한국산학기술학회논문지
Vol. 12, No. 1 pp. 318-326, 2011

객관 구조화 절차 기술 평가에서 채점자로서의 표준화환자의 신뢰도

손희정[1,5], 문중범[2], 이향아[3], 노혜린[4,5*]
[1]강원대학교 의학전문대학원 마취통증의학교실, [2]강원대학교 의학전문대학원 응급의학교실,
[3]강원대학교 의학전문대학원 산부인과학교실, [4]강원대학교 의학전문대학원 외과학교실,
[5]강원대학교 의학전문대학원 진료능력개발센터

Reliability of Standardized Patients as Raters in Objective Structured Clinical Examination

Hee Jeong Son[1,5], Joong Bum Moon[2], Hyang Ah Lee[3] and Hye Rin Roh[4,5*]
[1]**Department of Anesthesiology, School of Medicine, Kangwon National University**
[2]**Emergency Medicine, School of Medicine, Kangwon National University**
[3]**Obstetrics & Gynecology, School of Medicine, Kangwon National University**
[4]**Surgery, School of Medicine, Kangwon National University**
[5]**Clinical Performance Center, School of Medicine, Kangwon National University**

요 약 본 연구는 절차기술의 객관구조화 진료시험(Objective Structured Clinical Examination)에서 표준화환자가 평가자의 역할을 수행할 수 있는지 알아보기 위해 신뢰도를 평가하는데 그 목적이 있다. 시험의 주제는 남성 도뇨관 삽관과 창상드레싱 2가지로 정하고, 2년 이상의 객관구조화 진료시험 채점 경력이 있는 표준화환자와 교수 각 4명을 2명씩 짝을 지워 한 주제 당 표준화환자 그룹과 교수 그룹이 동시에 채점하게 하였다. 표준화환자들에게는 술기의 정의, 방법, 주의점, 후유증에 대한 교육이 이루어졌으며 동영상이 포함된 강의, 교수의 시연 후 표준화환자가 직접 실습해보고 교수로부터 되먹임을 받는 순서로 총 8시간(주제당 4시간)의 교육이 시행되었다. 8명의 평가자 모두 객관구조화 진료시험 전날 모여 기존의 동영상자료를 이용한 가상 채점으로 1시간동안 채점 표준화를 이루었다. 채점표는 체크리스트 14문항과 총괄평가 1문항으로 이루어졌다. 한 학생당, 주제당 5분간의 시험 후 2분간의 평가가 이루어졌다. 표준화환자와 교수간의 분석은 GENOVA program을 이용하였다. 연구 결과 주제 전체에서 G상수는 0.839, 평가자의 신뢰도는 0.946으로 매우 높았다. 표준화환자그룹과 교수그룹 사이의 평가자간 일치도는 체크리스트에서 0.949, 총괄평가에서 0.908이었다. 따라서 적절한 교육이 선행되어진다면 표준화환자도 절차기술의 객관화진료시험에 신뢰할 만한 평가자로 이용되어질 수 있을 것이다.

Abstract The purpose of this study is to investigate whether standardized patient(SP) can be used as a reliable examiner in Objective Structured Clinical Examination(OSCE).
4 SPs and 4 faculties who have more than 2 years experience of OSCE scoring were selected. For 1 assignment 2 members of faculty and 2 SPs were designated as raters. SPs were educated for assessing 2 technical skills, male Foley catheter insertion and wound dressing, for 8 hours (4 hours / day, each topic). The definition, method, cautions and complications for each of procedural skills were covered in the education. Theoretical lectures, video learning, faculty demonstration and practical training on mannequins were employed. The 8 raters were standardized for an hour with simulated OSCE scoring using previous videos on the day before the OSCE. Each assessment was composed of 14 checklists and 1 global rate. The allotted time for each assignment was 5minutes and for evaluation time 2 minutes per student. The evaluation from the faculty and SPs were compared and analyzed with the GENOVA program. The overall generalizability coefficient (G coefficient) was 0.839 from two cases of OASTS. The reliability of the raters was high, 0.946. The inter-rater agreement between faculty group and SP group was 0.949 for checklist and 0.908 for global rating. Therefore SPs can play a role of raters in OSCE for procedural skills, if they are given the appropriate training.
Key -words : OSCE, Procedural skill, Standardized patient, Generalizability

*교신저자 : 노혜린(hyerinr@kangwon.ac.kr)
접수일 10년 12월 01일 수정일 10년 12월 18일 게재확정일 11년 01월 13일

1. 서론

표준화환자란 환자의 질병을 표준화된, 일정한 형태로 표현하도록 주의 깊게 지도된 사람을 뜻한다[1]. 표준화환자는 Barrows[2]가 신경과 실습에서 학생들을 평가하기 위한 목적으로 사용하기 위해 처음 개발하였다. 표준화환자 프로그램은 발전을 거듭하여 현재는 의과대학에서의 교육과 평가, 전공의 및 전문의의 교육과 시험 등에서 다양하게 활용되고 있다[3].

표준화환자가 채점자로 활용되기 시작한 것은 1970년대 후반부터이다[4]. 이때 표준화환자는 학생으로부터 진찰을 받은 후, 객관화된 평가지에 체크를 하고 학생에게 즉각적인 되먹임을 주었다. 객관적인 체크리스트를 사용할 경우, 평가자는 단순히 학생의 행위를 관찰하고 해당하는 항목에 체크를 하면 된다. 체크리스트를 활용함으로써 평가자는 학생의 행동을 해석하고 판단하기보다 행동을 기록하는 역할을 담당하게 된다[5].

