Original Article

KME Review

Korean Medical Education Review 2013; 15(3): 105-111 | pISSN 2092-5603 | eISSN 2005-8160

의학교육 학생평가의 객관성에 대한 쟁점

민경석[1] · 양길석[2]

[1]세종대학교 인문과학대학 교육학과, [2]가톨릭대학교 교직과

Issues Related to the Objectivity of Student Assessment in Medical Education

Kyung-Seok Min[1] · Kil-Seok Yang[2]

[1]Department of Education, College of Liberal Arts, Sejong University; [2]Department of Education, The Catholic University of Korea, Seoul, Korea

This paper addressed various issues related to the objectivity of student assessment in medical education. The objectivity of assessment was related to all the steps of test development, administration, and results reporting in terms of reliability and validity. Specifically, the objectivity of item formats, representativeness of test content, standardization of test administration, consistency of scoring procedures, and appropriateness of reporting test results were discussed by comparing performance assessment with traditional paper-and-pencil tests. The conclusions were derived from current measurement theories such as standards-based assessment, evidence-based design, and outcome-based assessment. Further, based on Shepard's propositions (2006), the objectivity of student assessment could be achieved by improving the concordance between educational objectives and assessment components such as item types, test contents, and test administration, scoring, and reporting.

Keywords: Objectivity, Reliability, Student assessment, Validity

Corresponding author
Kyung-Seok Min
Department of Education, College of Liberal Arts, Sejong University,
209 Neungdong-ro, Gwangjin-gu,
Seoul 143-747, Korea
Tel: +82-2-3408-3128
Fax: +82-2-3408-4304
E-mail: minkyungseok@sejong.ac.kr

Received: September 28, 2013
Revised: October 17, 2013
Accepted: October 17, 2013

서 론

교육 및 심리검사에서 객관성(objectivity)이란 동일한 능력 혹은 특성을 갖는 피험자가 동일한 검사결과(예, 점수)를 획득함을 의미한다(Miller et al., 2009). 대학수학능력시험 혹은 전국단위 자격증 시험(예, 의사자격시험)과 같은 대규모 평가(large scale assessments)에서 활용되는 표준화 검사(standardized tests)는 주로 선다형 문항(multiple choice items) 혹은 단답형 문항(short answer items)으로 구성되며, 이에 따라 상대적으로 높은 객관성을 유지하고 있다. 특히, 선택형 문항은 점수를 할당하는 채점과정에서 채점자의 판단적 의사결정이 개입할 여지가 없다는 점에서 객관식 문항(objective type items)이라 불린다.

한편 강의자가 직접 제작한 학교평가(classroom assessment) 혹은 학생의 실기능력에 대한 수행평가(performance assessment)는 선택형 문항 중심의 표준화 검사와 비교하여 객관성이 상대적으로 낮은 것으로 취급된다(Miller et al., 2009). 예를 들어, 일반대학 학생평가에서 주로 활용되는 논술시험(혹은 보고서)은 채점자의 판단적 논리에 의해서 점수가 결정되며, 이에 따라 동일한 학생의 논술문에 대하여 서로 다른 채점자의 점수는 동일하지 않으며, 한 사람의 채점자 또한 여러 학생의 논술문을 채점하면서 일관된 채점기준을 적용하는 데 어려움을 겪는 경향을 보인다. 의학교육에서 학생의 임상실기능력을 평가하는 전통적 방법인 관찰평가, 임상증례보고 등은 논술시험과 유사한 특성을 보이며(Kogan et al., 2009), 채점자 내 혹은 채점자 간 점수의 차이는 평가점수의 비일관성을 의미하고, 이는 검사점수의 낮은 신뢰도로 이어진다.

또 다른 측면에서, 검사점수의 신뢰성을 높이고자 학생평가에서 선택형 문항 중심의 객관식 문항만을 활용하는 것은 검사가 측정하고자 하는 바를 측정하고 있는가를 의미하는 타당도에 문제점을 드러낸다. 의학교육에서 추구하는 교육목표에는 객관식 문항으로 측정될 수 있는 지식의 획득 여부뿐만 아니라 실제상황에서 지식내용을 적용하고 처치결과를 판단하는 임상능력이 중요한 요소로 포함될 때, 학생평가 또한 임상상황의 문제해결력, 의사소통능력, 비판적 사고력을 평가해야만 타당한 것이라고 할 수 있다.

의과대학과 의학전문대학원의 교육목표는 의학적 지식의 획득

과 임상능력의 배양으로 의료현장에서 문제해결능력과 전인적 치료자로서 의사의 태도를 강조한다(Miller, 1990). 이러한 교육목표에 근거하여 의과대학의 학생평가는 저학년의 지식 중심 교육과정과 고학년의 임상실기 중심의 교육과정으로 구분되며, 지식과 임상능력을 평가할 수 있는 다양한 학생평가 방식(선다형 문항, 구술, 임상사례, 관찰평가, 업무일지, 표준화 환자 등)이 활용된다. Mavis et al. (2001)은 126개 미국 의과대학 설문조사결과를 통하여 객관구조화진료시험(objective structured clinical examination, OSCE)과 같이 표준화 환자(standardized patients)를 활용한 학생평가방법이 지속적으로 확대되고 있음에도 불구하고, 이러한 모의 임상능력 평가결과는 진급, 졸업과 같은 중요한 의사결정에는 상대적으로 적게 활용되며, 전체적으로 선다형 문항과 관찰평가(preceptor rating)와 같은 전통적 평가방식이 보다 광범위하게 이용되고 있음을 밝힌다.

