

객관구조화진료시험(OSCE; Objective Structured Clinical Examination)의 임상적 활용

한양대학교 의과대학 가정의학교실

박 훈 기

Clinical Application of Objective Structured Clinical Examination (OSCE)

Hoonki Park, MD, PhD

Department of Family Medicine, Hanyang University Medical School

서 론

의과대학의 공통 교육목표는 졸업생들이 의학교육을 마친 후 일차의료를 수행하는 데 필요한 자질을 갖추게 하는 데 있다. 최근 의학교육의 추세는 능력 (competency)바탕 교과과정과 평가 방법을 강조하고 있다 (Small et al, 1993). 능력을 측정하는 한 가지 방법이 바로 객관구조화진료시험 (Objective structured clinical examination; OSCE)이다. Harden 등이 1975년 처음으로 지식, 태도, 수기 등을 동시에 측정할 수 있는 시험의 형태로 소개한 이후 OSCE는 의학교육에서 평가와 교육 방법으로서 발전해왔다 (Harden et al, 1975). LCME (Liaison Committee on Medical Education)의 2002~2003 미국 의대 교과과정 실태조사에 의하면 전체 126개 의대 중 77%가 임상의학입문과정 중 병력청취와 신체진찰 교육에서, 55%는 3, 4학년 최종임상종합시험의 형태로 OSCE / SP (Standardized patient)를 한 번 이상 시행하고 있다 (Barzansky & Etzel, 2003). 필수

임상 실습 교육 과정 중에 OSCE/SP의 활용정도는 교과목별로 보면 가정의학, 내과, 산부인과, 소아과, 정신과, 외과, 신경과의 순으로 높았다. 임상종합시험에서 스테이션의 수는 5개부터 15개 이상까지 다양하였고 측정영역은 병력청취, 신체진찰, 정보통합, 의사소통능력 등의 네 영역에서 모두 90% 이상으로 일정하게 높은 사용빈도를 보였다. 우리나라 의대의 OSCE이용 실태에 대한 정확한 보고는 아직 없지만 1994년 서울의대에서 OSCE를 실시한 이래 (안효섭, 1995) 새로운 임상수행평가방법으로 많은 의과대학이 이미 OSCE를 형성평가 혹은 총합사정으로 시행하고 있다. 캐나다에서는 2년간의 예비시험을 거쳐 1994년도부터 OSCE가 의사국가고시에서 이차실기시험으로 시행되고 있다 (Reznick et al, 1993). 평가는 교과과정의 변화를 주도하기 때문에 최근 의사국가고시에서 임상실기시험을 도입하려는 구체적인 계획은 앞으로 OSCE 형태의 시험이 우리나라에서도 더욱 확대될 전망이다 (김병수 등, 2001; 이영미 등, 2001).

교신저자: 박훈기, 한양대학교 의과대학 가정의학교실
 서울특별시 성동구 행당동 17번지
 Tel: 02)2290-8740, Fax: 02)2281-7279
 E-mail: hoonkp@hanyang.ac.kr

Table I. 객관구조화진료시험의 개발 단계

1. 평가하고자하는 기능 (skills)을 결정
2. 필요한 스테이션 수를 결정
3. 기능 평가
4. 채점 계획 수립
5. 시험 공간 확보
6. 필요 인력 확보: 평가자, 집행요원, 시간관리자, 표준화 환자
7. 시험 당일 배치
8. 결과 검토

이에 저자는 OSCE의 개념, 운영 전략, 신뢰도, 타당도, 점수 기준 설정을 중심으로 OSCE의 임상적 활용에 필요한 지식을 점검해 보고 아울러 OSCE의 제한점을 지적하고자 한다.

가. OSCE의 개념

OSCE는 임상능력평가시험의 한 형태로 수험생이 여러 개의 5~10분 단위 스테이션을 돌면서 각 스테이션에서 요구하는 특정한 수기 혹은 일련의 수기들을 수행하고, 1~2명의 평가자는 미리 정해진 체크리스트에 따라 객관적으로 채점을 하는 실기시험이다 (Harden et al, 1975; Newble, 2004). Fig. 1은 8개 방으로 구성된 OSCE의 한 예를 보여주고 있다. 하지만 이러한 고전적인 OSCE의 유형과 조금 다른 여러 가지 변형이 생겨났다. 예를 들면 스테이션 당 시간을 더 늘리거나 모의환자가 직접 채점을 하기도 한다. 선다형 문제나 지필고사를 OSCE에 삽입하는 것도 또 다른 변형이다 (Williams et al, 1999).

