보건의료교육평가
Journal of Educational Evaluation for Health Professions
2004. 제1권 제1호, 5-14

의사 국가시험 합격선 설정에 관한 측정학적 접근

이규민

계명대학교 교육학과

A Pychometric Approach to Setting a Passing Score on Korean National Medical Licensing Examination

Guemin Lee
Department of Education, Keimyung University

National Health Personnel Licensing Examination Board (hereafter NHPLEB) has used 60% correct responses of overall tests and 40% correct responses of each subject area test as a criterion to give physician licenses to satisfactory candidates. The 60%-40% criterion seems reasonable to laypersons without pychometric or measurement knowledge, but it may causes several severe problems on pychometrician' s perspective. This paper pointed out several problematic cases that can be encountered by using the 60%-40% criterion, and provided several pychometric alternatives that could overcome these problems. A fairly new approach, named Bookmark standard setting method, was introduced and explained in detail as an example. This paper concluded with five considerations when the NHPLEB decides to adopt a pychometric standard setting approach to set a cutscore for a licensure test like medical licensing examination.

Key Words: Licensure and Certificate Testing, Bookmark Standard Setting, Cutscore, Item Response Theory, Scaling, Equating, Validity, Pychometrics

책임저자 : 이규민
(우)704-701
대구시 달서구 신당동 1000번지
계명대학교 교육학과
Tel: 053-580-5964
Fax: 053-580-5162
E-mail: glee@kmu.ac.kr

서 론

한국보건의료인국가시험원(이하 국시원)이 제시한 『2002년도 제66회 의사 국가시험 응시 안내』를 보면 의학을 전공하는 대학을 졸업하고, 의학사 학위를 받은 자는 이 시험에 응시해서 합격하여야 의사 면허를 취득할 수 있는 것으로 명시하고 있다[1]. 시험은 '의학총론', '의학각론', '보건 의약 관계법' 의 세 과목으로 구성되어 있고, 문제는 450문항으로 이틀 동안 진행되며, 시험 시간은 총 660분이다. 합격자 결정은 전 과목 총점이 60% 이상, 매 과목 40% 이상 득점자로 규정하고 있다. 이 논문은 국시원이 제시한 60%, 40% 합격선 설정이 적절한 지를 측정학적으로 고찰하고, 다양한 합격선 설정 방안을 개괄하고, 앞으로 우리나라 의사 국가시험의 합격선 설정 시 고려해야 할 사항을 검토하고자 한다.

현행 합격선 설정의 적절성

국시원에서 사용하고 있는 전 과목 60%와 매 과목 40% 기준은 일반인들에게 매우 친숙한 수치인 것으로 보인다. 지금까지 받아온 평가에서 60% (또는 100점 만점에서 60점) 기준은 최소 능력(minimum competency)를 나타내는 보편적인 준거로 받아 들여져 왔다. 시험이 여러 과목으로 구성되어 있을 경우, 일

괄적으로 각 과목당 60% 기준을 적용하는 것이 너무 엄격한 기준으로 보일 경우, 과목 당 기준을 40%로 사용하는 것 또한 매우 일반적인 관행으로 보인다. 따라서 국시원에서 사용하는 60% - 40% 기준은 일반인의 관점에서 보면 그리 문제되는 합격선 설정은 아닌 것으로 보여질 것이다.

그러나 측정 전문가의 관점에서 보면 60% - 40% 합격선 설정은 합리적인 기준으로 설정되었다고 보기 어렵다. 왜냐하면, 의사 국가시험은 의사로서 갖추어야 할 기본적인 최소한의 능력을 보유하고 있는지를 판단하여, 의사 면허를 부여하는 목적으로 시행되는 시험이기 때문이다. 전 과목 문항의 60%, 각 과목 문항의 40%를 옳게 응답했다는 것이 의사로서 갖추어야 할 최소한의 기본적인 능력을 보유했음을 나타내 준다고 볼 수는 없을 것이다. 의사 국가시험의 합격선은 분명 능력 있는(competent) 의사 지원자와 능력 없는(incompetent) 의사 지원자를 구별해 줄 수 있도록 설정되어야 한다.

우리가 의사 면허를 부여한다는 것은 어떤 개인에게 의사직과 관련된 일련의 행위, 즉 의료행위를 할 수 있도록 권한(authority을 부여한다는 의미이다. 의사 면허제를 사용하는 것은 능력이 없는 자로 하여금 의사가 되어 받게 될 수 있는 가능한 불이익으로부터 일반 국민을 보호한다는 의도이다. 제대로 기능할 수 없는 자에게 의사 면허를 주는 것은 국민의 건강과 안정에 절대적인 위해가 된다. 이런 관점에서 의사 국가시험은 그 결과 활용의 중요도가 매우 높은 시험이고, 합격선 설정 또한 일반 대중에게 친숙하다는 이유로 쉽게 결정될 성질의 것이 아니다. 이런 관점에서 60% - 40% 기준은 의사로서 기능하기 위해 필수적인 최소 능력을 보증해 주는 기준으로 보기 힘들다.