현대 교육평가이론은 학생특성과 학업성취에 대한 정보를 수집하여 교육과정 개선을 위하여 활용하며, 이에 따른 교육성과를 높인다는 학생평가의 본원적 목적을 달성하기 위하여, 성취기준 기반평가(standards-based assessment, Stecher, 2010), 증거기반평가(evidence-centered design, Mislevy et al., 2003), 성과기반평가(outcome-based assessment, Dent & Harden, 2009) 등을 강조한다. 이러한 경향은 일부 강조점에서 차이를 보임에도 불구하고, 학생평가의 목적은 교육과정을 통한 교육목표의 달성 정도를 명확히 제시하는 것에 있으며, 이를 위하여 전통적인 선택형 문항 중심의 평가를 포함한 다양한 평가방식의 개발 및 활용을 제안한다. 특히, 1980년대를 전후로 미국을 중심으로 평가의 타당도를 중시하는 수행평가에 대한 논의가 활성화되었고, 나아가 수행평가의 객관성 즉, 학생평가결과의 타당도를 전제하면서도 그 신뢰성을 함께 높일 수 있는 방법이 다양하게 제안되어 왔다(Lane, 2010; Lane & Stone, 2006; Stecher, 2010).

학생평가를 포함한 교육평가는 측정의 양호도 판단기준으로 신뢰도와 타당도를 강조한다. 대규모의 평가 체제에서 평가시행 및 결과가 민감하게 작용하는 경우에는 무엇보다도 점수의 일관성을 의미하는 신뢰도를 우선시할 수밖에 없으며 그에 따른 타당도의 결여 측면을 검사의 설계, 문항내용의 충실성으로 보완하고자 노력한다. 반면에 실제적인 능력, 즉 수행능력을 제대로 측정하고자 하는 경우에는 다양한 수행과제를 활용하여 타당도를 먼저 확보하고자 하며, 그 방법들의 특성상 인간에 의한 판단이 개입될 수밖에 없기 때문에 보완적으로 신뢰도를 강화하려고 하는 노력을 수반한다. 즉, 학생평가의 객관성은 측정이론적 측면에서 신뢰도와 타당도의 문제를 의미하는 것으로서, 이는 평가계획, 실시, 결과보고 및 활용 등 평가의 전 과정과 관련된다. 이 논문에서는 학생평가과정에서 중요하게 고려해야 할 문항형식, 평가내용, 시행절차, 채점, 결과산출에 관하여 전통적인 지필시험과 수행평가를 대비하여 학생평가의 신뢰도와 타당도에 대하여 논의하고자 한다.

평가 문항형식의 객관성

학생평가에 활용되는 평가도구는 문항의 모둠으로 구성되며, 문항특성을 나타내는 문항형식은 평가상황에서 학생에게 요구하는 바가 무엇이며, 이에 따른 응답방식을 결정한다. 또한 문항형식은 학생 반응뿐만 아니라 평가내용, 시행환경과 절차, 채점방식 및 결과보고와 직접적으로 연관된 것으로(Allen & Yen, 1979), 평가의 목적에 근거하여 적절한 평가 문항형식을 결정하는 것은 학생평가의 객관성을 위한 가장 중요한 의사결정과정이라 할 수 있다.

학생평가에서 중요하게 대비되는 문항형식은 선택형(selected response items)과 구성형(constructed response items)이라 할 수 있다. 선택형 문항은 진위형(true-false form), 선다형(multiple choice form), 연결형(matching form) 등을 포함하며, 학생평가의 모든 분야에서 가장 자주 활용되고 객관적인 평가 문항형식으로 취급된다. 대표적인 선택형 문항인 선다형은 지문(stem)과 선택지(alternatives)로 구성되며, 선택지 중에서 지문이 요구하는 정답(key) 선택 여부에 따라 학생의 능력을 평가한다. 선다형 문항의 선택지에서 사전에 정답이 결정되어 있으며, 학생 응답과 정답의 일치 여부를 통하여 문항점수가 결정되기 때문에 채점자의 주관적 판단이 필요가 없다. 이에 따라 선다형 문항은 optical mark reader (OMR) 용지를 이용한 전산처리를 통하여 기계가 채점할 수 있다는 점에서 합리성, 공정성, 투명성을 강조하는 현대사회에 가장 대중화된 학생평가방법이라 할 수 있다.

구성형 문항은 단답형(short answer type), 완성형(completion type), 논술형(essay type)으로 구분되며, 지문에 대하여 간단한 단어나 문구를 제시하는 단답형이나 완성형에 비하여 논술형은 비교적 제한 없이 여러 개의 문장으로 학생이 응답하는 문항형태를 의미한다. 단답형과 완성형 문항은 선다형 문항과 유사하게 사전에 정답이 결정되어 선택형 문항수준에 가까운 채점의 일관성을 담보한다. 반면에 논술형은 피험자의 분석력, 비판력, 조직력, 종합력, 문제해결력, 창의력 등 고차원적 사고능력을 측정한다는 긍정적 특성에도 불구하고, 채점자의 판단에 의해 점수가 부여되며, 이에 따라 채점자 간 혹은 채점자 내 점수의 일관성이 선택형 문항에 비하여 낮아지는 특성을 보인다.