OSCE는 전통적인 임상 평가방법인 수술고사나 실제 환자를 통한 평가자 관찰시험과는 달리 채점 기준으로 구체적 체크리스트가 있기 때문에 평가자 별 차이를 없애고 객관적으로 평가할 수 있다는 강점이 있다 (Harden et al, 1975). 체크리스트를 만들 때 예상 답은 시험 시행 전에 미리 준비해야 하고 이는 가장 최소 단위의 하위기능 (subskills)인 문항 (items)으로 나뉘어야 한다. 각각의 문항은 중요도에 따라 가중치를 둘 수도 있다. 문항은 그 속성을 측정 영역 (domain)과 기능 (skills)으로 분류하여 행렬표 (matrix)를 만들어 볼 수 있다. 예를 들면 응급질

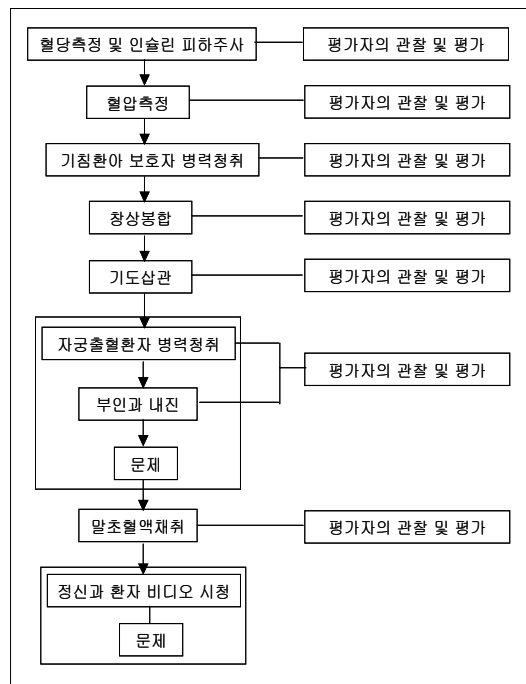


Fig. 1. 여덟 개 방으로 이루어진 객관구조화진료시험 구성

환의 의사소통기술을 측정하는 어떤 한 문항을 행렬표에 넣어 보면 어느 분야를 측정하고자 하는지 다른 문항과 더불어 전체 시험 내용의 균형을 점검해 볼 수 있다.

OSCE는 원칙대로 제대로 시행하면 학생의 수행 능력 정도와 실습교육의 질을 알 수 있게 해준다 (Sloan et al, 1995). OSCE 결과에 대한 피드백은 스테이션이 끝난 직후 중요한 점 한 두 가지를 평가자

Table II. 고관절 진찰의 시험 문항 초안(예)

수험생에게 지시사항: 환자의 왼쪽 고관절을 진찰하고 소견을 말하십시오.
준비 사항: 표준화환자

평가표 초안

정확한 시기	점수
1. 피부주름의 대칭성	1
2. 다리를 잡는 손의 자세	1
3. 고관절 외전	1
4. 후방 탈구 시도	1
5. 진단: 고관절 탈구	1

Table III. 의과대학 4학년의 임상수행능력을 평가하기 위한 OSCE의 시험영역 및 능력군

영역	병력 청취	신체 진찰	검사 요구, 시기
심혈관계 / 호흡기	X	X	X
이비인후과			X
산부인과	X		
정신건강	X		
비뇨기계			X
근골격계		X	
소화기계		X	
혈액종양계			X
신경계		X	

가 즉각적으로 학생에게 피드백해 주거나 나중에 성취도를 서면보고를 통해서 할 수 있다. 스테이션 수가 적고 OSCE를 교육적인 목적으로 활용할 때는 표준 모델을 담은 비디오를 통해 학생에게 스테이션이 끝난 후 바로 피드백을 줄 수 있다. OSCE 성적 자체는 그 문제에 해당하는 실습교육과정의 교육성파로 바로 피드백될 수 있으며 해당 교과목에서는 학생들의 성취도를 보고 실습교육을 짚을 개선할 수 있다.

나. OSCE 운영

OSCE를 임상적으로 어떻게 운영하는가는 학생 수, 지원의 정도, 교육과정의 특성에 따라 달라진다 (Selby et al, 1995; Smee, 2003). 지금부터는 OSCE를 임상실습교육에서 평가 목적으로 운영하고자 할

때 몇 가지 유의해야 할 사항을 단계별로 설명하고자 한다 (Tervo et al, 1997). Table I은 OSCE의 개발단계를 보여 주고 있다.

1) OSCE는 미리 측정하고자했던 문제를 출제해야 한다.

임상실습 교과과정 혹은 과목별 학습목표를 기준으로 중요한 것을 골라 시험 문항으로 선정해야 한다. 졸업 전 임상종합시험의 일환으로 시행하는 OSCE는 일차의료를 수행하는데 필수적으로 요구되는 임상능력을 평가해야 한다. 어느 한 교과목에 국한하지 않고 여러 과목을 종합하는 임상술기와 임무수행능력을 측정해야 한다. Table II는 OSCE시험의 한 스테이션의 시험 문항을 예로 들었다.

Table IV. OSCE의 문제 스테이션 예

A. 수기 스테이션: 복통	
관찰자 없음	
병력과 신체 진찰 소견이 적힌 종이	
수험생에게 지시한 사항: 앞의 스테이션에서 본 환자의 병력 및 신체 진찰 소견에서 빠진 것을 모두 나열하시오.	
B: 수기 스테이션: 요검사	
두 개의 완충 용액이 있음 - 하나는 혈액이 섞여 있고, 하나는 포도당과 단백질이 섞여 있음.	
장갑, 휴지통, 요검사 시험지가 있음	
평가자 유의 사항: 학생에게 수기를 시행하면서 자신이 무엇을 하고 있는지와 이상 소견을 큰 소리로 이야기하게 하시오.	
수험생 유의 사항: 이 두 가지 소변 검체를 검사해 보시오.	
	점수
전반적인 과정: 장갑을 끼고 사용한 뇨검사 시험지와 장갑을 제대로 버림	2, 1, 0
소변을 눈으로 관찰	1
요검사 시험지를 꺼낸 후 마개를 꼭 닫음	1
요검사 시험지를 제대로 사용. 판독 방향, 시간이 적절.	2, 1, 0
피, 당, 단백질의 존재 유무를 바르게 해석.	3

2) OSCE는 학습 성취도의 수준을 구별해 주어야 한다.

