교육평가연구 http://dx.doi.org/10.31158/JEEV.2018.31.3.631
Journal of Educational Evaluation
2018, 제31권 제3호, 631-651

문항구성 모의실험을 통한 온라인 학업성취도 평가 시스템 분할점수 자동설정 방안 탐색*

김 성 숙** 박 찬 호*** 김 미 경 김 창 환
한국교육과정평가원 계명대학교 한국교육과정평가원

이 연구의 목적은 학업성취도 평가의 기출 문제은행 시스템을 구축함에 있어 기존 문항정보를 사용하여 검사의 분할점수를 자동설정하기 위한 방안을 탐색하는 데 있다. 문항반응이론을 기반으로 2015년과 2016년 중학교 국어 평가 자료를 활용하여 모의실험을 수행하였다. 먼저 각 성취수준에서 1,000명을 무작위 추출한 응답 자료를 대상으로 분석하였다. 검사의 문항 수, 문항의 난이도, 그리고 문항유형을 어떻게 구성하는지 각각의 조건에 따라 분석 대상의 실제 성취수준 집단과 모의 분석을 통해 구분된 성취수준 집단 간의 일치정도를 산출함으로써 분류의 정확성을 판단한 것이다. 분석 결과 일부 모의분석 조건에서 낮은 정분류 비율이 나타났으나, 대체로 성취수준의 정분류 비율은 적절한 것으로 판단되었다. 문항 수와 문항 난이도의 영향이 더 크게 나타난 반면 문항유형 구성 영향 차이는 적은 것으로 나타났다. 모의실험 조건에 따른 비교를 통해 검사지 구성 시 고려해야 할 시사점을 제공하였다.

주제어 : 국가수준 학업성취도 평가, 성취수준, 분할점수, 문항반응이론 척도화

---

* 본 논문은 김미경, 김성숙, 김창환, 김도남, 최인숙(2017) '국가단위 평가의 수요자 중심 정보 활용 서비스 시스템 구축 및 운영(V): 맞춤형 학력진단검사 확대와 시스템 고도화'(한국교육과정평가원 연구보고서)의 연구 내용 일부를 보완 및 재구성한 것임.

** 주 저 자 : 김성숙, 한국교육과정평가원, sungs@kice.re.kr
*** 교신저자 : 박찬호, 계명대학교 교육학과, cpark@kmu.ac.kr

Ⅰ. 서 론

대표적 준거참조평가인 국가수준 학업성취도 평가(이하 학업성취도 평가)는 국가수준 교육과정에 기반한 준거로 지식과 기술의 이해 정도를 측정함으로써 학생들의 학업 성취수준을 판단한다. 학업성취도 평가에서 성취수준은 '교육목표에 대하여 개별 학생들이 알아야 할 것과 할 수 있는 것의 범위와 깊이를 구체적으로 제시함으로써 성취기준에 어느 정도 도달하였는가'를 의미한다. 일반적으로 성취수준을 판단하기 위해서는 성취수준 분할점수를 결정하는 준거설정이 필요하다. 학업성취도 평가의 성취수준 설정은 변형된 Angoff 방법(Angoff, 1971; Cizek & Bunch, 2007, pp. 87-92)을 적용하여 이루어지며, 이를 통해 우수, 보통, 기초, 기초미달을 가르는 세 개의 분할점수가 결정된다. 즉 성취수준 설정에 참가한 패널들은 우수, 보통, 기초 학력에 대한 개념적 정의에 대해 합의하고 각 수준에서의 최소능력 학생의 특성을 내면화하여 문항 단위로 각 성취수준에 대한 기대정답률을 판정함으로써 분할점수를 산출한다(Zieky, Perie, & Livingston, 2008). 즉 학생들의 원점수를 기반으로 성취수준 분할이 이루어지는 것이다(시기자, 김완수, 박인용, 박찬호, 구슬기, 2015).

학업성취도 평가에서 출제되었던 평가 문항들을 문제은행으로 재활용하기 위한 수요자 중심 정보 활용 서비스 시스템 연구가 단계적으로 진행되고 있다(김미경, 김성숙, 김창환, 김도남, 최인숙, 2017; 상경아, 최인봉, 김완수, 이은경, 2013; 최인봉, 김경희, 김미경, 김창완, 한선영, 2016; 최인봉, 박도영, 이은경, 2014; 최인봉, 이채희, 이은경, 박병기, 2015). 초등학교의 경우 학업성취도 평가는 2012년까지만 시행되었으나, 중・고등학교의 경우 현재 표집 평가로 전환되어 지속되고 있으므로 기출 문항이 상당 수 축적되어 있다. 따라서 학업성취도 평가 기출 문제은행 시스템을 구축하게 되면 중・고등학교 교사들이 원하는 난이도 수준과 평가 영역을 고려하여 자유롭게 문제지를 구성하여, 교수-학습 또는 진단이나 형성평가의 목적으로 다양하게 활용할 수 있을 것이다. 이때 준거참조평가인 학업성취도 평가를 제대로 활용하기 위해서는 개별 문항의 성취기준, 정답률 등 문항분석 정보와 함께 구성한 검사에 대한 성취수준 분할점수를 산출할 수 있어야 한다. 이를 통해 다른 검사나 평가 시스템과 차별화된 정보를 제공할 수 있을 것이다.

그러나 중・고등학교 학업성취도 평가 기출 문제은행을 위해 기존의 성취수준 설정 방법을 동일하게 적용하는 데에는 한계가 있다. 변형된 Angoff 방법으로 수백 개의 문항에 대하여 예상 기대 정답률로 우수, 보통, 기초학력의 분할점수를 설정하는 과정은 막대한 시간, 인력, 비용이 투입되어야 하기 때문이다. 뿐만 아니라 그 작업은 한 번에 끝나지 않으며, 매년 추가되는 문항에 대한 성취수준 설정도 고려해야 한다. 따라서 매년 추가해야 할 작업을 최소화하면서 문제은행에서 검사지를 구성할 때 성취수준 분할점수가 자동으로 제공되는 방

안을 강구할 필요가 있다.

분할점수를 자동으로 설정하기 위해서는 문항반응이론(item response theory)을 적용하여 문제은행에 탑재된 모든 문항이 단일한 척도상의 문항모수를 갖도록 하는 방안을 고려할 수 있다(Baker & Kim, 2004). 문항 모수가 동일한 척도에 놓여있도록 척도화하게 되면 문항반응이론의 불변성 원리가 적용되기 때문이다(Kolen & Brennan, 2004, pp. 437-444). 학업성취도 평가 문항은 이미 문항반응이론을 적용하여 분석한 문항특성 정보가 존재하고, 모든 문항의 모수 추정값이 동일한 척도 상에서 비교 가능하도록 척도화되어 있다. 또한 문항반응이론의 진점수를 이용하면 능력점수(Θ)를 원점수 척도로 변환할 수 있다.

따라서 본 연구의 목적은 학업성취도 평가 기출 문제은행 시스템을 효율적으로 구축하기 위해 개별문항의 기존 문항정보를 활용하여 자율적으로 구성된 검사의 성취수준 설정을 자동화할 수 있는 방법을 구안하는 것이다. 또한 학업성취도 평가 데이터를 활용하여 모의실험을 통해 자동설정 방안의 활용 가능성을 탐색할 것이다. 구체적인 연구문제는 다음과 같다. 첫째, 학업성취도 평가 결과 실제 분류된 성취수준 집단과 모의 분석 결과 분류된 성취수준 집단과의 일치정도, 즉 '정분류 비율'은 어떻게 나타나는가? 둘째, 문항 수, 문항 난이도, 문항 유형 등 조건을 다르게 설정할 때 어떤 차이가 나타나는가?