구성형 문항을 대표하는 논술은 주로 언어적 사고와 표현능력을 평가하는 것으로 수행평가의 한 방식이라 할 수 있다. 1980년대 이후 미국의 학생평가는 학생의 지식수준에서 수행능력을 강조함에 따라 전통적인 학생평가에서 수행과제 중심의 평가로 전환되었다. 즉, 학생 앎보다는 학생행동을 직접 측정하는 수행평가는 실제상황과 근접한 평가환경에서 시연되는 학생의 결과물과 성취과정에 중점을 두며, 이에 따라 단순한 지식에 대한 평가가 아닌, 수행과정에 중심을 두며 부가적으로 의사소통, 태도, 성실성 등의 정의적 특성을 평가한다(Lane, 2010; Lane & Stone, 2006).

학생행동을 직접 평가하는 수행평가는 교육내용에 대한 실제적 적용능력이라는 교육목표에 부합한다는 원론적 의미뿐만 아니라, 학생평가를 통한 교수-학습과정의 개선을 강조한다. 즉, 교육목표로 인지적 영역을 포함한 다양한 수행목표를 설정했음에도 불구하고, 학교 교육과정은 학생참여, 체험, 실기가 아닌 지식전달을 위한 강의자 중심의 수업이 이루어지며, 학생평가 또한 절차적 객관성을 강조하는 지식정보수준의 선다형 문항이 주로 활용되어지는 바, 수행평가를 통하여 교육과정과 성취결과에 대한 개선을 이루고자 하는 교육정책적 지향성을 내포하고 있다. 특히, 의학교육은 전문가 양성 및 직업교육의 성격을 포함하며, 학교교육을 통하여 양성된 신입 의사의 실제적인 임상능력검증에 대한 사회적 요구가 지속적으로 높아짐에 따라, 임상실기 중심의 수행평가를 통한 교육과정의 개선이 지속적으로 주장되어 왔다. 의학교육에서 학생평가의 수준을 구분한 Miller (1990)의 피라미드에 대응하는 평가 문항형식은 다음과 같다(Amin et al., 2006). 1) 지식과 방법(knows & know how): 구술시험(oral examination), 긴 논술(long essay question), 짧은 논술(short essay question), 선다형 문항(multiple choice question), 확장연결형문항(extended matching items, EMI), 핵심요소시험(key features examination); 2) 행동시연(show how): OSCE, 긴 사례(long case), 짧은 사례(short case); 3) 행동(does): 간편임상실습(mini clinical evaluation exercise), 진료과정관찰(direct observation of procedural skills), 점검표(checklist), 다면평가(360-degree evaluation), 진료일지(logbook), 포트폴리오(portfolio).

Miller (1990)의 4가지 평가목적에 대응하는 다양한 평가방식은 앞서 논의된 선택형, 구성형, 수행평가의 다양한 적용과 변형사례라고 할 수 있다. 예를 들어, EMI는 채점의 객관성을 유지한 채, 복잡한 지식 및 다양한 주제의 연계성을 평가할 수 있도록 선다형 문항의 확장된 형태라고 할 수 있다. 간편 논술은 비판적 사고능력을 측정하면서 채점의 객관성을 높이기 위한 구성형 문항의 변형이며, OSCE는 실제적 상황의 학생행동을 평가하면서 채점의 일관성을 높이는 수행평가의 한 형태라고 할 수 있다. 즉, 모든 문항형식은 객관성을 의미하는 신뢰도와 타당도라는 측면에서 강점과 약점을 가지며, 평가의 목적에 따라 문항형식의 변형과 개선을 통하여 지속적으로 새로운 문항형식을 활용하는 것이 필요할 것이다. 특히, 현대 컴퓨터기술(시뮬레이션, 네트워킹, 인공지능 등)의 발달에 따라(Drasgow et al., 2006) 기존 문항형식의 제한점을 극복하여 객관성과 현실 적용력이 높은 새로운 문항형식이 지속적으로 개발될 것이다.

평가내용의 객관성

학생평가 문항은 학생특성을 측정하기 위하여, 교육내용 혹은 교육목표를 대표하는 표본과제(sampled tasks)이며 평가도구(시험)는 이러한 문항의 모둠으로 정의된다(Allen & Yen, 1979). 즉, 학생평가는 한 학기 강의 혹은 학과목 내용을 모두 측정하는 것이 아니라 교육목표를 대표하는 내용을 반영한 평가 문항을 통하여, 학생의 이해, 적용능력을 추정(inference)하는 과정이라고 할 수 있다. 교육과정의 표본으로서 평가 문항이라는 논리는, 사회여론을 알기 위하여 모든 사람에게 의견을 묻는 것이 아니라 모집단을 대표할 수 있는 표본(일반적으로 1,000명 내외)을 조사하는 사회조사방법과 비유적으로 비교될 수 있다(Allen & Yen, 1979; Lohr, 1999). 즉, 공정하고 타당한 사회조사를 위하여 지역, 성별, 연령, 소득 등 다양한 요인을 복합적으로 고려하여 표집된 표본이 전체 모집단을 대표할 수 있으며, 모집단을 적절히 대표하는 표본의 조사결과가 모집단의 의견으로 추정된다. 유사하게, 학생평가에서 제한적으로 구성되는 평가 문항이 교육내용과 교육목표를 얼마나 적절히 대표하는가는 평가의 내용타당도(content validity)를 의미한다. 검사이론에서 학생평가의 내용적 대표성을 위하여 평가 문항의 구성을 위한 설계도(blueprints)로서 검사명세표(test specification)의 세밀한 설정을 중요하게 강조한다(Allen & Yen, 1979; Kane, 2006). 일반적으로 검사명세표는 내용영역과 행동영역이 교차하는 이원분류표로서 각 교차영역의 문항분포뿐만 아니라 문항형식, 난이도, 배점 등에 대한 상세한 정보를 포함한다.