시험의 타당도를 유지하려면 교육과 평가가 학습자 수준과 일치해야 한다 (Wilkes & Bligh, 1999). 예를 들어 임상종합시험에서는 병력청취, 진찰능력, 자료해석, 문제해결능력, 응급질환처치능력, 면담 및 의사소통기술 등 영역별로 의과대학 3~4학년에 교육했던 능력군이 OSCE 내용에 골고루 포함되어야 한다. 이 밖에도 OSCE는 학생의 의사로서 태도와 윤리적 자세도 측정할 수 있어야 한다. 스테이션의 난이도는 수험생의 수준에 따라 적절하게 조정되어야 한다.

3) OSCE 시험 내용을 포함하는 청사진 (test blueprint) 만들어 보아야 한다.

시험 청사진은 시험 영역 (content)을 제시하며 시험의 심도와 포괄성을 한눈에 알아 볼 수 있게 한다. Table III은 학습목표에 근거한 시험 청사진의 한 예를 보여주고 있다.

4) 시험 계획

a. 시험의 목적

OSCE 목적이 형성평가인가 총합평가인가에 따라 시험의 구성이 달라진다. 형성 평가인 경우 각 스테이션별로 구조화된 평가표가 있어야 하고 이에 따라 평가자가 학생 능력을 평가한다. 각 스테이션의 학생 임무가 끝나면 평가교수는 학생의 평가표를 검토해보고 이상적인 수준은 어느 정도여야 하는지를 학생에게 조언 해준다. 전체 점수는 교육자에게 다음 OSCE에서 어느 영역에 치중할 것인지와 어떤 스테이션을 또 출제하고 어떤 스테이션은 새로 만들어야 하는지를 결정하는데 기초 자료를 제공한다. 반면 총합평가에서는 즉각적인 피드백 보다는 학생의 수행능력의 당락판정과 프로그램의 질에 대한 평가가 주목적이 된다.

b. 시험 스테이션의 난이도

시험 결과 판정 기준은 시험 시행 전에 어떤 점수에서 통과를 시킬 것인가 미리 결정해 놓는 준거기준방법 (criterion-referenced)과 시험 성적의 분포에 따라서 학생들의 당락을 결정하는 규준기준방법 (norm-

referenced)으로 나눌 수 있다. 학생들의 예상득점 수준에 따라 스테이션별 난이도를 조정하는데, 예를 들어 의과대학 2학년 진단학에서 시험 문항은 신체 진찰에서 필요한 최소 요구 능력군을 위주로 하며, 의과대학 3학년 실습 후 OSCE에서는 주요 핵심과목의 실습을 마친 학생들이 보여 주어야 할 필수 수행능력을 중심으로 시험문항을 구성한다. 단순암기 수준을 넘어 실제 임상 문제나 새로운 상황에 배운 지식을 응용하게 하거나 기준바탕 점수산정 방법을 적용하면 OSCE의 난이도와 변별력이 올라간다.

c. 수험생 수와 스테이션 수

몇 명의 학생이 OSCE를 치르는가에 따라 스테이션 수, 문제 수, 문제 유형, 일정, 복제 수, 시험 장소가 결정된다. 125명의 학생을 대상으로 시행하는 OSCE의 경우 세군 데로 시험장을 복제하고 시간차를 두고 2부로 나누어 운영하려면 각 고사장 별로 21개 스테이션을 마련해야 하며 이 중 3~4개의 휴식 스테이션을 두면 적절한 시험문제 수는 15~17개 정도가 된다. 일반적으로 OSCE의 효율성이나 타당성을 적절히 유지하려면 15개 정도(범위: 10~20)의 스테이션은 되어야 한다(Selby et al, 1995). 스테이션 수가 10개 미만이면 시험의 주제를 충분히 고를 수가 없고 20개가 넘으면 운영이 힘들고 평가교수의 수가 많아져 동원이 어려워진다. 필요에 따라 중간에 휴식 스테이션(rest station)을 둘 수 있으며, 두 스테이션을 연결하여 학생들이 시간을 두 배로 사용하게 해주는 연결스테이션(link station)을 만들 수도 있다. 시험장을 두개 세트 이상으로 복제하는 경우에는 표준화 환자의 훈련 뿐 아니라 평가자의 훈련도 동시에 충분하게 이루어져야 신뢰도를 유지할 수 있다(박훈기 등, 1999).

d. 일정

공간, 운영요원, 평가자, 시간기록원, 학생 수 등이 시험 일정을 결정한다. 스테이션별 시간은 3~10분 범위 안에서 합리적인 선택을 해야 한다. 일반적인 교육과정에서 시험이나 심리적 측정의 신뢰도는 0.8을 넘어야 한다. 표준화환자를 이용한 시험의 신뢰도는 0.41~0.85정도의 범위에 있다고 보고하고 있

다. 전체 시험시간이 표준화환자를 이용한 시험의 신뢰도와 관련이 있는데 3~4시간이 적절한 시간이라고 제시한 연구가 있다(Colliver & Williams, 1993). 하지만 아직까지 OSCE 형식의 시험에서 정확하게 어느 정도의 스테이션 수와 스테이션당 시간이 적절한지 밝혀진 바는 없다.

