진료수행시험에서 표준화 환자 채점의 정확도

서울대학교 의과대학 의학교육실, 한양대학교 의과대학 의학교육학교실[1]

신좌섭 · 이성아 · 박훈기[1]

= Abstract =

Standardized Patients' Accuracy in Recording Checklist Items during Clinical Performance Examinations

Jwa-Seop Shin, MD, EdD, **Seong-A Lee,** MSc, **HoonKi Park[1],** MD, PhD

Office of Medical Education, Seoul National University College of Medicine, Seoul, Korea
Department of Medical Education, Hanyang University College of Medicine[1], Seoul, Korea

Purpose: Standardized patients participate in clinical performance examinations not only to simulate case scenarios but also to evaluate the performance of students using a checklist. The accuracy in checking off checklist items is one of the most important factors determining the reliability of this examination. The purposes of this study were to determine the SP's overall accuracy in recording checklist items, and whether their accuracy was affected by certain characteristics of checklist items.

Methods: Three professors, who have been fully involved in scenario development and SP training, reviewed videotapes of the examination and evaluated the performance of the students using the same checklist. SP's checklists were marked on this 'correct checklist'. The checklists and checklist guidelines of the items marked under the score of 50 out of 100 were analyzed.

Results: Results showed that the accuracy of the SP's in recording checklist items was 86.9% and was affected by certain characteristics, such as complexity or ambiguity of checklists and checklist guidelines.

Conclusion: In this study, the SP's accuracy in recording checklist items was good to very good, and the result suggested that the accuracy could be improved by the elaboration of checklists and checklist guidelines.

Key Words: Clinical performance, Observer variation, Educational measurement

교신저자: 신좌섭, 서울대학교 의과대학 의학교육실, 서울시 종로구 연건동 28번지
Tel: 02)740-8175, Fax: 02)740-8072, E-mail: hismed1@snu.ac.kr

서 론

표준화 환자(Standardized patient: 이하 SP)를 활용한 진료수행시험(Clinical performance examination: 이하 CPX)은 의대생이나 의사의 임상적 수행능력을 평가할 목적으로 널리 시행되고 있는 방법이다(Williams 등, 1987; Newble과 Swanson, 1988; Rothman 등, 1990). CPX에서 수험생은 의학적 문제를 갖고 있는 SP와 환자-의사 관계를 형성하고 병력청취와 신체검진을 통해 정보를 수집하며 예방-치료-재활에 관한 정보를 교환한다. SP는 임상의사가 작성한 대본에 근거하여 행동하고, 자신의 행동과 상호작용으로 일어나는 학생의 수행을 관찰하여 채점표를 작성한다. 채점표는 특정한 병력청취나 신체검진을 했는지, 예방-치료-재활과 관련하여 특정한 정보제공을 했는지, 환자-의사관계에 필요한 특정한 행동을 하거나 특정한 느낌을 주었는지를 체크하도록 되어 있다. 임상적 추론은 채점표에서 명시적으로 체크되지는 않지만, 채점표에 등장하는 특정한 수행 묶음 자체가 임상적 추론 능력을 체크하게 된다.