구체적인 평가계획으로서 검사명세표가 명확히 작성되었음에도 불구하고, 앞서 논의된 문항형식에 따라 검사의 내용 대표성은 이질적인 양태를 보인다. 문항당 풀이시간이 상대적으로 적은 선다형 문항의 경우, 제한된 평가시간 동안 많은 수의 문항이 시행될 수 있으며, 많은 수의 문항은 정해진 교육과정의 범위를 포괄하고 대표하는 데 강점으로 작용한다. 이에 반하여 수행평가에 포함되는 실기, 논술, 구술의 경우, 평가시행과 채점과정에서 많은 시간과 비용이 소요된다. 이에 따라 제한된 시험시간 동안 상대적으로 적은 문항이 출제되고, 결국 내용적 대표성을 확보하기 어려운 문제로 이어진다. 예를 들어, 동일한 임상능력을 측정하기 위하여 3시간 동안 500개의 선다형 문항을 출제하는 것과 5가지 임상사례에 기반한 표준화 환자를 활용한 평가를 비교할 때, 어느 방법이 보다 객관적인가의 문제는 문항 수에 따른 평가내용 대표성과 포괄성과 관련된 것이라 할 수 있다.

또한 문항 수는 내용 타당도뿐만 아니라 점수신뢰도와 관련되며, 일반적으로 문항 수가 많을수록 높은 신뢰도를 보인다(Allen & Yen, 1979). 표준화 검사의 신뢰도는 통상 0.9 이상이며 학교평가의 경우 0.7-0.8 수준임을 고려하여, 0.8 수준의 신뢰도를 위하여 약 10개 내외의 수행과제가 요구된다(Lane, 2010). 결국, 학생평가에서 내용적 대표성뿐만 아니라 평가결과점수의 일관성을 높이기 위해서는 수행과제를 분할하여, 여러 측면에서 학생 특성을 측정하는 것이 바람직하다고 할 수 있다.

평가내용의 대표성과 평가방식의 연관성, 이에 따른 평가결과의

신뢰도는 평가시행을 위한 현실적 조건(예, 시간, 비용, 장소 등)에 제약을 받는다. 고등교육의 목표가 단순 지식에서부터 문제해결력, 비판적 사고 등의 폭넓은 영역을 포괄하고 있음을 고려할 때, 학생평가는 어느 한 가지 평가방법을 선택하는 문제가 아니라 다양한 수준의 평가방법을 활용하여 학생의 특성에 대한 종합적 정보를 확보해 나갈 필요가 있다. 즉 임상의 기초가 되는 지식수준의 평가에는 선택형 문항이 강점을 가지며, 실제적 행위를 평가하기 위해서는 수행평가가 유용하게 적용될 수 있다. 이러한 다양한 방법을 포괄하여 전체적 학생특성을 평가할 수 있는 학생평가 체제를 구축하는 것이, 교육목표 달성 정도를 확인하고 교육과정을 개선하기 위한 학생평가 객관성에 중요하게 작용한다.

평가시행절차의 객관성

평가시행절차의 객관성은 검사가 모든 피험자에게 동일하게 시행되며, 채점되는 것을 의미한다. 즉, 표준화된 시행절차에 따라 검사결과는 시행시기, 검사유형(test forms)과 관계없이 모든 피험자에게 동일한 의미를 제공한다(Cohen & Wollack, 2006). 종종 표준화가 객관식 문항 혹은 표준점수로 산출되는 검사결과와 혼동되기도 하며, 검사의 표준화 절차가 규준참조검사에서만 필요한 것으로 오인되기도 한다(Kane, 2006).

측정이론적으로 검사의 표준화는 검사가 측정하고자 하는 바 이외에 검사점수에 영향을 미치는 외재요인(nuisance factors)을 최소화하고, 평가상황에서 모든 피험자에게 자신 능력 혹은 특성을 발휘할 수 있는 동등한, 공정한 기회를 제공하는 것을 의미한다. 이를 위하여 전통적인 지필검사에서는 모든 피험자에게 동일한 문항을 제시하고, 검사시간을 포함한 검사환경을 엄격히 통제한다. 또한 검사에서 측정하는 특성 이외의 요소에 의한 차별성을 배제하기 위하여 평가과목의 구성, 문항형태, 평가범위, 채점요소 및 절차에 대한 정보를 사전에 피험자에게 제공한다. 이러한 절차적 요소의 명확성을 위하여 표준화 검사의 경우, 과거 기출 문항을 공개하고, 모의시험과 같은 사전 연습 기회를 제공하여 검사가 측정하는 바가 무엇이며 검사상황에서 피험자가 해야 할 것과 하지 말아야 할 것에 대한 세부적 지침을 모든 피험자에게 공개적으로 제공한다.

학교 현장에서 오랫동안 실시되어 왔으며, 많은 선행연구가 이루어진 선택형 문항으로 구성된 지필평가는 이러한 시행절차의 표준화에 많은 장점을 갖는다. 즉, 검사내용과 형식에 대한 명확한 전달이 용이하며, 검사환경을 모든 피험자에게 동일하게 하는 절차가 비용과 시간적 측면에서 상대적으로 간편하다. 무엇보다 오랜 시행경험을 통하여 시행자와 피험자 모두에게 익숙한 평가절차라는 것은 지필검사의 큰 강점으로 작용한다.