예를 들어, 1999년은 의사 국가시험이 매우 쉽게 출제되어 평균이 총점의 80%였고, 2000년은 반대로 매우 어렵게 출제되어 평균이 총점의 65%였다면, 1999년 시험에 응시한 학생들이 2000년 시험에 응시한 학생들 보다 더 많이 합격했을 것이다. 물론 1999년과 2000년의 합격률이 다르다는 것이 문제는 아니다. 단지, 시험의 난이도에 의해서 의사 지원자의 합격 여부가 결정된다는 점이다. 결국 쉬운 시험을 친 능력 없는 지원자가 합격할 가능성과 어려운 시험을 쳐서 능력 있는 지원자가 불합격할 가능성이 있다는 말이다. 의사 면허 취득 지원자의 실제적인 능력 외의 시험 난이도가 합격과 불합격을 결정하는 하나의 주요한 요인이 될 수 있다는 점에서, 60% - 40% 기준은 합리적이지 못하다. 그렇다면, 보다 합리적인 합격선 설정 방법은 없는가? 이에 대한 측정학적 방법을 활용해서 합격선을 설정하는 접근을 개괄적으로 살펴보고자 한다.

스탠다드 설정 방법의 발전

합격선은 일종의 컷 스코어(cut score)로 이해될 수 있고, 그 컷 스코어를 기준으로 평가 점수가 넘은 지원자는 합격으로 넘지 못한 지원자로 불합격으로 처리하게 된다. 컷 스코어와 함께 스탠다드는 많은 학자들에 의해 구분되어 사용되기도 하고, 혼용되어 쓰이기도 한다. Zieky는 컷 스코어와 스탠다드를 구별하지 않고 같은 개념으로 사용하며[2], 스탠다드 보다는 컷 스코어가 이해하기 쉬운 개념임을 지적하였다. 반면, Kane은 컷 스코어와 스탠다드 사이에 개념적인 구분을 통해 좀더 명확한 이해에 도달할 수 있다고 주장하였다[3]. 그에 의하면 컷 스코어는 어떤 척도 위에 설정되는 특정 점수를 지칭하고, 스탠다드는 피험자가 알고 있거나 할 수 있는 수행능력 기준으로 일종의 능력 수준을 말한다. 예를 들어, 국어 능력 시험의 경우, 척도 위에 두 개의 컷 스코어가 설정된다면, 점수 척도는 세 부분으로 구분되고, 각각 기초 능력 수준, 중급 능력 수준, 또는 고급 능력 수준과 같은 스탠다드가 설정될 수 있다. 일반적으로 컷 스코어 설정이란 용어 보다는 스탠다드 설정이란 용어가 많이 사용된다.

측정 전문가들이 스탠다드 설정에 관심을 갖기 시작한 것은 20세기 중반부터인 것으로 보인다[2]. 1980년대 초반까지 몇 개의 스탠다드 설정 방법이 고안되어 일반적으로 사용되어 졌는데, 이러한 방법들은 Angoff 방법[4], Ebel 방법[5], Nedelsky 방법[6], 경계 집단 방법[7], 비교 집단 방법[8,9] 등으로 구분된다. 1990년대 초까지 가장 널리 사용된 것은 Angoff 방법이었다. 이 방법은 일반인에게 쉽게 설명될 수 있다는 장점과 선택형 문항

으로 구성되어 있는 검사에 적용된다는 특성을 갖고 있다. 그러나 Angoff 방법이 스탠다드 설정 참여자들이 문항에 옳게 응답할 확률을 제대로 평가할 수 없다는 근본적인 문제를 안고 있다.[10-12]. 반면, 아직까지도 Angoff 방법을 옹호하는 학자들도 있다[13, 14]. 최근에는 Angoff 방법의 문제점을 극복하는 방향으로 또는 새로운 접근 방법으로 스탠다드를 설정하려는 시도가 제안되어 사용되고 있다. 대표적인 것으로는 Bookmark 방법[15], Body of Work 방법[16,17], 군집분석법을 이용한 방법[18,19] 등을 들 수 있다.

스탠다드 설정 방법의 발전은 아이러니하게도 스탠다드 설정 방법이 갖고 있는 본질적인 문제와 연관된다. Glass는 모든 가능한 상황에서 스탠다드 설정은 "임의적"일 수밖에 없음을 지적하였고, 컷 스코어를 설정하는 것은 시간의 낭비라고 주장하였다[20]. Shepard 또한 가능한한 스탠다드 설정은 피해야 한다고 조언하였다[21]. 그러나, 많은 경우에 있어서 스탠다드 설정은 법으로 강제되고, 피할 수 없는 상황이 산재한다. 오히려, 최근에는 스탠다드에 기초한 평가의 중요성이 더욱 강조되었고, 이러한 흐름은 부시 행정부가 들어서며 주요한 교육정책으로 제시한 "No child left behind"라는 표어에도 잘 나타나 있다. 즉, 부시 행정부는 어떤 학생도 설정된 기준, 즉 스탠다드에 미치지 못하도록 남겨두지 않겠다는 큰 포부가 담긴 정책을 발표한 것이다. 결국, 측정 전문가들는 스탠다드 설정의 문제를 극복하기 위해 스탠다드 설정 방법을 발전시켜 온 것이다.

