taking, †Physical examination, ‡Clinical courtesy, §Patient-physician interaction, ‖Patient-physician interaction

공분산 분석을 이용한 영역점수 일치도 검증결과에 따르면 두통은 병력 (F=32.55, p>.01), 신체진찰 (F=10.28, p>.01), 환자-의사관계 (F=41.63, p>.01)에서 표준화 환자들의 점수가 일치하였으며, 설사는 신체진찰 (F=1.57, p>.01)과 임상예절 (F=1.37, p>.01) 점수가 일치하였다. 체중감소와 안면홍조는 모든 영역에서 표준화 환자들의 점수가 일치하지 않았다. 좌반신마비는 신체진찰 (F=2.29, p>.01) 점수만이 일치하였으며, 기침은 병력 (F=0.98, p>.01), 신체진찰 (F=1.17, p>.01) 그리고 임상예절 (F=2.55, p>.01) 점수가 일치하였다 (Table V).

6개 사례의 4개 영역점수와 사례점수의 일치도 여부를 Table VI에 한꺼번에 제시하였다. 영역점수 중에서 임상예절의 3개 점수와 신체진찰의 2개 점수는 공분산의 효과가 검증되지 않아 일치도 여부를 판단할 수 없었으며, 두통과 기침은 사례점수와 영역점수에서 고른 일치도를 보였고, 체중감소와 안면홍조는 전반적으로 일치하지 않은 결과를 보였다.

고 찰

연구결과에 따르면 표준화 환자의 채점자간 신뢰도가 확보된 사례는 6개 사례 중 두통과 기침에 불과하다. 이는 다양한 방법으로 채점자 간 일치도를 검증한 여러 연구에서 비교적 높은 일치율을 보고한 것과는 상반된 결과라 할 수 있다 (Neiman *et al*., 1988; Tamblyn *et al*., 1991; Martin *et al*., 1996; Wang *et al*., 1996; de Champlain *et al*., 1997). 각 사례의 영역별 분석을 살펴보면 두통을 연기한 표준화 환자들의 연기와 평가가 가장 안정적으로 '표준화'되었으며, 기침을 연기한 표준화 환자들도 환자-의사관계를 제외하면 적절한 표준화 수준에 도달하였다. 반면, 체중감소와 안면홍조를 연기한 표준화 환자들은 일치도를 확보한 영역이 전혀 없어 표준화에 실패하였다. 설사와 좌반신마비는 영역에 따라 부분적인 일치도를 보였다.

이 중에서 좌반신마비는 영역점수와 사례점수의 결과가 서로 다른 방향을 가리키고 있어 해석에 주의를 요한다. 영역점수 중 신체진찰에서만 채점자 간 결과가 일치하고, 나머지 3개 영역에서는 일치하지 않았음에도 불구하고 사례점수에서는 채점자 간 결과가 일치한 것으로 나타났다. Table I을 참조해 보면 좌반신마비는 전체 30 문항 중 신체진찰이 11문항으로 37%를 차지하고 있어 다른 영역에 비해 신체진찰의 비중이 높은 편이다. 신체진찰은 이 사례의 중요한 영역이며 이에 대한 표준화 환자의 훈련과 평가표 개발이 보다 신중하게 이루어졌을 것으로 짐작할 수 있다. 그러므로 영역점수와 사례점수에서의 상반된 결과는 잘 개발되고 준비된 하나의 중심영역이 전체 결과에 강력하게 영향을 주어 사례 점수의 일치도를 높인 것으로 이해할 수 있겠다.

사례점수의 일치도 여부와 상관없이 설사와 좌반신마비는 영역에 따라 부분적인 불일치를 보인 사

례이다. 이러한 사례에 대해서는 표준화를 이루지 못한 것으로 판정하고, 적절한 기준에 도달하지 못한 영역에 대한 수정, 보완이 이루어져야 한다. 그래야만 사례의 안정성을 확보할 수 있을 것이다.

연구결과에 근거해서 체중감소와 안면홍조의 전체 영역과 설사와 좌반신마비의 병력, 임상예절, 환자-의사관계 관련 부분의 시나리오와 표준화 환자 훈련지침을 내용 전문가가 다시 검토하고, 환자역할을 연기한 연기자와 시험에 참여한 학생들의 의견을 수렴하며, 사례를 재현한 표준화 환자들의 실제 학생대면 장면을 분석하여 각 영역의 표준화를 방해한 요소를 찾아내는 후속작업을 제언한다. 또한 분석자료의 범위를 넓혀 동일한 사례와 연기자를 활용하여 진료수행시험을 실시한 다른 대학들의 평가결과를 분석함으로써 6개 사례와 표준화 환자의 신뢰도에 대한 본 연구의 결과를 일반화하는 후속 연구도 함께 제언한다.