그러나 실기능력평가는 수행평가의 경우 상대적으로 시행절차의 표준화에 어려움을 보이며, 이러한 점 때문에, 학생 간 점수의 차이가 평가하고자 하는 능력의 차이에서 나타난 것인지 시행절차의 비표준화로 인한 외재요인에 따른 것인지에 대한 명확한 확인과 통제가 필요하다. 예를 들어, 임상능력을 평가하기 위하여 표준화 환자를 이용한 경우에서, 모든 피험자에게 동일한 표준화 환자가 활용될 수 있는가, 모든 피험자에게 동일한 환자가 제시될지라도 표준화 환자는 매번 동일한 양호도 수준에서 평가상황을 재현하는가, 만약 현실적 어려움으로 여러 명의 표준화 환자가 피험자 집단에 활용된다면, 서로 다른 표준화 환자의 수행은 학생 평가결과에 영향을 미치지 않는가, 또한 평가장소와 시기가 피험자마다 다른 경우 이러한 조건은 학생 평가점수에 영향을 미치지는 않는가 등의 문제는 평가결과의 신뢰성과 타당성에 대한 쟁점 사항이다(Epstein, 2007; Miller, 1990).

검사시행의 표준화는 또한 검사의 보안(test security)과 관련된다. 일반적으로 피험자가 검사문항을 사전에 입수하여 연습하거나 검사시행과정에서 부정한 방법으로 정답을 표기한다면, 검사점수는 피험자의 능력을 정확히 표시할 수 없을 것이다(Cohen & Wollack, 2006). 예를 들어, 표준화 환자에 대한 정보수집, 검사, 진단 등 다양한 절차를 통하여 피험자의 임상능력을 평가하는 수행평가에서, 피험자가 구체적 평가내용을 사전에 인지하였다면, 이는 임상능력을 평가하는 것이 아니라 단순 암기능력을 평가하게 된다(Epstein, 2007).

학생평가에서 표준화의 목적은 모든 피험자에게 동일한 평가조건과 기회를 부여하여 평가결과를 객관적으로 비교 가능하게 하는 것에 있다. 의학교육에서 임상실기능력평가를 위하여 전통적으로 활용된 직접관찰, 증례, 실습평가는 실기능력배양이라는 교육목적에 부합하는 평가임에도 불구하고 내용 대표성 및 절차의 표준화에 어려움을 갖는다. 이러한 점에서 표준화 환자를 이용한 임상평가는 실제와 유사한 상황에서 임상능력을 평가하고 평가의 객관성을 위한 수행평가 표준화의 선도적 방안이라고 할 수 있다. 특히, 국내외 의사자격시험에 포함된 OSCE는 임상사례 수, 단계(stations)의 할당시간, 표준화 환자의 훈련수준, 채점기준 등에 대한 다양한 개선을 통하여 수행평가 또한 선택형 문항수준의 객관성을 확보할 수 있음을 현실적으로 보여 준 사례라고 할 수 있다. 반면, 수행평가결과의 신뢰성에 영향을 미치는 요소인 채점자, 과제, 환경에 대한 선행연구에서 밝혀진 바와 같이(Cronbach et al., 1997), 평가 환경(occasions, 예, 표준화 환자 특성, 장소, 시간 등)이 중요한 요인 임에도 다른 두 요소에 비하여 상대적으로 소홀히 다루어져 왔기에, 이에 대한 지속적인 개선 노력이 필요할 것이다.

평가점수 산출의 객관성

학생평가의 객관성에서 가장 직접적인 단계로 논의되는 것이 평가점수를 산출하는 채점의 공정성, 투명성, 일관성이다. 선다형 문

항은 선택지 중에서 사전에 정답이 결정되어 있으며, 학생 응답과 정답을 비교하여 문항점수를 결정한다는 측면에서 채점자의 판단이 개입될 여지가 없다. 이에 반하여 학생이 응답/수행을 스스로 구성하는 수행평가에서는 상대적으로 자유로운 응답양식과 포괄적인 채점기준으로 인하여 채점자의 판단적 의사결정이 개입되며, 채점의 일관성과 타당성을 위하여 앞서 논의된 문항형식과 시행절차의 표준화와 함께 채점절차의 객관화가 요구된다. 의학교육에서 임상능력 측정을 위하여 표준화 환자를 이용한 평가의 필요성에 대한 대체적 동의가 이루어져 왔음에도 불구하고, 피험자의 어떤 행위/태도가 중요한 것이며, 동일한 피험자 행위에 대하여 복수의 채점자는 동일한 점수를 부여하는가, 채점자로서 표준화 환자가 포함되어야 하는가 등은 이러한 채점의 객관성과 관련된 사항이라 할 수 있다.

수행평가 채점의 객관성을 위한 절차로서 두 가지 단계가 제안된다. 첫째는 두 사람 이상의 채점자가 채점하며 채점자에 대한 사전훈련이 진행되어야 한다. 둘째, 사전에 채점기준을 명확히 제시한 채점기준표(scoring rubrics)를 활용해야 한다. 이때 수행평가의 채점방법은 크게 분석채점(analytic scoring)과 총괄채점(holistic scoring)으로 구분된다.

복수의 채점자를 활용하며, 채점자에 대한 사전훈련을 통하여 채점자 간, 채점자 내 점수의 일관성을 확보하여 학생평가 점수의 객관성을 높이는 절차는 대규모 학생평가 혹은 고부담평가(high-stake assessments)에서 엄격하게 적용되며, 많은 시간과 비용이 소요된다. 그러나 수행평가에서 평가자의 가치판단이 개입할 수 있음을 인정하는 전제에서 학생 응답에 의한 가치 판단이 아니라 채점자의 주관적 편견이 개입하는 것을 방지하기 위하여 두 사람 이상이 채점에 참여하고 또한 이러한 차이를 사전에 조정하는 채점자 훈련절차는 반드시 필요한 과정이라 할 수 있다. 복수채점의 수준은 시간과 비용이라는 현실적 여건을 고려하여, 일상적인 수업의 학생평가에서는 모든 피험자에 대한 복수채점보다는 일부 피험자 표본에 대한 복수채점을 통하여 평가의 객관성을 확인할 수 있다. 또한 채점자 훈련의 가장 중요한 과정은 채점기준표를 이해하고 실제 채점에서 이를 일관되게 적용하는 것이다. 즉, 채점에 임하기 이전에 채점기준표와 일치하는 혹은 일치하지 않는 학생 응답/수행을 명확히 확인하고, 각 점수수준을 대표하는 수행에 대한 명확한 설정이 이루어져야 한다(Lane & Stone, 2006).