교수를 객관구조화진료시험의 채점자로 혹은 채점 훈련을 위해 지속적으로 동원하는 것은 시간과 비용 면에서 큰 부담이 된다[5,6]. 이 대응책으로 표준화환자로 하여금 객관적인 체크리스트를 활용케 함으로써 교수 대신 채점자로 사용할 수 있다. 이렇게 표준화환자를 채점자로 이용하려는 시도는 여러 연구에서 진행된 바, Martin 등은 문헌고찰을 통해 연구자에 따라 표준화환자 채점자의 신뢰도가 0.46-0.75, 평가자간 신뢰도가 0.60-0.90로 교수채점자와 별 차이가 없는 것으로 기술 하였다[7]. 현재는 미국의사면허시험 실기시험에서도 표준화환자를 채점자로 활용하면서, 병력청취와 신체진찰은 체크리스트에서, 의사소통 영역은 총괄평가 문항으로 학생들을 평가하고 있다[6].

오늘 날 표준화환자는 병력청취, 신체검사, 환자교육, 의사소통기술 등에서 주로 채점하고 있다[6,8,9]. 반면 절차기술(procedural & technical skill) 영역은 아직까지 표준화환자가 채점한 연구보고는 없다. 절차기술 교육이나 평가 횟수가 늘어나면서 의료인 채점자를 동원하는 것은 다른 객관구조화진료시험에서 의료인 채점자를 동원하는 것과 마찬가지로 힘든 일이 된다.

이 연구의 목적은 의대 학생들을 대상으로 한 객관구조화 절차기술 평가에서 채점자로서의 표준화환자의 신뢰도를 알아보는데 있다. 이를 위해 표준화환자와 교수를 채점자로 투입한 후, 그 결과를 분석하여 표준화환자의 채점이 교수와 얼마나 일치하는지, 신뢰도와 일치도에 영향을 미치는 요인은 무엇인지 파악해보고자 하였다.

2. 연구 방법

2.1 연구대상

2.1.1 채점자

채점을 위해 표준화환자 4명, 교수 4명을 각각 선정하였다. 활동 경력이 2년 이상 되었고, 객관구조화진료시험에서 그동안 지속적으로 채점을 해오면서 일관성 있는 채점이 가능하였던 사람들로 정하였다. 표준화환자는 모두 여자였으며, 평균 연령은 43.75세(40-47)였다. 교수의 경우 여자가 3명, 남자가 1명이었으며, 평균 연령은 38.0세(34-44세)였다.

교수의 전공분야는 각각 외과,산부인과, 비뇨기과, 마취통증의학과로 모두 절차기술에 익숙한 계열이었다. 이들을 직종별로 2명씩 짝을 지어, 사례별로 총 4명이 채점하도록 하였다[표 1].

2.1.2 학생

절차기술 강좌를 모두 마친 의학과 3학년 학생 49명을 대상으로 하였다. 절차기술 강좌는 진료실습 직전 학생들을 대상으로 이루어지는 2주간의 집중 프로그램이다. 연구에 대한 설명을 한 후 동의를 얻었다.

2.2 연구방법

2.2.1 사례 개발

연구를 위해 사용할 사례는 남성 도뇨관 삽입(사례 1)과 창상 드레싱(사례 2)으로 하였다. 사례는 각각 15문항으로 이루어졌다. 이 중 14문항은 행위 하나하나에 대한 2점 척도 체크리스트 문항으로 하였다(1점 만점). 마지막

[표 1] 평가자의 인구학적 자료

	시험방				시험방 2			
	도뇨관 삽관				창상 드레싱			
평가자	교수 1	교수 2	표준화환자 1	표준화환자 2	교수 3	교수 4	표준화환자 3	표준화환자 4
성별	여	남	여	여	여	여	여	여
연령	35	34	47	41	44	39	47	40
전공과목	산부인과	비뇨기과			외과	마취통증의학과		

1문항은 전체적인 태도와 숙련도에 대한 항목으로, 5점 척도로 총괄 평가하였다. 사례는 해당 학과의 전문의가 개발한 것을 5인의 절차기술 강좌 운영위원들이 검토 및 수정한 것으로 지난 3년간 학생 평가에 여러번 사용되어 온 것이다. 이를 표준화환자가 이해하기 쉽도록 용어를 수정하는 작업또한 운영위원 5인이 모여 2차례에 걸쳐 하였다.

2.2.2 채점자 훈련

표준화환자를 대상으로 사례별로 각각 4시간의 술기 훈련을 실시하였다. 술기 훈련은 해당 주제의 사례를 개발하고 학생 강의를 담당하였던 교수가 직접 하였다. 강의와 함께 마네킹을 이용하여 실습을 하였고 이후 되먹임을 해주는 과정으로 진행하였다. 충분히 실습한 후에는 채점 문항과 채점 기준에 대해 설명하였다.

