ORIGINAL ARTICLE

Faculty Observer and Standardized Patient Accuracy in Recording Examinees' Behaviors Using Checklists in the Clinical Performance Examination

Jaehyun Park, Jinkyung Ko, Sunmi Kim and Hyobin Yoo

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

표준화 환자를 이용한 진료수행시험에서 교수와 표준화 환자의 채점 정확도

경희대학교 의학전문대학원 의학교육학교실

박재현, 고진경, 김선미, 유효빈

Purpose: The purpose of the study was to examine the recording accuracy of faculty observers and standardized patients (SPs) on a clinical performance examination (CPX).

Methods: This was a cross-sectional study of a fourth-year medical students' CPX that was held at a medical school in Seoul, Korea. The CPX consisted of 4 cases and was administered to 118 examinees, with the participation of 52 SP and 45 faculty observers. For the study we chose 15 examinees per case, and analyzed 60 student-SP encounters in total. To determine the recording accuracy level, 2 SP trainers developed an answer key for each encounter. First, we computed agreement rates (P) and kappa coefficient (K) values between the answer key–SPs and the answer key–faculty observers. Secondly, we analyzed variance (ANOVA) with repeated measures to determine whether the mean percentage of the correct checklist score differed as a function of the rater, the case, or the interaction between both factors.

Results: Mean P rates ranged from 0.72 to 0.86, while mean K values varied from 0.39 to 0.59. The SP checklist accuracy was higher than that of faculty observersat the level of item comparison. Results from ANOVA showed that there was no significant difference between the percentage of correct scores by the answer key, faculty observers and SPs. There was no significant interaction between rater and case factors.

Conclusion: Acceptable levels of recording accuracy were obtained in both rater groups. SP raters can replace faculty raters in a large-scale CPX with thorough preparation.

Key Words: Clinical competence, Undergraduate medical education, Observer variation, Educational measurement

Received: April 13, 2009 • Accepted: June 25, 2009

Corresponding Author: Jinkyung Ko
Department of Medical Education, Kyung Hee University School of Medicine, 1 Hoegi-dong, Dongdaemun-gu, Seoul 130-701, Korea
TEL) 02-961-9102 FAX) 02-969-0792 E-mail) michkay@khu.ac.kr

Korean J Med Educ 2009 Sep; 21(3): 287-297.
doi: 10.3946/kjme.2009.21.3.287.

서론

수행평가에서 채점자의 신뢰도에 관한 문제는 오래된 논쟁거리이다. 진료수행시험(clinical performance examination, CPX) 역시 예외가 아니어서 채점자신뢰도와 정확성을 확인하고, 이를 높이기 위한 방법을 모색하는 많은 연구들이 발표되었다[1]. 다양한 논의 가운데 주요 쟁점은 누가 가장 신뢰할 수 있는 채점자인지, 여러 명의 채점자가 참여하는 시험에서 이들의 채점이 일관적인지, 어떤 채점도구가 채점오류를 최소화 할 수 있는지 등이다. 채점자 신뢰도 문제는 실제 진료수행시험의 운영에서도 적절한 채점자의 자격요건을 결정하여 이들을 선정하고, 훈련하고 시험일정에 따라 배치하는 등 여러 가지 고려해야 할 사안들을 만든다. 특히 의사면허시험의 일환으로 국가수준의 대규모 진료수행시험을 준비하는 상황이라면 신뢰할 만한 채점자를 충분히 확보하고 훈련하는 것은 해결이 쉽지 않은 현실적 과제이다.

실제로 2009년 시작되는 의사실기시험에서 채점관을 결정하는 문제는 신중하게 판단해야 할 중요한 사안이다. 중요한 시험에 적합한, 신뢰할 만한 채점관은 누구일까? 첫 번째 후보는 임상교수들이다. 이들은 신뢰도 높은 채점자로서 진료에 관한 충분한 전문성과 경험을 가지고 있다. 그런데 임상교수 집단은 구성원 수가 적고, 시간적 제약도 매우 크다. 더구나 훈련과 시험에 투입되는 임상교수들의 전문성과 시간에 상응하는 비용지출의 문제도 간과하기 어려운 문제이다. 다음 후보는 진료수행시험에 참여하는 표준화 환자이다. 이들은 현실적 어려움이 많은 임상교수를 대신해서 '표준화된' 채점을 하기에 적합한 대상들이다. 또한 여러 CPX 컨소시엄이나 대학차원의 진료수행시험에서 이미 채점자로서 역할을 하고 있기도 하다. 그러나 자격시험과 같이 중요한 시험에서 채점관 역할을 하기에 충분한 역량을 갖추고 있는지, 혹시 시험의 신뢰도나 타당도를 떨어뜨리게 되지는 않을지 우려하는 목소리도 적지 않다.

진료수행시험을 먼저 도입한 북미와 유럽에서도 유사한 논란이 있었다. 임상교수와 표준화 환자 간의 진료수행평가의 채점 결과를 비교하는 연구논문이 여러 편 발표되었다. Martin et al. [2]은 의사와 표준화 환자의 채점 결과를 정답기준(golden standard)과 비교한 연구에서 의사채점자의 정확성이 더 높다는 결론을 내렸다. McLaughlin et al. [3]은 학생들은 표준화 환자 채점자에 대해 긍정적인 태도를 가지고 있으나 표준화 환자의 점수는 의사보다 전반적으로 높았고, 다른 측정(MCQ) 결과와의 관련성이 낮았다고 보고하였다. 반면, Kopp & Johnson [4]의 연구에 의하면 의사와 표준화 환자의 채점 결과의 일치도가 81~92% 사이였으며, 시험의 후반부로 갈수록 일치도가 높아졌다. 이들은 표준화 환자가 신뢰할만한 채점자의 역할을 할 수 있다는 결론을 내렸다. MacRae et al. [5]은 병력청취와 신체진찰에서 표준화 환자의 채점표 점수가 의사의 평가결과와 높은 상관관계를 나타냈음을 보고하였다. Kwon et al. [6]도 표준화 환자와 임상교수의 채점결과가 병력과 신체진찰 영역에서의 일치도가 높아 표준화 환자의 채점 결과를 임상교수의 것을 대신해 사용할 수 있다고 논의하였다. 그러나 환자-의사관계에서는 일치율이 크게 떨어졌음을 함께 보고하였다.