방법론적인 발전에도 불구하고, 스탠다드 설정의 임의성이라는 문제는 아직까지도 스탠다드 설정에 있어서 중요한 이슈가 되고 있다. Kane이나 Jaeger의 지적처럼, 어떤 특정한 컷 스코어를 선정해야 하는 간단하고 분명한 방법은 없다[3, 22]. 또한 설정된 컷 스코어 보다 조금 높거나 낮은 점수를 컷 스코어로 사용하면 안되는 이유도 없다. 컷 스코어 보다 조금 낮은 점수를 받은 피험자가 컷 스코어 보다 조금 높은 점수를 받은 피험자보다 능력 면에서 확실히 못하다고 볼 수도 없다. 그렇기 때문에 스탠다드 설정은 수학적인 계산으로 확정될 수 있는 것이 아닌 복잡한 정책적인 합의의 과정으로 이해되어야 한다. 결국, 스탠다드 설정에 있어서 해결하여야 할 문제는 가능한한 전문가에게나 일반 대중에게 납득될 수 있는 "명확한 의미를 갖는" 컷 스코어를 설정하는 것이다. 컷 스코어 설정이 모두 같은 정도의 불명확성을 갖는 것이 아니며, 어떤 방법을 사용할 경우는 다른 방법을 사용할 경우 보다 더 분명한 의미를 지닌 스탠다드 설정이 가능하다. 이와 같은 관점에서 적절한 방법을 사용하여 컷 스코어를 설정하고, 설정된 스탠다드를 타당화하는 작업이 필요하다.

스탠다드 설정 방법 -Bookmark 방법을 중심으로-

이제 Bookmark 스탠다드 설정 방법을 설명함으로써, 측정학적 스탠다드 설정 방법이 60% - 40%와 같은 임의적인 합격선 설정과 어떻게 다른지, 상대적으로 명확한 의미를 지닌 스탠다드 설정이 어떻게 가능한지를 설명하고자 한다. Bookmark 방법은 Lewis, Mitzel과 Green에 의해 개발된 이후[15], 미국에서 28개 주 이상에서 교육성취도 수준을 설정하기 위하여 사용된 방법으로, 현재 가장 보편적으로 사용되고 있는 스탠다드 설정 방법이다[23,24]. Bookmark 설정 방법은 스탠다드 설정을 위해 참여하는 패널들로 하여금 검사의 본질과 결과 활용에 익숙해지도록 하고, 패널들로 하여금 합격선에 대한 그들의 기대치를 표현할 수 있는 수단을 제공하는 방법이다. 스탠다드 설정 패널은 그 검사와 관련된 전문가로 구성되는데, 예를 들어, 의사 국가시험 합격선 설정을 위한 스탠다드 설정 패널은 의학 교육을 담당하는 교수, 실무 경험이 풍부한 전문의, 의료법 관련 법조인, 보건복지부 관계자 등 의료 관련 전문가가 될 것이다. Bookmark 설정 방법은 다음과 같은 절차를 거쳐 스탠다드를 설정하게 된다.

1 단계 : 패널들에게 난이도별로 쉬운 문항부터 어려운 문항으로 구성된 "순서화된 문제집"(ordered item bookelt: OIB)을 나누어 주고, 문제집 구성에 대해 설명한다. 난이도로 순서화된 문제집을 재구성하기 위해, 측정이론 중의 하나인 문항반응이론(item response theory)이 사용되며, 각 문항에 옳게 응답할 확률이 2/3

가 되는데 필요한 능력 점수를 그 문항의 척도 점수로 사용한다.

2 단계 : 패널을 몇 개의 작은 소그룹으로 나눈다. 예를 들어, 총 참여자가 24명일 경우, 6명씩 4개의 소그룹을 만들고 각 그룹별로 피험자들이 각 문항에 옳게 응답하기 위해 필요한 지식과 기능이 무엇인지 토의하도록 유도한다.

3 단계 : 각 소그룹 별로 토의가 마무리 되면, 각 패널로 하여금 의사 면허를 부여하기 위해 반드시 맞추어야 할 문항 중 난이도 별로 볼 때, 가장 어려운 문항에 표시하도록 한다. 이때 표시된 문항의 척도 점수가 그 패널이 기대하는 첫 번째 컷 스코어이다.

