객관구조화진료시험(OSCE)에서 교수와 표준화환자 사이의 점검표 채점의 일치도

한양대학교 의과대학 가정의학과, 내과[2], 핵의학과[3], 흉부외과[4], 신경정신과[5]
성균관대학교 의과대학 가정의학과[1]

박훈기 · 이정권[1] · 황환식 · 이재웅[2] · 최윤영[3] · 김 혁[4] · 안동현[5]

= Abstract =

The Agreement of Checklist Recordings Between Faculties and Standardized Patients in an Objective Structured Clinical Examination (OSCE)

Hoonki Park, MD, PhD, **Jungkwon Lee[1],** MD, PhD, **Hwansik Hwang,** MD, PhD, **Jaeung Lee[2],** MD, PhD, **Yunyoung Choi[3],** MD, PhD, **Hyuck Kim[4],** MD, PhD, **Dong-Hyun Ahn[5]**, MD, PhD

Department of Family Medicine, Internal Medicine[2], Nuclear Medicine[3], Thoracic Surgery[4], Neuropsychiatry[5], Hanyang University College of Medicine, Department of Family Medicine[1], Sungkyunkwan University School of Medicine

Purpose: A high degree of agreement between standardized patients(SP) check-list recordings and those of faculty will be necessary if SPs are to eventually replace faculties in the OSCE evaluaton process. This study was conducted to know to what degree SPs' checklist recordings agree with those of faculties during an OSCE.

Methods: One hundred and twenty one fourth-year medical students of Hanyang University College of Medicine took an OSCE. In each of two study stations, a student saw an SP for four minutes and the SP recorded the same checklists as a faculty examiner did, for the following fifty seconds.

Results: For the 'bad news delivery' station, SP evaluations were more lenient compared to those of faculties (56 vs 45, $p<0.01$), but in the case of 'chest pain', there was no significant difference. Pearson correlation coefficients for the 'bad news delivery' station and for the 'chest pain' case were 0.60 and 0.65, respectively. The mean percentages of agreement for the 'bad news delivery' and the 'chest pain' checklists were 71% and 82%, respectively. The mean kappa statistics for the 'bad news delivery' and the 'chest pain' check-lists were 0.19 and 0.49, respectively.

Conclusion: The ratings by SPs were found to be consistent with those of faculties only in moderate degree. The exactness of scoring criteria, and the optimal SP training are to be the premise for the replacement of faculties by SPs during OSCE checklist recordings.

Key Words: Clinical competence, Observer variation, Educational measurement, Patient simulation

교신저자: 이정권, 삼성서울병원 가정의학과
서울시 강남구 일원동 50번지
Tel: 02)3410-2441, Fax: 02)3410-0388, E-mail: jkwonl@smc.samsung.co.kr
* 이 연구는 2001년도 한양대학교 교내 연구비 수혜로 이루어졌음

서 론

현재 의학교육 평가분야에서 혁신적인 변화 중 하나는 표준화 환자 (standardized patient; SP)를 이용한 시험이다 (Harden 등, 1975; Miller, 1990; Vu 등, 1992a). 표준화환자를 이용하는 시험은 단편적인 (segmental) 임상수기 (procedure) 중심의 객관구조화진료시험 (Objective structured clinical examination; OSCE)과 좀 더 포괄적이고 진료현실에 가깝게 구성된 진료수행시험 (clinical performance examination; CPX)의 두 형태로 대별할 수 있다 (Miller, 1990; Ferrel, 1995). 국내에서는 도입단계부터 두 방법이 맞물려 소개되어 스테이션이 어느 형태에 속하는지 구분이 잘 안 될 때가 많지만 스테이션 당 허용 시간이나 문제의 구성 내용을 보면 객관구조화진료시험이 여러 의과대학에서 종합평가방법으로서 비교적 활발히 시행되고 있다 (이병국, 1997; 박훈기 등, 1998; 서보양 등, 1998; 박훈기 등, 1999).

임상종합평가 목적으로 시행한 객관구조화진료시험에서 시험문제노출에 대한 걱정 (Colliver 등, 1992)과 운영상의 효율성을 고려하여 학생정원이 많은 의과대학에서 객관구조화진료시험을 한 날짜에 시행하려고 도입한 방법이 고사장 복제다 (박훈기 등, 1999). 그러나 고사장 복제는 점검표 상의 평가자 간의 차이로 객관구조화진료시험의 신뢰도를 떨어뜨리고 시험의 타당도까지 위협하게 된다. 특히 임상교수의 바쁜 일정과 학생교육에 대한 가치관, 열정, 실행의지의 다양성 등은 객관구조화진료시험 운영에서 필요한 교수인력을 확보하는데 커다란 걸림돌로 지적되는 바 (박훈기 등, 1998; 서보양 등, 1998; 박훈기 등, 1999), 시험운영의 효율성만을 고려해서 표준화환자를 학생 평가에 활용하여 필요한 교수수를 최소로 줄이는 방법이 있다 (Colliver와 Williams, 1993; 박훈기 등, 1999). 표준화환자를 평가에 활용하면 그만큼 교수요원의 평가부담을 덜어주는 효과가 있고 시험 전 점검표 적용의 신뢰도 향상 훈련을 비교적 쉽게 시킬 수 있어 평가 자체의 신뢰도를 높일 수 있다.