이러한 연구문제에 답하기 위해 먼저 이 문항반응이론 기반의 분할점수 자동설정 방법을 제안하고 모의실험을 통해 이 방법의 적절성을 검증하고자 한다. 모의실험에서는 검사지 구성 시 문항 수, 문항 난이도, 문항 유형 등의 조건에 따라 학생들의 성취수준 분류에 어떠한 차이가 나타나는지 살펴볼 것이다. 모의실험 조건에 따른 비교를 통해 검사지 구성 시 학생들의 분류를 위한 방안을 제안하고자 한다.

Ⅱ. 이론적 배경

1. 문항반응이론과 분할점수 자동설정

학업성취도 평가 문항이 탑재되어 일선 학교에서 활용하게 될 정보 활용 서비스 시스템에는 기실시된 초·중·고 학업성취도 평가 문항이 탑재되어 있으며, 더 이상 평가가 실시되지 않고 있는 초등학교를 제외하고 중학교와 고등학교의 평가 문항은 매년 추가로 탑재될 예정이다. 이 문항들은 학교에서 교사들이 자유로이 활용할 수 있도록 제공되는데, 정답률과 같은 메타데이터가 함께 제공되므로 교사들은 희망하는 난이도 수준과 내용을 고려하여 기출문항들로 검사지를 구성할 수 있게 되어있다. 예를 들어 교사들은 내용 범위를 몇

개의 대단원으로 한정한 후 희망하는 난이도 수준에 맞추어 문제지를 구성할 수 있다. 이를 통해 시스템의 현장 활용도를 제고할 수 있을 것이다.

학업성취도 평가는 평가 결과로 4개의 수준(우수, 보통, 기초, 기초 미달)을 제공하는 준거참조평가(criterion-referenced evaluation)이다. 교사들이 학업성취도 평가 문항을 이용하여 검사지를 구성할 때 평가 결과로 성취수준에 대한 정보를 함께 제공받기를 원한다. 이 경우 매년 모든 학생이 동일한 검사형(test form)으로 평가를 받게 되는 학업성취도 평가와 달리 교사의 재량에 의해 검사형이 새로이 구성되게 되므로 성취수준 부여를 위한 분할점수를 어떻게 설정할 것인가가 관건이 된다. 검사형이 저난도 문항으로 구성될 경우 분할점수가 그에 맞춰 올라가야 하며 고난도 문항일 경우에는 맞춰서 낮아져야 하기 때문이다.

검사형의 구성에 따라 분할점수를 정하는 방법으로는 시스템에 탑재된 모든 문항에 수준별 최소능력자(minimally competent person, 이하 MCP)의 기대 정답률을 미리 정하는 방법이 있다(Berk, 1996; Cizek & Bunch, 2007)). 준거설정 방법으로 널리 사용되는 변형된 Angoff 방법(Angoff, 1971)은 문항별 MCP의 기대 정답률을 합산하여 분할점수로 설정하기 때문이다. 이렇게 문항별 MCP의 기대 정답률을 각 문항의 메타 데이터로 시스템에 탑재하게 되면 사용이 편리하며 추가 작업이 필요하지 않다는 장점이 있다. 그러나 문항 당 세 개의 기대 정답률(분할점수별)을 설정하는 작업이 필요하다는 것은 단점이다. 학교급과 과목을 고려하면 매년 시스템에 탑재될 수백 개의 문항에 대해 준거설정 작업을 실시하기 위해서는 매년 막대한 시간과 비용이 필요하므로 현실적으로 받아들이기 어려운 안이다.

따라서 매년 추가해야 할 작업을 최소화하면서 시스템에서 검사형이 구설될 때 분할점수를 자동으로 설정할 수 있는 방안을 강구할 필요가 있다. 분할점수를 자동으로 설정하기 위해서는 시스템에 탑재된 모든 문항이 단일한 척도상의 값을 지니도록 척도화하는 것이 일차적으로 필요하다. 이를 위해 문항반응이론(de Ayala, 2009)의 불변성 원리를 활용할 필요가 있다. 이때 문항반응이론을 기반으로 하는 준거설정 방법인 Bookmark 방법(Lewis, Mitzel, & Green, 1996)을 적용하거나 또는 변형된 Angoff 방법으로 설정된 기존의 분할점수를 Bookmark 방법으로 설정된 분할점수처럼 취급하여 문항반응이론을 적용하는 방안도 고려할 수 있다.

2. Bookmark 준거설정 방법

Bookmark 방법은 문항분석에 기반을 두고 있는 준거설정 방법이며 문항반응이론을 기반으로 한다(Lewis, Mitzel, & Green, 1996; Mitzel, Lewis, Patz, & Green, 2001). 즉 시스템 안의 모든 문항이 문항반응이론으로 사전에 분석되어 있다는 것을 가정한다. 또한 모든 문항의 모수(parameter) 추정치가 동일한 척도상에서 서로 비교가 가능하도록 척도화될 필요가 있다

(Kolen & Brennan, 2004). 문항 모수가 동일한 척도에 놓여있지 않을 경우 문항반응이론의 불변성 원리가 적용되지 않기 때문이다.

Bookmark 방법을 적용하기 위해서는 순서화된 문항집(ordered item booklet, 이하 OIB)을 구성해야 한다(Cizek & Bunch, 2007). 일반적으로 Bookmark 방법을 적용할 때 MCP가 문항의 답을 맞힐 수 있는 능력이 있는지 여부를 판정하기 위해 2/3의 응답확률(response probability, 이하 RP)을 기준으로 하기 때문에 동일 척도상에 놓인 문항 모수를 이용하여 RP=2/3에 해당하는 문항반응이론 능력점수(Θ)의 값에 따라 OIB를 구성하게 된다. Rasch 모형과 2모수 모형에 대해 RP=2/3에 대한 Θ 값을 구하는 방법은 Cizek과 Bunch가 소개하고 있으나 3모수 모형에 대해서는 소개되어 있지 않다. 3모수 모형의 공식을 이용하여 $\Theta_{RP=2/3}$ 값을 구해보면 다음과 같다.

$$
\begin{aligned}
& c+(1-c)p^{*}=\frac{2}{3} \\
\Rightarrow & (1-c)p^{*}=\frac{2-3c}{3} \\
\Rightarrow & p^{*}=\frac{2-3c}{3-3c}
\end{aligned}
\qquad \text{(식 1)}
$$

여기에서 p*는 2모수 모형의 정답확률로 다음과 같다.

$$
p^{*}=\frac{\exp[1.7a(\theta-b)]}{1+\exp[1.7a(\theta-b)]} \qquad \text{(식 2)}
$$

(식 1)과 (식 2)를 결합하여 다음 식이 유도된다.

$$
\begin{aligned}
& \frac{\exp[1.7a(\theta-b)]}{1+\exp[1.7a(\theta-b)]}=\frac{2-3c}{3-3c} \\
\Rightarrow & \exp[1.7a(\theta-b)]=2-3c \\
\Rightarrow & 1.7a(\theta-b)=\log(2-3c) \\
\Rightarrow & \theta=b+\frac{\log(2-3c)}{1.7a}
\end{aligned}
\qquad \text{(식 3)}
$$

즉, $b+\log(2-3c)/1.7a$로 계산된 $\Theta_{RP=2/3}$ 값에 따라 OIB를 구성할 수 있다. OIB를 이용하여 세 개의 분할점수에 해당하는 문항만 찾으면 되기 때문에 OIB에는 시스템 안의 모든 문항

을 포함할 필요가 없다. 또한 원하는 정확도에 따라 $\Theta_{RP}=2/3$ 값이 .1 또는 .5의 간격이 되도록 OIB를 구성할 수 있다.