OSCE의 시험 시간에는 스테이션 간 이동시간, 휴식 시간, 수험생의 등록/오리엔테이션 시간 등이 함께 고려되어야 한다. 스테이션을 이동하는 시간은 10~30초 정도를 할당하는 것이 좋고 이동시간이 특히 길어지는 구간은 중간에 휴식 스테이션을 두는 것이 바람직하다.

e. 시험의 비밀유지

OSCE에서 동일한 문제를 연속하여 사용하는 경우 시험성적에 미치는 영향을 본 연구에서 적어도 1년 이상 간격을 두고 동일한 문제를 사용했을 때 성적에 미치는 영향은 거의 없었다(Cohen et al, 1993). OSCE의 문제에 대한 학생들 사이의 정보교환과 이에 따르는 시험성적의 변화는 거의 없는 것으로 알려져 있다(Rutala et al, 1991; Colliver et al, 1991; Vu et al, 1992). 환자의 문제를 알고 있더라도 체크리스트를 모르면 기능향상에는 큰 도움이 되지 않는다. 그러나 자세한 정보교환이 학생들 사이에 일어나면 시험 성적에 어떤 효과를 나타내는 지 아직 밝혀진 바는 없다. OSCE에 대한 시험공부는 성적향상에 큰 도움은 되지 않고 오히려 이전까지 축적된 학업성적이 OSCE와 상관성을 보인다고 알려져 있다(Mavis, 2000). OSCE의 문제 내용이 학생실습교육의 학습목표와 일치한다면 사전실습교육시간은 OSCE 성적에 큰 영향이 없다고 알려져 있다(Thomas & Shatzer, 2000). 임상의학입문과정에서의 경험 역시 OSCE 성적과 상관성이 0.2이하로 매우 낮은 것으로 알려져 있다(Jolly et al, 1996). 하지만 외과영역의 수기에 관한 경험과 피드백의 양은 OSCE 성적을 향상시킨다는 보고도 있다(Chatenay et al, 1996). 수험생간 문제의 사전 유출을 해결하기 위해 OSCE와 슬라이드 시험 혹은 지필고사를 직렬적으로 연결하여 운영하기도 한다. 남일리노이의대에서

는 매년 50~70%씩 새로운 문제로 바꾸어 임상실습종합평가로서 표준화환자를 이용한 수행시험을 시행하고 있다(Vu et al, 1992).

5) OSCE 문제(question)의 유형

OSCE 문제는 객관적인 문항과 구성적인 문항(constructive items)으로 만들 수 있다. 객관적인 문항은 여러 가지 답가지에서 옳은 것을 고르기, 단답형 혹은 완성형으로 만들 수 있다. 구성형 문항은 단답형/완성형, 논술형, 문제해결형으로 만들 수 있다. Table IV는 객관식과 주관식의 문항의 예를 보여 주고 있다.

6) OSCE 시험 당일

학생과 평가 교수에게 각각 오리엔테이션을 실시해야 한다. 학생 개인 별로 시험장의 지도를 나누어 주고 스테이션별로 주어진 시간, 메모 기록 가능여부, 스테이션별 점수, 수험 요령 등을 간단명료하게 설명해 주어야 한다. 학생이 필히 지참해야 할 물품과 가지고 오면 안 되는 것을 미리 알려 주고 명찰을 배부한다. 스테이션의 끝을 알리는 종과 예령의 규칙을 알려 준다. 이동 시간을 30초 정도 주고 시작과 끝의 종만을 울려주는 운영방법도 있다(Tervo et al, 1997). 학생들이 모의환자를 실제 환자로 대할 것을 강조해 준다. 만약 피드백이 예정되어 있으면 언제 어떻게 제공되는지를 알려 준다.

7) OSCE 결과 피드백

OSCE는 학생 뿐 아니라 교육자에게 형성평가로서 이용될 수 있다. 학생 각자에게 관찰자인 교수가 직접 각 방마다 피드백을 줄 수 있다. 즉각적인 피드백은 다음에 이어지는 스테이션에서 학생능력을 향상시킬 뿐 아니라 수험생의 학습경험의 질을 높이는 효과가 있다(Hodder et al, 1989). 그룹 별 수행능력에 대한 피드백은 교육프로그램에서 부족한 점을 찾아 낼 수 있다(Elliot et al, 1994).