의학교육에서 임상평가의 채점은 주로 교수자 한 사람에 의하여 실행된다는 점을 고려할 때, 채점 공정성을 위하여 무엇보다 중요한 과정은 채점기준표를 명확히 작성하는 것이라 할 수 있다. 즉, 검사가 측정하고자 하는 바를 실제 측정하기 위하여 검사를 제작하기 이전에 검사명세표를 세밀하게 작성하여 기준으로 활용하는 것과 동일하게, 채점기준표는 채점의 일관성과 주관적 요소를 배제하기 위하여 필수 과정이다. 채점기준표 설정의 근거는 교육목표에서 달성하고자 하는 성취기준이며, 평가도구를 통하여 측정하고자 하는 학생의 지식, 기술수준을 세밀하게 나열하고, 이러한 평가내용에 대한 수행수준에 따라 점수를 할당하는 것이라 할 수 있다. 그러므로 채점기준표는 채점을 위한 수행요소를 구체적으로 설정함에 따라, 학생이 수행해야 할 핵심내용을 보다 명확하게 하여, 수행과제 자체의 타당도를 높이는 데 기여한다. 또한, 채점기준표는 평가자의 채점 일관성뿐만 아니라 피험자에게 자신의 점수가 무엇에 근거한 것인가를 확인하는 기회를 제공하여 평가를 통한 학생 성취에 대한 피드백을 제공할 수 있다는 점에서 평가의 객관성을 높이는 데 중요한 역할을 한다.

마지막으로, 채점의 객관성에 영향을 미치는 중요한 요인은 채점방식이다. 수행평가에서는 분석채점(analytic scoring)과 총괄채점(holistic scoring) 등 크게 두 가지 방식이 활용된다. 분석채점은 수행과제를 구성하는 여러 요소(예, 문진, 검사, 진단, 처치 등)를 구분하여 각 영역에 대한 점수를 부여하고, 이를 합산하여 전체 수행점수를 산출한다. 반면에, 총괄채점은 피험자의 수행에 대한 전체적 수준에 대하여 하나의 종합점수를 부여하는 방식이다. 주로 논술시험의 채점방법에 관한 선행연구는 분석채점이 높은 채점자 신뢰도를 보이며, 전체 점수뿐만 아니라 세부 영역에 대한 학생수행정보를 제공하는 장점을 갖는 반면, 총괄채점은 개별요소보다는 이들이 모여 종합된 성취수준을 평가하는 장점을 보인다. 음악회에서 공연되는 오케스트라 연주의 질은 관악기, 타악기, 현악기 등 각 파트 연주의 탁월함으로 평가될 수 없다는 Mullis (1984)의 비유처럼, 진단, 검사, 처치로 이어지는 임상과정은 세부 영역의 정확성과 함께 전체 과정의 효율성, 효과성 등이 동시에 중요하게 평가될 수 있다. 즉, 임상능력평가를 위한 채점방법은 평가의 목적과 채점기준표의 구성, 평가결과의 활용에 따라 분석채점과 총괄채점이 선택적, 종합적으로 활용되어야 할 것이다. 또 다른 측면에서는, 현대 컴퓨터기술의 발달에 따라 다양한 문항형식의 조합, 개선이 이루어지는 것과 유사하게, 인공지능을 활용한 정보탐색기능을 활용하여 수행평가의 채점에서 사람을 대신한 기계 채점의 도입은(Lane, 2010) 객관성 향상을 위한 지속적 노력의 과제라고 할 수 있다.

평가결과보고의 객관성

학생평가의 최종 단계는 평가결과를 학생, 학부모, 및 교육기관에 보고하는 것이다. 앞서 논의된 문항형식과 내용, 시행절차, 채점 등의 과정이 평가의 목적에 부합하도록 적절히 설정되어야 하는 것과 동일하게 평가결과의 보고 또한 평가의 목적과 활용에 의하여 결정된다. 예를 들어, 학과목 내용에 기반하여 평가가 실시되었다면, 평가결과는 교수학습과정을 개선하기 위한 중요한 정보로 적절한 시간에 제시되어야 하며, 평가의 목적이 모든 학생의 능력수준을 구분하는 서열화에 있다면, 학생 전체의 능력수준과 개인의 위

치정보를 제공할 수 있는 점수 척도(예, Z점수, T점수)가 활용되어야 할 것이다.