Bookmark 방법의 장점은 선다형과 같은 선택형 문항뿐만 아니라 구성형의 다분문항으로 구성된 검사, 선택형과 서답형 문항이 혼재하는 혼합형 검사에도 쉽게 적용할 수 있다는 점을 들 수 있다. 또한 모든 문항에 대해 판정을 내려야 하는 변형된 Angoff와는 달리 분할점수의 개수만큼만 문항을 선정하면 되므로 비교적 간단하며, 문항반응이론의 불변성을 기반으로 하기 때문에 서로 다른 대상에게 실시된 문항들도 동일한 기준으로 비교할 수 있다. 그리고 오차를 포함한 관찰점수를 이용하는 변형된 Angoff 방법과는 달리 오차를 배제한(error-free) Θ 점수를 이용하며, 이 Θ 점수는 또한 연속변수이므로 원점수보다 정밀한 측정이 가능하다는 장점을 들 수 있다(Embretson & Reise, 2000).

반면 Bookmark 방법을 적용하기 위해서는 문항반응이론에 따른 문항분석과 척도화(scaling) 작업을 거쳐야 하므로 준비가 복잡하고 사용자가 이해하기 어렵다는 단점이 있다. 다만 비전문가는 이해하기 어려운 Θ점수 척도의 문제를 해결하기 위해 문항반응이론의 진점수(true score)를 이용하여 Θ점수를 원점수 척도로 변환하여 사용하는 방법을 고려할 수 있다. 그러나 점수 변환 과정에서 오차가 추가로 개입될 수 있으며 원점수와 진점수의 의미가 서로 다를 수 있다는 단점이 존재한다. 그리고 3모수 모형을 사용할 경우 검사형에 포함된 모든 문항의 c-모수 합이 진점수의 최솟값이 되므로 낮은 수준에 대한 분할점수는 설정이 어려울 수 있다(Kolen & Brennan, 2004).

Ⅲ. 연구 방법

1. 분석 자료

본 연구에서 실시한 모의실험은 2015년과 2016년 학업성취도 평가 국어 과목의 중학생 자료를 기반으로 하였다. 2년 간의 학업성취도 평가 자료에서 우수, 보통, 기초, 기초미달 등 성취수준별로 각 1,000명의 채점결과 자료를 무작위로 추출하였다. 피험자의 능력점수(Θ)는 문항반응 자료와 동등화된 문항 모수를 이용하여 문항반응이론 3모수 모형에 따른 최대우도 추정값(maximum likelihood estimate)을 이용하였다. 그리고 모의실험 과정에서 문항반응이론 척도상의 성취수준 분할점수를 원점수로 변환하여 사용하게 되므로 원점수를 기준으로 성취수준과 문항반응이론 척도 구분에 의한 성취수준이 동일한 학생 자료만을 사용하였다. 예를 들어, 2015년 중학교 국어에서 우수 성취수준 학생 1,000명의 문항반응 자료를 이용하여 각

각의 문항반응이론 척도 상 능력점수(Θ)를 추정하고 문항반응이론 척도에서도 동일하게 우수 성취수준으로 구분되는 학생의 자료만 이용한 것이다. 사전 분석을 통해 연도 간 차이가 없는 것을 확인하고, 2015년과 2016년 학생자료를 통합하여 분석하였으며, 성취수준별로 학생 수가 다르게 나타난 것을 감안하여 각 성취수준에서 1,000명을 무작위 추출한 후 모의실험을 진행하였다. 평가 문항의 경우 2015년과 2016년 문항을 통합하여 문제은행을 구성하였으며 유형별 문항 수는 <표 1>과 같다.

〈표 1〉 모의실험에서 사용한 문항 수와 유형

선다형			서답형			총계
2015	2016	소계	2015	2016	소계	
40	40	80	8	6	14	94

2. 모의실험 조건

모의실험을 진행하는 과정에서 고려해야 할 조건은 문항의 수, 문항의 난이도 수준 그리고 문항유형 등 세 가지로 설정하였다. 즉 문항의 수가 많고 적음에 따라 어떤 결과를 나타내는지, 문항의 난이도가 골고루 섞인 경우와 쉬운 문항을 또는 어려운 문항을 주로 사용한 경우 어떤 결과를 나타내는지, 그리고 문항 유형을 선다형 문항만 사용한 경우와 선다형과 서답형을 문항을 함께 사용한 경우에 어떤 결과를 나타내는지 각각 분석하였다. 문항수는 커질수록 더 정확한 결과를 기대할 수 있으나 학교에서 40문항 이상으로 구성된 검사를 사용하지 않는다는 교사들의 의견을 반영하여 최대 30문항으로 설정하였다. 또한 난이도의 단순 정답률을 기준으로 쉽고 어려운 정도를 의미하며, 정답률 기준으로 쉬운 문항(정답률 상위 50%)에서 선택하는 것을 '쉬운 문항'으로, 어려운 문항(정답률 하위 50%)에서 선택하는

〈표 2〉 분할점수 자동설정 모의실험을 위한 문항 조건 요소

문항	조건 요소		
문항 수	① 10개 문항	② 20개 문항	③ 30개 문항
난이도	① 골고루 선택 ② 쉬운 문항 선택(정답률 기준 상위 50% 문항 선택) ③ 어려운 문항 선택 (정답률 기준 하위 50% 문항 선택)		
문항 유형	① 선다형 서답형 혼합 구성		② 선다형 구성

것을 '어려운 문항' 조건으로 구분하였다. 각각의 조건은 사전 분석 결과를 고려하여 설정하였다. 모의실험을 위한 분석 조건 요소를 정리하면 <표 2>와 같다. 세 가지 조건의 각 요소를 교차하면 총 18개의 경우가 나오며 경우 당 1,000번씩 반복하여 분석한 결과의 평균값을 분석하였다.

3. 모의실험 절차

모의실험 절차는 다음과 같다. 첫째, 2015년과 2016년 중학교 국어 학업성취도 평가에서 우수, 보통, 기초, 기초미달의 성취수준별로 1,000명의 자료를 무작위로 추출하였다. 둘째, 학업성취도 평가에서 성취수준 분할을 위해 사용하고 있는 원점수 상 분할점수(변형된 Angoff 방법으로 산출)를 문항반응이론 척도로 변환하였다. 셋째, 모의실험 조건에 따라 검사지를 구성하였다. 넷째, 구성된 검사지별로 문항반응이론 척도 상의 분할점수를 다시 원점수로 변환하였다. 원점수로 변환한 이유는 문제은행 시스템을 사용하게 될 교사들이 문항반응이론 척도상의 피험자 능력치(일명 θ)를 이해하기 어렵기 때문이다. 다섯째, 피험자의 원점수와 원점수 척도로 변환된 성취수준 분할점수를 이용하여 피험자의 성취수준을 판정하였다. 이때 피험자가 원래 속한 성취수준 집단과 모의실험을 통해 분류된 성취수준 집단이 동일한 경우 정분류, 다른 경우 오분류로 판정하였다.