채점 훈련은 객관구조화 절차기술 평가가 있기 하루 전날 이루어졌다. 이때는 채점을 담당한 교수와 표준화환자가 함께 모여 진행하였다. 채점 훈련은 그전에 시행되었던 의학과 학생들의 술기 동영상 4개를 차례로 보면서 채점표에 근거하여 실제 채점을 해보는 것으로 하였다. 채점 후에는 사례를 개발한 교수의 채점 기준을 기준으로 하여 표준화를 하였다. 채점 훈련은 1시간 정도 이루어졌다[그림 1].

2.2.3 객관구조화 절차기술 평가 진행

한 학생 당 주어진 시간은 사례별로 5분이었다. 사례와 사례 사이에는 2분을 주었다. 시험방 1에서는 사례 1을 하였고, 시험방 2에서는 사례 2를 하였다. 채점자들은 시험방 안에서 학생을 관찰하며 채점하였다. 채점자들은 맡은 사례를 처음부터 끝까지 모두 채점하였다. 8명의 학생이 끝날 때마다 9분씩 휴식하였다. 점심 휴식시간은 1시간이었다. 휴식 및 점심시간을 포함한 전체 시험시간은 총 8시간이었다. 채점자들은 모두 각각 독립적으로 채점하였다.

2.2.4 분석 방법

총점과 항목별 점수는 체크리스트 채점 문항은 1점 만점, 총괄 채점 문항은 4점 만점으로 코드화하였다. 시험 전체의 신뢰도는 McMaster 대학에서 개발한 urGENOVA의 Windows 프로그램인 g string 4.1.0을 이용하여 일반화가능도계수로 구하였다. 오차 요인은 시험방(Station, S), 채점자(Rater, R), 문항 종류(체크리스트 또는 총괄)(Domain, D), 문항(Item, I) 등 4가지로 하였다.

채점자 간 일치도는 분산성분에서 시험관 요소를 분석하여 구하였다. 채점자 간 일치도는 전체, 교수 간, 표준화환자 간 일치도를 각각 구하였고, 이를 각각 비교하여 교수와 표준화환자 간 채점의 차이가 있는지 분석하였다. 체크리스트 채점 문항과 총괄 채점 문항 각각의 신뢰도는 GENOVA for PC로 구하였다. 오차 요인은 시험방(Station, S), 채점자(Rater, R), 채점자그룹(교수, 표준화환자)(Group, G), 문항(Item, I) 등 4가지로 하였다.

체크리스트 채점 문항과 총괄 채점 문항 각각에서 채점자그룹 간 일치도를 분산성분에서 채점자그룹 요소를 분석하여 구하였다.

3. 연구 결과

3.1 전체 신뢰도

3.1.1 신뢰도와 일치도

교수 4명과 표준화환자 4명이 채점한 객관구조화 절차기술 평가의 전체 신뢰도는 0.839이었다. 교수의 전체 신뢰도는 0.634였다. 표준화환자의 전체 신뢰도는 0.635였다.

채점자 간의 일치도는 0.946 (=2.4306-0.1324/2.4306)이었다. 교수 간의 일치도는 0.917 (=2.8034-0.2331/2.8034)이었고, 표준화환자간의 일치도는 0.900 (=2.4774-0.2484/2.4774)이었다.

3.1.2 분산성분

문항 종류(체크리스트 또는 총괄)에 대한 분산성분이 상대적으로 가장 컸다(2.0738, 85.32%) [표 2]. 학생과 문항 종류와의 상호작용에 대한 분산성분(0.1458, 6.00%)이 두 번째로 컸고, 학생과 채점자와의 상호작용에 대한 분

[그림 1] 평가자 훈련 과정

[표 2] 객관 구조화 절차 기술 평가의 분산성분

효과	자유도	분산성분 (%)		
		전체	교수	표준화환자
P	48	0.0025 (0.10)	0.0030 (0.11)	0.0023 (0.09)
S	1	0.0000 (0.00)	0.0000 (0.00)	0.0000 (0.00)
R:S	2	0.0000 (0.00)	0.0000 (0.00)	0.0000 (0.00)
D:S	2	2.0738 (85.32)	2.2670 (80.87)	1.8765 (75.75)
I:D:S	26	0.0275 (1.13)	0.0304 (1.08)	0.0290 (1.17)
P:S	48	0.0000 (0.00)	0.0000 (0.00)	0.0000 (0.00)
PR:S	96	0.0570 (2.35)	0.1391 (4.96)	0.1673 (6.75)
PD:S	96	0.1458 (6.00)	0.2321 (8.28)	0.2822 (11.39)
PI:D:S	1248	0.0486 (2.00)	0.0378 (1.35)	0.0390 (1.57)
RD:S	4	0.0286 (1.18)	0.0374 (1.33)	0.0237 (0.96)
RI:D:S	52	0.0000 (0.00)	0.0000 (0.00)	0.0000 (0.00)
PRD:S	192	0.0000 (0.00)	0.0000 (0.00)	0.0000 (0.00)
PRI:D:S	2496	0.0468 (1.93)	0.0566 (2.02)	0.0574 (2.32)
Total	4311	2.4306 (100.00)	2.8034 (100.00)	2.4774 (100.00)

P: 학생(Person), S: 시험방(Station), R: 채점자(Rater),
D: 문항종류(Domain of items ; Checklist or global rating), I: 문항(Items)

산성분(0.0570, 2.35%)이 세 번째로 컸다. 시험방, 채점자, 채점자와 문항의 상호작용, 학생과 채점자와 문항 종류와의 상호작용의 분산성분은 0.000 미만이었다. 교수(2.2670, 80.87%)나 표준화환자(1.8765,75.75%) 별로 분석한 분산성분도 문항의 종류가 가장 컸다. 교수별, 표준화환자별 분산성분에서도, 그 다음은 학생과 문항 종류와의 상호작용, 학생과 채점자와의 상호작용의 순이었다.

