교육과정평가연구
The Journal of Curriculum and Evaluation
2023, Vol. 26, No. 1, pp. 163~184
DOI: https://doi.org/10.29221/jce.2023.26.1.163

검사 길이 변화에 따른 국가수준 학업성취도 평가의 안정성 분석[1)]

김희경 (한국교육과정평가원 선임연구위원)*
김성훈 (한양대학교 교수)**

요약

본 연구는 대규모 평가에서 일어나는 문항 수 감축 상황이 점수척도의 안정성에 어떠한 영향을 끼치는지 점검하는 것을 목적으로 하였다. 따라서 "대규모 표준화 검사도구가 이전과 비교하여 축소된 문항 수로 구성됨에도 불구하고 척도의 점수대별 오차가 안정적인가?"하는 문제를 확인하기 위한 모의실험을 수행하였다. 이를 위해 검사 길이가 축소되기 이전의 교과별(국어, 수학, 영어) 실제 검사를 기반으로 모의 검사를 구성하였다. 또한 문항반응이론에 기반하여 동일한 능력점수(θ)를 가진 학생별로 1,000명의 모의 검사에 대한 응답 자료를 생산하였다. 국가수준 학업성취도 평가와 유사한 목적을 가진 대규모 학생평가에서는 특히 '기초학력 미달' 학생 비율에 대한 연도 간 변화에 사회적인 관심이 크게 집중되므로 척도의 안정성은 '기초학력'과 '기초학력 미달' 판별의 정밀함이 관건이다. 본 연구를 통해 산출된 모의실험 결과를 요약하면, 기초학력과 기초학력 미달의 경계선 부근 학생을 판별하기에 적당할 정도로 난이도가 낮으면서(문항반응이론 난이도 $b < -1.5$)이면서 변별력도 어느 정도 갖춘 (문항반응이론 변별도 $a \geq 0.5$) 문항을 교과별 검사에서 최소한 3개 이상 갖춘 경우, 기초학력 미달 분할점수 부근 학생의 성취도(능력점수) 추정의 평균오차(RMSE)가 0.05를 초과하지 않을 정도로 안정적임을 확인하였다.

주제어: 대규모 평가, 국가수준 학업성취도 평가, 검사 길이, 점수척도 안정성, 점수 체제

1) 이 논문은 '국가수준 학업성취도 평가 점수 체제 개선 및 결과 활용도 제고 방안'(김희경 외, 2019) 보고서의 일부 내용을 수정·보완한 것임.

* 제1저자, heekyoung@kice.re.kr
** 교신저자, seonghoonkim@hanyang.ac.kr

I. 서 론

최근 국제 학업성취도 평가인 PISA에서도 협력적 문제해결력, 글로벌 역량 등을 다루는 역량 기반의 평가를 도입하고 이를 강화하고 있는 추세이다(동효관 외, 2017, p.208). 우리나라에서도 2015 개정 교육과정을 통하여 역량을 중시하고자 하였고, 2017년 초등학교 1~2학년부터 본격적으로 적용되기 시작하여 점차 교과역량을 반영한 지필 형태의 평가도구에 대한 요구가 늘고 있으나 마땅한 방향 설정이 이루어지지 못한 상태였다(동효관 외, 2017, p.208).