문항반응이론을 적용하기 위한 학업성취도 평가 문항들은 선다형(이분문항)과 서답형 문항(이분 또는 다분문항)이 모두 포함되어 있으므로 이분문항과 다분문항의 혼합형 검사 형태임을 고려해야 한다. 선다형 이분문항에는 3모수 모형, 서답형 이분문항에는 2모수 모형, 서답형 다분문항에는 일반화부분점수 모형을 적용하였다(Baker & Kim, 2004, pp. 14-21; de Ayala, 2009, p.100, p.124, p.209). 선다형의 이분문항에는 다음 식 (1)과 같이 문항의 변별도(a-모수), 난이도(b-모수), 추측도(c-모수)를 고려하는 3모수 모형을 적용한다. 서답형 문항 중 이분문항의 경우 추측으로 답을 맞힐 가능성이 없으므로 식 (2)와 같이 문항의 변별도(a-모수), 난이도(b-모수)만을 고려하는 2모수 모형이 적합하다. 그리고 서답형의 다분문항(2점 이상의 점수이며 부분점수가 허용되는 문항)은 2모수 모형을 확장하여 문항 변별도(a-모수)와 난이도(b-모수) 외에 문항의 범주 수(예를 들어, 2점 만점 문항의 경우 0, 1, 2점 등 세 개의 범주)보다 하나 적은 수의 단계모수(s-모수)를 고려하여 일반화부분점수 모형을 적용한다. 각 식에서 공통적으로 사용되는 D값은 로지스틱 척도를 정규 오자이브 척도로 변환하기 위한 교정상수 1.702 를 의미한다(Baker & Kim, 2004, p.15; de Ayala, 2009, p.368).

$$\text{3모수 모형:}\quad P_i(\theta)= c_i + \frac{(1-c_i)}{1+e^{-Da_i(\theta-b_i)}}$$

$$\text{2모수 모형:}\quad P_i(\theta)= \frac{1}{1+e^{-Da_i(\theta-b_i)}}$$

$$\text{일반화부분점수 모형:}\quad P_{ik}(\theta)= \frac{\exp\sum_{g=0}^{k}\left[Da_i(\theta-\{b_i-s_{ih}\})\right]}{\sum_{h=0}^{K}\exp\sum_{g=0}^{h}\left[Da_i(\theta-\{b_i-s_{ih}\})\right]}$$

<표 3>에서 모의분석에 사용된 학생 집단의 기술통계에서 사례수 비율을 보면 원점수 기반 척도점수를 이용한 성취수준 분류와 문항반응이론 기반 성취수준 분류 간 일치율을 확인할 수 있다. 일치율, 즉 정분류 비율은 '기초미달' 성취수준이 가장 낮고, '보통'이 가장 높은 것으로 나타났다. 이때 오분류의 원인은 측정오차, 척도 변환에 따른 오차(척도연계 오차), 구분점수 올림에 따른 오차 등에 기인하는 것으로 볼 수 있다. 연속변수인 문항반응이론 척도 점수는 이산변수인 원점수로 변환하는 과정에서 가장 가까운 상위 점수로 올림이 이루어지기 때문이다.

〈표 3〉 분석 대상 학생 집단의 기술통계

성취수준	사례 수(비율*)	$\widehat{\theta_{MLE}}$		척도점수	
		평균	표준편차	평균	표준편차
우수	1,789(89.45)	1.26	0.57	238.09	15.05
보통	1,836(91.80)	-0.11	0.38	197.24	11.71
기초	1,834(91.70)	-1.31	0.30	160.26	9.38
기초미달	1,460(73.00)	-2.91	0.70	108.04	25.20

* 원점수 기반 척도점수를 이용한 성취수준 분류와 문항반응이론 기반 성취수준 분류 간 일치율을 의미함.

Ⅳ. 연구 결과

1. 성취수준 분할 원점수

분석대상 학생이 학업성취도 평가 결과 실제 포함되어있는 성취수준 집단과 모의분석 결과 분류된 성취수준 집단과의 일치정도를 '정분류 비율'이라고 칭하였다. 모의 분석을 위해 사용된 각 조건 즉 문항 수, 문항 난이도, 문항유형 조건에 따른 정분류 비율을 산출하였다. 그리고 원점수 척도상에서 성취수준을 구분하는 분할점수를 '성취수준 분할 원점수'라고 명명하였다. 중학교 국어에 대한 모의분석 결과 산출된 성취수준 구분을 위한 분할 원점수는 아래 <표 4>와 같다. 모의분석에서 비교를 위해 서답형 문항의 경우를 포함한 모든 문항을 1점으로 가정하여 계산하였다. 예를 들어, 검사를 선다형 및 서답형 문항의 혼합으로 구성하고 난이도는 고르게 유지하며 문항 수를 10개로 하였을 때 '우수'와 '보통' 을 구분하는 분할 원점수는 8.53점으로 산출되었다. 이는 1,000번의 반복 과정에서 최소 6점, 최대 10점(즉 주어진 10문항을 모두 맞혀야 '우수'가 된다는 의미)이었으며 평균 8.53점이고 표준편차는 .60으로 나타난 것을 의미한다.

문항유형을 어떻게 구성하는지, 난이도를 어떤 방식으로 선택하여 구성하는지, 그리고 문항 수 증가에 따라 어떻게 분할 원점수가 산출되는지를 정리하였다. 문항 수 조건을 보면 대체로 문항 수가 증가하는 데 비례하여 분할점수가 올라가지만 문항 수 증가와 정확하게 비례하여 분할점수가 오르지는 않는 것으로 나타났다. 예를 들어 혼합 문항구성과 고른 난이도 조건에서 '기초'와 '기초미달'을 구분하는 점수는 10문항일 때 평균 3.38점, 20문항일 때 평균 6.27점, 30문항일 때 평균 9.14점으로 나타났다. 10문항의 평균 3.38점을 기준으로, 20문항과 30문항에 대한 단순 비례로 예측할 수 있는 분할 점수는 각각 6.76점, 10.14점이었다. 즉 문항 수가 증가함에 따라 비례식으로 예측된 값보다 분할점수가 작게 나타나고 이는 문항 수가 증가함에 따라 오차가 줄어드는 것으로 해석할 수 있다. 따라서 문항 수가 적으면 성취수준 분할 원점수가 과대 추정될 수도 있음을 시사한다.