또한 교수별, 표준화환자별 분석에서도 시험방, 채점자, 채점자와 문항의 상호작용, 학생과 채점자와 문항 종류와의 상호작용의 분산성분 등은 모두 0.000 미만이었다.

3.2 체크리스트 문항의 신뢰도

3.2.1 신뢰도와 채점자그룹 간 일치도

체크리스트 문항의 전체 신뢰도는 0.727이었다. 교수 그룹과 표준화환자 그룹 간의 일치도는 0.949 (=0.1237-0.0063/0.1237)이었다[표 3].

3.2.2 분산성분

학생과 문항, 그리고 채점자의 상호작용(0.0669, 54.08%)이 가장 큰 분산 성분이었다. 두 번째로 큰 분산성분은 문항과 채점자의 상호작용(0.0224, 18.11%)이었다. 문항, 채점자그룹, 학생과 문항의 상호작용, 학생과 채점자그룹의 상호작용, 시험방과 채점자그룹의 상호작용, 학생과 시험방과 채점자그룹의 상호작용 등의 분산성분 등은 모두 0.000미만이었다.

[표 3] 체크리스트의 분산성분

효과	자유도	분산성분 (%)
P	48	0.0010 (0.81)
S	1	0.0066 (5.34)
I:S	26	0.0000 (0.00)
G	1	0.0000 (0.00)
R:G	2	0.0055 (4.45)
PS	48	0.0063 (5.09)
PI:S	1248	0.0000 (0.00)
PG	48	0.0000 (0.00)
PR:G	96	0.0016 (1.29)
SG	1	0.0000 (0.00)
SR:G	2	0.0021 (1.70)
IG:S	26	0.0026 (2.10)
IR:SG	52	0.0224 (18.11)
PSG	48	0.0000 (0.00)
PSR:G	96	0.0050 (4.04)
PIG:S	1248	0.0037 (2.99)
PIR:SG	2496	0.0669 (54.08)
Total	5487	0.1237 (100.00)

P: 학생(Person),
S: 시험방(Station),
I: 문항(Items)
R: 채점자(Rater),
G: 채점자그룹(Rater group; Faculty or Standardized Patient)

3.3 총괄평가 문항의 신뢰도

3.3.1 신뢰도와 채점자그룹 간 일치도

총괄평가 문항의 전체 신뢰도는 0.921이었다. 교수 그룹과 표준화환자 그룹 간의 일치도는 0.908(=0.9485-0.0842/0.9485)이었다[표 4].

[표 4] 총괄평가의 분산성분

효과	자유도	분산성분 (%)
P	48	0.0659(6.95)
S	1	0.2352(24.80)
G	1	0.0165(1.74)
R:G	2	0.0000(0.00)
PS	48	0.2520(26.57)
PG	48	0.0677(7.14)
PR:G	96	0.0000(0.00)
SG	1	0.0000(0.00)
SR:G	2	0.0557(5.87)
PSG	48	0.0000(0.00)
PSR:G	96	0.2555(26.94)
Total	391	0.9485(100.00)

P: 학생(Person),
S: 시험방(Station),
R: 채점자(Rater),
G: 채점자그룹(Rater group; Faculty or Standardized Patient)

3.3.2 분산성분

학생, 시험방, 채점자의 상호작용(0.2555, 26.94%)이 가장 큰 분산성분이었다. 두 번째로 큰 분산성분은 학생과 시험방의 상호작용(0.2520, 26.57%)이었고, 세 번째로 큰 분산성분은 시험방(0.2352, 24.80%)이었다. 채점자, 학생과 채점자의 상호작용, 시험방과 채점자 그룹의 상호작용, 학생과 시험방과 채점자그룹의 상호작용 등의 분산성분은 모두 0.000미만이었다.

4. 고 찰

Harden 등[10]이 1970년대 중반에 처음 제안한 객관구조화진료시험은 병력청취나 신체진찰, 검사결과 판독 등의 과제를 학생들이 시험방 각각에서 하나씩 차례차례 수행하도록 하고 이를 채점자가 객관적인 채점표에 의거하여 평가하는 것이었다. 절차기술이 객관구조화진료시험 형태로 평가되기 시작한 것은 1990년대 후반부터이다. Martin 등[11]은 외과 전공의를 대상으로 객관구조화 절차기술 평가를 특수 제작된 모형과 동물에서 각각 시행한 결과 동등한 결과를 얻었다며, 절차기술을 평가하는 데 객관구조화진료시험을 활용할 수 있다고 하였다.