4 단계 : 소그룹 단위로 서로 표시한 문항을 가지고 토의하도록 한다. 이 토의를 통해, 서로 간에 이견을 좁힐 수 있고, 좀더 합의된 컷 스코어에 이를 수 있다. 소그룹 토의가 끝난 후에, 패널은 각자 두 번째로 "순서화된 문제집"에 자신의 선택을 표시하도록 한다. 이것이 그 패널의 두 번째 컷 스코어 기대치가 된다.

5 단계 : 소그룹을 중그룹으로 묶어 각자의 의견을 가지고 토의하도록 한다. 소그룹 토의에서는 특정 개인의 의해 전체 의견이 주도될 가능성이 있으므로, 소그룹을 둘씩 묶은 중그룹 토의를 통해 의견을 교환하고 개인의 선택을 바꿀 수 있는 기회를 준다. 토의가 끝나면, 패널은 각자 다시 순서화된 문제집에 자신의 의견을 표시토록 한다.

6 단계 : 마지막으로 패널 전체를 한 자리에 모아 전체 그룹 토의에서 패널 각자가 선택한 문항을 중심으로 토의를 진행하도록 한다. 이제 중그룹이 아닌 전체 그룹 토의를 통해 패널은 전체적인 의견을 들을 수 있고, 자신의 선택을 바꿀 수 있는 기회를 갖는다. 토의가 마무리되면, 패널은 순서화된 문제집에 자신의 의견을 표시한다.

7 단계 : 스탠다드 설정팀은 최종적으로 표시된 문항의 척도 점수를 수합, 이를 이용하여 컷 스코어를 결정한다. 일반적으로 많이 사용되는 방법은 패널이 지정한 문항들의 척도 점수들의 중앙값을 계산하여 사용하는 방법이지만, 평균값을 사용할 수도 있다.

8 단계 : 스탠다드 설정팀은 설정된 컷 스코어를 중심으로 문항의 척도 점수가 컷 스코어 보다 낮은 문항을 분석함으로써 합격한 사람이 수행할 수 있는 지식이나 기능을 열거하여 설정된 수행 지표 목록을 작성한다. 즉, 해당 스탠다드에 포함되는 점수를 획득한 지원자는 이러한 지식과 능력을 소유하고 있는 것으로 제시된다.

다음의 Table 1은 미국의 한 주에서 초등학교 4학년을 대상으로 Bookmark 방법을 이용하여 스탠다드 설정하고 각 수준별 수행 지표 목록을 제시한 것이다.

지금까지 설명한 Bookmark 스탠다드 설정 방법을 사용해서 의사 국가시험의 합격선을 설정할 경우, 기존의 60% - 40% 기준과 비교해 볼 때, 먼저 의료 관련 전문가들의 합의에 의해 합격선을 설정한다는 점이 가장 큰 다른 점일 것이다. 또한, 그러한 합의의 과정이 단순한 토의를 통해서가 아니라, 실제로 의료 지원자들이 치른 시험 문항을 중심으로 분석된다는 점에서 객관성을 확보할 수 있다. Bookmark 방법은 의료 관련 전문가들이 각각의 문항 난이도를 추정할 수 없다는 Angoff 방

Table 1. Grade 4 mathematics performance levels and descriptions for a state in United States

Performance levels	Performance descriptions
Level 4 (Advanced)	Students use estimation, probabilistic prediction, and graphical representations, and identify equivalence and complex measures. They order decimals and identify, create, and describe combinations and patterns. They also analyze situations, apply and explain reasoning, and draw conclusions.
Level 3 (Proficient)	Students consistently understand probabilities, percents, and relationships among fractions, and identify patterns and parts of various figures. They work with and interpret real-world data. They also solve multi-step problems and present reasonable solutions with justifications.

Level 2 (Progressing)	Students generally are able to use all basic operations and demonstrate an understanding of whole-numbers. They use manipulatives to solve for an unknown and to model simple fractional relationships. They also identify various shapes and patterns and interpret data.
Level 1 (Basic)	Students may use some of basic operations and show some understanding of simple concepts, data, and figures. They may use manipulatives to explore patterns and represent whole-number relationships.

법의 문제점을 보완하였고, 비교적 측정학에 대한 전문적인 지식이 없는 일반인에게 사용될 수 있다는 장점을 갖고 있다. 또한, 합격자가 최소한도로 지녀야 할 의료 지식과 능력을 지표화하여 만듬으로써, 의료 전문가와 일반 대중이 이해하고 납득할 수 있는 합격선을 설정할 수 있게 된다. 이런 측면에서 Bookmark 방법을 이용한 스탠다드 설정 방법은 기존의 60% - 40% 기준 보다는 합리적인 합격선을 제시할 수 있을 것이다.