OSCE에서 피드백은 스테이션을 마친 후 즉각적으로 시행하는 것이 가장 효과적이라고 알려져 있다. 이상적인 피드백 과정은 전문가의 진료 시범을

녹화한 모델 비디오를 보고 스스로 판단하게 하거나 아니면 평가자가 직접 평가표를 근거로 장점과 발전방향을 동시에 조언해 줄 수 있다. OSCE 도중 각 스테이션별로 짧게 피드백을 해주는 경우에는 스테이션 종료 30초~1분 전에 해당 학생에게 평가자가 직접 관찰한 사항을 근거로 중요한 점 한 두 가지를 피드백 해준다.

다. OSCE의 타당도

OSCE의 내용 타당도를 높이려면 다음 세 가지 단계로 접근해야 한다(Newble, 2004). 제 1단계는 수험생이 제대로 수행할 수 있어야 하는 문제(problems) 혹은 상황(conditions)을 찾아내는 일이다. 전문가 집단의 의견을 종합하거나 좀 더 체계적인 연구를 통해서 이 작업을 수행할 수 있다. 한 명의 출제자 보다는 여러 명의 전문가가 평가항목을 선정하는 것이 중요한 문항 선정에 더 바람직하다(Valentino et al, 1998). 두 번째 단계는 앞의 문제 혹은 상황별로 수험생이 꼭 할 수 있어야 하는 임무(tasks)를 결정하는 단계이다. 예를 들어 가슴 통증이라는 문제를 놓고 협심증 환자로부터 병력청취, 심전도를 찍고 해석하기, 심폐소생술 시행하기, 협심증약 복용이나 식이요법에 대한 환자교육하기 등이 임무에 해당한다. 시험청사진 혹은 격자를 작성해 보는 것이 제 3단계 작업이다. 이 과정은 시험에 포함시킬 문항(items)의 포본을 정의하는 단계이다. 한 축에는 고유능력(generic competencies, 예; 병력청취, 의사소통기술, 신체진찰, 검사, 처치), 다른 한 축에는 수험생이 능력을 보여 주어야 하는 문제 혹은 상황을 담은 이차원 행렬표를 작성한다. 영국에서 외국 졸업자를 대상으로 시행한 의사자격시험에서 세로축에는 병력청취, 신체진찰, 의사소통술, 진료수기, 응급진료의 5개 영역이 있고 가로축에는 장기시스템별로 15개 분야로 나누어 좀 더 세분화된 시험 문항의 청사진을 만들어 사용했다(Tombsleson et al, 2000).

OSCE의 타당도를 논할 때 시험에 관련된 수 있는 문화, 종교, 경제, 혹은 다른 중요한 행위동기유발요인을 고려하는 것이 바람직하다(Hodges 2003). 예를 들어 우울증 환자의 10분 소요 증례를 다루는

문제에서 내용 타당도 (content validity)의 정도를 말할 때 가정의학과 임상실습 학생과 전공의 혹은 정신과 실습학생과 전공의의 능력정도를 비교하는 시험인가에 따라 OSCE의 타당도는 다른 의미를 가진다. 그래서 타당도를 논할 때는 누구를 대상으로 하는 시험이며 어떤 배경에서 시행하는 시험인지를 알고 시작해야 한다. OSCE의 스테이션은 대개 팀 접근 진료를 평가하지는 않는다. OSCE는 수험생이 기본적인 단순한 병력 청취, 신체 진찰을 잘 하면 실제 임상 상황에서도 훌륭한 능력을 발휘할 것이라는 가정을 하고 시작하는 시험이라고 할 수 있다.

의과대학에서 OSCE를 도입하고 의사국가시험에서 OSCE를 계획하고 있는 것 자체만으로 의과대학의 교육과정에 지대한 변화를 가져올 수 있다. 그렇다면 이미 출제된 문제를 그대로 다음 번에 출제하는 경우 과연 원래 의도했던 대로의 능력군을 변함 없이 측정할 수 있는가가 의문시된다. OSCE의 스테이션이 무슨 과에서 출제되었느냐 혹은 어떤 과목의 교수가 평가자인가에 따라 똑같은 표준환자와 면담 과정 중 선택하는 행동반응 유형이 달라질 수 있다. 반면 OSCE에서 보여준 반응 행동이 다음에 임상에서 일정한 행동으로 계속 유지될 것인지 확인할 수가 없다. 따라서 OSCE의 타당도 연구에서는 기존의 정량적인 연구를 넘어서 질적 연구가 추가적으로 더 필요하다.

라. OSCE 신뢰도 확보

OSCE처럼 임상수행능력을 평가하는 시험은 증례와 문제 수를 증가시켜야 하고 적어도 4~8시간 정도로 시험시간이 길어져야 OSCE가 신뢰도를 유지할 수 있다고 한다 (Vu et al, 1992; Colliver & Williams, 1993). OSCE의 객관성이란 평가자가 적용할 수 있는 평가표에 나열된 임무와 채점 기준이 표준화가 되었다는데서 비롯된다 (Harden et al, 1975). 이러한 객관적 채점 기준은 평가자간 신뢰도를 향상시킨다. 표준화환자가 채점을 할 때 채점자간 일치도는 90%까지 보고된바 있으며 일반적으로는 80%이상의 수준을 보인다고 한다 (Vu et al, 1992). OSCE에서 평가자간 채점 일치도 Kappa 지수는 기도 삼관 문제

에서 0.65, 통합문제에서는 0.71로 나왔다 (O'connor & McGraw, 1997). 한편 채점자간의 훈련부족은 신뢰도를 저하시키는 중요한 요인으로 지적받고 있다 (박훈기 등, 2003).