표준화환자는 충분한 훈련을 받고 어느 정도 경험이 쌓이면 표준화환자연기 뿐 아니라 정해진 점검표에 따라 평가까지 할 수 있다 (Norman 등 1985; Williams 등, 1987; Tamblyn 등, 1990; Vu 등 1992b; Colliver와 Williams, 1993; De Champlain 등 1997; Regehr 등 1999b). 이런 연구에서 병력청취, 신체진찰, 환자교육 영역에 대해 표준화환자의 점검표 채점은 임상의사 혹은 임상의사가 아닌 관찰자의 채점과 비교하여 80～100%의 높은 일치도를 보였다.

의학교육에서 누가 평가를 해야 하는가는 평가의 목적에 따라 달라질 수 있는데 (Westberg와 Jason, 1993), 종합시험으로서 객관구조화진료시험 평가 목적이 학생의 진료수행능력 성숙정도를 확인하는 데 있다면 환자 시각에서 바라본 평가가 교수의 평가 결과만큼 중요하다고 할 수 있다. 이론상으로 객관구조화진료시험의 점검표는 누가 평가자가 되더라도 일정 수준 이상의 신뢰성을 유지할 수 있을 때만 제대로 개발되었다고 인정받을 수 있다 (Newble과 Swanson, 1988; Cohen 등 1991). 객관구조화진료시험에서 표준화환자가 교수 평가를 대신할 수 있으려면 최소한 '제대로 개발된 점검표'를 통해서 교수의 평가와 표준화환자의 평가가 큰 차이가 없다는 것을 사전에 입증할 수 있어야 한다 (Newble과 Swanson, 1988; Colliver와 Williams, 1993). 본 연구는 일 개 대학에서 시행한 객관구조화진료시험에서 표준화환자평가와 교수평가의 일치도를 분석하여 앞으로 우리나라에서 표준화환자의 역할을 확대하여 평가자로 이용할 수 있을지 그 가능성을 점검하고자 하였다.

대상 및 방법

가. 전체 객관구조화진료시험 구성 및 운영

한양대학교 의과대학 4학년 학생 121명을 대상으로 2001년 2월 24일 임상종합평가의 한 축으로 학점을 부여하는 객관구조화진료시험을 시행하였다. 객관구조화진료시험은 총 20개 스테이션으로 구성되었으며 2개의 연결형 스테이션과 2개의 휴식 스테이션을 포함하고 있어 문제는 총 16문제가 출제되었다. 이 중 13개 스테이션에서 표준화환자를 활

용하였다. 전체적으로 각 스테이션 당 허용시간은 5분이었고 이 중 30~50초는 피드백 제공시간이었다. 전체 121명이 하루에 시험을 마칠 수 있도록 하기 위해 시험장을 두 군데로 복제하였으며 학생은 3부로 나뉘어 시험을 치렀다. 점검표의 항목의 전체적인 구성은 병력청취 27%, 신체진찰 7%, 진단 2%, 결과해석 10%, 향후계획수립 17%, 수기 3%, 환자교육 11%, 태도 11%, 총괄평가 11%, 보너스 점수가 1%를 차지하였다.