객관 구조화된 절차기술 평가에서 체크리스트와 총괄평가는 모두 높은 신뢰도를 보였으며[12,13], Bould 등은 절차기술 평가에서 사용될 수 있는 10가지 평가 방법을 나열하며 그 중 체크리스트, 총괄평가, Multi-station bench testing을 신뢰도와 타당도가 가장 우수한 평가방법으로 꼽았다[14]. 여러 연구에서 총괄평가와 체크리스트를 이용한 평가가 서로 비교된 바 있으나 어느 한가지 방법이 다른 방법보다 특별히 신뢰도가 높은 것으로 판명되지는 않았다. Friedlich 등은[15] 의학과 학생을 대상으로 8개의 시험방에서 시행한 객관구조화 절차기술 평가의 신뢰도가 체크리스트 문항의 경우 0.71, 총괄평가문항의 경우 0.65였다고 보고하였으나, 본 연구에서는 체크리스트 문항의 신뢰도는 0.727, 총괄평가 문항의 신뢰도는 0.921로, 총괄평가 문항의 신뢰도가 더 높았다. 절차기술은 시행 내용이 순차적이며 비교적 예측 가능하므로 총괄평가보다는 체크리스트가 더 구조적인 평가에 적합하다는 의견도 있으나[16], 8가지 절차기술을 대상으로 시행한 평가에서 체크리스트 단독 혹은 총괄평가 단독의 신뢰도 보다 둘을 겸용한 평가에서 신뢰도가 더 높이 나왔다[17]. 따라서 Friedman의 주장대로 체크리스트와 총괄평가를 겸용하여 상호보완 할 수 있도록 쓰는 것이 평가의 전체적인 신뢰도를 높일 수 있는 길이라 하겠다[18].

그동안 절차기술에 대한 평가는 의사나 간호사 등 의료인, 특히 각각의 절차기술 항목에 전문적인 의료인이 채점해야 하는 것으로 인식되어 왔다[15,17]. 절차기술에서 의료인이 총괄평가를 하는 경우 최소한 체크리스트로 채점하는 것과 동등한 신뢰도를 보이는 것으로 연구되고 있다[13,17]. 그러나 교육 외에 진료와 연구도 소홀히 할 수 없는 교수를 객관구조화진료시험의 채점자로 혹은 채점 훈련을 위해 지속적으로 동원하는 것은 경제적으로 큰 부담이 될 뿐 아니라[5,6] 교수라 하더라도 내재된 신뢰도가 낮은 교수의 경우 훈련을 통해서도 쉽게 나아지기 어렵다는 연구 결과도 보고된 바 있다[19]. 이에 비해 표준화환자는 교수보다 쉽게 동원이 가능하며, 채점 훈련을 하기도 더 용이하다. 또한 반복하여 동원되고 훈련함에 따라 표준화환자의 채점의 신뢰도가 더 높아질 수 있다. 체크리스트로 채점한 시험에서 표준화환자는 병력청취에서 94.63% 신체검사에서 96.09%의 높은 정확도를 나타내었다[20]. 또, 의사채점자와 훈련된 비의사 채점자의 평가에 대한 비교연구에서 체크리스트를 이용한 평가는 그 신뢰도가 인정되어 국가고시 같은 고비용 OSCE에도 비의사 채점자가 채점자로 쓰일 수 있다는 연구결과도 제시된 바 있다[21]. 그렇다면 다른 영역과 마찬가지로 절차기술의 영역에서도 의료인이 아닌 표준화환자가 평가의 채점자로 동원될 수 있을 것인가? 표준화환자는 체크리스트 문항이나 총괄평가문항 모두에서 의료인과 일치된

채점이 가능할 것인가? 저자들은 구조화된 절차기술 채점표를 사용하여 표준화환자에게 적절한 채점 훈련을 할 경우 체크리스트 문항에서든, 총괄평가 문항에서든, 표준화환자가 의사에 못지않은 신뢰도와 일치도를 나타낼 수 있을 것으로 가설을 세웠다.

채점자 간 신뢰도의 분석은 퍼센트 일치도나 Cohen의 kappa 값 혹은 Pearson 상관계수를 사용하는 등 다양한 방법으로 구할 수 있다[20,22-24].

Martin 등은[7] 세 명의 의사가 비디오를 보고 채점한 값으로 기준으로 하여, 의사와 표준화환자가 각각 91개의 같은 비디오를 채점한 평균을 분산 분석함으로써 일치도를 구하였다. 이러한 방법들은 그 시험에서 채점자간의 점수가 어느 정도 일치하였는지를 분석하는 수준에서 활용이 가능하다. 그러나 의사와 표준화환자 간의 일치도를 일반화하기에는 제약이 따르며, 수행평가에서 흔한, 다수의 채점자, 다수의 문항, 다수의 사례에서의 일치도를 체계적으로 분석하기에는 무리가 있다.

여러 가지 방법 중 채점자 간 신뢰도를 분석하기에 가장 체계적이고 정교한 방법은 일반화가능도 이론에 따르는 것이다[22,24]. 일반화가능도 이론을 이용하면 채점자 간의 영향 뿐 아니라 다른 요인들을 종합적으로 분석하여 오차 점수의 분산 원인을 규명할 수 있는 장점이 있다[23]. 일반화가능도 이론 연구에 의해 구해진 분산성분은 체크리스트 항목, 채점자, 사례 등 변수 각각에 대한 분산이 어느 정도인지 알려준다. 이들을 계산함으로써 채점자의 오류가 어느 정도인지를 계산할 수 있다. 또한 채점자의 오류로부터 채점자의 일관성과 채점자 간의 일치도를 분석할 수 있다[22]. 채점자와 관련된 분산성분의 합을 채점자의 오류라고 할 수 있다. 전체 분산성분의 합에서 채점자와 관련된 분산성분의 합을 뺀 나머지의 값이 전체 분산성분에서 차지하는 비중이 바로 채점자의 일치도가 된다. 이에 본 연구에서는 가장 체계적인 채점자 간 분석 방법인 일반화가능도 이론을 이용하여 일치도를 분석하였다.