정답률 기준 문항 난이도가 골고루 구성된 것을 기준으로 할 때 문항 난이도가 낮으면 성취수준 분할 원점수는 높아지고, 난이도가 높으면 분할 원점수는 낮아지는 것으로 나타났다. 문항 난이도가 낮을 경우 10문항일 때 '우수'와 '보통' 의 분할점수가 10점으로 나타났으며, 특히 선다형만으로 구성할 때에는 대부분 항상 '우수'와 '보통' 분할점수가 10점으로 산출되었다. 그리고 문항 유형을 선다형만으로 검사지를 구성할 경우 혼합형인 경우보다 성취수준 분할 원점수 평균은 약간 올라가지만 차이가 크지 않으므로, 10문항의 선다형 문항만으로 '우수'와 '보통' 수준을 정밀하게 구분해야 하는 상황이 아니라면 문항 유형은 크게

〈표 4〉 성취수준 분할 원점수 산출 결과

문항유형	난이도	문항수	성취수준 구분	평균	표준편차	최솟값	최댓값
혼합	고름	10	우수/보통	8.53	0.60	6	10
			보통/기초	5.59	0.72	3	8
			기초/기초미달	3.38	0.56	2	5
		20	우수/보통	16.59	0.78	14	19
			보통/기초	10.65	0.90	8	13
			기초/기초미달	6.27	0.66	5	8
		30	우수/보통	24.62	0.86	22	27
			보통/기초	15.69	1.00	13	19
			기초/기초미달	9.14	0.76	7	11
	낮음	10	우수/보통	9.85	0.36	9	10
			보통/기초	7.18	0.53	6	9
			기초/기초미달	4.21	0.54	3	6
		20	우수/보통	18.97	0.20	18	20
			보통/기초	13.88	0.59	12	16
			기초/기초미달	7.87	0.64	6	10
		30	우수/보통	28.05	0.21	28	29
			보통/기초	20.55	0.56	19	22
			기초/기초미달	11.57	0.61	10	13
	높음	10	우수/보통	7.41	0.58	6	9
			보통/기초	4.00	0.34	3	5
			기초/기초미달	2.66	0.47	2	3
		20	우수/보통	14.33	0.70	12	16
			보통/기초	7.49	0.52	6	9
			기초/기초미달	4.75	0.43	4	5
		30	우수/보통	21.22	0.67	20	24
			보통/기초	10.97	0.42	10	12
			기초/기초미달	6.84	0.37	6	7

〈표 4〉 성취수준 분할 원점수 산출 결과 (계속)

문항유형	난이도	문항수	성취수준 구분	평균	표준편차	최솟값	최댓값
선다	고름	10	우수/보통	8.60	0.63	7	10
			보통/기초	5.68	0.67	4	8
			기초/기초미달	3.56	0.55	2	5
		20	우수/보통	16.76	0.76	14	19
			보통/기초	10.86	0.87	8	14
			기초/기초미달	6.65	0.66	5	9
		30	우수/보통	24.84	0.86	22	27
			보통/기초	16.01	0.96	13	19
			기초/기초미달	9.68	0.72	8	12
	낮음	10	우수/보통	10.00	0.03	9	10
			보통/기초	7.33	0.51	6	9
			기초/기초미달	4.37	0.54	3	6
		20	우수/보통	19.01	0.10	19	20
			보통/기초	14.14	0.54	13	15
			기초/기초미달	8.28	0.61	7	10
		30	우수/보통	28.59	0.49	28	29
			보통/기초	20.94	0.48	20	22
			기초/기초미달	12.13	0.56	11	13
	높음	10	우수/보통	7.41	0.61	6	9
			보통/기초	4.03	0.32	3	5
			기초/기초미달	2.94	0.23	2	3
		20	우수/보통	14.30	0.65	12	16
			보통/기초	7.60	0.50	7	9
			기초/기초미달	5.01	0.09	4	6
		30	우수/보통	21.22	0.58	20	23
			보통/기초	11.13	0.39	10	12
			기초/기초미달	7.12	0.32	7	8

고려하지 않아도 되는 조건임을 알 수 있다.

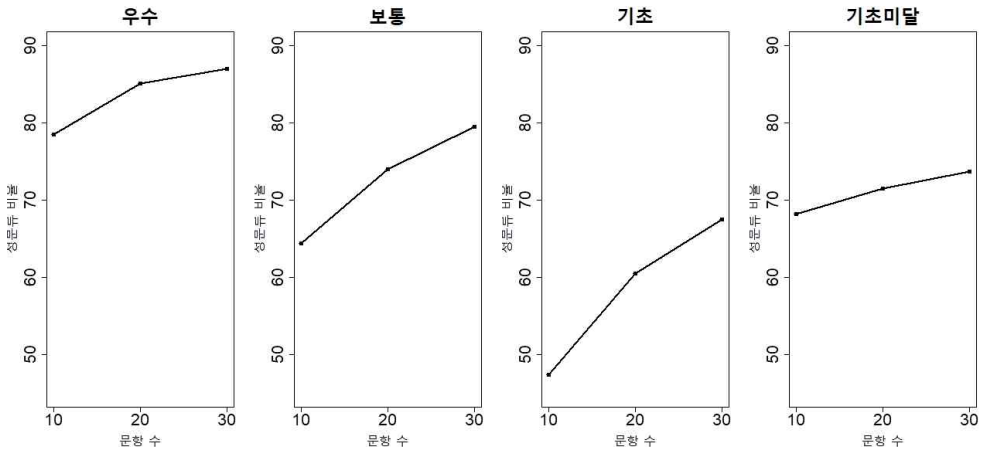
2. 조건 별 정분류 비율 결과

본 연구에서 분석대상 학생이 학업성취도 평가 결과 실제 포함되어있는 성취수준 집단과 모의분석 결과 분류된 성취수준 집단과의 일치 정도를 '정분류 비율'이라고 칭하였다. 모의분석을 위해 사용된 각 조건 즉 문항 수, 문항 난이도, 문항유형 조건에 따른 정분류 비율을 산출하였다. 또한 다른 조건에 대한 평균을 이용하여 문항 수에 의한 조건을 비교한 결과는 <표 5>와 같다. 모든 경우에 문항 수가 증가함에 따라 정분류 비율이 증가함을 확인할 수 있다.

그리고 성취수준에 따른 집단별로 정분류 비율만을 그래프로 나타낸 결과는 [그림 1]과 같다. 정분류 비율은 10문항일 경우 '기초' 집단이 가장 낮고 '우수' 집단이 가장 높게 나타났으며 문항 수가 증가함에 따라 '기초' 집단의 정분류 비율이 가장 빠르게 증가하는 것을 알 수 있다. 아울러 다른 조건에 의한 차이는 고려하지 않고 문항 난이도에 의한 조건만을

〈표 5〉 문항 수에 따른 성취수준별 정분류 비율 (단위: %)

문항 수	실제 성취수준	분류 결과							
		우수		보통		기초		기초미달	
		평균	표준 편차	평균	표준 편차	평균	표준 편차	평균	표준 편차
10	우수	78.58	6.53	21.29	6.15	0.73	0.74	0.03	0.07
	보통	21.26	6.52	64.45	7.78	28.56	6.28	5.15	2.56
	기초	0.16	0.25	12.16	4.08	47.35	9.10	26.61	7.71
	기초미달	0.01	0.02	2.10	1.33	23.36	6.38	68.22	7.54
20	우수	85.12	3.74	16.59	3.56	0.04	0.08	0.00	0.00
	보통	14.88	3.74	73.98	4.26	23.67	4.35	1.84	1.03
	기초	0.00	0.02	8.97	2.15	60.56	5.93	26.66	4.92
	기초미달	0.00	0.00	0.47	0.30	15.73	3.59	71.50	4.81
30	우수	87.03	3.39	13.14	2.85	0.00	0.02	0.00	0.00
	보통	12.97	3.39	79.57	3.28	20.50	3.09	0.77	0.36
	기초	0.00	0.00	7.14	1.42	67.52	3.98	25.55	3.89
	기초미달	0.00	0.00	0.15	0.11	11.98	2.36	73.68	3.93