나. 채점자간 점검표 채점 비교 대상 스테이션

표준화환자와 교수의 점검표 적용의 차이를 알아보기 위하여 심장내과에서 개발한 '흉통' 스테이션과 가정의학과에서 개발한 '나쁜 소식 전하기'의 두 스테이션을 연구대상으로 하였다. 이 두 스테이션에는 훈련된 표준화환자뿐 아니라 훈련된 평가자로서 교수가 한 명씩 추가로 배정되었다. 한 학생이 4분 동안의 표준화환자진료를 끝내면 표준화환자는 50초 동안 교수가 사용하는 것과 똑 같은 점검표에 따라 그 학생의 진료내용과 질을 평가했다. 교수는 면담 전반에 걸쳐 4분 동안 학생의 행동 표현과 동시에 점검표의 항목에 따라 평가를 했다. 표준화환자가 교수가 사용한 것과 같은 점검표를 채점하는 동안, 교수는 학생이 그 스테이션에서 보여준 두드러진 장단점에 대하여 짧게 되먹임을 제공하였다. '흉통'의 점검표는 총 18개의 병력 청취 항목과 1개의 보너스 항목으로 구성되었으며 이 중 9개 항목은 주소(chief complaints)관련 항목이었고 5개 항목은 심혈관질환의 위험인자의 질문여부를 평가하는 항목이었다. 기타 항목으로 인사하기, 과거병력, 추정진단 등이 포함되었다. 마지막 항목은 총괄 수행능력 평가로 상, 중, 하의 3구간 척도를 사용했고 다른 항목은 모두 '시행했다', '시행하지 않았다'의 이분척도를 사용했다. '나쁜 소식 전하기' 점검표는 일반적인 환자면담의 질을 평가하는 항목 5개와 증례-특이 (case-specific) 항목 8개와 총괄수행능력평가 한 항목으로 구성되었다. 총괄평가만 3구간 척도였고 나머지는 이분척도였는데 두 항목(진단의심에 대한 대처, 다른 의사와 연결)은 '했다', '하지 않았다', '해당 없음'의 3개의 응답키 (key)가 있었다. 표준화환자의 훈련은 1회 1~2시간씩 총 3회를 실시했고 두 번의 표준화환자 점검표 채점연습과 한 번의 교수 점검표 채점 연습이 포함되었다. '나쁜 소식 전하기' 표준화환자 훈련에는 전공의, 교수가 학생 역할을 하여 한 표준화환자 당 세 번씩 모의학생면접 실습 기회를 제공하였다. 채점자간 점검표 적용 일치도를 높이는 연습에는 직접평가와 녹화비디오를 통한 평가방법을 동시에 사용했고 교수와 표준화환자가 함께 참여하였다. 표준화환자는 모두 현역 전문 연극배우들로 표준화환자 경험은 이번 시험이 두 번째였다. 한 명의 표준화환자가 대면한 학생 수는 총 60~61명이었다.

다. 통계적 방법

스테이션별, 성별 채점자간 평가점수의 차이를 보기 위해 student t-검정을 시행하고 시험순서별 차이 검정은 분산분석을 시행하였다. 스테이션, 시험순서, 성별로 채점자간의 스테이션 점수의 연관성을 보기 위해 피어슨 (Pearson) 상관분석을 시행하였다. 스테이션별 채점자간 점수의 차이를 비교하기 위해 회귀분석을 시행하고 그 결과를 회귀식으로 나타내었다. 이상의 분석에서 유의 수준 α, 0.05를 기준으로 통계적 유의성 여부를 판단하였다. 문항별로 전체 평가 시도 중 두 채점자간 정확히 일치하는 평가쌍을 백분율로 나타내어 퍼센트 일치도 (percent of agreement)로 제시하였다. 또 다른 방법으로 채점자간 일치도를 보기 위해 이분 척도 문항은 전반적 일치도 카파 (kappa measurement of agreement)를 구했고 3분 척도에 대해서는 가중치를 둔 카파 (weighted kappa)를 구하였다 (Armitage와 Berry, 1994). 모든 통계적 처리는 의학적 통계전문프로그램인 SAS 6.12판을 이용하였다.

결 과

전체 121명 학생 중 남학생이 94명으로 77.7%를 차지하였다. 전체 객관구조화진료시험의 표준화 Cronbach α는 0.70이었다.

Table I. The Inter-rater Comparisons of Station Scores according to Examination Site, Examination Order, and Sex

Variables			Bad news delivery			Chest pain		
			Mean±SD	Range	p-value*	Mean±SD	Range	p-value*
Site	1st	Faculty	36±16	0-69	<0.01	57±13	25-85	0.10
		SP	47±15	13-75		53±15	20-80	
	2nd	Faculty	54±12	25-75	<0.01	53±11	29±77	0.03
		SP	63±13	25-81		59±15	29±88	
Order	1st	Faculty	42±17	0-73	0.02	52±12	25-75	0.25
		SP	52±18	13-75		55±16	20-88	
	2nd	Faculty	44±16	13-75	<0.01	55±14	30-85	0.65
		SP	56±15	27-81		53±15	20-80	
	3rd	Faculty	47±18	6-69	<0.01	59±10	41-80	0.79
		SP	59±15	25-81		59±13	29-88	
Sex	M	Faculty	43±17	0-73	<0.01	54±12	25-85	0.84
		SP	55±16	13-81		55±15	20-88	
	F	Faculty	51±16	13-75	0.04	59±11	40-77	0.82
		SP	60±16	13-81		60±13	35-82	
All		Faculty	45±17	0-75	<0.01	55±12	25-85	0.79
		SP	56±16	13-81		56±15	20-88	

* P-values were obtained by student t-test or ANOVA test

가. 채점자간 스테이션 점수 비교

'나쁜 소식 전하기' 스테이션에서는 표준화환자평가 점수가 56점으로 교수 평가 점수 45점보다 통계적으로 유의하게 높았다 (p<0.01). '흉통' 스테이션 점수는 표준화환자평가와 교수평가 간에 통계적으로 유의한 차이가 없었다 (Table Ⅰ).

고사장별로 보면 '나쁜 소식 전하기' 스테이션 점수는 두 시험장소 모두에서 표준화환자평가 점수가 교수 평가 점수보다 통계적으로 유의하게 높았다 (p<0.01). '흉통' 스테이션 점수는 제 2고사장에서만 표준화환자평가점수가 교수평가점수보다 높았다 (p=0.03).