일반적으로 학생평가에서 강의자는 다양한 평가방법을 활용한다. 예를 들어, 객관식 시험, 퀴즈, 임상실습, 출석 등과 같이 네 가지 방법으로 평가를 실시하였다면, 최종 학생평가결과를 산출하기 위하여 네 가지 점수를 어떤 식으로 종합할 것이다. 가장 간단한 방법으로, 각 시험의 만점을 25점으로 설정하여 합산하면 100만점의 최종점수가 결정될 것이다. 이때 만약 출석과 퀴즈에서 모든 학생이 동일한 점수를 받았다면, 실제적으로 최종점수는 객관식 시험과 임상실습에 의해서 결정되는 것이라 할 수 있다. 즉 이 경우, 형식적으로 네 가지 평가요소가 각 25%로 동일한 비중을 가짐에도 불구하고 학생 변별을 위한 실제적 요소는 객관식 시험과 임상사례 토의에만 해당되며, 출석과 퀴즈의 실제적 평가 가중치는 0%가 된다. 의학교육은 매우 복잡한 교육과정을 포함하며, 이에 따라 다양한 평가방법을 활용하여 학생정보를 수합한다. 그러므로 교육적 의사결정의 객관성을 위하여, 각 평가요소에 대한 명목 가중치와 실제 가중치에 대한 계획이 명확히 설정될 필요가 있다.

두 번째 고려할 사항으로는, 학생평가가 학생들을 서열화하여 세부적으로 변별하는 것에 목적이 있는가, 혹은 준거(criterion)에 의하여 기본필수능력을 성취했는가를 중시하는가에 따라 평가 결과의 산출과 보고방식이 다르게 설정된다. 규준참조평가(norm referenced tests)의 경우 상위, 중위, 하위 모든 능력수준을 세부적으로 구분할 수 있는 평가의 구성과 점수산출이 요구되는 반면, 준거참조평가(criterion referenced test)는 비/통과를 결정하는 기준선의 객관성이 주요한 관심 대상이 된다. 의학교육이 전문가 및 직업교육의 특성을 갖는다는 점에서, 기본필수능력의 습득 여부가 서열적 정보보다 학생평가에서 중요하게 다루어질 필요가 있다. 이러한 점에서 준거참조검사의 기준점수를 결정하는 준거설정(standard setting)은 합격과 불합격이라는 의사결정의 객관성을 확보하기 위한 중요한 절차로 다루어져야 한다. 구체적인 준거설정에는 매우 다양한 방법(예, Bookmark 방법, 수정된 Angoff 방법 등)이 있음에도 불구하고, 가장 중요한 것은 준거점수가 의미하는 피험자의 지식, 능력수준이 명확히 정의되어야 한다는 것이다(Kane, 2006). 즉, 준거점수에 해당하는 지식과 능력수준이 구체적으로 정의되고, 이에 대하여 전문가, 교육자들의 합의가 이루어질 때, 준거점수를 기준으로 한 교육적, 행정적 의사결정은 객관성을 담보한다.

마지막으로 평가정보의 내용과 명세화 수준은 교육과정에서 평가결과를 활용하는 목적에 따라 형성평가(formative assessment)와 총괄평가(summative assessment)로 구분된다. 형성평가는 교수학습 과정에서 학생 및 강의자에게 수시로 피드백(feedback)을 제공하여 교육과정 및 수업을 개선시키는 평가를 의미한다. 또한, 형성평가의 평가결과는 학생에게 학업동기를 유발하고, 자기주도적 학습능력을 함양하게 하고, 사고능력을 배양하는 피드백의 역할을 한다. 교수자와 학습자의 의사소통이라는 피드백으로서 평가결과는 학생의 서열뿐만 아니라, 시간 흐름에 따른 발전 정도, 또한 학업에 대한 정의적 태도 등의 정보를 포함할 수 있다. 즉, 매번의 학생평가가 졸업과 진급과 같이 합격/불합격의 결정에 제한된 것이 아니라면, 평가의 교육적 활용(교수학습의 개선, 학생 학업동기 배양 등)이 결과 보고에 고려될 때 결과타당도(consequential validity, Kane, 2006)라는 측면에서 평가의 객관성이 확보된다.

요약하면, 평가의 마지막 단계인 결과보고에서는 다양한 평가요소의 합산을 위한 실제 가중치 수준, 준거참조검사에서 준거기준에 해당하는 학생수행의 수준에 대한 명확한 설정, 형성평가와 총괄평가와 같은 평가 목적의 구분이 필요하다.

결 론

이 논문에서는 의학교육 학생평가의 문항형식, 문항내용, 시행절차, 채점, 결과산출에 관하여 전통적인 지필시험과 수행평가를 대비하여, 신뢰도와 타당도를 중심으로 학생평가의 객관성에 대하여 논의하였다.

문항형식이라는 측면에서 전통적으로 수행능력을 강조하는 의학교육은 수행평가를 선도하고 있다. 특히, OSCE는 전국 규모의 자격시험에 활용될 정도의 표준화가 마련된 대표적 사례라고 할 수 있다. 또한, 현대 컴퓨터기술(시뮬레이션, 네트워킹, 인공지능 등)의 발달은 표준화 환자의 수행 일관성을 높이기 위한 방안으로 활용될 수 있을 것이다. 문항내용은 평가의 교육과정 대표성과 관련된 것으로, 임상평가의 과제 수, 시행시간, 내용적 포괄성과 관련된 문항형식의 고려를 통하여 지속적 개선이 필요할 것이다. 평가 실시 절차 표준화의 목적은 모든 피험자에게 동일한 평가 조건과 기회를 부여하여 평가결과를 객관적으로 비교 가능하게 하는 것에 있다. 수행평가결과의 신뢰성에 중요하게 영향을 미치는 요소로 채점자, 과제의 일관성을 위하여 많은 노력이 투입된 반면, 평가환경의 영향은 상대적으로 소홀히 다루어져 왔다. 평가결과 산출을 위한 채점기준표는 학생수행요소를 명시함에 따라, 학생이 수행해야 할 핵심내용을 보다 명확하게 하여, 수행과제 자체의 타당도를 높이는 데 기여한다. 또한, 채점기준표는 평가자의 채점 일관성뿐만 아니라 피험자에게 자신의 점수가 무엇에 근거한 것인가를 확인하는 기회를 제공하여 평가를 통한 학생성취에 대한 피드백을 제공할 수 있다는 점에서 평가의 객관성을 높이는 데 중요한 역할을 한다. 평가의 마지막 단계인 결과보고에서는 다양한 평가요소의 합산을 위한 실제 가중치 수준, 형성평가와 총괄평가와 같은 평가목적의 구분, 준거참조검사에서 준거기준에 해당하는 학생수행의 수준에 대한 명확한 설정이 필요하다.