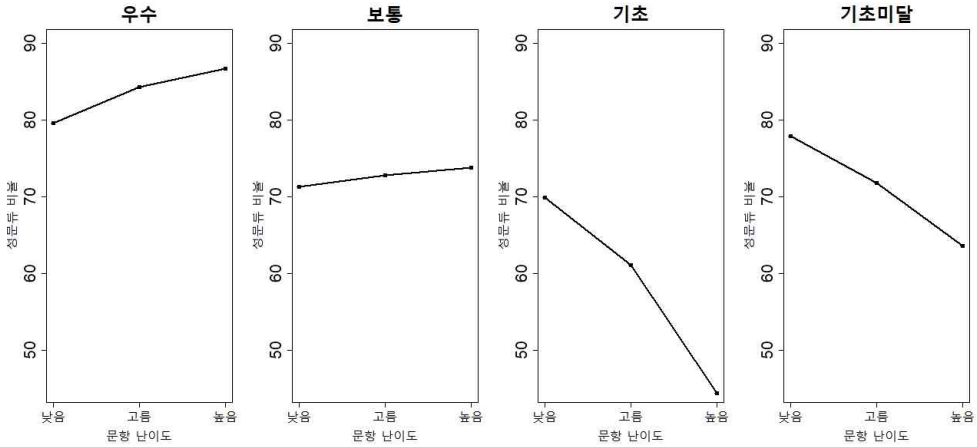


〔그림 1〕 문항 수에 따른 성취수준별 정분류 비율 그래프

비교한 결과는 <표 6>과 같다. 모의 분석에서 산출된 정분류 비율은 '우수'와 '보통' 집단에서는 문항 난이도가 높을수록 정분류 비율이 낮아지고, '기초'와 '기초미달' 집단에서는 문항 난이도가 낮을수록 높게 나타나고 있다.

〈표 6〉 문항 난이도에 따른 성취수준별 정분류 비율 (단위: %)

문항 난이도	실제 성취수준	분류 결과							
		우수		보통		기초		기초미달	
		평균	표준편차	평균	표준편차	평균	표준편차	평균	표준편차
고름	우수	84.33	4.77	16.80	4.35	0.30	0.33	0.01	0.03
	보통	15.61	4.76	72.82	5.22	23.80	4.68	1.41	1.13
	기초	0.06	0.11	9.95	2.70	61.09	6.71	26.77	5.71
	기초미달	0.00	0.00	0.43	0.45	14.81	4.31	71.81	5.84
낮음	우수	79.64	5.55	20.00	4.95	0.25	0.26	0.00	0.00
	보통	20.31	5.55	71.31	5.64	20.69	4.10	0.28	0.24
	기초	0.05	0.08	8.62	2.32	69.94	4.71	21.79	4.33
	기초미달	0.00	0.00	0.08	0.09	9.11	2.47	77.93	4.36
높음	우수	86.75	3.33	14.22	3.25	0.22	0.24	0.02	0.05
	보통	13.20	3.33	73.87	4.46	28.25	4.94	6.07	2.58
	기초	0.05	0.08	9.69	2.63	44.39	7.59	30.25	6.48
	기초미달	0.01	0.02	2.21	1.20	27.14	5.54	63.65	6.08



〔그림 2〕 문항 난이도에 따른 성취수준별 정분류 비율 그래프

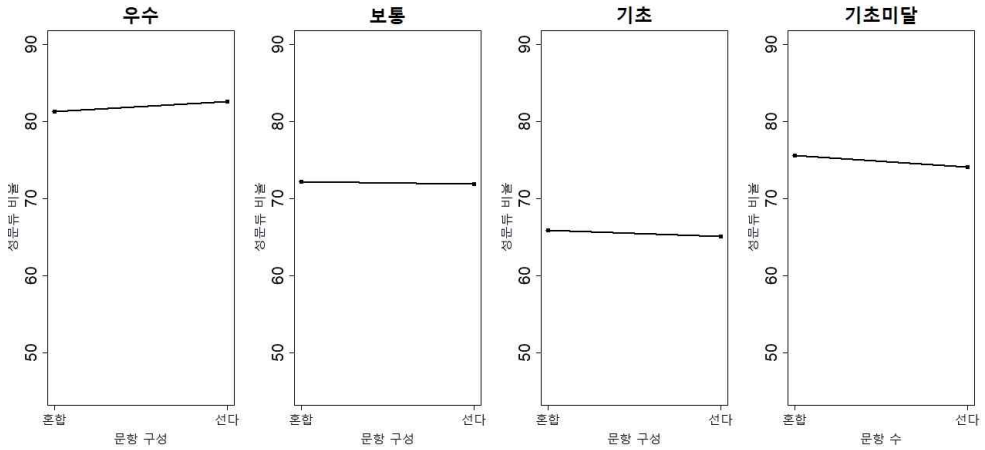
성취수준별로 정분류 비율을 '낮음-고름-높음' 순으로 산출하여 그래프로 나타낸 결과는 [그림 2]와 같다. 문항 난이도가 낮은 경우의 정분류 비율은 네 종류의 성취수준 집단 간 차이가 크지 않는 것으로 나타났다. 그러나 문항 난이도가 높아짐에 따라 '기초미달'과 '기초' 집단의 정분류 비율은 떨어지고, '보통'과 '우수' 집단에서의 정분류 비율은 올라가는 것을 알 수 있다. 즉 문항 난이도는 '기초' 집단에서 가장 큰 변화를 나타내고 있으므로 영향력이 크다는 것을 시사한다.

〈표 7〉 문항 구성에 따른 성취수준별 분류 비율 (단위: %)

문항 유형	실제 성취수준	분류 결과							
		우수		보통		기초		기초미달	
		평균	표준 편차	평균	표준 편차	평균	표준 편차	평균	표준 편차
혼합	우수	81.38	5.42	18.22	4.75	0.28	0.32	0.00	0.01
	보통	18.55	5.42	72.18	5.50	22.10	4.39	0.75	0.65
	기초	0.06	0.11	9.36	2.56	65.96	5.76	23.65	4.99
	기초미달	0.00	0.00	0.24	0.26	11.65	3.37	75.60	5.07
선다	우수	82.59	4.90	18.58	4.56	0.27	0.28	0.00	0.01
	보통	17.36	4.90	71.94	5.36	22.38	4.39	0.94	0.72
	기초	0.05	0.09	9.21	2.45	65.08	5.66	24.91	5.05
	기초미달	0.00	0.00	0.27	0.28	12.27	3.42	74.15	5.13

문항 수와 난이도 등 다른 조건에 대한 평균을 이용하여 문항 유형 구성에 의한 조건을 비교한 결과는 <표 7>에서 볼 수 있듯이 모든 성취수준 집단에서 혼합형과 선다형 간 정분류 비율 차이는 크게 나타나지 않았다. 문항 유형을 어떻게 구성하는지에 따라 전반적으로 '보통'과 '우수' 집단이 '기초'와 '기초 미달' 집단보다 정분류 비율이 높음을 알 수 있다.