이에 2015 개정 교육과정을 첫 적용하게 된 2019년 국가수준 학업성취도 평가(이하 학업성취도 평가)의 고등학교 검사의 평가틀은 교과별(국어, 수학, 영어) 역량에 대한 학생들의 성취도를 타당하게 측정하기 위한 신유형의 문항 도입을 고려하여 새롭게 설계되었다(동효관 외, 2018, pp.121-125). 2015 개정 교육과정이 적용된 2019년 학업성취도 평가의 특징은 복합형 문항의 도입이라고 할 수 있는데, 복합형 문항의 정의는 2개 이상의 성취기준을 측정하기 위한 복합적인 자료를 제시하고, 종합적인 분석 능력 등의 교과역량을 반영한 문항 형태라고 볼 수 있다(이재봉 외, 2020, p.115) 이러한 복합적인 자료를 종합적으로 해석한 후 문제를 인식하고 해결해야 하는, 단순 지식을 넘는 고차원적인 문항에 응답하기 위해 학생이 소요하는 시간은 이전에 비해 늘어날 수밖에 없다. 한편 문항의 난이도가 높아졌다고 해서 교과별 검사 시간을 늘리기는 어려운 시행 상에 문제점이 발생한다. 즉 학업성취도 평가 시행은 학교의 협조를 받아 학교 수업 시간(50분)을 고려하여 시간제한을 두어야 하는 상황에서 개별 학생에게 주어지는 문항 수를 감축할 수밖에 없는 현실에 직면하였다.

이와 같이 2019년 학업성취도 평가의 고등학교 검사에서는 교과역량 측정을 위한 문항 변화와 함께, 문항 수의 축소라는 조치가 수반될 수밖에 없었다. 개별 학생에게는 시간제한에 적절한 문항 수로 감축 구성된 평가를 시행하는 대신에, 학생 집단군에 따라 서로 다른 검사 유형을 배분받는 시행 체제를 설계하게 되었다. 이를 통해 전체 학생이 치르는 평가도구에 포함된 문항을 모두 종합하면 교육과정상의 교과별 성취기준 대표성을 충분히 보장할 수 있도록 행렬표집 시행 체제의 도입을 고려하였다.

동효관 외(2018, pp.121-168)에서는 2019년 고등학교 학업성취도 평가 예비시행 연구를 통해 2015 개정 교육과정에서 중요시하는 역량에 대한 학생의 성취를 타당하게 측정하기 위해 학생이 문제를 푸는 과정에서 드러내는 지식, 기능, 능력을 종합적으로 평가할 수 있는 문항을 개발하기 위한 평가틀을 설계하였다. 이러한 예비시행 연구에서 역량을 반영한 평가를 풀기 위해 소요되는 시간은 늘어날 것으로 예측하여 개별 학생이 응답해야 하는 문항 수를 감축할 수밖에 없었고, 〈표 1〉과 같이 개별 학생이 교과별 검사를 50분 내에 풀 수 있도록 적절하게 축소된 문항 수를 제시하였다. 〈표 1〉과 같이 개별 학생용 문제지에 포함되는 문항 수는 기존의 30문항 내외에서 20문항 내외로 축소하되, 4종의 검사 유형을 전체 학생 집단에게 배포하도록 설계하여, 교육과정상의 교과별 성취기준 대표성을 보장할 수 있도록 효율적인 행렬표집 시행 체제를 도입하고자 하였다.

〈표 1〉 2018년 대비 2019년 검사 길이

교과	2018년			교과	2019년		
	총 문항 수	선다형 문항 수	서답형 문항 수		총 문항 수	선다형 문항 수	서답형 문항 수
국어	32	26	6	국어	20	12	8
수학	29	25	4	수학	20	15	5
영어	36	30	6	영어	21	16	5

학업성취도 평가에서 행렬표집 시행 체제를 도입하면, 서로 다른 검사 유형을 학생 집단별로 배분하므로 개별 학생은 부담 없는 개수의 문항들로 구성된 한 가지 유형의 검사만을 실시하되, 전체 학생에게 실시되었던 각각의 검사 유형에 포함된 문항 수를 전체적으로 합산하면 개별 학생에게 노출된 문항 수보다 최소한 1.5~2배 정도로 문항 수를 확장할 수 있다. PISA(Programme for International Student Assessment), TIMSS(Trends in International Mathematics and Science Study), NAEP(National Assessment of Educational Progress)과 같은 국외 대규모 학업성취도 평가에서도 행렬표집 설계를 활용하고 있는데, 이러한 행렬표집 설계에서는 일반적으로 개별 학생에게 제공되는 평가결과의 정확성은 떨어지더라도 국가수준의 평가결과(연도별 학생의 성취수준별 비율 추이)는 안정적으로 산출할 수 있다(김경희 외, 2003, p.42).