교육 분야에서 평가의 역할과 기능에 대한 논의는 평가주도 교육과정(test driven curriculum)과 교육과정주도 평가(curriculum

driven test)로 대별된다. 평가주도 교육과정은 학생평가내용이 학생들이 이수해야 할 교육과정을 규정하고, 이에 따라 교육개혁을 위한 효율적 정책방향으로 지지되어 왔다. 물론 평가주도 교육과정으로 인한 교육과정의 협소화, 평가 만능화 등에 대한 다양한 비판이 있어 왔음에도 불구하고, 현대 증거기반 교육연구, 성과기반 교육정책과 같은 객관주의적 관점에서 학생평가의 결과는 교육의 성과를 판단하고 교육개혁을 위한 주도적인 역할을 한다. 이에 따라 학생평가의 객관성은 학생 개인뿐만 아니라 교육기관의 책무성, 국가 교육정책의 효과성을 판단하기 위하여 매우 중요하게 다루어지고 있다. 특히 의학교육은 일반 교육의 공공성뿐만 아니라 의료 인력양성이라는 사회적 책무성을 포함함에 따라 학생평가의 객관성이 더욱 강조된다. 학생평가의 객관성은 평가계획의 수립에서 결과보고에 이르는 전 과정의 내용적, 절차적 타당성에 근거한 것으로, Shepard (2000)는 학생평가에 대하여 다음과 같이 제안한다.

바람직한 학생평가를 위해서는 첫째, 학생의 사고능력 및 실제 수행능력을 향상시킬 수 있는 과제가 주어져야 하며, 둘째, 학습결과뿐만 아니라 학습 과정을 다루어야 하며, 셋째, 수업과 통합된 지속적 활동이어야 하며, 넷째, 학생학습을 지원할 수 있도록 형성적 평가가 이루어져야 하며, 다섯째, 학생들에게 무엇이 기대되는지 명확히 확인시킬 수 있어야 하며, 여섯째, 학생들이 자신의 수행을 평가하는 데 능동적으로 참여하게 하며, 일곱째, 학생학습뿐만 아니라 수업개선을 위하여 평가결과가 활용되어야 한다.

Shepard의 제안은 전통적 지필평가와 수행평가 모두에 적용되는 것으로, 학생평가의 교육적 활용을 강조한다. 평가상황에서 제한적으로 수집된 학생의 말, 행동, 반응 등은 그 학생이 보다 넓은 범위에서 무엇을 알고, 할 수 있으며, 어떤 능력을 갖는지에 대한 추정의 근거가 된다. 이러한 추정의 정확성이 학생평가의 객관성을 의미하며, 이는 학생평가를 구성하는 문항형식, 문항내용, 시행절차, 채점, 결과산출 과정이 논리적, 실천적으로 평가목적에 부합하였는가로 귀결된다.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Long Grove: Waveland Press.

Amin, Z., Chong, Y. S., & Khoo, H. E. (2006). *Practical guide to medical student assessment.* Hackensack: World Scientific.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring and reporting. In R. B. Brennan (Ed.). *Educational measurement* (4th ed., pp. 355-386). Westport: Praeger Publishers.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educ Psychol Meas, 57*(3), 373-399.

Dent, J. A., & Harden, R. M. (2009). *A practical guide for medical teachers* (3rd ed.). Edinburgh: Churchill Livingstone Elsevier.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. B. Brennan (Ed.). *Educational measurement* (4th ed., pp. 471-515). Westport: Praeger Publishers.

Epstein, R. M. (2007). Assessment in medical education. *N Engl J Med, 356*(4), 387-396.

Kane, M. T. (2006). Validation. In R. B. Brennan (Ed.). *Educational measurement* (4th ed., pp. 17-64). Westport: Praeger Publishers.

Kogan, J. R., Holmboe, E. S., & Hauer, K. E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA, 302*(12), 1316-1326.

Lane, S. (2010). *Performance assessment: the state of the art. (SCOPE student performance Assessment Series).* Stanford: Stanford University, Stanford Center for Opportunity Policy in Education.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. B. Brennan (Ed.). *Educational measurement* (4th ed., pp. 387-431). Westport: Praeger Publishers.

Lohr, S. L. (1999). *Sampling: design and analysis*. Pacific Grove: Duxbury Press.

Mavis, B. E., Cole, B. L., & Hoppe, R. B. (2001). A survey of student assessment in US medical schools: the balance of breadth versus fidelity. *Teach Learn Med, 13*(2), 74-79.

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

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River: Pearson Education International Co.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design (RR-03-16).* Princeton: Educational Testing Service.

Mullis, I. V. (1984). Scoring direct writing assessment: what are the alternatives? *Educ Meas: Issues Pract, 3*(1), 16-18.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educ Res, 78*(1), 153-188.

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability.* Stanford: Stanford University, Stanford Center for Opportunity Policy in Education.