성취수준별로 문항유형 구성에 따른 정분류 비율만을 그래프로 나타낸 결과는 [그림 3]과 같다. [그림 3]에서 보면 선다형 문항으로만 검사지를 구성할 경우 혼합형 보다 다소 정분류 비율이 낮아지지만 유의한 차이를 나타내지 않는다. 따라서 문항 유형을 어떻게 구성하는지 여부는 정분류 비율에 크게 영향을 주지 않는다고 할 수 있다.



〔그림 3〕 문항 구성에 따른 성취수준별 정분류 비율 그래프

Ⅴ. 결론 및 제언

본 연구에서는 맞춤형 학력진단검사 시스템을 중학교와 고등학교 대상으로 확대 적용할 계획을 세우면서, 그동안 기출문항에 대하여 문항별 성취수준 설정 과정을 생략하고 기존의 문항정보를 활용하여 분할점수를 자동으로 설정하기 위한 방안을 탐색하고자 하였다. 분할점수를 자동으로 설정하기 위해 문항반응이론을 기본으로, 시스템에 탑재된 모든 문항이 단일한 척도상의 값을 전제로 하여 기 산출된 능력점수(Θ)를 활용하여 모의실험을 실시하였다. 즉 불변적으로 유지되는 문항반응이론 척도상의 분할점수를 교사가 구성하는 검사형에 따라 원점수 척도로 변환하여 적용하는 모의실험을 실시하였으며, 문항 수, 문항 난이도, 문항 유형에 따른 정분류 비율(학생의 성취수준과 동일하게 분류되는 비율)을 비교하였다.

조건별로 살펴보면 문항 수가 증가할수록 정분류 비율이 증가하였다. 문항 난이도별로는 집단에 따라 다른 결과가 나타났는데 우수와 보통 집단에서는 문항 난이도가 증가할수록 정분류 비율이 증가하였으며, 기초와 기초 미달 집단에서는 반대되는 결과가 나타났다. 문항 구성에 따른 차이는 크지 않아 검사를 선다형으로만 구성하여도 선다형과 서답형을 혼합하는 경우에 비해 결과가 크게 나빠지지 않았으며 우수 집단에서는 소폭이나마 선다형의 정분류 비율이 높았다.

연구의 분석 결과에 따라 다음과 같은 결론을 제시할 수 있다. 첫째, 일부 모의실험 조건에서 낮은 정분류 비율이 나타났으나, 대체로 성취수준의 정분류 비율은 적절한 것으로 받아들일 수 있다. 정분류 비율은 전체적으로 분류 정확도 관련 선행 연구보다 다소 낮게 산출되었으나, 본 연구에서는 현장 활용성을 위하여 문항반응이론 척도 점수를 원점수로 변환하여 성취수준을 분류하였음을 감안하여야 한다. 문항반응이론 척도 점수를 원점수로 변환하는 과정에서 나타나는 변환 오차, 원점수 상의 분할점수를 적용하기 위하여 값을 올림(rounding up)하는 과정에서 나타나는 오차(예를 들어 분할점수가 10.01로 산출되어도 실제로는 11점이라는 분할점수가 적용됨)를 감안하면 문항반응이론을 기반으로 분류 정확도를 살펴 본 선행연구에 비하면 정분류의 비율이 낮게 산출될 수밖에 없다. 이러한 단점에도 불구하고 학교 현장의 활용도 제고를 위해 분할점수를 원점수 척도로 제공하는 것으로 정책적 결정이 내려졌으며 본 연구는 그러한 의사 결정의 근거 자료가 될 것이다. 이상의 원인으로 오분류가 발생하는 경우에도 성취수준이 두 단계 이상 달라지는 경우는 비율이 매우 낮았다.

둘째, 모의실험을 위한 조건별 분석 결과를 보면, 문항 수와 문항 난이도의 영향이 더 크게 나타난 반면 문항유형 구성으로 인한 차이는 적었다. 즉 문항 수를 늘리고 난이도를 어떤 수준으로 구성하는지의 여부가 문항 유형보다 중요한 조건이라 하겠다. 셋째, 성취수준 집단별 결과를 보면 '기초' 성취수준 집단의 정분류 비율이 가장 낮고 다음으로 '기초미달', '보통', '우수' 순으로 나타났다. '기초' 성취수준 집단에서 특히 정분류 비율이 낮게 나타난 것은 다른 성취수준에 비하여 '기초' 성취수준에 적합한 문항이 적기 때문으로 분석되며, '우수' 성취수준은 어떠한 조건에도 결과가 좋게 나타나 학업성취도 평가 문항이 전반적으로 쉬운 편이지만 '우수' 분할 점수도 높지 않으므로 '우수' 집단을 분류하기 위한 문항 수는 충분하다는 것을 시사한다. 넷째, 문항 수, 문항 난이도, 문항 유형 조건에 대한 분석 결과를 통해 '보통', '우수' 성취수준 집단을 위해서는 충분한 수의 고난도 문항을 지닌 혼합형 검사를 만드는 것이 바람직하다고 안내할 수 있다. 반면 '기초미달', '기초' 집단에는 충분한 수의 저난도 문항으로 혼합형 검사를 구성하는 것을 권장한다. 다섯째, '기초' 집단은 상대적으로 학생들의 학력 구분이 어려운 것으로 나타나 주의가 요구되며 특히 수학에 대

해서는 '기초' 성취수준 학생을 대상으로 할 때 고난도 문항이 포함되지 않도록 각별하게 유의하여야 함을 안내할 필요가 있다.

모의실험 결과에서 시사하는 바는 특정 조건에서 집단 분류의 일치 정도가 떨어지는 경우도 있지만 일반적인 신뢰도 수준(.7)으로 판단의 근거를 설정한다면 대부분의 결과에서 집단 분류의 일치 정도는 수긍할 만하였다. 조건별 분석 결과에 의하면 문항 수와 문항 난이도의 영향이 문항 유형 구성보다 의미 있게 나타났다. 성취수준 집단 별 결과를 참조하면 기초 집단의 정분류 비율이 가장 낮았고 기초 미달, 보통, 우수 순으로 나타났다. 특히 우수 성취수준의 경우 어떤 조건에서도 정분류 비율이 높게 나왔으며 이는 학업성취도 문항의 난이도 특성으로 해석할 수 있다. 이러한 결과를 토대로 단위학교에서 진단검사를 구성할 때 학생 수준이나 평가 목적에 따라 어떻게 구성하는 것이 최적의 방법인지 제시하고자 한다. 모의분석 결과를 기준으로 정분류 비율 .7 이상을 보장하려면 성취수준별 집단에 따라 요구하는 문항 구성의 조건이 차이가 있다.

이와 같은 모의실험 분석을 통해 성취수준 구분에 대한 분할 점수가 부정확해질 수 있는 일부 조건에 대해서는 검사지 구성 시 경고 메시지나 부가적인 설명을 안내하는 방안 등도 고려할 만하다. 분할점수를 제공하는 검사인 경우 최소 20문항 이상으로 구성하는 것을 권고하고 있으며 검사지 마법사에서 자동 구성하는 경우도 최소 20문항으로 설정되어있다. 물론 능력 추정이나 수준 설정 등에 대한 측정의 정확성은 검사 문항의 수가 증가함에 따라 개선되는 것이 일반적이다. 문항의 난이도와 문항유형 구성을 고려하여 다양한 경우의 분석 결과를 제시함으로써 각각의 문항이 전체 검사 정보함수에 어떠한 영향을 긍정적으로 미치고 있는지 살펴본 것은 의미 있는 과정이라 하겠다.