일반화 (generalization) 이론에 입각한 연구결과를 보면 평가자간 일치도보다 더 중요한 것이 바로 증례 특이성 (case specificity)이다 (Newble, 2004). 이러한 증례 특이성이란 현실적 문제를 줄이기 위해 몇 가지 대안이 제시되었다. OSCE와 다른 형식의 시험을 접목하는 것이 한 가지 방법이다 (Wass et al, 2001; Verhoeven & Hamers, 2000). 예를 들어 90분 분량 OSCE는 신뢰도가 0.6으로 낮았지만 여기에 90분짜리 필기시험을 추가하면 신뢰도가 0.8로 올라갔다.

다른 한 돌파구는 평가 척도를 혼합하는 방법이다 (Newble, 2004). OSCE 평가표를 만들다 보면 체크리스트가 신뢰성 있게 자세한 항목으로 만들어지긴 하지만 중간에 측정이 곤란하지만 중요한 항목은 고의적으로 빠져 실제로 수행능력을 평가하는데 있어서는 질이 떨어지는 평가표로 전락할 수 있다. 체크리스트를 이용한 평가에 비하여 총괄채점 (global rating)의 우수성이 최근 각광을 받고 있다 (Cohen et al, 1996; Regehr et al, 1998; Hodges & McIlloy, 2003). 일반적으로 기술적인 요소가 많이 가미된 수기 스테이션은 체크리스트 척도가 적당하고 의사소통기술이나 진단 해결형 스테이션의 경우에는 총괄채점이 적절하다고 한다. 체크리스트 점수는 스테이션 사이에 그다지 큰 연관성을 보이지 않지만 총괄채점은 스테이션간에 서로 중등도의 높은 상관성을 보였다 (Regehr et al, 1999). 특정한 수기를 평가할 때는 체크리스트를 적용하고 환자교육이나 의사의 임무에 전반적인 접근방법에 관해서는 총괄채점척도를 사용하여 이 두 가지 척도를 적절하게 섞어서 사용하는 것이 OSCE의 신뢰도를 높이는 방법이다.

마. 기준 설정 (standard setting)

일반적으로 기준 설정에는 상대 평가인 분포근거 방법 (norm-referenced)과 절대 평가인 준거기준방법

(criterion-referenced)의 두 가지가 있다(Cusimano, 1996). 수행능력평가에서는 절대평가방법이 더 바람직하다. 이는 경계선 집단 방법(borderline group method)과 Angoff 방법의 두 가지로 다시 나뉜다(Kaufman et al, 2000). Angoff 방법에서는 증례 개발자 혹은 일정 수의 전문가가 경계선상에 있는 수험생의 점수를 미리 예상해 보게 하고 이를 기준으로 합격기준을 정하는 방법이다. 경계선 집단 방법은 대규모 OSCE나 국가시험 등에서 뿐 아니라 소규모 OSCE에서도 널리 이용하고 있는 방법이다. 평가자는 체크리스트 채점과 상관없이 해당 학생의 수행능력을 전반적평가척도(global scale)에 예를 들면 통과/경계선/낙제의 기준으로 평가한다. 경계선의 점수를 받은 수험생의 점수의 평균값을 그 스테이션의 통과 기준으로 잡고 전체 OSCE의 당락평정은 모든 스테이션의 경계선에 걸린 학생들의 총 평균값을 기준으로 결정한다. 이 방법은 Angoff 방법에 비하여 비교적 쉽고 시간이 덜 들며 직접관찰을 통하여 기준을 설정했다는 측면에서 신뢰성이 더 높은 것으로 알려져 있다. 변형경계군 방법(modified borderline-group method)에 의한 기준점수는 증례개발자가 제시한 기준점수(Case-author method)에 비하여 더 높게 나왔으며 전자의 경우 더 많은 학생이 탈락하는 것으로 나타났다(Humphrey-Murto & Mac Fadyen, 2002). 기준 이하 성적을 보이는 수험생의 특성으로는 환자 문제에 집중하지 못하는 점, 불충분한 지식, 의사소통기술의 부족 등이 주를 이루었다(Cusimano, 1998).

바. OSCE의 제한점

OSCE는 전체 스테이션이 현실적인 임상 상황을 얼마나 재현할 수 있느냐에 있어서 세 가지의 제한점을 갖고 있다(Smee, 2003). 첫째, 각 스테이션별로 정해진 시간만 사용해야 하므로 의사-환자 접촉을 임의로 분리하여 임상환자면담을 고립된 단면으로 만들어 수험생 수행능력을 평가한다는 한계를 갖고 있다. 따라서 형성평가로서는 취약한 점이 있지만 스테이션 수를 늘리면 신뢰도가 올라가고 총괄평가로서 이용하기에는 오히려 유리해진다. 둘째,

OSCE는 임무 단위의 문항으로 구성된 평가표를 사용하는데 이는 의사환자간의 접촉이 행동의 나열로 설명될 수 있다는 것을 전제로 하고 있다. 따라서 평가표는 수험생의 완벽한 행동을 요구하지만 수험생의 수준이 올라가면 이런 완벽성에 입각한 기준은 중요성이 상대적으로 감소한다. 마지막으로 시뮬레이션에 한계가 있기 때문에 실제로 시험에 포함시킬 수 있는 환자 문제가 제한적이라는 점이다. 이 또한 수험생의 경험 및 교육 수준이 올라가면 더욱 제한점으로 부각된다.