의사국가시험 합격선 설정시 고려 사항

국시원에서 우리 나라 의사 국가시험의 합격선을 측정학적 방법을 동원하여 설정한다면, 어떤 점을 고려해야 할 것인지에 대해 검토하고자 한다. 스탠다드 설정은 단순히 스탠다드 설정 방법을 동원하여 몇 개의 컷 스코어를 만들어 내는 것으로 종결되는 간단한 문제가 아닌, 전체 검사의 개발과 척도화, 검사 점수의 동등화 등과 관련되어 논의되어야 하는 문제임을 염두에 두어야 할 것이다.

이 연구를 통해 제시되는 고려 사항에 대해, 미국의 경우를 분석함으로써, 그 제안의 타당성을 검토하고자 한다. 미국에서는 우리 나라의 의사 국가시험과 유사한 United States Medical Licensing Examination (USMLE)를 시행하고 있다[25]. 이 시험은 National Board of Medical Examiners(NBME)에서 주관하여 시행하고 있고, 모두 세 스텝으로 나뉘어 시행된다. USMLE는 의료 지식, 개념, 원리, 기능을 적용할 수 있는 종합적인 능력을 측정하여, 안전하고 효율적인 환자 관리가 이루어 질 수 있도록 유도한다. 세 스텝을 모두 통과하여야 하며, 각각의 스텝에 대해 각각의 합격선이 주어진다. 스텝 (1)은 대략 350개의 선택형 문항으로 구성되고, 총 420분의 시험 시간과 쉬는 시간을 합쳐 8시간이 소요된다. 스텝 (2)는 대략 370개의 선택형 문항으로 구성되고, 480분의 시험 시간과 쉬는 시간을 합쳐 9시간이 소요된다. 스텝 (3)는 대략 480개의 선택형 문항과 9개의 컴퓨터 시뮬레이션 케이스 처치 문항이 주어진다. 컴퓨터 시뮬레이션 처치 문항은 컴퓨터 상황에서 환자가 주어지고 지원자는 처치를 함으로써 의사로서의 수행 능력을 평가하게 된다. 스텝 (3)는 하루 8시간 씩, 이틀 동안 치러진다. 이 연구를 위한 분석에는 NBME가 제시한 USMLE Bulletin이 사용되었다.

1. 척도 개발(scaling)이 우선되어야 한다.

컷 스코어는 척도 위에 존재하는 특정 척도 점수이다. 따라서, 척도의 개발이 컷 스코어를 개발하기 위한 전제 조건이 된다. 60% - 40% 기준은 척도의 개발에 기초하고 있지 않다. 단지, 전체 시험 문항의 60%, 또는 과목 시험 문항의 40%만 맞추면 되는 것이다. 척도의 개발이 없으면, 검사 점수의 의미는 찾을 수 없다. 예를 들어, 전체 문항의 60%를 맞추었다는 것이 무엇을 의미하는가? 평균이 전체 문항의 80%일 때와 평균이 전체 문항의 30%일 때, 60%의 의미는 전혀 다르다. 척도의 개발이 없다면 컷 스코어를 설정한다는 것이 불가능하고, 무의미해 진다. 따라서, 의사 국가시험의 합격선 설정을 위해 먼저, 의사 국가시험의 척도를 개발해야 함을 제안하고 싶다.

"The number of test items you answer correctly is converted to two equivalent score, one on a three-digit score scale and one on a two-digit score sclae. Both scales are used for score reporting purposes." (USMLE Bulletin 중에서)

위의 문장에서 알 수 있듯이 USMLE는 두 개의 척도를 가지고 있다. 기본적인 척도는 3자리수 척도로 보통 160점에서 240점 사에에 모든 학생이 위치하게 되고, 평균은 대략 200점에서 220점 정도, 표준편차는 20점 정도라고 한다. 두 자리수 척도는 세 자리수 척도로부터 도출되는데, "75점 합격이라"는 요구를 충족시키기 위해 개발되었다고 한다. 결과적으로, USMLE의 기본 척도는 세 자리수 척도임을 알 수 있다. 단지 합격선의 쉬운 이해를 위해 두 자리수 척도로 변환된 변환 점수 척도도 사용하고 있다. 우리 나라 의사 국가시험도 60% 기준을 사용하고 싶다면, 척도를 개발하고, 다시 두 자리수 척도로 변형하여 설정된 컷 스코어가 두 자리수 척도에서 60에 해당하도록 변환하는 방법을 적용할 수 있을 것이다. 위의 예문을 통해 USMLE는 자체 척도를 개발하여 사용하고 있음을 알 수 있다.

2. 검사 점수의 동등화 (equating)과정이 있어야 한다.