임상교수와 비교하는 방법을 택하지 않고 표준화 환자의 채점 결과의 신뢰성만 집중적으로 조사한 연구들도 있다. Vu et al. [7]에 의하면 표준화 환자의 채점정확도는 '좋음'에서 '매우 좋음' 수준을 보였다. 또한 이들의 채점정확도는 채점표의 길이와 문항의 유형 및 명료함에 영향 받았으나 하루 중의 시기나 시험기간 중 날짜에는 영향 받지 않았다고 보고하였다. De Champlain et al. [8]은 표준화 환자와 관찰자의 채점 결과를 정답기준과 비교하고, 표준화 환자와 정답기준과의 높은 일치율을 제시하였다. Heine et al. [9]은 정답기준을 작성하여 비교함으로써 표준화 환자의 채점오류의 유형(누락오류[omission]/첨가오류[commission])과 빈도를 분석하였다. 표준화 환자들은 신체진찰보다는 병력청취에서 더 많은 오류를 범했고, 누락오류보다는 첨가오류를 더 자주 범하였다. 전반적으로 표준화 환자의 채점 정확성은 높은 수준이었고, 발생된 오류는 대부분 학생들에게 호의적인 결과로 나타났다.

위에서 제시한 선행연구들의 연구방법과 결과를 비교분석해 보면 임상교수와 표준화 환자의 채점 정확성을 분석하기 위한 최선의 방법을 선택할 수 있다. 먼저 정확성 판정을 위한 기준을 살펴보면, 먼저 임상교수와 표준화 환자의 채점 결과를 비교하여 일치도를 구하는 방법이 있다. 이때 기준은 임상

교수의 채점 결과이고, 이에 일치하거나 관련성이 높은 정도에 의해 표준화 환자 채점 결과의 정확성을 판정하였다[3,4,5]. 또는 따로 개발한 정답기준과 비교하여 채점 결과의 정확성을 파악하는 방법도 있다[2,8,9,10]. 방법적인 측면에서 볼 때 오류가능성을 안고 있는 임상교수의 결과를 기준으로 하여 일치도나 상관을 구해 표준화 환자의 채점 정확성을 판정하는 방법보다는 정답기준을 작성하여 이를 근거로 채점 정확성을 판단하는 것이 더 명확한 연구방법으로 판단된다. 물론 합당한 정답기준을 개발하기 위해서는 다수의 전문가들이 참여하여, 정해진 절차와 기준에 따라 연구에 포함된 모든 대면에 대한 정답기준을 마련해야 하는 지난한 과정을 필요로 한다.

분석의 수준도 결정해야 한다. 각 문항점수를 분석하는 방법과 사례별로 합산하여 분석하는 방법이 있다. 분석의 수준이 높아질수록 연구자가 원하는 결과를 얻기 쉽다. 즉, 문항점수보다는 사례점수를, 사례점수보다는 여러 사례의 총점을 비교할 때 신뢰도는 높아지고, 동시에 다면적인 정보를 얻을 수도 있다[11]. 그러나 문항수준의 분석은 점수의 합산을 통해 조정되지 않은 실제 채점 결과의 정확성을 알려준다. 그러므로 번거로울 뿐 아니라 이미 시행된 시험에 대해 낮은 신뢰도를 보고하게 될 수도 있지만 문항수준의 분석은 확인해 볼 만한 가치가 있다.

이 연구는 진료수행시험에서 표준화 환자와 임상교수의 채점 정확성을 확인하고자 한다.

많은 선행연구가 임상교수의 채점 결과를 기준으로 표준화 환자의 채점 정확성을 판정하고 있으나 연구팀은 임상교수의 전문성이 채점표와 같이 단순화된 평가에서는 오히려 오류를 높일 가능성이 있다는 아이디어에 주목하였다. 반면, 잘 훈련된 표준화 환자의 명료한 판단기준은 주어진 채점표의 한도 내에서 더 정확한 채점 결과로 나타날 수 있다. 만약 표준화 환자의 채점 정확성이 충분히 높게 나타난다면 이 연구는 국가고시와 같은 대규모 진료수행시험에서 표준화 환자를 채점관으로 선정하는 것에 대한 근거를 제공할 수 있을 것이다. 이 연구는 문항차원의 일치도 분석과 사례점수 간의 분산분석을 통해 정답기준과 임상교수 및 표준화 환자의 진료수행시험의 채점 결과의 정확성을 검증하였다.

대상 및 방법

1. 표본과 연구도구

이 연구는 2008년 7월 K 의학전문대학원에서 시행한 진료수행시험결과를 분석하였다. K 의학전문대학원은 진료수행시험을 위해 4개의 사례를 개발하고, 하루에 두 사례씩 이틀에 걸쳐 시험을 시행하였다. 4개의 사례와 채점표는 각각 체중감소, 복통, 가슴 두근거림, 손저림을 주 증상으로 내분비내과, 산부인과, 순환기내과, 재활의학과 교수들에 의해 개발되었으며, 임상실습 책임교수들로 구성된 임상수기위원회에서 검증하였다. 진료수행시험의 한 사례는 20분 동안 진행된다. 학생들은 처음 1분 동안 시험장 앞에서 준비된 상황소개와 지침을 숙지하고, 입실 후 12분 동안 표준화 환자를 진료한다. 진료가 끝나면 퇴실하여 7분간 사이시험을 보고 다음 시험장으로 이동한다.

진료수행시험의 대상자는 4학년 학생 118명이었으며, 시험결과는 임상수기능력의 일부로 성적에 반영되었다. 연구를 위해 전체 자료 중 일부를 표본으로 추출하였다. 118명의 학생들이 4개 사례의 표준화 환자들을 대면하는 472개의 장면 중 사례별로 15개를 무선표집하여 모두 60개의 대면장면과 이를 평가한 임상교수와 표준화 환자의 채점표 작성결과를 분석하였다.