연구 결과 표준화환자의 신뢰도는 높았으며, 채점자와 문항의 상호작용이 0.000미만이었다. 이것은 채점자가 문항에 따라 오차가 나는 평가를 하지 않고, 일관성 있는 평가가 가능하였음을 뜻한다. 표준화환자의 채점 점수는 교수의 점수와 비교하였을 때 그 일치도도 매우 높았다. 채점자에 의한 오차는 적었다. 채점자 간의 분산성분은 0.000미만이었다. 이러한 결과는 교수 그룹 또는 표준화환자 그룹에서만 분석한 결과에서도 마찬가지였다. 따라서 표준화환자가 채점자로서 교수와 같은 기준으로 학생을 평가할 수 있었음을 알 수 있다. 또한 체크리스트 문항과 총괄평가 문항에서 표준화환자와 의료인의 채점 일치도를 비교하였을 때, 체크리스트 문항에서 평가자간 일치도가 약간 더 높았으나, 둘 다 모두 0.9 이상의 높은 수치를 보였다. 이 결과는 표준화환자라 할지라도, 절차기술 채점에 있어 총괄평가에 의한 채점이 체크리스트 채점에 비해 결코 뒤지지 않을 수 있음을 시사한다. Bullock 등은[25] 의과대학 1, 2학년에 재학 중으로, 아직 절차기술을 배우지는 않았지만, 무균조작에 대해서는 배운 학생들에게 2시간 동안의 채점 훈련을 한 후 응급의학 전공 의사가 평가한 것과 비교한 결과, 일치도가 높았다며, 채점자가 전문 의료인이 아니더라도 가능하였다고 보고하였다.

교수 간 일치도나 표준화환자 간 일치도보다 채점자 간 일치도가 높았다. 이 결과는 직업 간 차이보다 같은 직업군 내에서 구성원 간 차이가 컸다는 것이며, 이는 채점의 차이가 채점자의 전공에 따른 것이 아니라는 것을 뜻한다. 따라서 채점자 간 일치도를 높이기 위해서는 채점자를 의사나 표준화환자로 제한하는 것 이외의 방법이 필요함을 이 연구결과는 제시하고 있다.

일반화가능도 이론은 신뢰도에 영향을 미치는 오차를 분석하기에 적합하다. 본 연구에서는 각각의 분산성분을 분석하여 신뢰도에 미치는 영향을 파악해보았다. 이 결과로, 어떤 분산성분에 주의를 기울임으로써 신뢰도를 높일 수 있는지 알 수 있다.

분산성분이 높았던 것은 문항 종류, 문항, 시험방 등 채점자와는 관련 없는 분야들이었다. 이는 채점의 오차가 주로 문항 종류나 문항, 시험방 사례 등에 의해 발생하였음을 의미한다. 즉, 채점자 자체에 의해 발생한 오차보다는 문항과 시험방 사례와 관련된 오차가 더 컸다. 이 연구결과는 문항 종류를 다양화 하고, 문항 수를 적절히 배정하며, 시험방 수를 늘린다면, 절차기술 객관구조화 절차기술 평가에서도 표준화환자를 채점자로 활용할 수 있음을 알려준다. 채점자가 절차기술의 전문가인지 아닌지는 신뢰도에 큰 영향을 미치지 못하였다. 표준화환자가 채점하기 쉬우려면 채점표가 잘 개발되어야 하며 적절히 훈련되어야 한다. 전문가들은 객관적이고 명확하며 행동 중심적이며, 의료인이 아닌 사람이 이해하기 쉬운 용어로 쓰인, 사례 특이적인(case specific) 채점표를 사용할 것을 권하고 있다[24]. 본 연구에서도 표준화환자의 용이한 채점을 위해 이해하기 쉬우며, 객관적이고 명확한 용어를 사용하여 행동 중심적인 채점표를 구성하였다. 본 연구의 신뢰도와 일치도가 높다는 것은 본 연구에서의 채점표 개발이 성공하였음을 말해준다.

전체 신뢰도 분석 결과, 채점자에 관계없이 문항 종류에 대한 분산성분이 여러 분산성분 중 가장 많은 비중을 차지하였다. 이는 체크리스트 문항의 점수와 총괄평가 문항의 점수 간에 차이가 많이 있음을 뜻한다. 이 결과로부터 우리는 체크리스트 문항과 총괄평가 문항을 조합함으로써 보다 다

양한 측면에서 평가를 할 수 있음을 알 수 있다. 즉, 총괄 평가는 사례 특이적이지는 않지만, 체크리스트에서 빠질 수 있는, 학생들의 전반적인 능력을 평가하는데 주로 사용될 수 있다[5].