SP가 채점표 기입을 얼마나 정확하게 하는가는 CPX의 신뢰도를 결정하는 가장 기본적인 요건이다. 물론 CPX에서 SP의 역할은 측정도구만이 아니라 학생의 수행을 '연기'를 통해 상호작용적으로 촉발하는 촉발도구이기도 하므로 '연기'의 표준화도 마찬가지로 중요한 요건이다. SP가 수행한 채점표 기입의 정확도는 동일한 수험생의 수행을 별도의 채점자가 채점하여 '정답표'를 만들고 이것과의 일치도를 보는 방법, SP 자신의 검사-재검사 신뢰도를 보는 방법 등이 있을 수 있다. 환자-의사관계를 제외한 병력청취와 신체검진, 정보교환의 3개 영역에 있어서 SP의 채점과 다른 관찰자(직접관찰 혹은 비디오 관찰)의 채점표 기입 일치도를 본 결과는 80~100%로 보고되었다(Norman 등, 1985; Rethans와 Van Boven, 1987; Williams 등, 1987; Tamblyn, 1991; Vu 등, 1992). 검사-재검사 상황에서도 검사-재검사 일치도는 82~85%로 높게 나타났다(Rethans와 van Boven, 1987; Tamblyn, 1991). Vu 등(1992)은 SP 채점의 전반적 정확도를 검증하고 채점표의 여러 특성이 정확도에 영향을 미친다는 것을 밝혔다. 이 연구에서는 사전에 훈련된 2명의 관찰자가 CPX 장면을 녹화한 동일한 비디오를 각각 채점한 후, 불일치하는 항목에 대해서 2명의 관찰자가 합의하도록 하여 이른바 '정답표'를 만들고 이것과 SP 채점표의 일치도를 구하였다. 국내의 경우 박훈기 등(2003)은 객관구조화진료시험에서 SP 채점과 사전에 훈련된 교수 채점의 일치도를 외국보다 다소 낮은 71~82%로 보고하였으며, 김주자 등(2004)은 CPX에서 SP 채점과 교수 채점의 일치도가 신체검진, 정보교환에서는 비교적 높았으나 병력청취에서는 낮게 나타났다고 보고하고 있다. 김주자 등(2004)의 경우 교수 채점자는 채점에 관한 별도의 훈련을 받지 않은 것으로 보인다. SP의 채점과 훈련된 교수의 채점을 비교하여 채점의 정확도를 측정하는 연구의 경우 '정답'으로 간주되는 교수 채점자의 훈련 정도, 교수 채점자간의 표준화가 결정적으로 중요한 문제가 된다.

본 연구는 CPX에서 SP가 수행한 채점이 충분히 훈련된 교수평가자의 채점과 비교하여 얼마나 정확한지를 밝히고, 정확도가 낮은 항목의 채점지침에 어떤 문제가 없는지를 살펴보고자 하였다.

대상 및 방법

2004년 8월 16일부터 5일 동안 한양대학교 4학년 학생 124명을 대상으로 실시한 CPX에서 SP가 작성한 채점표와 녹화 비디오를 비교분석 대상으로 하였다. 수험생의 수행능력이 SP의 채점에 미치는 영향을 배제하기 위해 성적 분포별로 상위 1인, 중위 2인, 하위 1인 총 4인의 녹화 비디오를 표집하였다. 학생별로 총 8개 시험방을 돌았고 각 시험방은 12분이었으므로 녹화 비디오의 분량은 학생당 96분, 총 384분이었다. 한양대학교에서는 SP 한명이 하루 12명 이하의 학생을 대면했으므로 SP의 피로도는 영향을 미치지 않을 것이라고 가정하였다.

비디오 채점은 교수 2인과 SP 트레이너 1인이 담당했다. 이들은 '서울 CPX 컨소시엄'의 센터를 담

Table I. Correlation of Students' Performance Scores by Professors (Pearson R)

Case	N	Between professors A and B	Between professors B and C	Between professor A and C
IDA	4	.973*	.976*	1.000†
HBV	4	.817	.952	.866
Cough	4	.995†	.948	.948
DM	4	.990*	.953*	.986*
Bad news	4	.997†	.944	.961
Depression	4	.892	.921	.956*
SAH	4	1.000†	.966*	.964*
IBS	4	.927	.894	.997†
Total	32	.929†	.914†	.935†

* p<.05, † p<.01

IDA: iron deficiency anemia, HBV: hepatitis B virus infection, Cough: chronic cough, DM: diabetes mellitus, Bad news: bad news delivery, Depression: depression, SAH: sub-arachnoid hemorrhage, IBS: irritable bowel syndrome