학생들의 진료수행능력 평가는 임상교수와 표준화 환자에 의해 이원적으로 이루어졌다. 임상교수는 시험이 진행되는 동안 시험장 내에 머물면서 학생들의 진료과정을 직접 관찰하고 평가하였다. 그러나 학생의 진료과정에 개입하거나 표준화 환자에게 임의적인 피드백을 주지 않고 객관적인 관찰자 입장을 유지하였다. 표준화 환자는 학생이 퇴실한 후 학생과의 상호작용을 기억하여 결과를 기록하였다. 임상교수와 표준화 환자는 동일한 채점표를 활용하여 학생을 평가하였으나 평가내용에 대하여 논의하거나 정보를 공유하지 않았다.

채점표는 병력청취, 신체진찰, 임상예절, 환자-의사관계를 평가하는 문항으로 구성되었으며, 사례에 따라 환자교육이나 진단계획 등에 관한 문항이 추가되었다. 각 영역별 문항 수는 사례에 따라 달랐고, 평정척도도 평가영역에 따라 각각 달랐

Table 1. Subcomponents and Items Numbers for 4 Cases

Case	Total	Hx	PE	CC	PPI	Edu	DxP
1	28	12 (43%)	2 (7%)	4 (14%)	7 (25%)	3 (11%)	
2	27	10 (37%)	6 (22%)	4 (15%)	7 (26%)		
3	31	12 (39%)	3 (10%)	4 (13%)	7 (23%)	2 (6%)	3 (10%)
4	26	10 (38%)	5 (19%)	4 (15%)	7 (27%)		

Hx: History taking, PE: Physical examination, CC: Clinical courtesy, PPI: Patient-physician interaction, Edu: Patient education, DxP: Diagnostic plan.

다. 병력청취와 임상예절은 '했음/하지 않았음'의 2단계로, 신체진찰이나 환자교육은 '했음/제대로 하지 않았음/하지 않았음'의 3단계로 표시하도록 하였다. 한편 환자-의사관계는 '매우 동의함'부터 '전혀 동의하지 않음' 사이의 5점 척도 내에서 학생의 수준을 평정하도록 하였다. 각 사례별 채점표의 영역별 문항 수는 Table 1에 제시하였다. 4사례 모두 병력청취의 비중이 높고, 사례 2, 4의 신체진찰 문항이 비교적 많은 편이다.

표준화 환자는 각 사례별로 13명을 모집하여 모두 52명의 표준화 환자를 훈련시켰으며, 훈련은 3차에 걸쳐 총 9시간이 소요되었다. 마지막 훈련세션에서는 사례개발자와 각 과의 인턴들이 참여하여 표준화 환자의 사례구현과 채점의 정확성을 검증하였다. 채점관으로 참여한 45명의 임상교수들은 각 사례별로 1시간 내외의 훈련세션을 통해 사례특징과 채점기준을 숙지하였다.

학생피드백과 연구를 위해 학생과 임상교수, 그리고 표준화 환자의 동의하에 모든 시험장면을 녹화하였다.

2. 정답기준(answer key)

임상교수와 표준화 환자가 기록한 점수의 정확성을 판정하는 정답기준은 두 명(SK, HY)의 표준화 환자 트레이너가 개발하였다. 이들은 학생 60명이 표준화 환자를 대면하는 장면을 보고 각각 60개의 평가기록표를 작성하였다. 두 사람의 채점자는 각기 다른 장소에서 녹화장면을 보면서 독립적으로 평가기록표를 작성하였으며, 평가를 마친 후 평가기록표의 각 문항을 비교하여 일치하지 않은 결과에 대해서는 대면장면을 다시 확인하고 토의를 통해 합일된 의견을 구하였다. 판정이 모호한 몇 개의 문항에 대해서는 제3의 채점자(JK)가 확인하여 최종판정을 내렸다. 이러한 과정을 통해 개발된 정답기준은 각 학생이 표준화 환자 대면과정에서 보인 행동에 대한 가장 적절한 판정을 담고 있다.

Table 2. Agreement Rates and Kappa Coefficients between Trainers 1 & 2

Case	Trainer 1 – Trainer 2		
	Mean±SD	Max	Min
1	0.93±0.12 (0.85±0.17)[a]	1.00 (1.00)[a]	0.47 (0.44)[a]
2	0.91±0.18 (0.78±0.20)	1.00 (1.00)	0.67 (0.35)
3	0.93±0.10 (0.79±0.25)	1.00 (1.00)	0.67 (0.19)
4	0.93±0.08 (0.83±0.15)	1.00 (1.00)	0.73 (0.53)

SD: Standard deviation.
[a]Kappa coefficients values.

정답기준 개발과정에서 두 사람의 트레이너가 작성한 평가표 간의 일치도는 Table 2, 3에 제시하였다. 환자-의사관계영역을 제외한 문항들의 일치율과 Kappa 계수의 평균은 사례에 따라 각각 91~93%와 0.78~0.85로 나타나 채점자 간 신뢰도가 매우 높았다. 그러나 평정척도가 다른 환자-의사관계 영역의 채점자 점수 간 상관은 낮게 나타났다. 채점자 간 일치도와 상관에 관한 설명은 분석방법과 결과에 제시하였다.

3. 문항수준의 분석

채점정확도를 확인하기 위하여 먼저 임상교수와 표준화 환

Table 3. Estimates for Intra-class Correlation Coefficient for PPI Items between Trainers 1 & 2

PPI	Mean±SD		Trainer 1 – Trainer 2	
No	Trainer 1	Trainer 2	F	ICC (95% CI)
1	3.43±0.62	3.43±0.54	0.00	-
2	3.42±0.62	3.67±0.51	14.75[a]	0.31 (0.02 ~ 0.52)[a]
3	3.13±0.74	3.48±0.58	20.58[a]	0.34 (0.04 ~ 0.55)[a]
4	3.46±0.66	3.69±0.55	10.88[a]	0.29 (0.00 ~ 0.50)[a]
5	3.54±0.55	3.73±0.51	9.27[a]	0.26 (-0.04 ~ 0.48)[a]
6	3.52±0.63	3.78±0.54	17.33[a]	0.47 (0.21 ~ 0.64)[a]
7	3.54±0.62	3.80±0.54	15.13[a]	0.34 (0.05 ~ 0.54)[a]