이에 학업성취도 평가에서는 2015 개정 교육과정에서 중시하는 역량을 반영한 평가를 도입함에 따라 기존의 교과별 문항 수에 비해 감축(30문항 내외 → 20문항 내외)하게 되고, 행렬표집 시행 체제를 도입하는 상황에서 국가수준의 평가결과뿐 아니라 개별 학생에게 제공되는 평가결과도 모두 안정적으로 산출할 수 있도록 검사구성 방안을 마련할 필요가 있다. 학업성취도 평가에서는 학생의 성취도를 4수준(우수학력, 보통학력, 기초학력, 기초학력 미달)으로 세분하여 보고하며 사회의 관심은 기초학력 미달 학생 비율의 연도간 추이에 집중된다. 따라서 이 연구의 목적은 가장 주목받는 이슈인 “2019년 학업성취도 평가가 이전과 비교하여, 축소된 문항 수로 구성됨에도 불구하고 척도의 점수대별 오차를 안정적으로 유지하고, 기초학력과 기초학력 미달 학생을 판별하는 데 문제가 없는가?” 라는 연구문제에 대한 해답을 찾는 것이다. 이러한 질문에 답을 찾기 위해, 2015년부터 학업성취도 평가에서 활용하고 있는 ‘문항반응이론 진점수(item response theory true score)에 기반한 척도’를 2019년에도 지속적으로 적용할 것을 전제로 하였다. 또한 2019년 시행을 위해 계획된 검사구성안(총 문항 수, 선다형 및 서답형 문항 수, 배점 등)을 반영하여 모의 검사 구성 및 학생 응답 자료를 생산한 후, 학생의 능력점수(θ)를 추정하고 오차를 점검하였다.

본 연구는 2019년 국가수준 학업성취도 평가의 검사 길이가 2018년에 비해 교과별로 약 10개 문항 정도 감축되어야 하는 시점에서 수행된 연구이므로, 2018년 자료와 2019년 검사구성 계획을 활용하여 모의실험을 수행하였다. 또한 모의실험 방법에 있어서 3모수 문항반응이론에 기초하였으며, 척도점수 산출을 위한 능력모수 추정을 위해 문항반응이론 진점수에 기반한 척도화 방법을 사용하였는데, 이는 학업성취도 평가의 척도화를 위한 기술적 방법을 그대로 적용하기 위함이다. 연구 결과에서 도출되는 시사점은 자료의 시점이나 척도화를 위한 기술적 방법에 상관없이 기존 검사 길이를 감축해

야 하는 평가도구에 있어서 평가 결과를 이전과 마찬가지로 안정적으로 활용할 수 있는지를 판단하고자 할 때 수행할 수 있는 연구 방법을 제시할 수 있을 것이다.

II. 이론적 배경

원점수(raw score)란 학생이 부여받은 문항 점수의 총합으로 '변환을 거치기 이전의 그대로의 점수'를 의미한다. 원점수와 유사한 점수에는 100점 환산점수(percent-correct score)가 있는데, 이는 100점 총점을 가정하여 획득한 문항 점수 합계의 백분율을 산출하는 것이다. 교사별로 계획하여 단위학교 또는 학급 규모로 실시하는 소규모 평가가 아니라 표준화된 대규모 평가의 경우, 평가결과 제공 시 원점수 척도를 그대로 활용하기보다는 평가의 목적에 맞도록 결과 해석이 용이한 변환점수 척도를 개발하여 활용한다. 검사동등화 문헌에서, 변환점수 척도는 종종 점수척도 혹은 더 간단히 척도라고 불린다. 변환점수 척도의 점수는 척도점수라고 불린다(Kolen & Brennan, 2004, p.329).