당하는 인력으로서 8개 시험방의 시나리오와 채점표, 채점표 작성지침에 직접 관여하여 그 내용을 충분히 숙지하고 있고 SP의 훈련도 직접 담당하였으므로 SP의 정확도를 평가하기 위한 '정답표'를 만들기에는 가장 적합할 것이라고 가정하였다. 3인이 동일한 비디오를 함께 보면서 학생의 수행이 발생할 때마다 그때그때 채점표에 기입했으며, Vu 등 (1992)의 경우와 달리 평가자 간의 일치도를 보기 위해 서로 의견을 교환하지 않았고 일치하지 않는 부분을 다시 검토하여 수정하는 일도 하지 않았다.

교수채점자간의 일치도를 보기 위해 시험장별로 채점표의 각 문항을 1점 만점으로 환산하여 평정자간 신뢰도 (Pearson R)를 구하였다. SP 채점의 정확도를 평가할 기준을 만들기 위해 각 학생, 각 시험장, 각 문항별로 교수채점자가 매긴 점수의 평균을 구하여 32건의 '정답표'를 만들었다. 교수채점자와 SP채점자 중 어느 한쪽이 점수를 후하게 주거나 박하게 주는 경향이 없는지를 보기 위해 교수채점자가 준 점수의 평균과 SP가 준 점수를 시험장별로 비교해 보았다. 또한 이 정답표의 점수와 SP가 매긴 점수간의 상관을 각 시험장별로 비교하여 시험장별로 채점의 정확도가 다른지를 알아보고자 하였다. 다음으로는 각 학생, 각 시험장, 각 문항별로 만들어진 '정답표'를 기준으로 SP의 채점표를 '산술적으로 채점'하여 SP 채점의 정확도를 산술 점수로 구하였다. '2, 3 Scale (명명척도)'에서는 완전히 일치할 경우 맞는 것으로 간주하여 1점을 주고, '6 Scale (서열척도)'에서는 2점 이하의 차이는 맞는 것으로 간주하여 1점을 주었다. 이 같은 방식으로 4명의 학생, 8개 시험장, 총 32건의 SP 수행을 채점하였다. 또 각 시험장별 4건의 SP 수행점수를 문항별로 합산하여 (만점 4점) 2점 이하인 문항 (정답률 50점 이하)을 가려내고 이 문항들의 채점표나 채점기준표, 시나리오에 정확도를 낮출만한 어떤 문제가 없는지를 살펴보았다. 또한 병력청취, 신체검진 등 측정영역별로도 채점의 정확도를 산출하여 영역별로 채점의 정확도에 차이가 있는지를 보았다. 통계에는 SPSS 10.0을 사용하였다.

결 과

채점표의 각 문항을 1점 만점으로 환산하여 계산한 교수채점자간의 신뢰도는 8개 시험장에서 모두 상관계수가 0.8 이상이었으며 대부분의 시험장에서

Table II. Difference and Correlation of the Scores Marked by SPs and the Mean of Scores Given by Professors

Case	Difference (professor-SP)	Standard error	t	df	p	R	p
IDA	-.400	.569	-.703	3	.533	.954	.046
HBV	-.850	.273	-3.117	3	.053	.997	.003
Cough	-4.058	1.351	-3.004	3	.057	.917	.083
DM	-.408	.312	-1.309	3	.282	.994	.006
Bad news	-.525	1.100	-.477	3	.666	.867	.133
Depression	-1.983	.931	-2.129	3	.123	.931	.069
SAH	-.225	.838	-.268	3	.806	.917	.083
IBS	-.992	.528	-1.879	3	.157	.951	.049