시험순서별로 보면 '나쁜 소식 전하기' 스테이션 점수는 세 시기 모두에서 표준화환자평가 점수가 교수 평가 점수보다 통계적으로 유의하게 높았다 (p<0.05). '흉통' 스테이션 점수는 세 시기 모두에서 채점자간에 통계적으로 유의한 차이가 나지 않았다. 교수평가나 표준화환자평가 둘 다 시험순서가 후반으로 갈수록 '나쁜 소식 전하기' 스테이션 점수가 증가추세를 보였고 '흉통' 스테이션 점수는 이러한 추세가 교수평가에서만 나타났으나 통계적으로 유의한 차이는 아니었다.

성별로 보면 '나쁜 소식 전하기' 스테이션 점수는 남녀 모두에서 표준화환자평가 점수가 교수 평가 점수보다 통계적으로 유의하게 높았으나 (p<0.05), '흉통' 스테이션 점수에서는 남녀 모두 채점자간 차이가 없었다.

나. 표준화환자평가와 교수평가의 상관성

표준화환자평가와 교수평가간의 피어슨 상관계수는 '나쁜 소식 전하기' 스테이션 점수가 0.60, '흉통' 스테이션 점수는 0.65였다 (Table Ⅱ). 두 스테이션

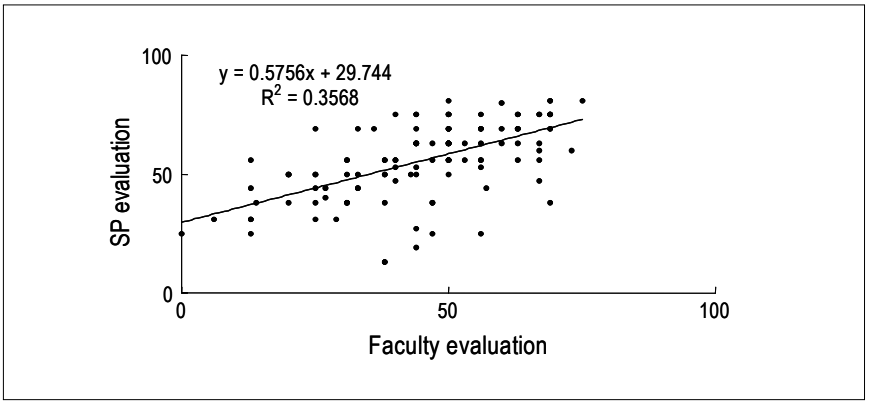

Fig. 1. The relationship between faculty and standardized patient evaluation for the station of 'Bad news delivery'. As faculty scores increase, standardized patient (SP) scores increase linearly with the slope of 0.58. The explainability of faculty scores for SP evaluation scores is 36%.

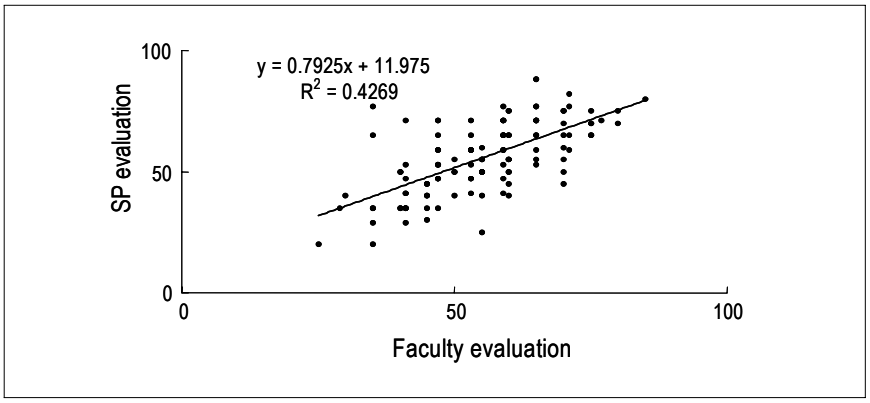

Fig. 2. The relationship between faculty and standardized patient evaluation for the station of 'Chest pain'. As faculty scores increase, standardized patient (SP) scores increase linearly with the slope of 0.79. The explainability of faculty scores for SP scores is 43%.

모두 고사장별 채점자간 상관계수가 차이가 있었고 남학생보다 여학생의 경우 채점자간 스테이션 점수 상관계수가 더 컸다. 시험시기별로는 '나쁜 소식 전하기' 스테이션에서는 제 2부 수험생에 대한 채점자간 점수 상관계수가 가장 낮았고, '흉통' 스테이션의 경우는 제 3부 수험생에 대한 채점자간 점수 상관계수가 가장 낮았다. 표준화환자평가와 교수평가의 관계를 회귀식으로 표현하였다 (Fig. 1, 2).