진료수행시험은 병력, 신체진찰, 임상예절, 환자-의사관계 증 4개 영역을 주축으로 평가한다. 연구결과를 영역별로 구분해서 살펴보면 신체진찰은 4개 사례에서 채점자간 점수가 일치했으나 병력은 2개 사례에서, 환자-의사관계는 단지 1개 사례에서만 채점자 간 점수가 일치하였다. 신체진찰이 병력보다 더 많은 사례에서 일치도를 보인 결과는 선행연구와 같다. Heine *et al.* (2003)는 신체진찰에서 96%, 병력청취에서 94%의 일치도를 보고하면서 표준화 환자가 학생의 신체진찰 행위의 정확성에 대해서 판단하기가 쉽지 않음에도 불구하고 병력청취보다 높은 일치도를 보인 이유는 신체진찰이 더 기억하기 쉽기 때문이라고 설명하였다. 연구자가 실제 표준화 환자 훈련에 참석한 경험을 떠올려 보면 표준화 환자들은 학생의 수행여부를 기록하는 신체진찰보다는 병력이나 환자-의사관계에서 학생의사의 말과 행동에 대한 해석의 범위를 어떻게 정해야 하는지에 대한 질문과 논의를 더 많이 하였다. 이 연구에서 신체진찰 점수가 병력이나 환자-의사관계 점수보다 많은 일치한 것은 신체진찰이 더 기억하기 쉽고, 주관적 해석의 여지가 적기 때문인 것으로 해석할 수 있다.

환자-의사관계 점수의 낮은 일치도에 대해서는 Pangaro *et al.* (1997)도 의사소통 영역의 점수가 다른 영역에 비해 일치도가 떨어지는 연구결과를 보고한 바 있다. 환자-의사관계는 의사소통 및 대인관계 능력을 평가하는 문항들로 구성되어 있어 병력청취보다 더 주관성이 개입되기 쉬운 특성이 있다. 그러나 환자-의사관계 문항들은 모든 사례에 공통문항으로 사용되며, 진료수행시험뿐 아니라 향후 우리나라 환자-의사관계 교육훈련의 기준으로 활용될 가능성이 높으므로 평가문항에 대한 충분한 분석이 필요하다.

임상예절에 대해서는 전면적인 논의가 필요하다. 6개 사례 중 3개 사례에서 공분산 분석이 적절하지 않았고, 1개 사례는 평가자 간 평가결과가 일치하지 않았다. 공분산 분석이 적절하지 않다는 것은 관계변수인 전체 진료수행점수와 종속변수인 해당 사례의 임상예절 점수 간 상관이 없어 공분산 분석이 무의미함을 나타낸다. 즉, 임상예절 영역의 문항들이 학생들의 진료수행능력과 상관이 없는 다른 능력을 측정하고 있거나 심각한 오류를 가지고 있을 수 있음을 암시한다. 임상예절은 모든 사례에 동일한 문항이 적용되며 신체진찰을 할 때의 예절을 평가한다. 문항의 내용은 1) 손 씻기 2) 환자의 신체를 적절히 가려주기 3) 미리 신체검진에 대해 설명하기 4) 진찰에 대한 불편여부에 대해 질문하기 등이다. 이 문항들은 신체진찰에서의 수험자의 행동과 점수와 관련성이 매우 크다. 신체진찰이 필요 없는 사례이거나 수험자가 임의로 신체진찰을 하지 않을 때 임상예절 영역의 채점은 문제가 될 수 있다. 이러한 문항의 내재적인 모순 때문에 각 문항은 했음/안했음 이외에 '해당 없음' 항목을 두고 있다. 그러나 '해당 없음' 항목은 표준화 환자의 판단에 방해가 될 뿐 아니라 점수산출에서도 문제를 야기한다. 전체 점수 산출을 위해서는 '해당 없음' 항목에 점수를 배정하는 것이 적합한가 아닌가에 대한 판단이 필요한데 이에 대해서도 논란의 여지가 있다. 예를 들어 필요한 신체진찰을 하지 않은 수험자가 있다고 가정할 때, '해당 없음'을 0점으로 처리한다면 임상예절 문항은 신체진찰에서의 실수를 배가시키는

함정이 될 수 있다. 반대로 점수를 준다면 마땅히 평가를 해야 하는 능력에 대해서 판단을 유보하고 무조건 기본점수를 주는 무의미한 문항이 될 수 있다. 이 연구에서는 '해당 없음'을 '했음'과 같은 점수를 주어 합산하였다. 임상예절 문항이 가지고 있는 문제점은 연구결과 3개 사례의 임상예절 점수에 대해 공분산 분석이 부적절하다는 통계적 표현으로 나타났다. 이에 대한 논의가 요망된다.