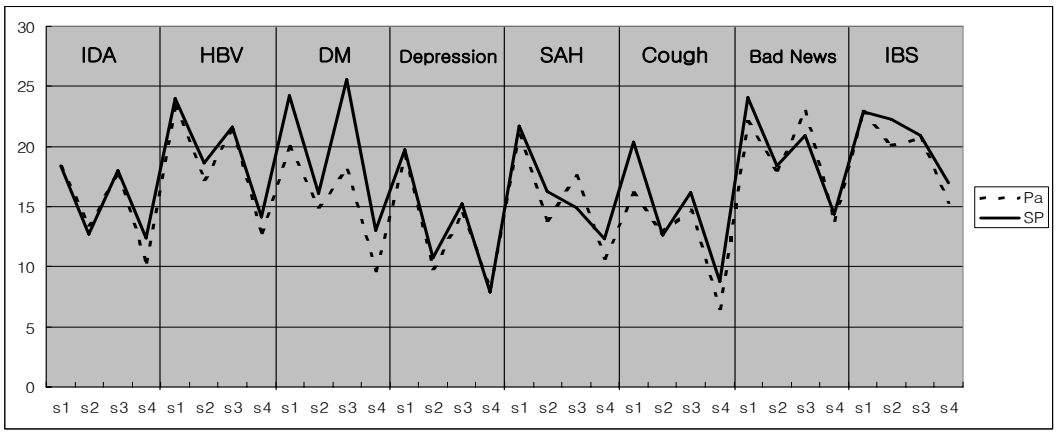

Fig. 76. Correlation of the Scores Marked by SPs and the Mean Scores Given by Professors.
* Pa: the average of scores measured by professors, SP: the score by SP
s1-s4: student number

0.9 이상으로 높게 나타났다 (Table I).

SP 채점의 정확도를 평가할 기준을 만들기 위해 각 학생, 각 시험장, 각 문항별로 교수채점자가 매긴 점수의 평균을 구하여 32건의 '정답표'를 만들었다. 이 정답표의 점수와 SP 채점표의 점수를 시험장별로 비교해본 결과 SP가 교수에 비해 다소 점수를 후하게 주는 경향이 있었으나 (Table II) 그 차이가 통계적으로 유의하지는 않았다. 정답표의 점수와 SP가 준 점수의 상관관계는 시험장별로 차이가 있으나 전반적으로 일치하는 경향을 보여주고 있으며 (Fig. 1), 상관계수도 0.85 이상으로 높게 나타났다 (Table II). 95% 신뢰수준에서 통계적으로 상관이 유의하지 않은 경우는 'Cough, Bad News, Depression, SAH'의 4개 시험장으로 나타났다.

32건의 정답표를 기준으로 SP의 채점표를 '산술적으로 채점'하여 SP 채점의 정확도를 산술 점수로 구한 결과 (Table III) 전반적으로 86.9점의 정확도 점수를 보였고, 평균 74점을 보인 'Bad news' 시험장을 제외하고 83점 이상의 정확도 점수를 보였다. 사례별로 보면 'Cough, Bad News, Depression, SAH'에서 낮은 점수를 보여 정답표 점수와 SP가 준 점수의 상관관계와 일치하는 결과를 보여주었다

Table III. Number of Items Correctly Checked per Case Number (%)

Case	Items	Student 1	Student 2	Student 3	Student 4	Average
IDA	34	31 (91.2)	30 (88.2)	34 (100.0)	31 (91.2)	31.5 (92.6)
HBV	31	27 (87.1)	28 (90.3)	25 (80.6)	30 (96.8)	27.5 (88.7)
Cough	32	27 (84.4)	28 (87.5)	28 (87.5)	24 (75.0)	26.8 (83.8)
DM	27	26 (96.3)	26 (96.3)	20 (74.1)	23 (85.2)	23.8 (88.1)
Bad news	23	16 (69.6)	20 (87.0)	17 (73.9)	15 (65.2)	17.0 (73.9)
Depression	24	22 (91.7)	20 (83.3)	21 (87.5)	19 (79.2)	20.5 (85.4)
SAH	28	25 (89.3)	24 (85.7)	25 (89.3)	21 (75.0)	23.8 (85.0)
IBS	33	31 (93.9)	28 (84.8)	32 (97.0)	31 (93.9)	30.5 (92.4)
Average	29	25.6 (88.3)	25.5 (87.9)	25.3 (87.2)	4.3 (83.8)	25.2 (86.9)