다. 표준화환자평가와 교수평가의 문항별 일치도

'나쁜 소식 전하기'의 경우 표준화환자평가와 교수평가의 퍼센트 일치도 (percent of agreement) 전체 평균은 71%, 범위는 48～99%였다 (Table Ⅲ). 영역별로 증례-특이 (case-specific) 문항의 평균은 68%, 범위 48～99%, 일반 (general)면담평가문항의 평균은 81%, 범위 72～93%였고 총괄 (global) 평가항목의 퍼센트 일치도는 55%였다. 문항별 일치도 카파

Table II. The Pearson Correlation Coefficient (R) between Faculty Evaluation and Standardized Patient Evaluation according to Exam Site, Order of Examination and Student Gender

Variables		Bad news delivery		Chest Pain	
		R	p-value	R	p-value
Site	Faculty 1	0.44	<0.00	0.80	<0.00
	Faculty 2	0.48	<0.01	0.60	<0.01
Order	First Part	0.72	<0.01	0.65	<0.01
	Second Part	0.44	<0.01	0.75	<0.01
	Third Part	0.62	<0.01	0.53	<0.01
Student	Male	0.57	<0.01	0.62	<0.01
Gender	Female	0.64	<0.01	0.78	<0.01
All		0.60	<0.01	0.65	<0.01

Table III. The Proportion of Agreement Rates and Kappa Coefficients between Faculty and Standardized Patient(SP) Evaluation for 'Bad News Delivery'

Items	Percent of agreement	Kappa
Verification of biopsy results	64	0.03
Exploring patient's history of coping with big challenge	99	0.00
Check patient's readiness to bad news	53	0.17
Asking experience of significant others with cancer	97	0.65
Notification of the fact with simple and clear words	78	0.34
Continuation of eye contact more than 50%	76	0.05
Empathy with patient's embarrassment	56	0.12
Check patient's response right after bad news delivery	51	0.03
Positive coping with patient's doubt about diagnosis	48	0.10
Check patient's understanding of doctor's explanation	93	0.17
Giving patient opportunities of inquiry	82	0.36
Avoidance of medical jargon	72	-0.06
Trying to refer smoothly with other doctors	63	0.28
Following patient's cue	82	0.36
Global satisfaction with the encounter	55	0.31
All	71	0.19

(kappa)는 전체 평균이 0.19였고, 의학적 용어 사용 여부 문항이 최소치 -0.06, 주위 사람의 비슷한 질병 경험 묻기 항목이 최대치 0.65로 나왔으며 총괄평가 항목의 카파는 0.31이었다.

'흉통'의 경우 표준화환자평가와 교수평가의 퍼센트 일치도 전체 평균은 82%, 범위는 53～98%였다 (Table IV). 영역별로 주소관련문항의 평균은 82%, 범위 70～98%, 심혈관질환위험요인문항의 평균은 93%, 범위 91～98%였다. 그 외 퍼센트 일치도는 인사하기 항목이 79%, 과거력 질문항목이 67%, 진단추정항목이 80%, 보너스 항목이 73%, 총괄평가항목이 53%를 보였다. 문항별 일치도 카파는 전체 평균은 0.49였고, 흉통의 성격 질문 문항이 최소치 -0.05, 흡연력 질문항목이 최대치 0.84로 나왔으며 총괄평가항목의 카파는 0.20이었다.

Table IV. The Proportion of Agreement Rates and Kappa Coefficients between Faculty and SP Evaluation for 'Chest Pain'

Items	Percent of agreement	Kappa
Introduce with greetings	79	0.58
Chief complaints: Onset	98	0.49
Character	89	-0.05
Duration	85	0.36
Interval	73	0.45
Location	77	0.43
Radiation of pain	82	0.62
Aggravating factors	70	0.09
Effect of body posture	86	0.48
Other related symptoms	82	0.62
Risk factors: Diabetes mellitus	91	0.75
Hypertension	91	0.80
Smoking	92	0.84
Hyperlipidemia	98	0.66
Family of premature cardiac death	91	0.81
Past medical history	67	0.35
Diagnostic impression	80	0.60
Bonus	73	0.20
Global rating	53	0.20
All	82	0.49

고 찰

이번 연구에서 교수 평가와 표준화환자평가의 퍼센트 일치도는 71～82%로 다른 기존의 연구의 80～100%에 비하면 상대적으로 낮은 수준이었다 (Norman 등 1985; Williams 등 1987; Tamblyn 등 1990; Vu 등 1992a). 두 채점자간의 점검표 적용 차이의 원인으로는 점검표 항목 자체의 애매한 채점 기준, 채점자간 일치도 향상 훈련 부족, 평가자의 피로도 등을 들 수 있다 (Miller, 1990; Tamblyn, 1991; De Champlain 등 1997). 이 외에도 점검표 항목 수와 항목의 척도 종류가 점검표 적용의 정확성에 영향을 줄 수 있다. 점검표의 채점자간 일치도가 낮은 경우 일치도를 높이려면 문항별로 명확한 채점 기준표를 만들어 각 척도 구간의 정의가 누가 보아도 확실하게 해석될 수 있게 하고, 또 이 채점 기준표를 갖고 해당 평가자 모두가 사전에 충분한 적용훈련을 받게 해야 한다. 시험 실시 시기별로 채점자간 점수의 상관성에 특별한 변화 경향은 없었지만 시험시기가 뒤로 갈수록 두 평가자의 점수가 공통적으로 높아진 것으로 미루어 보아 평가자의 피로도가 채점자간 일치도에 영향을 주었을 가능성이 있다.