학업성취도 평가에서는 2003년에 최초로 변환점수 척도를 개발하였으며, 그 이후 2차 재척도화는 2010년, 3차 재척도화는 2015년에 수행하여 학업성취도 평가결과를 보고하는 기본 척도로 활용하고 있다(시기자 외, 2015, pp.4-5). 교육평가에서 활용하는 '원점수-척도점수 변환' 방식은 선형 변환, 정규분포에 의한 변환, 아크사인(arcsine) 변환, 문항반응이론에 기반한 변환 등으로 분류할 수 있다(Kolen & Brennan, 2004, pp.336-351; 김경희 외, 2003, pp.92-108).

1. 학업성취도 평가 점수척도의 변천

학업성취도 평가는 2003년부터 4단계 성취수준 및 동등화를 도입하였으며, 연도별 원점수를 동일한 척도에 표현할 수 있도록 점수척도(이하 척도)를 개발하였다(김경희 외, 2003, pp.89-114). 〈표 2〉에는 학업성취도 평가 결과로 나타난 학생의 성취도를 4수준으로 구분하기 시작한 '2003년', 전수평가 도입에 맞춰 새로운 점수척도를 정비한 '2010년', 2009 개정 교육과정이 이루어진 '2015년'의 3개 시점으로 구분하여 점수척도의 변천을 정리하였다.

〈표 2〉 학업성취도 평가의 점수척도 변천 과정

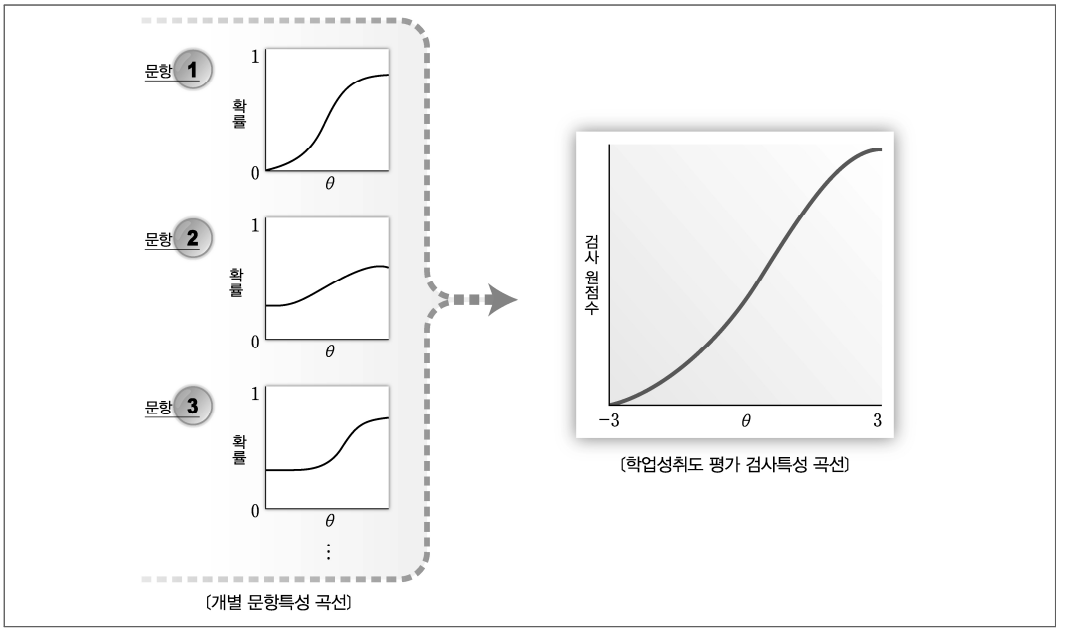
시점	대상	교과	특성	비고
2003	초6 (표집 1%) 중3 (표집 1%) 고1 (표집 1%)	국어, 사회, 수학, 과학, 영어	**〈원점수 → 척도점수 변환 방법〉** 아크사인(arcsine) 변환 **〈척도점수 범위〉** 초6 (점수 범위 130~190, 증분 1) 중3 (점수 범위 230~290, 증분 1) 고1 (점수 범위 330~390, 증분 1)	• 비교적 간단한 변환 방식을 사용하여 일반인들의 이해가 쉬움 • 점수대별 조건적 측정오차 크기를 동일하게 유지 • 원점수의 구분점에 비해 척도점수의 구분점이 적음 • 학교급별(초·중·고) 척도점수 범위가 차별