Table IV. Number of Items Correctly Checked per Checklist Domain

Domain	Number of items	Percentage
Global	8	100%
Clinical courtesy	28	87.5%
History taking	76	85.9%
Physical examination	31	80.7%
Information sharing	33	68.2%
Patient-physician interaction	56	86.2%
Total	176	84.8%

(Table II). 각 학생별 점수 평균도 83점 이상의 정확도를 보였다. 학생수준별로는 가장 높은 점수를 얻은 학생 1에서 SP도 높은 정확도를 보였고, 가장 낮은 점수를 얻은 학생 4에서 SP도 가장 낮은 정확도를 보였다. 병력청취, 신체검진 등 측정영역별 정확도를 구한 결과 환자교육, 치료지침 전달 등을 포함하고 있는 정보공유 영역에서 정확도가 68.2%로 나타났고 나머지 영역에서는 80% 이상의 정확도를 보였다 (Table IV).

각 스테이션별 4건의 SP 수행점수를 문항별로 합산하여 (만점 4점) 2점 이하인 문항 (정답률 50점 이하)을 가려낸 결과 'IDA 1문항, HBV 2문항, DM 3문항, Depression 5문항, SAH 4문항, Cough 6문항, Bad News 7문항, IBS 1문항'의 총 29문항으로 나타났다. Bad News의 7문항은 모두 환자교육 (정보공유)에 해당하는 문항들이었다. 이 문항들의 채점표나 채점기준표를 살펴본 결과 (1) '앞가슴 심장 부위를 네 곳에서 정확히 청진하였나?'는 질문에 '제대로 했음 (네 곳 모두 정확히), 제대로 못했음 (두 곳이나 세 곳만 제대로), 하지 않았음 (청진하지 않았거나 정확히 청진한 곳이 한 곳 뿐)'으로 체크하도록 되어있는 식으로 채점의 체계가 복잡한 문항, (2) '가족에 대한 B형 전염성을 설명해 주는지: "항원 (e항원)이 아직도 있기 때문에 가족에게 전염시킬 가능성", "가족 중에 B형 간염에 대한 면역 항체가 없는 분이 있다면 예방 접종을 하는 게 원칙": 두 가지를 모두 말해야 점수를 받을 수 있다. 전염력이 있다고만 얘기하고 가족의 예방 대책을 설명하지 않거나, 가족에게 무조건 전염 안 된다고 설명한 경우는 0점' 식으로 복수 질문을 담고 있는 문항, (3) '다른 데 아픈데 있는지, 병원에 다녀온 적 있는지 등…' 식으로 판단의 준거를 여럿 나열하여 판단의

준거가 복잡한 문항 등이었다.

고 찰

이 연구에서는 SP 채점의 정확도를 평가하는 데 있어서 교수채점자와 SP 채점자의 일치도를 구하는 방식의 한계를 극복하기 위하여 충분히 훈련된 교수채점자의 채점을 '정답'으로 간주하고 SP 채점의 정확도를 검증하여 보았다. SP 채점과 교수 채점의 일치도를 보는 연구는 일치도가 낮게 나올 경우 과연 그것이 교수, SP 중 어느 쪽의 훈련부족으로 일어난 현상인지 구분하기 어렵기 때문이다(박훈기 등, 2003). 이번 연구의 경우 3인의 전문적 교수 평가자가 채점을 하고 그중 일치하는 것을 정답으로 간주하였으므로 이 같은 문제점은 어느 정도 극복되었다고 볼 수 있겠다. 교수 채점자간에 서로 의논을 하지 않았음에도 채점자간 상관계수가 대부분의 시험장에서 0.9 이상으로 높게 나타난 것도 '정답'의 타당성을 반증한다.