PPI: Patient-physician interaction, ICC: Intra-class correlation coefficient, CI: Confidence interval.
[a]$p<0.05$

Table 4. Proportion of Agreement Rates (and Kappa Coefficients) by Rater Comparison

	Faculty – Key			SP – Key		
Case	Mean±SD	Max	Min	Mean±SD	Max	Min
1	0.83±0.16	1.00	0.40	0.83±0.16	0.93	0.40
	(0.58±0.35)	(1.00)	(-0.07)	(0.59±0.37)	(1.00)	(-0.19)
2	0.73±0.12	0.93	0.27	0.72±0.23	1.00	0.13
	(0.39±0.27)	(0.76)	(-0.09)	(0.52±0.34)	(1.00)	(-0.09)
3	0.78±0.17	1.00	0.40	0.86±0.12	1.00	0.67
	(0.45±0.39)	(1.00)	(-0.13)	(0.54±0.37)	(1.00)	(-0.12)
4	0.79±0.19	1.00	0.40	0.79±0.15	0.93	0.60
	(0.41±0.30)	(1.00)	(0.02)	(0.52±0.31)	(1.00)	(0.17)

SP: Standardized patients, SD: Standard deviation.

자가 작성한 채점표를 정답기준과 비교하여 문항별 일치도를 분석하였다. 문항수준의 일치도는 일치도 통계(agreement statistics)와 Kappa 계수를 구하여 분석하였다. 일치도 통계는 채점자가 피험자의 수행을 어떤 유목이나 범주로 분류할 때 사용하는 방법으로, 두 채점자가 일치하게 채점한 피험자의 비율을 산출한다. Kappa 계수는 일치도 통계가 우연에 의해서 일치하게 평정된 경우를 포함하고 있어 두 채점자 간의 일치도가 과대 추정되는 문제점을 해결하기 위하여 제안된 방법이다. Kappa 계수는 일치도 통계에서 우연에 의하여 두 채점자의 평정결과가 일치하는 확률을 제거한 수치이므로 일치도 통계보다 항상 낮게 추정된다[12,13].

채점표 영역 중 환자-의사관계는 다른 영역과 달리 5점 척도상에서 문항에 동의하는 정도를 평정하도록 되어있어 위와 같은 일치도(agreement) 분석이 적합하지 않다. 그래서 환자-의사관계 문항들은 서열척도로 평가된 자료에서 채점자 간 신뢰도를 측정하는 대표적인 방법인 급내상관계수(intra-class correlation coefficient, ICC)를 구하여 분석하였다[14]. ICC는 자료의 분포가 분산이 아주 적은 경우 결과가 무의미하게 나타나므로 비교하는 두 집단의 분산분석을 선행한 후 해석하였으며, 이차원변량모형(two-way random effects)과 절대동의서(absolute agreement) 방식으로 계산하였다[15]. 환자-의사관계 영역의 7개 문항은 4개 사례에 동일하게

적용되므로 사례구분을 하지 않고 한번에 분석하였다.

4. 사례수준의 분석

각 사례에서 학생들이 획득한 점수를 비교하여 임상교수와 표준화 환자, 그리고 정답기준의 사례별 총점이 어떻게 다른지 분석하기 위하여 사례(4개)와 채점자(3수준: 임상교수, 표준화 환자, 정답기준)를 요인으로 하여 반복측정에 의한 이원분산분석을 시행하였다. 이 때 사례점수는 학생들이 획득한 문항 수의 합을 채점표의 전체 문항 수로 나누어 계산한 정답 퍼센트점수(percentage of correct scores)이다.

결과

1. 문항수준의 일치도 분석

Table 4는 임상교수와 정답기준을 비교한 결과와 표준화 환자와 정답기준을 비교한 결과이다. 분석한 문항은 환자-의사 관계를 제외한 문항들이며, 일치율과 Kappa 계수를 제시하였다. 일치율의 경우 0.85 이상이면 높은 것으로 해석하고[12], Kappa 계수의 경우 0.41~0.60이면 보통 수준(moderate agreement), 0.61~0.80이면 높은 수준(substantial agreement), 그리고 0.81~1.0이면 거의 완벽하게 일치(almost perfect agreement)하는 것으로 해석한다[16]. 분석결과에 따르면, 표준화 환자와 임상교수는 정답기준과 유사한 일치율을 보였다. 그러나 Kappa 계수를 참조하면 두 평가그룹은 정답기준과 일치도가 달랐다. 표준화 환자는 모든 사례에서 정답기준과 보통 수준의 일치도를 보였다. 즉, 표준화 환자는 보통 수준의 채점 정확성에 도달하였다. 반면, 임상교수는 사례 2에서 정답과의 일치도가 크게 떨어졌다. 전반적으로 표준화 환자의 채점정확도가 임상교수보다 높은 편이다.

채점표에서 비교적 비중이 큰 병력청취와 신체진찰 영역의 정확도를 Kappa 계수로 비교한 결과를 Table 5에 제시하였다. 사례 1의 신체진찰은 두 문항 모두 평정점수의 분포가 편중되어 Kappa 계수가 계산되지 않아 제외하고, 3개의 사례를 비교하면 임상교수와 표준화 환자 모두 신체진찰의 정확성이 병력청취보다 높은 편이다. 특히, 사례 2에서 표준화 환자의 신체진찰 채점 정확성은 현저하게 높다. 사례 2는 신체진찰의 비중이 가장 큰 사례이기도 하다(22%). 사례 4는 표준화 환자의 신체진찰 정확성은 높은 반면 임상교수의 신체진찰 정확성은 낮았다. 사례 4 병력청취의 채점 정확성은 두 채점자 모

Table 5. Kappa Coefficients by Rater and Subcomponent Comparison

	Faculty − Key		SP − Key	
Case	Hx	PE	Hx	PE
1	0.68±0.23	-	0.73±0.22	-
2	0.37±0.32	0.46±0.14	0.55±0.38	0.63±0.25
3	0.42±0.36	0.50±0.32	0.52±0.36	0.48±0.32
4	0.34±0.11	0.32±0.44	0.35±0.23	0.57±0.33

SP: Standardized patients, Hx: History taking, PE: Physical examination.