또한 최근 과정 중심 평가가 대두되면서 평가 결과를 이용하여 학생들에게 제공되는 형성적 피드백의 중요성이 강조되고 있다(Black & Wiliam, 2998. McMillan, 2014). 궁극적으로 피드백은 교사 주도로 실시되는 교실평가를 통해 제공되어야 하지만 학업성취도 평가 문항을 탑재한 수요자 맞춤형 정보 활용 서비스 시스템의 활용도를 제고하기 위해서는 시스템을 통해 구성되는 검사지가 준거참조평가로 기능할 수 있어야 한다. 이러한 점에서 본 연구에서 확인한 준거참조평가 도구로서의 온라인 학업성취도 평가 시스템은 교수-학습을 위한 보조도구로도 활용될 수 있을 것으로 기대된다.

참고문헌

김미경, 김성숙, 김창환, 김도남, 최인숙(2017). **국가단위 평가의 수요자 중심 정보 활용 서비**

스 시스템 구축 및 운영(V): 맞춤형 학력진단검사 확대와 시스템 고도화(RRE 2017-1). 한국교육과정평가원.

상경아, 최인봉, 김완수, 이은경(2013). **국가단위 평가의 수요자 중심 정보 활용 서비스 시스템 구축 및 운영: 시스템 구축 방안 수립 및 시범 사이트 개발・운영**(RRE 2013-16). 한국교육과정평가원.

시기자, 김완수, 박인용, 박찬호, 구슬기(2015). **국가수준 학업성취도 평가 기술보고서**(ORM 2015-106). 한국교육과정평가원. (비공개).

최인봉, 김경희, 김미경, 김창환, 한선영(2016). 국**가단위 평가의 수요자 중심 정보활용 서비스 시스템 구축 및 운영: iNAEA 시스템 고도화를 통한 수요자 중심 정보 활용 서비스 확대**(RRE 2016-7). 한국교육과정평가원.

최인봉, 박도영, 이은경(2014). **국가단위 평가의 수요자 중심 정보 활용 서비스 시스템 구축 및 운영: 시스템 개선 및 고도화 방안 수립**(RRE 2014-10). 한국교육과정평가원.

최인봉, 이채희, 이은경, 박병기(2015). **국가단위 평가의 수요자 중심 정보 활용 서비스 시스템 구축 및 운영: 학업성취도 평가 결과 활용 연구 동향 및 시스템 고도화**(RRE 2015-3). 한국교육과정평가원.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning, ***Assessment in Education: Principles, Policy and Practice, 5***(1), 7-73.

Angoff, W. H. (1971). Scales, Norms and Equivalent scores. In R. L. Thorndike (Ed.), ***Educational measurement*** (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Berk, R. A. (1996). Standard setting: The next generation (Where few psychometricians have gone before!). ***Applied Measurement in Education, 9***, 215-235.

Baker, F. B., & Kim, S.-H. (2004). ***Item response theory: Parameter estimation techniques*** (2nd ed.). New York: Marcel Dekker.

Cizek, G. J. & Bunch, M. B. (2007). ***Standard setting: A guide to establishing and evaluating performance standards on tests***. Thousand Oaks, CA: Sage.

de Ayala, R. J. (2009). ***The theory and practice of item response theory***. New York: Guilford.

Embretson, S. E., & Reise, S. P. (2000). ***Item response theory for psychologists***. New York: Lawrence Erlbaum Associates

Kolen, M. J. & Brennan, R. L. (2004). T***est equating, scaling, and linking: Methods and practices*** (2nd ed.). New York: Springer.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), ***IRT-based standard-setting procedures utilizing behavioral anchoring***.

Symposium conducted at the Council of Chief State School Officers National Conference on Large-scale Assessment, Phoenix, AZ.

McMillan, J. H. (2014). ***Classroom Assessment: Principles and Practice for Effective Instruction*** (6th ed.). Boston: Pearson.

Mitzel, H. C, Lewis, D. M., Patz, R.J. & Green, D. R. (2001). The Bookmark Procedure: Psychological Perspectives. In G. Cizek, (ed.), ***Setting performance standards*** (pp. 249-281). Lawrence Erlbaum Associates.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). ***Cutscores: A manual for setting standards of performance on educational and occupational tests***. Princeton, NJ: Educational Testing Service.

◎ 논문접수: 2018. 7. 31 / 수정본 접수: 2018. 9. 9 / 게재승인: 2018. 9. 15

저 자 소 개

· 김성숙 : University of Virginia에서 교육측정·평가 전공 박사학위를 취득하였으며, 현재 한국교육과정평가원 선임연구위원으로 재직. 연구 관심분야는 측정의 양호도 개선, 학업성취도 평가, 일반화가능도 이론과 적용 등임. sungs@kice.re.kr

· 박찬호 : University of Wisconsin-Madison에서 양적방법론·교육측정 박사학위를 취득하였으며, ACT, Inc.와 한국교육과정평가원을 거쳐 현재 계명대학교 교육학과 교수로 재직중임. 주요 관심분야는 문항반응이론 모형의 확장과 모수 추정, 진단평가 등임. cpark@kmu.ac.kr

· 김미경 : University of Texas, Austin에서 TEFL 박사학위를 취득하였고, 현재 한국교육과정평가원 연구위원으로 재직중임. 연구 관심분야는 영어 쓰기교육, 영어 성취능력 특성, 영어과 교육과정임. mikyung32@kice.re.kr

· 김창환 : 고려대학교에서 교육학 박사학위를 취득하였고 현재 한국교육과정평가원에서 부연구위원으로 재직중임. 최근 연구 관심분야는 교육훈련기관 인증평가 등 기관평가 분야임. kimch@kice.re.kr

〈ABSTRACT〉

An Exploratory Study of Automatic Standard Setting for iNAEA Through a Monte Carlo Simulation of Test Construction

Sungsook Kim (KICE) Chanho Park (Keimyung University) Mikyung Kim (KICE) Changhwan Kim (KICE)

The NAEA is a criterion-referenced assessment that assesses the degree of achieving the national curriculum, and the student's achievement in the assessment subject is categorized in accordance with the degree of achieving the curriculum into the following achievement levels: 'advanced', 'proficient', 'basic' and 'below-basic.' This study has focused on developing a customized diagnostic test system with the previously-administered NAEA test items and the related information. Based on item response theory, a simulation study was carried out by using all the items in the system, which were calibrated to be on one scale. The response data of 1,000 randomly selected students in the middle school for the subject of the Korean language in 2015 and 2016 were analyzed. The accuracy of the classification was determined by calculating the degree of agreement between the actual achievement level of the subject and the achievement level obtained by the simulation analysis according to the respective conditions of the number of the items of the test, the difficulty of the item. As a result, acceptable results were obtained overall, whereas lower classification ratios were found in some conditions. The effect of the number of items and item difficulty was larger, while the effect of item type composition was smaller. This study provides suggestions to consider when constructing a test form by comparing according to the simulation conditions.

Keywords : NAEA, achievement standard, cut-off score, item response theory scaling