이 연구의 연구방법은 동일한 사례를 한 사람 이상의 표준화 환자가 재현해야 하는 평가의 특수상황을 이용하여 진료수행시험에서 강조되고 있는 '표준화'를 검증하기 위해 채택되었다. 표준화 환자라는 용어에서도 나타나듯이 진료수행시험은 각 학생들에게 제시되는 표준화된 자극과 상황이 평가의 신뢰도에 막대한 영향을 미치는 평가이다. 이 표준화를 위해서 여러 내용전문가와 평가전문가들이 사례와 평가표를 여러 번 검사하고 확인하지만 실제로 어떤 수준까지 표준화를 이루고 있는지 확인하기는 어려운 것이 사실이다. 왜냐하면 사례에 따라, 표준화 환자에 따라, 그리고 평가받는 학생에 따라 여러 요소들이 한꺼번에 변하는 평가 상황에서 무엇을 기준으로 점수를 비교하여 최소한의 일관성을 검증해야 할지 모호하기 때문이다. 이에 본 연구는 사례가 동일할 때 학생들의 능력수준을 통제한 후, 다수의 표준화 환자들에 의해 만들어지는 점수의 분산을 분석하는 방법을 사용하였다 (Boulet *et al.*, 2003).

이 연구에서 사용한 공변량분석은 통계패키지를 이용하여 간단한 과정을 거쳐 결과를 낼 수 있다. 물론 평가의 오차요인을 탐색하여 그 영향력과 이들 간의 상호관계를 밝히는 다른 분석방법이 많이 있지만 복잡한 절차와 어려운 해석을 요구하는 방법은 효용성이 떨어질 수 있다. 이 연구에서 제시한 방법은 표준화 환자의 표준화 수준을 검증하는 데 초점을 두고 각 수행평가 사후에 간단한 분석작업을 통해 사례의 재사용 여부나 수정해야 할 영역에 대한 즉각적인 정보를 얻을 수 있다는 측면에서 진료수행시험의 질관리에 도움을 줄 수 있으리라 생각된다.

이 연구는 각 사례 및 영역에 대한 표준화 환자 간 평가결과 일치도 여부를 검정하였다. 그러나 연구가 관심을 두고 있는 표준화 연기자가 유발하는 오차의 이면에는 정확하지 않은 평가표, 연기자의 사례재현의 오류, 채점자의 평가표 이해 부족, 혹은 공정하지 않은 평가 등 다양한 원인이 있을 수 있다. 그럼에도 불구하고 연구결과는 구체적인 오차요인을 지적하지 못하였다. 이는 본 연구의 한계로서 후속연구를 통해 밝혀져야 할 중요한 사항이다. 구체적인 오차의 소재를 밝히기 위해서는 표준화 환자와 학생의사의 대면 장면을 분석하고, 표준화 환자의 행동과 평가결과를 비교할 수 있는 질적 연구방법이 필요하다. 이를 통해 잠재적인 오차요인들이 밝혀진다면 오차요인들을 체계적으로 배열한 적절한 평가 상황을 설계하고, 그 결과에 대해 일반화가능도 이론 (generalizability theory)을 적용함으로써 각 오차요인의 영향력과 최선의 평가 상황을 구성하는 세부 요소들을 결정할 수도 있다. 이러한 후속연구들은 진료수행평가의 안정성을 높이는 데 크게 기여할 수 있을 것이다. 아울러 이 연구는 일개 의과대학에서 시행한 진료수행시험만을 분석한 결과이므로 일반화해서 적용하기에는 부족하다는 점도 연구의 한계로 밝혀 둔다.

참 고 문 헌

Ahn, D. & Im, H.(2001). Standard setting in student assessment by criterion referenced evaluation. *Korean Journal of Medical Education, 13(1)*, 41-45.

Auewarakul, C., Dowing, S.M., Praditsuwan, R. & Jaturatamrong, U.(2005). Item analysis to improve reliability for an internal medicine undergraduate OSCE. *Adv Health Sci Educ Theory Pract, 10(2),* 105-113.

Boulet, J.R., McKinley, D.W., Whelan, G.P., & Hambleton, R.K.(2003). Quality assurance methods for performance-based assessments. *Adv Health Sci Educ Theory Pract, 8(1),* 27-47.

Brennan, R.(2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24(4)*, 339-353.