일반적으로 컷 스코어의 설정, 또는 스탠다드의 설정은 검사가 이루어지는 매 해, 매 번 시행되지 않는다. 대신에 서로 다른 해에 개발된 검사의 점수 척도가 서로 비교할 수 있도록 통계적 기법을 이용하여 조정되어 진다. 예를 들어, 1999년의 검사 난이도가 2000년의 검사 난이도 보다 어렵웠다면, 그 만큼 통계적 조정의 작업을 거치게 된다. 이러한 통계적 조정의 과정을 검사 점수 동등화(equating)라고 부른다[26,27]. 검사 점수가 동등화 되면, 한번 설정된 컷 스코어가 계속 사용될 수 있다. 즉, 1999년에 설정된 합격선이 350점이었다면, 2000년 검사 점수와 1999년 검사 점수가 동등화의 과정을 거친 후, 2000년에도 350점의 합격선이 사용될 수 있다. 검사 점수 동등화는 스탠다드 설정 보다 쉽게 성취될 수 있기 때문에, 검사 제작 기관은 스탠다드를 한 번 설정하고, 매 해 검사 점수 동등화 기법을 적용하는 방법을 일반적으로 사용하고 있다.

"A statistical procedure ensures that the performance required to pass each test form is equivalent to that needed to pass other forms; this process also places scores from different forms on a common scale." (USMLE Bulletin 중에서)

위의 문장에서 알 수 있듯이, 한 검사형에서 합격하는데 필요한 수행 능력이 다른 검사형에서 합격하는데 필요한 수행 능력과 상응하도록 만들기 위한 통계적 조정의 과정이 있었음을 말해 주고 있다. 이 과정은 앞서 지적한 것처럼 검사 점수 동등화의 과정이다. 또한 이 과정은 서로 다른 검사형으로부터 나온 점수를 같은 공통의 척도 위로 올려 놓아 준다고 말하고 있는데, 이는 문항반응이론을 이용한 검사 점수 동등화에서 자주 사용되는 용어이다[26,27]. 결과적으로 이 문장을 통해, NBME는 USMLE 척도 개발을 위해 문항반응이론을 사용했고, 매해 검사 점수 동등화 과정을 거치고 있음을 알 수 있다.

3. 적절한 스탠다드 설정 방법이 결정되어야 한다.

앞 절에서 설명한 것처럼, 상당히 많은 수의 스탠다드 설정 방법이 개발되어 사용되고 있다. 어떤 한 방법이 절대적으로 옳거나 정확한 합격선을 제공해 주는 것은 아니다. 그렇다고 아무 방법이나 사용해도 되는 것은 아니다. 가장 먼저 고려해야 할 사항은 검사의 척도 개발과 관련된다. 즉, 척도를 어떤 이론에 기반을 두고 개발하였는가에 따라 스탠다드 설정 방법도 달라져야 할 것이다. 예를 들어, 척도가 고전 검사 이론(classcial test theory)에 기반을 두고 개발되었다면, 전통적인 방법의 스탠다드 설정 방법이 유용할 것이고, 척도가 문항반응이론을 이용하여 개발되었다면 문항반응이론을 이용한 스탠다드 설정 방법이 더 적합할 것이다. 또한 검사가 지필로 이루어지는 검사인지 아니면 컴퓨터로 이루어진 검사인지도 스탠다드 설정 방법에 영향을 미칠 수 있다[28]. 결국, 검사의 전반적인 특성을 감안하여, 적절한 스탠다드 설정 방법을 선정, 사용하여야 할 것이다.

"The USMLE program recommends a minimum passing score for each Step. Recommended performance standards for the USMLE are based on a specified level of proficiency." (USMLE Bulletin 중에서)

USMLE가 보고한 Bulletin에는 어떤 방법을 이용하

여 합격선을 설정하였는지 밝히고 있지 않다. 단지, USMLE는 최소 합격선을 제안하고 있고, 이 제안을 받아들일 것인지 아닌지는 각각의 의료 면허 부여 기관에서 결정하도록 하고 있다. 위의 예문에 나타난 것처럼, USMLE의 스탠다드는 구체화된 "수행 능력 수준"을 바탕으로 하고 있다는 말에서, 앞서 살펴본 측정학적 방법을 이용하여 스탠다드를 설정하였음을 추론할 수 있다. 또한, 척도가 문항반응이론으로 개발되었고, USMLE 검사가 컴퓨터로 시행되는 점을 감안할 때, 그에 상응하는 스탠다드 설정 방법이 사용되었을 것으로 추정된다.

4. 설정된 스탠다드에 대한 주기적인 평가가 이루어져야 한다.

스탠다드가 일단 설정되고 나면, 그 다음에 검사가 개발되어 시행될 때, 검사 점수의 동등화 과정을 밟고, 다시 스탠다드 설정의 절차를 거치지 않음을 지적하였다. 그러나, 다양한 방법을 동원하여 설정된 스탠다드에 대한 타당화의 작업을 시도하여야 한다. 설정된 컷 스코어의 정확성에 관한 정보를 수집한다든지[23], 외부의 준거를 이용하여 설정된 스탠다드가 적절한 지도 평가할 수 있다[29]. 또는 다양한 스탠다드 설정 방법을 동원하여 스탠다드를 설정하고 각각의 방법을 비교하는 접근이나, 서로 다른 검사에 적용된 스탠다드를 통해 나타난 합격-불합격 비율을 비교할 수도 있다[3]. 어떤 방법을 동원하든지 설정된 스탠다드를 주기적으로 평가하는 작업이 동반되어야, 그 스탠다드 사용의 타당도를 높힐 수 있다.