Table 6. Estimates for Intraclass Correlation Coefficient for PPI Items by Rater Comparison

Item	Mean±SD			Faculty − Key		SP − Key	
No	Key	Faculty	SP	F	ICC (95% CI)	F	ICC (95% CI)
1	3.50±0.57	3.33±0.66	3.13±0.77	3.04	-	11.35[a]	0.33 (-0.09 ~ 0.59)[a]
2	3.63±0.52	3.28±0.67	3.12±0.80	12.89[a]	0.30 (-0.14 ~ 0.57)	17.84[a]	0.03 (-0.41 ~ 0.37)
3	3.43±0.59	2.85±0.78	2.62±0.80	27.02[a]	0.27 (-0.20 ~ 0.56)	52.49[a]	0.25 (-0.28 ~ 0.57)[a]
4	3.68±0.50	3.30±0.70	2.98±0.81	15.22[a]	0.31 (-0.13 ~ 0.58)	40.72[a]	0.24 (-0.25 ~ 0.55)
5	3.73±0.45	3.50±0.57	3.43±0.67	7.79[a]	0.30 (-0.13 ~ 0.58)	7.85[a]	-0.10 (-0.75 ~ 0.32)
6	3.72±0.58	3.38±0.67	3.20±0.86	11.80[a]	0.40 (0.00 ~ 0.64)[a]	17.19[a]	0.20 (-0.24 ~ 0.50)
7	3.82±0.47	3.33±0.68	2.97±0.88	25.07[a]	0.24 (-0.22 ~ 0.54)	51.51[a]	0.17 (-0.26 ~ 0.48)

PPI: Patient-physician interaction, SP: Standardized patients, ICC: Intra-class correlation coefficient, CI: Confidence interval.
[a]$p<0.05$.

두 낮았다.

Table 6은 환자-의사관계 문항에서 임상교수, 표준화 환자의 평가결과와 정답기준 간의 ICC를 분석한 결과이다. ICC는 일반적으로 0.75 이상이면 높은 것으로 해석한다[17]. 연구결과를 살펴보면 임상교수가 채점한 첫 번째 문항의 F 통계량이 유의하지 않은 것을 제외하고 모든 문항이 정답기준과 분산에 차이가 있으므로 ICC의 해석이 가능하다. 그러나 임상교수와 표준화 환자 모두 모든 문항의 ICC가 0.75에 근접하지 않았다. 문항수준의 분석에서 환자-의사관계의 채점 정확성은 채점자를 막론하고 매우 낮았다.

2. 사례수준의 일치도 분석

사례와 평가자를 요인으로 사례점수를 반복측정 분산분석한 결과를 Table 7에 제시하였다. 먼저, Mauchly 구형성검정(test of sphericity) 결과를 살펴보면 자료의 구형성을 만족하지 못하므로(chi-square estimate=15.30, df=2, p<0.001) 일변량분석결과를 참조하였다. 일반적인 반복측정 분산분석에서는 개체 간 효과가 연구자의 주요 관심사이나 이 연구는 채점자 간 신뢰도를 검정하고자 하기 때문에 동일한 피험자에 대한 채점자 간 차이를 분석한 개체 내 효과에 주목해야 한다. Table 7에 따르면 채점자 간에는 통계적으로 유의미한 차이가 없으며(F=3.28, p>0.05), 채점자와 사례 사이에 상호작용도 없었다(F=1.76, p>0.05). 즉, 임상교수와 표준화 환자의 사례점수(percent-correct scores)는 정답기준의 사례점수와 차이가 없으므로 두 채점자 그룹 모두 정확하게 채점한 것이다. 또한 이들의 채점 정확성은 사례의 특성에 영향으로 특정 사례에서 높아지거나 낮아지지 않고 안정적이었다. Table 8의 기술통계를 살펴보면 정답기준(63.12±12.77)에 비해 임상교수(59.18±15.63)와 표준화 환자(60.68±17.84)의 사례점수가 더 낮았고, 교수에 비해 표준화 환자의 점수가 다소 높음을 알 수 있다. 즉, 교수와 표준화 환자 모두 엄격하게 채점하는 경향(error of harshness)이 있고, 표준화 환자

Table 7. Analysis of Variance (ANOVA) with Repeated Measures for Percentage-Correct Checklist Scores on the Rater Factor

	SS	df	MS	F	p-value
Rater	473.69	1.61	294.35	3.28	0.05
Rater X case	761.97	4.83	157.83	1.76	0.13
Error (rater)	8083.14	90.12	89.69		

Mauchly's test of sphericity:

Mauchly's w=0.76, chi-square estimate=15.30, df=2, p<0.001

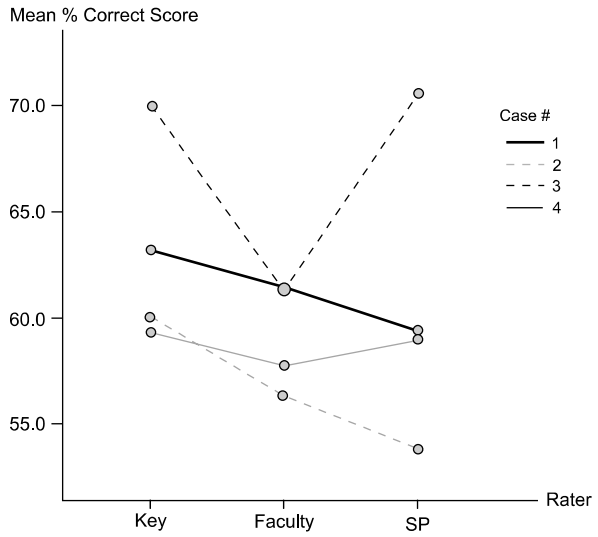
SS: Sum of square, MS: Mean square.