Chesser, A.M., Laing, M.R., Miedzybrodzka, Z.H., Brittenden, J. & Heys, S.D.(2004). Factor analysis can be a useful standard setting tool in a high stakes OSCE assessment. *Med Educ, 38(8)*, 825-831.

CPX Seoul-Gyeonggi Consortium http://www.cpx.or.kr/ (search date: 2007.9.15)

Crossley, J., Davis, H., Humphris, G., & Jolly, B.(2002). Generalisability: a key to unlock professional assessment. *Med Educ, 36(10)*, 972-978.

de Champlain, A.F., Margolis, M.J., King, A., & Klass, D.J.(1997). Standardised patients' accuracy in recording examinees' behaviours using checklists. *Acad Med, 72(10 Suppl 1)*, S85-S87.

Downing, S.(2004). Reliability: on the reproducibility if assessment data. *Med Educ, 38*, 1006-1012.

Heine, N., Garman, K., Wallace, P., Bartos, R., & Richards, A.(2003). An analysis of standardized patient checklist errors and their effect on student scores. *Med Educ, 37*, 99-104.

Im, H., & Kim, S.(2005). A study of investigating error sources and reliability for clinical performance examination (CPX). *Journal of Educational Evaluation, 18(1)*, 27-46.

Kim, J., Lee, K., Choi, K., & Lee, D.(2004). Analysis of the evaluation for clinical performance examination using standardized patients in one medical school. *Korean Journal of Medical Education, 16(1)*, 51-61.

Kim, K., & Song, M.(2001). Comparison of inter-rater consistence and accuracy of estimation for ability parameters among multiple scoring scales. *Journal of Educational Evaluation, 14(1)*, 327-347.

Kim, S., Park, S., Hur, Y., & Lee, S.(2005). The appropriateness of using standardized patients (SPs) assessment scores in clinical performance examination (CPX). *Korean Journal of Medical Education, 17(2)*, 163-172.

Kwon, I., Kim, N., Lee, S., Eo, E., Park, H., & Lee, D.(2005). Comparison of the evaluation results of faculty with those of standardized patients in a clinical performance examination experience. *Korean Journal of Medical Education, 17(2)*, 173-183.

Martin, J.A., Reznick, R.K., Rothman, A., Tamblyn, R.M., & Regehr, G.(1996). Who should rate candidates in an objective structured clinical examination? *Acad Med, 71(2)*, 170-175.

Mazor, K.M., Ockene, J.K., Rogers, H.J., Carlin, M.M., & Quirk, M.E.(2005). The relationship between checklist scores on a communication OSCE and analogue patients' perceptions of communication. *Adv Health Sci Educ Theory Pract, 10(1)*, 37-51.

Nieman, L.Z., Vernon, M.S., Holbert, D., & Boyett, L.(1988). Training and validating the use of geriatric simulated patients. *Res Med Educ, 27*, 154-159.

Pangaro, L.N., Worth-Dickstein, H., Macmillan, M.K., Klass, D.J., & Shatzer, J.H.(1997). Performance of "standardized examinees" in a standardized-patient examination of clinical skills. *Acad Med, 72(11)*, 1008-1011.

Park, H., & Kwon, O.(2005). Sharing of information among students and its effect on the scores of clinical performance examination. *Korean Journal of Medical Education, 17(2)*, 185-195.

Park, W., Lee, S., Kim, E., Kim, Y., Kim, S., & Shin, J.(2005). Correlation of CPX scores with the scores of the clinical clerkship assessments and written examinations. *Korean Journal of Medical Education, 17(3)*, 297-303.

Solomon, D.J., & Ferenchick, G.(2004). Source of measurement error in an ECG examination:

implications for performance-based assessments. *Adv Health Sci Educ Theory Pract, 9(4)*, 283-290.

Solomon, D.J., Szauter, K., Rosebraugh, C.J., & Callaway, M.R.(2000). Global rating of student performance in a standardized patient examination: is the whole more than the sum of the parts? *Adv Health Sci Educ Theory Pract, 5(2)*, 131-140.

Sung, T.(1994). Nonsulhyung gosawa yecheneuinggye silgigosareul wihan chejumjagan silroedo chujeong. *Journal of Educational Evaluation, 7(1)*, 43-56.

Tamblyn, R., Klass, D., Schnabl, G., & Kopelow, M.(1991). Sources of unreliability and bias in standardised patient rating. *Teaching Learning Med, 3*, 74-85.

Wang, W., Stillman, P., Stunick, A., Ben-David, M., & Williams, R.(1996). The effect of fatigue on the accuracy of standardised patients' checklist recording. *Teaching Learning Med, 8*, 148-151.