체크리스트 문항에서의 분산 성분을 분석함으로써 체크리스트 채점의 신뢰도를 더 높일 방법을 강구할 수 있다. 체크리스트 문항의 분산성분에 영향을 미치는 것은 학생, 문항, 그리고 채점자 간의 상호작용이 컸다. 반면 문항 자체, 채점자(표준화환자인지 의사인지), 학생과의 상호작용, 시험방과의 상호작용 등은 영향을 미치지 못하였다. 즉 체크리스트 문항을 개발할 때는, 채점자 각자가 학생을 평가함에 있어 문항에 대해 다르게 생각할 수 있다는 것을 알고 이에 대비하는 것이 중요하겠다. 이는 문항 자체가 잘 개발되었다 하더라도, 학생들이 절차 기술을 수행하면서 채점자가 문항에서 미리 예측하지 못한 다양한 행위를 하게 되고, 이를 채점에 반영하는 것이 쉽지 않다는 것을 뜻한다. 즉, 절차기술 문항을 개발함에 있어서는 문항 자체의 명확한 기술도 필요하지만, 그에 못지않게 명확하고 자세한 채점 기준과 돌발 상황에서 어떻게 대처하고 채점기준을 적용할 것인지에 대한 구체적인 설명이 있어야한다는 것이다. 또한 돌발 상황에 익숙해지기 위해서는 다양한 상황을 예측해보고 미리 채점 훈련을 해보는 것이 도움이 될 것으로 생각된다. 연구 결과, 채점자그룹 간에는 차이가 없었다는 것은, 이러한 채점의 오류가 의사냐, 표준화환자냐에 따라 달라지는 것이 아니라, 개개의 채점자의 경험이나 속성에 따라 달라질 수 있음을 뜻한다. 따라서 채점자를 선정함에 있어 돌발 상황에서의 대처나 채점기준에 대한 적용에 유연한 특성을 가진 채점자를 선택하는 것도 신뢰도를 높이는 한 방법일 것으로 생각된다.

총괄평가 문항에서의 분산성분에 영향을 미치는 것은 체크리스트와 약간 달랐다. 총괄평가 문항에서는 전반적으로 시험방과 관련된 요인들이 주요 분산성분에 배치되었다. 이는 총괄평가 문항이 시험방 별로 1개씩만 주어진 것도 영향을 미쳤다고 생각된다. 이 결과는 시험방 수 또는 총괄평가 문항 수를 늘림으로써 총괄평가의 신뢰도를 높일 수 있음을 의미한다. 그러나 전체 신뢰도 분석에서는 시험방의 분산성분이 0.000 미만이었다. 즉, 총괄평가의 신뢰도를 높이기 위해서는 시험방 수를 늘리는 것이 필요하지만, 전체 신뢰도를 높이는 데는 영향을 미치지 않음을 알 수 있었다.

본 연구의 제한점은 첫째, 총괄평가 문항을 1개만 사용하여, 다수의 총괄평가 문항이 있을 때의 경우를 분석할 수 없었다는 점이다. 둘째, 체크리스트 문항과 총괄평가 문항이 함께 있었기에 D 연구는 시행하지 못하였다. D연구(Design study)란 일반화 가능도 이론에서, 이미 계산된 분산성분을 이용하여 역으로 적정 수준(0.80)의 일반화가능도 상수를 얻을 수 있는 조건을 유추해보는 연구를 의미한다[26]. 즉, 적절한 문항의 개수나 시험방의 수, 채점자의 수에 대한 연구는 하지 못하였다. 셋째, 2개의 사례만을 이용하여 객관구조화 평가를 시행하였기에 사례 자체의 특성이 결과에 영향을 미쳤을 수 있다. 앞으로 더 다양한 사례를 이용하여 시행한 객관구조화 절차기술 평가에서 체크리스트나 총괄평가 문항의 세부적인 분석을 통해 결과를 일반화하는 것이 타당할 것이다. 넷째, 경험이 부족한 표준화환자와 비교하지는 못하였다. 이 역시 차후 연구가 더 필요할 것으로 생각된다.

결론적으로 객관구조화 절차기술 평가에서 잘 훈련된 표준화환자는 교수의 채점과 비교하여 비슷한 수준의 신뢰도와 일치도를 보였다. 따라서 표준화환자는 객관구조화 절차기술 평가에서 채점자 역할이 가능할 것으로 조심스럽게 판단된다. 채점자 그룹이 교수인지의 여부보다는 문항, 시험방 등이 신뢰도에 영향을 미쳤으므로 객관구조화 절차기술 평가의 신뢰도를 높이기 위해서는 채점자를 전공자로 선정하는 것보다 문항을 잘 개발하고 다수의 시험방을 활용하는 것이 더 필요한 것으로 생각된다.

참고문헌

[1] Barrows HS, "An overview of the uses of standardized patients for teaching and evaluating clinical skills", Acad Med, 68(6), pp. 443-451, 1993.

[2] Barrows HS, Abrahamson S, "The programmed patient: a technique for appraising student performance in clinical neurology", J Med Educ, 39, pp. 802-805, 1964.

[3] Lee B, "Recent world trend in performance-based assessments and application of the standardized patient program in Korean medical education", Korean J Med Educ, 12(2), pp. 377-392, 2000.

[4] Stillman PL, Ruggill JS, Sabers DL, "The use of practical instructors to evaluate a complete physical examination", Eval Heal Prof, 1, pp. 49-54, 1978.

[5] Regehr G, Freeman R, Robb A, Misshiha N, Heisey R, "OSCE performance evaluations made by standardized patients: comparing checklist and global rating scores", Acad Med, 10(Oct suppl), S135-S137, 1999.