단순하게 얼마나 많은 평가의 쌍이 서로 일치하는가를 나타내는 퍼센트 일치도만으로 채점자간 일치도를 판단하는 것은 옳지 않다 (Armitage와 Berry, 1994). 일치도 카파의 산정은 우연에 의한 일치확률을 고려하면서 두 채점자간에 얼마나 평가가 일치하는가를 판단하는데 도움이 된다. 문항에 대한 평가가 극단적으로 거의 모두 '시행했다'로 채점되었거나 혹은 '시행하지 않았다'로 채점되었을 경우 퍼센트 일치도는 매우 높은 값을 보이나 카파는 오히려 낮게 나올 수 있다. 한 예로 '나쁜 소식 전하기'의 '어려운 과거 극복경험'질문은 전체 118명의 학생 중 1명을 제외하고 117명에서 '시행하지 않았다'

로 체크되어 퍼센트 일치도는 100%에 가깝지만 카파는 매우 낮은 값을 보였다. 퍼센트 일치도와 카파가 동시에 높게 나온 항목은 일치도가 높은 것으로 해석할 수 있다 (Vu 등, 1992b; Armitage와 Berry, 1994; De Champlain 등, 1997). 이 기준을 적용하면 '나쁜 소식 전하기' 점검표에서는 '주위 사람의 비슷한 질병경험 묻기'가 상대적으로 일치도가 높은 항목이고 '소식전달 후 환자감정체크하기', '진단의심대처하기', '환자감정에 공감하기' 등의 항목에서 일치도가 낮았다고 판단할 수 있다. '흉통' 점검표에서는 '악화요인질문' 항목과 '과거병력질문' 항목이 상대적으로 일치도가 낮은 항목이라고 판단할 수 있다. 총괄평가 항목과 '해당 없다'의 구간을 가진 '진단의심 대처하기'와 '다른 의사와 연결해주기' 항목의 일치도가 상대적으로 낮은 것은 3구간 척도가 이분척도보다 정확하게 체크하기가 어렵다는 것을 의미한다 (De Champlain 등, 1997).

주소관련항목의 일치도가 위험요인질문항목의 일치도보다 낮은 이유를 개방형 질문과 폐쇄형 질문의 구성비 차이로 해석할 수 있다 (Westberg와 Jason, 1993). 현재병력을 파악할 때 표준화환자의 대꾸는 종종 2개 이상의 점검표 속 대답을 한꺼번에 말할 수 있다. 이 경우 채점항목의 지나치게 상세한 구분은 점검표 적용 시 평가자의 판단착오를 초래하기 쉽다 (Colliver와 Williams, 1993; De Champlain 등, 1997). 특히 교수는 학생의 행동발생과 동시에 점검표에 체크를 할 수 있어 여유가 있지만 표준화환자는 4분 후에 자신의 기억에만 의존하여 점검표를 50초 동안에 작성해야 했기 때문에 그 만큼 정확도가 떨어졌을 가능성이 있다.

'나쁜 소식 전하기'는 가정의학과 강의시간에 부분적으로 다루어졌지만 한양의대 객관구조화진료시험에서는 처음 출제된 문제였다. 이 문제에서는 많은 학생들이 전체적으로 낮은 점수를 얻었는데 주로 증례-특이항목에서 낮은 점수를 보였다. 따라서 증례-특이항목의 일치도가 일반면담평가항목의 일치도보다 낮은 것은 증례-특이항목의 경우 해당 항목을 시행하지 않은 학생이 시행한 학생보다 훨씬 많아 표준화환자나 교수 공통적으로 채점상 위음성 (false negative) 반응이 많았던 것을 하나의 원인으로 설명할 수 있다 (Reznick 등, 1998).

총괄평가문항의 퍼센트 일치도는 이분척도를 가진 다른 문항에 비하여 일치도가 53~55%로 낮게 나왔다. 이는 총괄평가의 속성상 구간별로 정확한 채점기준을 마련하기가 어려워 평가자 주관이 상대적으로 많이 개입되었다는 데서 그 원인을 찾을 수 있다 (Newble과 Swanson, 1988; Cohen 등 1991; Regehr 등 1998; Regehr 등 1999b). 3구간 척도가 아니고 5구간 혹은 7구간 척도인 경우 이 문제가 더욱 심각해질 수 있는 데 이런 경우 일반적으로는 척도에서 한 구간 정도의 차이는 두 평가가 서로 일치하는 것으로 인정하고 있다 (Westberg와 Jason, 1993). 하지만 원칙적으로 총괄 평가의 척도 각 구간별로 가능한 한 확실한 채점 기준을 마련하는 것이 평가의 객관성을 유지하는 데 중요하다.