OSCE에 처음 참가하는 학생들은 평가교수가 자신의 행동을 지켜보고 있다는 사실에 스트레스를 받을 수 있으며 정해진 시간에 일정한 수기를 해내야 한다는 강박관념으로 긴장하고 불안에 빠지기 쉽다(Zartman et al, 2002). OSCE의 스테이션별 점수를 시험 본 순서대로 나열하면 순서가 뒤로 갈수록 스테이션별 점수가 향상된다는 보고도 있다(Ross & Syal, 1987). OSCE에 있어서 시험 시간이 수험생에게 주는 스트레스를 본 연구에서 적어도 3.5시간 동안에는 수험생의 피로도에 의한 시험 성적 저하는 없었다(Petrusa, 1987, 박훈기 등, 1999). OSCE 본 평가 전에 축소형 연습 OSCE를 시행하면 학생들의 시험 스트레스를 줄이고 적응능력을 향상시킬 수 있다(Tervo et al, 1997). 이처럼 학생들이 미리 준비하고 마음을 편하게 실제 시험에 임하면 더 좋은 점수를 얻을 수 있다.

결 론

객관구조화진료시험은 도입된 지 거의 30년이 되면서 이제는 총합시험으로서 의사자격시험에 활용되고 있으며 피드백을 기초로 하여 형성평가로도 널리 활용되고 있다. OSCE가 전통적인 임상교육평가방법의 제한점을 수용하는 대안으로 발전해 온 것은 사실이지만 임상실기 성취도를 완벽하게 평가할 수 있는 황금물은 아니고 오히려 다른 임상실습 시험과 서로 보완적으로 활용될 때 그 효과가 더 크다고 할 수 있다(Carraccio & Englander, 2000). OSCE를 성공적으로 계획하고 운영하려면 단계별로

지침에 따라 적절하게 준비를 해야 하고 원칙에 따라 시행을 해야 한다. OSCE의 타당도는 임상교육의 학습목표를 바탕으로 하여 내용 타당도를 유지하는 것도 중요하지만 수험생의 배경을 중요시하는 타당도를 고려할 필요도 있다. OSCE의 신뢰도 유지에는 평가자의 훈련이 중요하고 시험 시간과 시험의 구성은 신뢰도를 결정하는 또 다른 요인이다. OSCE 점수를 바탕으로 당락판정을 할 때는 준거기 준방법을 사용하는 것이 가장 합리적이고 실용적이다. OSCE는 그 자체의 약점을 충분히 고려하고 사용하는 것이 바람직하다.

참 고 문 헌

김병수, 이영미, 안덕선, 박정률(2001). 임상의학입문(Introduction to clinical medicine)평가를 위한 객관적임상실기시험(objective structured clinical examination) 경험. *한국의학교육*, 13, 289-298.

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

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

박훈기, 이정권, 황환식, 이재웅, 최윤영, 김혁 등(2003). 객관구조화진료시험(OSCE)에서 교수와 표준화환자 사이의 점검표 채점의 일치도. *한국의학교육*, 15, 141-150.

안효섭(1995). *임상교육평가를 위한 OSCE 운영의 경험*. 서울: 서울대학교 출판부.

이영미, 황건, 백상호, 홍명호, 최윤선, 김수현. 의사국가시험에 임상수행능력측정의 도입을 위한 객관적 임상실기시험(objective structured clinical examination) 경험. *한국의학교육*, 13, 47-57.

Barzansky B, Etzel SI(2003). Educational programs in US medical schools, 2002-2003. *JAMA*, 290, 1190-1196.

Carraccio C, Englander R(2000). The objective

structured clinical examination: a step in the direction of competency-based evaluation. *Arch Pediatr Adolesc Med*, 154, 736-741.

Chatenay M, Maguire T, Skakun E, Chang G, Cook D, Warnock GL(1996). Does volume of clinical experience affect performance of clinical clerks on surgery exit examinations? *Am J Surg*, 172, 366-372.

Cohen DS, Colliver JA, Marcy MS, Fried ED, Swartz MH(1996). Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills. *Acad Med*, 71(1 Suppl), s87-s89.

Cohen R, Rothman AI, Ross J, Poldre P(1993). Impact of repeated use of objective structured clinical examination stations. *Acad Med*, 68(10 Suppl), s73-s75.

Colliver JA, Barrows HS, Vu NV, Verhulst SJ, Mast TA, Travis T(1991). Test security in examinations that use standardized-patient cases at one medical school. *Acad Med*, 66, 279-282.

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

Cusimano MD(1996). Standard setting in medical education. *Acad Med*, 71(10 Suppl), s112-s120.