"The recommended minimum passing level is reviewed periodically and may be adjusted at any time." (USMLE Bulletin 중에서)

NBME는 제안된 합격선은 정기적으로 평가됨을 보고하고 있고, 언제라도 필요에 의해 조정될 수 있음을 이야기 하고 있다. 구체적으로 어떤 방법으로 주기적인 평가가 이루어지고 있는지에 대한 정보를 제시하고 있지는 않다.

5. 문제 구성에 있어서 문항 정보를 사용할 수 있다.

일단 합격선이 설정되고 나면, 전체 의사 국가시험 문제의 구성도 그에 맞게 변화될 수 있다. 예를 들어, 100문제로 구성된 수리력 시험에서, A라는 사람의 능력이 70점 정도에 해당한다고 가정하자. 100문제 중 대략 60점 정도에 해당하는 것보다 쉬운 문항은 거의 모두 맞출 것이고, 80점에 해당하는 것보다 어려운 문항은 거의 모두 틀릴 것이다. 결국, A라는 사람의 수리력을 측정하는데 60점에서 80점에 해당하는 몇 몇 문항을 이용하는 것이다. 만일 100문항이 모두 60점에서 80점 사이에 해당하는 문항으로 구성한다면, A라는 사람의 수리력을 보다 정확히 측정할 수 있다. 같은 논리가 의사 국가시험 문제 구성에도 적용될 수 있다. 예를 들어, 의사 국가시험의 평균이 200점이고, 표준편차가 30점이며, 합격선이 150점으로 설정되었다고 가정하자. 이 경우, 문항이 100점에서 300점에 해당하는 문항으로 구성하는 것보다, 150점 주위의 난이도를 가진 (대략 130점에서 170점에 해당하는) 문항으로 구성하는 것이 합격자와 불합격자를 판별하는 데, 더 효율적일 것이다. 즉, 쉽고 어려운 문항을 고르게 썩거어서 지원자를 골고루 변별하는 것이 목적이 아니고, 개별 지원자가 최소한의 기본적인 능력을 보유하고 있는지를 판별하는 것이 중요한 만큼, 합격선 근처에 해당하는 문항들로 검사를 구성하는 것이 바람직하다. 이를 위한, 문항 분석 정보를 이용하는 다양한 방법이 제시되어 있다[30-31].

"While the percentage of correctly answered multiple-choice items required to pass varies from form to form, typically you must answer 60 to 70 percent of items correctly to achieve a passing score." (USMLE Bulletin 중에서)

위의 문장에서 알 수 있듯이, USMLE의 경우, 합격점을 넘기 위해서 옳게 응답해야 하는 문항의 비율은 다르다. 이는 NBME가 국시원에서 사용하는 기준을 사용하고 있지 않다는 증거이다. 문항을 작성하는데 NBME가 위에 제안된 방법을 사용한다는 명시적 근거는 없지만, 합격점을 획득하기 위해 대략 60% 또는 70%의 문항에 옳게 응답해야 한다는 지적에서, 문제가 합격선 근처에서 비교적 쉽게 출제되었다고 볼 수 있다.

결 론

필자는 지금까지 의사 지원자에게 의사 면허를 부여하기 위한 합격선을 중심으로, 그 적절성에 관해 논의하고, 보다 명확하고 합의된 합격선 설정을 위한 측정학적 스탠다드 설정 방안을 개괄했다. 또한 국시원이 측정학적 스탠다스 설정 방안을 채택하여 실시할 경우, 고려해야 할 사항을 검토하고, 미국 NBME의 USMLE의 경우를 분석함으로써 그 타당성을 확보하려고 하였다. 그러나 역시 어떤 합격선 설정 방법을 채택하느냐 또한 정책적 결정의 과정인 점을 고려할 때, 국시원이 앞으로 어떤 결정을 내려 합격선을 설정할 것인지 측정 전문가로서 관심을 갖게 된다. 부디 국시원에서 보다 명확한 합격선 설정에 대한 중요성을 인식하여, 능력 있는 의사 지원자와 능력이 부족한 지원자와 판별할 수 있는 합격선 설정의 방식을 채택하기를 기대해 본다.