교수 채점자의 평균을 채점의 '정답'으로 인정하다면 본 연구에서 SP의 채점은 100점 만점을 기준으로 86.9점을 얻은 셈이 되어 높은 정확도를 보이고 있으며 SP 채점과 다른 평가자 채점의 일치도를 80~100%로 보고한 기존의 연구결과들(Norman 등, 1985; Rethans와 Van Boven, 1987; Williams 등, 1987; Tamblyn, 1991; Vu 등, 1992)과 일치한다.

또한 SP가 교수에 비해 다소 후하게 점수를 주는 경향이 있는 것으로 나타났는데, 이는 표준화 환자가 교수와 같은 채점기준표를 사용하더라도 점수를 후하게 주는 경향이 있다는 기존의 연구 결과와 일치한다(박훈기 등, 2003; Westberg와 Jason, 1993). SP가 다소 관대한 평가를 한다는 점을 제외하고는 충분한 훈련을 거친다면 구태여 교수가 평가할 필요가 없다는 것을 시사하는 결과로 해석된다.

SP 채점의 정확도가 'Cough, Bad News, Depression, SAH'에서 낮게 나타난 것은 각 시험장별 이들 4개 시험장에서 문제 문항이 많이 발견된 것과 일치하며, 환자교육, 치료지침 전달 등을 포함하고 있는 정보공유 영역에서 SP 채점의 정확도가 낮게 나타난 것은 문제 문항으로 분류된 Bad News의 7문항이 정보공유 영역의 문항이었던 것을 반영한다. 이번 연구에서 정확도의 저하는 이 시험장과 측정영역의 채점표와 채점기준표가 정확도를 낮출 수 있는 문항들을 다수 포함하고 있는데서 비롯된 것으로 보아도 좋을 것이다.

채점표와 채점기준표의 문제들은 채점의 체계가 복잡하거나, 복수질문을 담고 있거나, 판단의 준거가 복잡한 경우들이었다. 본 연구에서 SP 채점의 정확도는 86.9점으로 나왔지만, 이 같은 문항들을 보다 명료하고 판단하기 쉽게 수정한다면 채점의 정확도는 더 향상될 수 있음을 알 수 있다.

참 고 문 헌

김주자, 이경재, 최규연, 이동환(2004). 일개 의과대학에서 실시한 표준화 환자를 이용한 임상수행능력평가시험 결과 분석. *한국의학교육*, 16(1), 51-61.

박훈기, 이정권, 황환식, 이재웅, 최윤영, 김혁, 안동현(2003). 객관구조화진료시험에서 교수와 표준화 환자 사이의 점검표 채점의 일치도. *한국의학교육*, 15(2), 141-150

Newble DI, & Swanson DB(1988). Psychometric characteristics of the objective structured clinical examination. *Medical Education*, 22, 335-41

Norman GR, Neufeld VR, Walsh A, Woodward CA, & McConvey GA(1985). Measuring physician performance by using simulated patients. *Journal of Medical Education*, 60, 925-34.

Rethans JJE, & van Boven CPA(1987). Simulated patients in general practice: a different look at the consultation. *British Medical Journal*, 294, 809-12

Rothman AI, Cohen R, & Ross J(1990). Evaluating the clinical skills of foreign medical school graduates participating in an internship preparation program. *Academic Medicine*, 65, 391-4

Tamblyn R(1991). Sources of unreliability and bias in standardized patient rating. *Teaching and Lear-*

ning in Medicine, 3, 74-85

Vu NV, Marcy MM, Colliver JA, Verhulst SJ, Travis TA, & Barrows HS(1992). Standardized (simulated) patients' accuracy in recording clinical performance check-list items. *Medical Education*, 26, 99-104

Westberg J, Jason H(1993). *Collaborative clinical education.* New York: Springer Publishing Company.

Williams RG, Barrows HS, Vu NV, Verhulst SJ, Colliver JA, Marcy MM, & Steward D(1987). Direct, standardized assessment of clinical competence. *Medical Education*, 21, 482-9