Table 8. Descriptive Statistics for Percentage-Correct Checklist Scores by Three Rater Groups

Case	n	Mean±SD		
		Key	Faculty	SP
1	15	63.17±14.02	61.43±17.84	59.37±16.63
2	15	60.00±13.63	56.33±17.62	53.83±18.27
3	15	70.00±11.65	61.25±15.06	70.56±22.13
4	15	59.30±9.58	57.72±12.36	58.95±9.05
Total	60	63.12±12.77	59.18±15.63	60.68±17.84

SD: Standard deviation, SP: Standardized patients.

Fig. 1. Rater-by-Case Interaction Plot of Mean Percent-Correct Checklist Scores



SP: Standardized patients.

는 교수보다는 더 호의적으로 채점하였다.

단, 위에서 보고한 결과는 환자-의사관계 문항을 제외한 사례점수이다. 환자-의사관계 문항을 포함한 점수로 같은 분석을 해보면 위와는 매우 다른 결과가 나타난다. 채점자 간 차이가 통계적으로 인정되어($F=8.59$, $p=0.001$) 임상교수(60.39±13.02)와 표준화 환자(60.37±14.56)의 채점 결과는 정답기준(65.08±10.18)과 다르므로 두 그룹 모두 채점 정확성을 담보할 수 없다는 결론에 이르게 된다. 그러나 앞서 문항수준의 분석에서 나타난 환자-의사관계 문항과 평정척도의 차별성과 잠재적 문제점을 인정하여 본 연구에서는 환자-의사관계 문항을 제외하고 분석한 결과를 최종결과로 보고하였다.

사례에 따른 임상교수와 표준화 환자, 그리고 정답기준의 점수분포를 Fig. 1에 제시하였다. 앞서 제시한 바와 같이 평균적으로 볼 때에는 임상교수가 표준화 환자보다 더 엄격하게 채점을 했으나, 사례에 따라 경향성이 다소 다르다. 사례 1, 2에서는 임상교수가, 사례 3, 4에서는 표준화 환자가 학생에게 더 호의적으로 채점을 했으며, 채점자가 호의적으로 채점할수록 정답기준에 근접하는 것으로 나타났다.

고찰

이 연구에서도 선행연구와 마찬가지로 사례수준의 분석은 문항수준의 분석보다 신뢰도 판정이 후하게 내려졌다[8]. 평가의 신뢰도를 분석할 때에는 분석자료가 더 많은 정보를 포함할수록, 예를 들어 더 많은 문항을 가진 시험지이거나, 더 많은 피험자를 가진 시험일수록 신뢰도가 높게 평가된다. 같은 원리로 진료수행시험에서 스테이션의 수를 늘리거나 다수의 채점자를 배정하면 시험의 신뢰도가 높아진다[18]. 그래서 이러한 조건을 수행시험의 신뢰도를 높이기 위한 전략으로 채택하기도 한다. 그러나 시험의 신뢰도 및 정확성을 분석하는 연구자로서 여러 수준의 자료 중에서 신뢰도를 높게 추정하는 자료를 분석하여 방어적인 연구결과를 내는 것 보다는 현실을 엄격하게 반영하는 자료를 분석하여 개선과 변화를 위한 계기를 만드는 것이 옳은 태도라 생각한다.

문항수준의 분석결과를 살펴보면 교수 채점자들의 정확성이 표준화 환자들보다 떨어진다(Table 4 Kappa 계수 참조). 특히 사례 2에서 교수의 채점 정확성이 매우 낮다. Table 5에서 사례 2의 영역별 정확성을 살펴보면, 교수의 병력청취 정확성이 크게 떨어진다. 그러나 표준화 환자의 병력청취 정확성은 보통수준을 유지하였다. 이러한 결과는 사례 2 병력청취 채점표의 문항에 대한 교수 채점자들의 해석과 판단기준이 사례개발자의 의도와 달랐던 것이 아닌가 의심해 볼 수 있다. 사례 4의 신체진찰에서도 유사한 패턴이 나타난다. 사례 4 신체진찰은 결막검사, 심장청진, 맥박수 측정, 목 정맥 관찰, 갑상선 촉지 등 일반적인 검사로 이루어져있다. 그런데 임상교수들에게 익숙한 검사일수록 오히려 검사를 '제대로' 한 경우와 그렇지 않은 경우를 판단하는 기준의 일관성이 떨어지기 쉽다. 왜냐하면 임상교수들은 각자의 필요나 습관에 따른 검사방법에 익숙하기 때문이다. 반면 표준화 환자들은 제대로 한 검사를 판단하는 한정적인 기준을 숙지했으므로 모든 채점자가 일관된 기준에 의해 판단하게 된다. 이러한 현상에 대해서 De Champlain et al. [8]은 일반적으로 신체진찰의 채점 정확성이 병력청취보다 높게 나타나나 '수행을 제대로 한 경우'와 '수행을 하긴 했으나 정확히 하지 못한 경우'를 구분해서 평정하도록 되어 있는 채점표를 쓰는 경우 사례에 따라

신체진찰의 채점 정확성이 크게 떨어질 수도 있다고 논의한 바 있다. 이 연구의 사례 4 교수채점 결과에서 이러한 현상이 나타난 것으로 보인다. 사례 4는 병력청취 채점 결과도 주의해서 볼 필요가 있다. 두 채점자 모두 정확성이 매우 낮았다. 이는 채점자의 특성보다는 채점표에서 기인한 현상일 가능성이 높으므로 채점표 문항의 수정을 고려해 보아야 한다. 해당 문항이 판단이 명확하지 않은 내용을 담고 있거나 문장이 모호하게 기술되어 있을 수 있다. 이상의 문항수준의 분석결과는 사례와 영역에 따라 세부적인 특성들이 나타나나 전반적으로는 신체진찰의 채점정확도가 병력청취보다 높아 선행연구와 동일한 결과를 보였다[5,6].