이번 연구에서는 교수평가를 황금률 (gold standard)로 하여 표준화환자평가가 어느 정도 이에 접근하고 있는가를 알아보았다. 채점 기준이 분명한 '흉통'의 경우 두 채점자간의 점수차이가 없었지만 일치도가 낮았던 '나쁜 소식 전하기' 스테이션에서는 표준화환자의 평가점수가 교수 평가 점수에 비하여 24% 높았다. 이는 표준화환자가 교수와 같은 채점기준표를 사용하더라도 점수를 후하게 주는 경향이 있다고 해석할 수 있다 (Westberg와 Jason, 1993).

표준화환자평가점수와 교수 평가점수의 상관계수는 0.60~0.65로 설명력으로 표현하면 36~42% 정도이다. 표준화환자평가점수와 교수평가점수는 '흉통'의 경우 '나쁜 소식 전하기' 증례에 비하여 좀 더 높은 선형관계를 보이고 있다 (Fig. 1, 2). 중요한 점은 전체 문항 일치도가 71~82% 수준일 때 표준화환자평가점수는 교수평가점수의 36~42%만을 반영한다는 것이다.

'흉통'의 채점자간 점수 차이와 채점자간 점수의 상관성이 고사장별로 다르게 나타나는 것은 표준화환자나 교수가 바뀌면 채점의 일치도가 달라진다는 것을 의미하여 시험 전 채점훈련의 중요성을 제기하고 있다 (Colliver와 Williams, 1993). '나쁜 소식 전하기'의 경우 고사장간에 표준화환자평가나 교수

평가 점수차이가 있고 채점자간 상관계수에서도 차이를 보인다. 이는 문제에 따라서는 점검표 적용 훈련이 3회만으로는 부족하고 이처럼 채점기준이 애매한 경우에는 별도로 일정 수준에 도달할 때까지 필요한 만큼 채점연습을 더 해야함을 시사하고 있다 (Vu 등, 1992b; Westberg와 Jason, 1993; De Champlain 등, 1997).

이 연구는 학점을 부여하는 임상종합시험으로서 객관구조화진료시험 실시 중에 시험의 흐름을 방해하지 않고 수행되었기 때문에 연구방법에 있어서 몇 가지 제한점을 갖고 있었다. 첫째 황금률로 사용한 교수평가의 타당성 여부이다. 교수평가의 정확성 저하는 이 연구에서 채점자간 일치도 하락에 중요한 기여를 할 수 있다 (박훈기 등, 1998; 박훈기 등, 1999). 일반적으로 표준화환자의 훈련보다 교수채점 훈련이 더 어려운 게 사실이지만 과연 일치도의 저하가 교수, 표준화환자 중 어느 쪽의 훈련부족으로 일어난 현상인 지 이 연구에서 밝히기는 어렵다. 둘째, 채점 기준표의 명확성을 확실하게 보장할 수 없었다는 점이다. '흉통'의 경우 5년 이상 반복 출제된 문제여서 다년간 되먹임을 통해 비교적 명확한 채점 기준을 갖고 있었지만 '나쁜 소식 전하기'는 처음 도입한 문제이기 때문에 문항별 채점기준의 정확성을 실제 사용을 통해서는 한번도 검증 받지 못했다. 채점기준의 불명확성은 훈련의 효율성을 떨어뜨리고 그에 따라 일치도 향상 효과도 그 만큼 떨어지게 된다 (Colliver와 Williams, 1993; 박훈기 등, 1999). 셋째, 표준화환자와 교수의 평가 과정이 서로 차이가 있었다는 점이다. 즉, 표준화환자는 학생 면담이 끝난 후에 평가했고 교수는 면담진행과 더불어 학생을 평가했다. 표준화환자가 채점하는 동안 교수는 학생에게 되먹임을 제공하고 있었으므로 이로 인해 표준화환자의 판단 정확도가 더 떨어질 수도 있었다. 또한 표준화환자는 연기도 해야 하고 평가도 해야 했으므로 채점만 하면 되는 교수보다 더 피로가 누적되었을 가능성이 있다. 이러한 채점과정의 환경적 차이가 점수에 어떤 영향을 주는 지는 이 연구에서는 구분할 수는 없었다. 좀 더 이상적인 비교를 하려면 교수 역시 학생면담이 다 끝난 후에 표준화환자와 같은 시각에 점검표를 작성해야 했다. 하지만 학생에게 되먹임을 제공하는 것이 객관구조화진료시험의 또 다른 중요한 목적이기도 했기 때문에 이 연구의 제한점으로 받아들일 수밖에 없었다. 또 다른 방법으로는 학생면담을 녹화해 놓고 교수는 이 비디오를 시청한 후에 50초 동안 표준화환자와 동일한 조건으로 평가를 하고 그 결과를 비교하는 것이다 (Tamblyn, 1991; Vu 등, 1992b; De Champlain 등, 1997). 이 번 연구에서처럼 교수 한 명에 의존한 평가가 황금률로서는 부적합하므로 2인 이상의 교수가 동시에 같은 학생면담을 평가하여 서로 비교하여 일치하는 채점결과를 황금률로 사용하는 것이 더 바람직하다.