[6] Whelan GP, Boulet JR, McKinley DW, Norcini JJ, van Zanten M, Hambleton RK, Burdick WP,

Peitzman SJ, "Scoring standardized patient examinations: lessons learned from the development and administration of the ECFMG Clinical Skills Assessment (CSA□)", Med Teach, 27(3), pp. 200-20, 2005.

[7] Martin JA, Reznick RK, Rothman A, Tamblyn RM, Regehr G, "Who should rate candidates in an objective structured clinical examination?", Acad Med, 71, pp. 170-175, 1996.

[8] Stillman PL, "Technical issues: Logistics", Acad Med, 68, pp. 464-468, 1993.

[9] Peggy Wallace, "Coaching Standardized Patients for Use in the Assessment of Clinical Competence", Springer Publishing Company. 2007.

[10] Harden RM, Stevenson M, Downie WW, Wilson GM, "Assessment of clinical competence using objective structured examination", BMJ, 1, pp. 447-451, 1975.

[11] Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M, "Objective structured assessment of technical skill (OSATS) for surgical residents", Br J Surg, 84, pp. 273-278, 1997.

[12] Goff BA, Lentz GM, Lee D, Houmard B, Mandel LS, "Development of an objective structured assessment of technical skills for obstetric and gynecology residents", Obstet Gynecol, 96, pp. 146-150, 2000.

[13] Morgan PJ, Cleave-Hogg D, Guest CB, "A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator", Acad Med, 76(10), pp. 1053-1055, 2001.

[14] Bould MD, Crabtree NA, Naik VN, "Assessment of procedural skills in anaesthesia", Br J Anaesth, 103, pp. 472-483, 2009.

[15] Friedlich M, Wood T, Regehr G, Hurst C, Shamji F, "Structured assessment of minor surgical skills (SAMSS) for clinical clerks", Acad Med, 77(10), S39-S41, 2002.

[16] Lammers RL, Davenport M, Korley F, et al, "Teaching and assessing procedural skills using simulation: metrics and methodology", Acad Emerg Med, 15, pp. 1079-87, 2008.

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

[18] Friedman Z, Siddiqui N, Katznelson R, Devito I, Davies S, "Experience is not enough: repeated breaches in epidural anesthesia aseptic technique by novice operators despite improved skill", Anesthesiology, 108, pp. 914-920, 2008.

[19] Newble DI, Hoare J, Sheldrake PF, "The selection and training of examiners for clinical examinations",. Med Edu, 14(5), pp. 345-349, 1980.

[20] Heine N, Garman K, Wallace P, Bartos R, Richards A, "An analysis of standardized patient checklist errors and their effect on student scores", Med Educ, 37, pp. 99-104, 2003.

[21] Humphrey-Murto S, Smee S, Touchie C, Wood TJ, Blackmore DE, "A comparison of physician examiners and trained assessors in a high-stakes OSCE setting", Acad Med, 80(10), s59-s62, 2005.

[22] Downing SM, "Reliability: on the reproducibility of assessment data", Med Educ, 38, pp. 1006-1012, 2004.

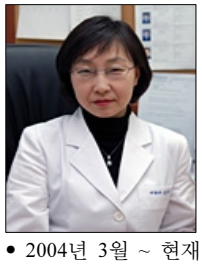
[23] 성태제, "타당도와 신뢰도", 학지사, pp 137-163, 2002.

[24] 강 승호, "신뢰도", 교육과학사, pp 73-80, 2004.

[25] Bullock G, kovacs G, Macdonald K, Story BA, "Evaluating procedural skills competence: inter-rater reliability of expert and non-expert observers", Acad Med, 74(1), pp. 76-78, 1999.

[26] Crick JE, Brennan RL, "Manual of GENOVA: A generalized analysis of variance system", American College Testing Program, USA, 1983.

손 희 정(Hee-Jeong Son) **[정회원]**



- 1993년 2월 : 이화여자대학교 의과대학교 대학원 의학과 (의학석사)
- 1996년 8월 : 이화여자대학교 의과대학교 대학원 의학과 (의학박사-마취통증의학)
- 2000년 ~ 2001년 : 강원대학병원 임상교수
- 2004년 3월 ~ 현재 : 강원대학교 의전원 마취통증의학과 교수

<관심분야>
소아마취, 기도관리, 의학교육

노 혜 린(Hye-Rin Rho) **[정회원]**

- 1993년 2월 : 서울대학교 의과대학(의학사)
- 2000년 2월 : 서울대학교 의과대학(의학석사)
- 2009년 2월 : 서울대학교 의과대학(의학박사)
- 2000년 3월 ~ 현재 : 강원대학교 의전원 외과학 교수

<관심분야>
간이식, 의학교육

문 중 범(Joong-Bum Moon) **[정회원]**

- 1997년 2월 : 연세대학교 원주의과대학 의학과 (의학사)
- 2006년 3월 ~ 현재 : 강원대학교 의과대학 의학과 교수

<관심분야>
소생술, 노인의학

이 향 아(Hyang-Ah Lee) **[정회원]**

- 1999년 2월 : 연세대학교 원주의과대학 의학과 (의학사)
- 2005년 8월 : 연세대학교 의과대학 의학과 석사 (의학석사)
- 2008년 8월 : 울산대학교 의과대학 의학과 박사(의학박사)
- 2007년 3월 ~ 현재 : 강원대학교 의과대학 의학과 교수

<관심분야>
불임, 성폭력에방