문항수준의 분석결과에 따르면 표준화 환자의 채점 정확성이 임상교수보다 더 높다. 이는 정답기준과 비교해서 임상교수의 채점 정확성이 표준화 환자보다 더 높았다고 보고한 Martin et al.[2]의 연구와 상반된 결과이다. 이 결과에 대해서 표준화 환자의 훈련수준의 차이 때문으로 단순하게 설명할 수도 있겠다. 그러나 선행연구를 이 연구와 면밀히 비교한 결과, 연구팀은 중요한 차이점을 발견할 수 있었다. Martin et al.의 연구에서는 정답기준이 임상교수들에 의해 개발된 반면, 이 연구에서는 표준화 환자 트레이너들에 의해 개발되었다. 임상교수가 만든 기준에 임상교수의 채점 결과는 더 잘 들어맞았고, 표준화 환자 트레이너가 만든 기준에는 표준화 환자의 채점 결과가 더 일치하였다. 결국 두 연구에서 만든 정답기준 모두 완벽한 정답으로 인정하기 어려워진다. 정답기준의 문제는 수행평가의 주관적인 본질이 반영된 것으로 이 연구의 한계로 지적할 수 있으며, 채점자 정확성 판정을 위한 정답기준 개발에 다양한 그룹에 속한 다수의 전문가가 포함되어야 함을 후속연구자들에게 제언한다.

환자-의사관계 문항은 채점자에 상관없이 매우 낮은 채점 정확도를 보였다. 이는 Kwon et al. [6]이 환자-의사관계 문항에서 채점자 간 낮은 일치도를 보고한 것과 동일한 결과이다. 환자-의사관계 문항의 낮은 신뢰도는 선행연구뿐 아니라 이 연구 내에서 정답기준을 개발한 두 트레이너 간 일치도 분석에서도 재확인할 수 있다. 다른 영역의 문항에서는 매우 높은 수준의 일치도를 보인 트레이너들은 환자-의사관계 문항에서는 매우 낮은 상관을 보였다(Table 3). 이 영역의 문항들이 5점 평정척도로 이루어진 것을 감안하여 다른 영역과는 달리 ICC 분석을 했음에도 불구하고 상관은 보통수준에도 이르지 못하였다. 이는 환자-의사관계영역의 주관적인 성향에 기인하는 결과이기도 하나, 한편으로는 채점표의 문항이 모호하게 진술되어 있거나 한 문항이 하나 이상의 수행을 진술하고 있는 것이 아닌지 의심하게 한다[19].

한편, 임상교수와 표준화 환자의 환자-의사관계 채점 정확성을 비교해 보면 다른 영역과는 상반된 경향성이 나타난다. 즉, 임상교수와 정답기준의 일치도가 표준화 환자와 정답기준의 일치도보다 더 높은 것이다. 여기에서 연구팀은 또 하나의 중요한 채점자 특성을 발견하였다. 임상교수들과 정답기준을 개발한 트레이너들은 학생들의 진료장면을 직접, 혹은 화면을 통해 관찰한 반면, 표준화 환자들은 진료의 대상자로 학생과의 상호작용에 직접 관여하였다. 즉, 이들은 관찰자와 참여자로서 시험장면을 보는 시점이 달랐다. 이러한 결과는 진료수행시험의 환자-의사관계에서 채점자의 시점에 따라 채점 결과가 얼마나 달라질 수 있는지 보여주는 결과이다. 임상교수 채점자와 표준화 환자 채점자 간의 차이를 고려해야 한다면 의학지식과 경험의 차이뿐 아니라 시점에 따라 시험장면에서 얻는 정보의 특성이 달라진다는 사실도 반드시 기억해야 한다. 이 연구결과에 따르면 표준화 환자의 환자-의사관계 영역에 대한 채점 정확성이 임상교수보다 훨씬 더 낮다. 그러나 연구팀은 상호작용의 주체였던 이들의 채점 결과가 더 부정확하였다고 결론을 내리지 못하였다.

사례점수를 가지고 정답기준, 임상교수 채점, 표준화 환자 채점을 개체 내 요인으로 두고, 4개 사례를 개체 간 요인으로 두어 반복측정 분산분석을 한 결과, 정답기준, 임상교수 채점, 표준화 환자 채점 간에는 점수의 차이가 없었다. 즉, 오류로 점수를 더하거나 뺀 문항을 모두 합산한 결과(net score)에 따르면 임상교수와 표준화 환자가 채점한 점수는 정답기준과 점수가 같았다. 이는 표준화 환자 채점과 임상교수 채점의 일치도에 대해서 긍정적인 결론을 내린 많은 선행연구를 지지하는 결과이다[4,5,6,7,8,9].

한편, 개체 간 요인인 사례에 따른 차이에 대해서는 이 연구의 범위를 벗어나므로 결과를 제시하지 않았다. 단지 사례에 따라 채점자 간 정확성이 달라지지 않았으므로 사례의 영향을 받지 않고 안정적으로 채점하였다는 결과만을 제시하였다. 사실 사례점수는 채점의 정확성보다는 사례의 난이도를 논의

하기에 더 적합한 수치이다[8]. Fig. 1의 그래프에서 4개 사례의 정답기준 사례점수를 비교해 보면, 사례 1, 3이 상대적으로 쉬운 사례이고, 사례 2, 4가 어려운 사례이다. 그러나 사례의 난이도와는 상관없이 사례 1, 2는 임상교수가 더 정답기준과 근사한 결과를 냈고, 사례 3, 4는 표준화 환자가 정답과 더 비슷한 점수를 산출하였다. 이 연구에서 사례의 난이도는 채점의 정확성에 영향을 주는 특성이 아니었다. 그러나 이 연구의 분석방법으로는 채점에 영향을 주는 또 다른 사례특성을 밝혀내지 못하였다. 또한 표준화 환자들이 누락오류보다는 첨가오류를 더 많이 범하므로 정답보다 높은 점수를 내는 경향이 있다는 선행연구와도 다른 결과가 나타났다[9]. 사례별로 나누어 분석했을 때 표준화 환자와 임상교수의 점수에서 특정한 패턴을 찾기 어려웠다.