Cusimano MD, Rothman AI, Keystone J(1998). Defining standards of competent performance on an OSCE. *Acad Med*, 73(10 Suppl), s112-s113.

Elliot DL, Fields SA, Keenen TL, Jaffe AC, Toffler WL(1994). Use of a group objective structured clinical examination with first-year medical students. *Acad Med*, 69, 990-992.

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

Hodder RV, Rivington RN, Calcutt LE, Hart IR(1989). The effectiveness of immediate feedback during the objective structured clinical examination.

- Med Educ*, 23, 184-188.
- Hodges B(2003). Validity and the OSCE. *Med Teach*, 25, 250-254.
- Hodges B, McIlroy JH(2003). Analytic global OSCE ratings are sensitive to level of training. *Med Educ*, 37, 1102-1106.
- Humphrey-Murto S, MacFadyen JC(2002). Standard setting: a comparison of case-author and modified borderline-group methods in a small-scale OSCE. *Acad Med*, 77, 729-732.
- Jolly BC, Jones A, Dacre JE, Elzubeir M, Kopelman P, Hitman G(1996). Relationships between students' clinical experiences in introductory clinical courses and their performances on an objective structured clinical examination (OSCE). *Acad Med*, 71, 909-916.
- Kaufman DM, Mann KV, Muijtjens AM, van der Vleuten CP(2000). A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Acad Med*, 75, 267-271.
- Mavis BE(2000). Does studying for an objective structured clinical examination make a difference? *Med Educ*, 2000, 34, 808-812.
- Newble D(2004). Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ*, 38, 199-203.
- O'Connor HM, McGraw RC(1997). Clinical skills training: developing objective assessment instruments. *Med Educ*, 31, 359-363.
- Petrusa ER(1987). The effect of number of cases on performance on a standardized multiple-station clinical examination. *J Med Educ*, 62, 859-860.
- Regehr G, Freeman R, Hodges B, Russel L(1999). Assessing the generalizability of OSCE measures across content domains. *Acad Med*, 74, 1320-1322.
- 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, 993-997.
- Reznick RK, Blackmore D, Cohen R, Baumber J, Rothman A, Smee S, et al(1993). An objective structured clinical examination for the licentiate of the Medical Council of Canada: from research to reality. *Acad Med*, 68, s4-s6.
- Ross JR, Syal S, Hutcheon MA, Cohen R(1987). Second-year students' score improvement during an objective structured clinical examination. *J Med Educ*, 857-858.
- Rulata PJ, Witzke DB, Leko EO, Fulginiti JV, Taylor PJ(1991). Sharing of information by students in an objective structured clinical examination. *Arch Intern Med*, 151, 541-544.
- Selby C, Osman L, Davis M, Lee M(1995). Set up and run an objective structured clinical exam. *BMJ*, 310, 1187-1190.
- Sloan DA, Donnelly MB, Schwartz RW, Strodel WE(1995). The objective structured clinical examination. The new gold standard for evaluating postgraduate clinical performance. *Ann Surg*, 222, 735-742.
- Small PA Jr, Stevens CB, Duerson MC(1993). Issues in medical education: basic problems and solutions. *Acad Med*, 68(10 Suppl), s89-s98.
- Smee S(2003). Skill based assessment. *BMJ*, 326, 703-706.
- Tervo RC, Dimitrievich E, Trujillo AL, Whittle K, Redinius P, Wellman L(1997). The objective structured clinical examination (OSCE) in the clinical clerkship: an overview. *S D J Med*, 50, 153-156.
- Thomas PA, Shatzer JH(2000). Standardized patient assessment of ambulatory clerks: effect of timing and order of the clerkship. *Teach Learn Med*, 12(4), 183-188.
- Tombleson P, Fox RA, Dacre JA(2000). Defining the content for the objective structured clinical examination component of the professional and linguistic assessments board examination: development

- of a blueprint. *Med Educ*, 34, 566-572.
- Valentino J, Donnelly MB, Sloan DA, Schwartz RW, Haydon RC 3rd(1998). The reliability of six faculty members in identifying important OSCE items. *Acad Med*, 73, 204-205.
- Verhoeven BH, Hamers JG, Scherpbier AJ, Hoogenboom RJ, van der Vleuten CP(2000). The effect on reliability of adding a separate written assessment component to an objective structured clinical examination. *Med Educ*, 34, 525-529.
- Vu NV, Barrows HS, Marcy ML, Verhulst SJ, Colliver JA, Travis T(1992). Six years of comprehensive, clinical, performance-based assessment using standardized patients at the Southern Illinois University School of Medicine. *Acad Med* 67, 42-50.
- Wass V, McGibbon D, Van der Vleuten C(2001). Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? *Med Educ*, 35, 326-330.
- Wilkes M, Bligh J(1999). Evaluating educational interventions. *Br Med J*, 318, 1269-1272.
- Williams RG, McLaughlin MA, Eulenberg B, Hurm M, Nendaz MR(1999). The patient findings questionnaire: one solution to an important standardized examination problems. *Acad Med*, 74, 1118-1124.
- Zartman RR, McWhorter AG, Seale NS, Boone WJ(2002). Using OSCE-based evaluation: Curricular impact over time. *J Dent Educ*, 66, 1323-1330.