결 론

객관구조화진료시험에서 점검표를 이용한 표준화환자의 학생평가는 교수평가와 중등도의 일치도를 보였다. 객관구조화진료시험에서 표준화환자가 교수를 대신하여 평가자 역할을 정확하게 수행하기 위해서는 점검표 채점기준의 명확성과 철저한 표준화환자 채점훈련이 선행되어야 한다.

참 고 문 헌

박훈기, 김동원, 김덕언, 최호순, 김경태(1998). 의학과 4학년 종합평가로서의 객관적-구조적 임상능력평가(OSCE)의 경험. *한국의학교육*, 10(1), 43-56.

박훈기, 이정권, 김승룡, 김경태, 박해영(1999). 시험장 복제(Duplication)가 객관적 구조적 임상 시험(OSCE)의 신뢰도에 미치는 영향. *한국의학교육*, 11(1), 37-52.

서보양, 이두진, 권굉보, 강복수(1998). 객관적으로 구조화된 임상시험의 시행경험. *한국의학교육*, 10(2), 363-381.

이병국(1997). 외과 객관적-구조화-임상시험에서의 표준화 환자 및 의사의 채점에 대한 비교 고찰. *한국의학교육*, 9(부록 1), 61.

Armitage P, Berry G(1994). *Statistical methods in*

medical research(3rd ed.). London: Blackwell Science Ltd.

Cohen R, Rothman AI, Poldre P, Ross J(1991). Validity and generalizability of global ratings in an objective structured clinical examination. *Acad Med*, 66(9), 545-548.

Colliver JA, Travis TA, Robbs RS, Barnhart AJ, Shirar LE, Vu NV(1992). Test security in standardized-patient examinations: analysis with scores on working diagnosis and final diagnosis. *Acad Med*, 67(10 Suppl), S7-S9.

Colliver JA, Williams RG(1993). Technical issues: test application. AAMC. *Acad Med*, 68(6), 454-460.

De Champlain AF, Margolis MJ, King A, Klass DJ(1997). Standardized patients' accuracy in recording examinees' behaviors using checklists. *Acad Med*, 72(10 Suppl 1), S85-S87.

Ferrell BG(1995). Clinical performance assessment using standardized patients: a primer. *Fam Med*, 27(1), 14-19.

Harden RM, Stevenson M, Downie WW, Wilson GM(1975). Assessment of clinical competence using objective structured examination. *Br Med J*, 22:1(5455), 447-451.

Miller GE(1990). The assessment of clinical skills/ competence/performance. *Acad Med*, 65(9 Suppl), S63-S67.

Newble DI, Swanson DB(1988). Psychometric characteristics of the objective structured clinical examination. *Med Educ*, 22(4), 325-334.

Norman GR, Neufeld VR, Walsh A, Woodward CA, McConvey GA(1985). Measuring physicians' performances by using simulated patients. *J Med Educ*, 60(12), 925-934.

Regehr G, Freeman R, Hodges B, Russell L(1999a). Assessing the generalizability of OSCE measures across content domains. *Acad Med*, 74(12), 1320-1322.

Regehr G, Freeman R, Robb A, Missiha N, Heisey R(1999b). OSCE performance evaluations made by standardized patients: comparing checklist and global rating scores. *Acad Med*, 74(10 Suppl), S135-S137.

Regehr G, MacRae H, Reznick RK, Szalay D(1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med*, 73(9), 993-997.

Reznick RK, Regehr G, Yee G, Rothman A, Blackmore D, Dauphinee D(1998). Process-rating forms versus task-specific checklists in an OSCE for medical licensure. Medical Council of Canada. *Acad Med*, 73(10 Suppl), S97-S99.

Tamblyn RM, Klass DK, Schanbl GK, Kopelow ML(1990). Factors associated with the accuracy of standardized patient presentation. *Acad Med*, 65(9 Suppl), S55-S56.

Tamblyn R(1991). Sources of unreliability and biases in standardized patient rating. *Teach Learn Med*, 3(1), 74-85.

Vu NV, Barrows HS, Marcy ML, Verhulst SJ, Colliver JA, Travis T(1992a). Six years of comprehensive, clinical, performance-based assessment using standardized patients at the Southern Illinois University School of Medicine. *Acad Med*, 67(1):42-50.

Vu NV, Marcy MM, Colliver JA, Verhulst SJ, Travis TA, Barrows HS(1992b). Standardized (simulated) patient's accuracy in recording clinical performance check-list items. *Med Educ*, 26(2), 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 M, Steward D(1987). Direct, standardized assessment of clinical competence. *Med Educ*, 21(6), 482-489.