위의 논의를 종합하면 표준화 환자와 임상교수의 채점 정확성이 확보되었다는 결론을 내릴 수 있다. 이러한 결론을 통해 대규모 진료수행시험에서 표준화 환자가 임상교수를 대신해 학생들의 진료수행능력을 신뢰성 있게 평가할 수 있으리라는 실제적 판단을 내릴 수 있다. 철저한 훈련이 주어진다면 표준화 환자들은 사례개발자가 평가하고자 의도한 행동이나 지식을 일관성 있게 평가할 수 있는, 신뢰할 만한 채점관의 역할을 수행할 수 있을 것이다. 그러나 이러한 주장은 충분한 훈련과 명료한 채점표가 제공된다는 전제 하에 설득력이 있음을 기억해야 한다. 채점자의 오류를 줄이기 위한 채점표와 훈련의 조건에 대해서 Williams et al. [19]은 채점표를 간결하게 구성하고, 평정단계도 가능한 줄여야 하며, 채점자가 사례와 채점문항을 이해하는 것뿐 아니라 수행평가가 가지는 한계점까지도 이해하도록 충분히 훈련할 것을 제안하였다. 또한 채점자에게 시험이 진행되는 과정에서 적절한 피드백을 줄 것도 함께 권고하였다. 실제로 Wallace et al. [10]의 연구에 의하면 표준화 환자에게 무작위(random) 피드백을 주었을 경우에 피드백을 주지 않거나 계획된 피드백을 주었을 때보다 높은 채점 정확성을 보였다.

한편, 표준화 환자 채점자의 신뢰도를 높이는 것과 관련해서 주의해야 할 점도 있다. 표준화 환자 채점의 신뢰도를 높이기 위해 채점표를 단순화시켜, 학생들의 행동 여부에 대한 피상적인 판단을 묻는 문항으로만 구성한다면 학생들의 수행에 대한 판단이 전체 맥락으로부터 분리될 뿐 아니라 통합적인 진료능력을 평가할 수 없게 된다. 그러므로 진료수행시험의 성공을 위해서는 평가의 기본원칙으로 돌아가 신뢰도와 타당도라는 양날의 칼을 잘 조절하는 균형 있는 판단력이 필요하다.

이 연구의 제한점으로는 표본 수가 적은 것을 지적할 수 있다. 각 표본에 따라 각기 다른 정답기준을 마련해야 하는 연구방법의 특성상 많은 표본을 연구에 포함시키기 어려웠다. 또한 한 학교의 시험만을 분석한 점 역시 연구 결과의 일반화에 제한이 될 수 있다. 결국 표집의 한계가 연구의 가장 큰 제한점이라 하겠다. 두 번째는 채점 정확성 분석의 중요한 도구인 정답기준에 대한 타당성과 신뢰성에 대한 한계도 지적할 수 있다. 앞서 논의한 바와 같이 이 연구에서 개발한 정답기준을 완벽한 정답으로 간주하기는 어렵다. 그러나 문항수준과 사례점수의 수준으로 구분하여 채점 정확성을 분석함으로써 엄격한 분석수준과 실제적인 적용수준이라는 두 차원에서 채점자의 신뢰도를 논의한 점과 정답기준을 개발하여 채점의 정확성의 근거를 제시한 점, 그리고 의사국가고시에서 진료수행시험을 처음으로 적용하는 시점에서 채점자 선정에 대한 근거를 제공하였다는 점에서 이 연구의 의의를 찾을 수 있다.

REFERENCES

1. Ko J, Yoon TY, Park J. Inter-rater reliability in a clinical performance examination using multiple standardized patients for the same case. Korean J Med Educ 2008; 20: 61-72.
2. Martin JA, Reznick RK, Rothman A, Tamblyn RM, Regehr G. Who should rate candidates in an objective structured clinical examination? Acad Med 1996; 71: 170-175.
3. McLaughlin K, Gregor L, Jones A, Coderre S. Can standardized patients replace physicians as OSCE examiners? BMC Med Educ 2006; 6: 12.
4. Kopp KC, Johnson JA. Checklist agreement between standardized patients and faculty. J Dent Educ 1995; 59: 824-829.

5. MacRae HM, Vu NV, Graham B, Word-Sims M, Colliver JA, Robbs RS. Comparing checklists and databases with physicians' ratings as measures of students' history and physical-examination skills. Acad Med 1995; 70: 313-317.
6. Kwon I, Kim N, Lee SN, Eo E, Park H, Lee DH, et al. Comparison of the evaluation results of faculty with those of standardized patients in a clinical performance examination experience. Korean J Med Educ 2005; 17: 173-183.
7. Vu NV, Marcy MM, Colliver JA, Verhulst SJ, Travis TA, Barrows HS. Standardized (simulated) patients' accuracy in recording clinical performance check-list items. Med Educ 1992; 26: 99-104.
8. De Champlain AF, Margolis MJ, King A, Klass DJ. Standardized patients' accuracy in recording examinees' behaviors using checklists. Acad Med 1997; 72: S85-S87.
9. Heine N, Garman K, Wallace P, Bartos R, Richards A. An analysis of standardised patient checklist errors and their effect on student scores. Med Educ 2003; 37: 99-104.
10. Wallace P, Heine N, Garman K, Bartos R, Richards A. Effect of varying amounts of feedback on standardized patient checklist accuracy in clinical practice examinations. Teach Learn Med 1999; 11: 148-152.
11. Colliver JA, Robbs RS, Vu NV. Effects of using two or more standardized patients to simulate the same case on case means and case failure rates. Acad Med 1991; 66: 616-8.
12. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20: 37-46.
13. Kim KH, Song MY. Comparison of inter-rater consistence and accuracy of estimation for ability parameters among multiple scoring scales. J Educ Eval 2001; 14: 327-347.
14. Janjua NZ, Khan MI, Clemens JD. Estimates of intraclass correlation coefficient and design effect for surveys and cluster randomized trials on injection use in Pakistan and developing countries. Trop Med Int Health 2006; 11: 1832-1840.
15. Cha JS, Kim YB. An analytical review of interrater reliability & agreement. J Bus Res 1994; 23: 75-102.
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174.
17. Fleiss JL. Design and analysis of clinical experiments. New York, USA: Wiley; 1986.
18. Vu NV, Barrows HS. Use of standardized patients in clinical assessments: recent developments and measurement findings. Educational Researcher 1994; 23: 23-30.
19. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med 2003; 15: 270-292.