핵심 단어 면허 시험, 자격시험, Bookmark 수준 설정, 분할점수, 문항반응이론, 동등화, 타당도

참고문헌

1. 한국보건의료인국가시험원. 2002년도 제66회 의사 국가시험 응시 안내. 서울: 저자. 2002.
2. Zieky MJ. So much has changed: How the setting of cutscores has evolved since the 1980s. In: Cizek GJ, ed. Setting performance standards: concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates, 2001; 19-51.
3. Kane MT. So much remains the same: Comception and status of validation in setting standards. In: Cizek GJ, ed. Setting performance standards: concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates, 2001; 53-88.
4. Angoff WH. Scales, norms, and equivalent scores. In: Thorndike RL, ed. Educational measurement (2nd ed.), Washington, DC: American Council on Education, 1971; 508-600.
5. Ebel RL. Essentials of educational measurement (2nd ed). Englewood Cliffs, NJ: Prentice-Hall, 1972.
6. Nedelsky L. Absolute grading standards for objective tests. Educational and Psychological Measurement 1957; 14: 3-19.
7. Gulliksen H. Theory of mental tests. Hillsdale, NJ: Lawrence Erlbaum Associates, 1987 (Original work was published in 1950).
8. Berk RA. Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education 1976; 45:4-9.
9. Livingston SA. Choosing minimum passing scores by stochastic approximation techniques. Princeton, NJ: Educational Testing Service, 1976.
10. Shepard LA., Glaser R, Linn R, Bohrnstedt G. Setting performance standards for student achievement. Stanford University, Stanford, CA: National Academy of Education, 1993.
11. Kane MT. Examinee-centered vs. task-centered standard setting. In the Proceedings of the Joint Conference on Standard Setting for Large-scale Assessment, Vol. II . Washington, DC: National Assessment Governing Board and the National Center for Educational Statistics. 1995; 119-141.
12. American College Testing. Setting achievement levels on the 1992 national assessment of educational progress in mathematics, reading and writing: A technical report on reliability and validity. Iowa City, IA: Author, 1993.
13. Mehrens WA. Methodological issues in standard setting for educational exams. In the Proceedings of the Joint Conference on Standard Setting for Large-scale Assessment, Vol. II . Washington, DC: National Assessment

Governing Board and the National Center for Educational Statistics. 1995 ; 221-263.
14. Cizek GJ. Reactions to National Academy of Education report, Setting performance standards for student achievement. Washington, DC: National Assessment Governing Board. 1993.
15. Lewis DM, Mitzel HC, & Green DR. Standard setting: A Bookmark approach. Paper presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Boulder, CO. 1996.
16. Kahl SR, Crockett TJ, DePascale CA, & Rindfleisch SL. Using actual student work to determine cutscores for proficiency levels: New methods for new tests. Paper presented at the National Conference on Large-Scale Assessment, Albuquerque, NM. 1994.
17. Kahl SR, Crockett TJ, DePascale CA, & Rindfleisch SL. Setting standards for performance levels using the student-based constructed-response method. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. 1995.
18. Sireci SG. Using cluster analysis to solve the problem of standard setting. Paper presented at the annual meeting of the American Psychological Association, New York. 1995.
19. Sireci SG, Robin F, & Patelis T. Using cluster analysis to facilitate standard setting. Applied Measurement in Education; 1999 ; 12: 301-325.
20. Glass GV. Standards and criteria. Journal of Educational Measurement 1978 ; 15 : 237-261.
21. Shepard LA. Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), Practices and problems in competency-based measurement Washington, DC: National Council on Measurement in Education. 1979 ; 72-88.
22. Jaeger RM. Establishing standards for teacher certification tests. Educational Measurement: Issues and Practice 1992 ; 9: 15-20.
23. Lee G, & Lewis DM. A generalizability theory approach toward estimating standard errors of cut scores set using the Bookmark standard setting procedure. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA, 2001.
24. Hanson B, Lewis DM, Egan K, Lee G, & Patz R. Classification inconsistency when using cut scores set on one form for an alternate form. Paper presented at the 2002 Annual Meeting of the American Educational Research Association, New Orleans, LA, 2002.
25. National Board of Medical Examiners. 2002 United states medical licensing examination (USMLE) Bulletin. Philadelphia, PA: Author. 2002.
26. Kolen MJ, & Brennan RL. Test equating: Methods and practices. New York: Springer-Verlag, 1995.
27. Lee G, Kolen MJ, Frisbie DA, & Ankenmann RD. A comparison of the performance of dichotomous and polytomous IRT models in equating scores from tests composed of testlets. Applied Psychological Measurement 2001; 25: 357-372.
28. Sireci SG, & Clauser BE. Practical issues in setting standards on computerized adaptive tests. In Cizek GJ. (Ed.), Setting performance standards: concepts,methods, and perspectives. Mahwah, NJ: Lwarence Erlbaum Associates, 2001 ; 355-369.
29. Kane MT. Choosing between examinee-centered and task-centered standard-setting methods. Educational Assessment 1998 ; 5: 129-

145.
30. Hambleton RK. Principles and selected applications of item response theory. In Linn RL. (Ed.), Educational measurement (3rd ed.). Phoenix, AZ: Oryz Press, 1989.
31. Hambleton RK, Swaminathan H, & Rogers HJ. Fundamentals of item response theory. Boston: Kluwer-Nijhoff Publishing, 1991.