시점	대상	교과	특성	비고
				적으로 설정되어 있어 학년 간 비교 해석 가능하다는 오해 발생
2010	초6 (전수) 중3 (전수) 고1 (전수)	국어, 수학, 영어 (초6·중3 : 국어, 사회, 수학, 과학, 영어)	**〈원점수 → 척도점수 변환 방법〉** 아크사인(arcsine) 변환 **〈척도점수 범위〉** 초6 (점수 범위 100~300, 증분 1) 중3 (점수 범위 100~300, 증분 1) 고2 (점수 범위 100~300, 증분 1)	• 비교적 간단한 변환 방식을 사용하여 일반인들의 이해가 쉬움 • 점수대별 조건적 측정오차 크기를 동일하게 유지 • 척도점수의 구분점 개수를 61점 → 201점으로 확장함으로써 각 성취수준을 판별하는 분할점수의 간격을 넓힘 • 학교급(초·중·고) 구분 없이 척도점수 범위를 동일하도록 개선하여 학년 간 비교 가능한 척도로 오해하기 쉬운 문제 해소
2015	중3 (전수) 고2 (전수)	국어, 수학, 영어 (중3: 국어, 사회, 수학, 과학, 영어)	**〈원점수 → 척도점수 변환 방법〉** 문항반응이론 진점수 변환 **〈척도점수 범위〉** 중3 (점수 범위 50~350, 증분 1) 고2 (점수 범위 50~350, 증분 1)	• PISA, TIMSS, NAEP 등 해외 대규모 평가에서도 자주 활용하고 있는 변환 방식으로, 아크사인 변환 같은 전통적인 방식보다 수리적으로 정교한 척도화 방법임 • 검사 길이(문항 수)가 감축되는 상황에서 학생의 성취도를 측정하는 경우에도 영향을 덜 받음. • 척도점수의 구분점 개수를 201점 → 301점으로 추가적으로 확장함으로써 각 성취수준을 판별하는 분할점수의 간격을 더욱 넓힘

학업성취도 평가에서는 2003년에 원점수-아크사인 변환 방법으로 비선형 변환한 척도를 개발하여 2009년까지 사용되었으며, 2010년에 이를 일부 개선한 척도를 적용하였다(박정 외, 2006, pp.82-88; 김경희 외, 2011, pp.49-73). 따라서 2003년부터 2014년에 사용된 학업성취도 평가의 척도점수는 원점수를 아크사인(arcsine) 함수를 통해 전환하는 전통적인 척도화 방법을 적용한 것이다. 2015년에는 2009개정 교육과정 도입을 위한 평가틀 변화에 따라 학업성취도 평가의 재척도화가 필요하였고, 고전검사이론에 기반한 척도 대신 문항반응이론 진점수에 기반한 척도를 개발하였다(박인용 외, 2017, pp.71-75).

문항반응이론은 학생의 능력(θ)과 문항에 대한 학생의 반응 간 관계를 확률 모형으로 나타내며, 따라서 학생의 능력 수준에 따라 문항의 정답을 맞힐 확률을 추정하여 개별 문항마다 고유한 문항특성곡선(item characteristic curve: ICC)을 생성한다(Kolen & Brennan, 2004, pp.157-158). 현행 학업성취도 평가는 이러한 문항반응이론 진점수에 기반한 척도화 방법을 사용하고 있다(박인용 외, 2017, pp.71-75). 진점수 척도화는 각 문항에 대한 문항특성곡선 추정 및 검사를 구성하는 모든 문항의 문항특선곡선을 통합한 검사특성곡선(test characteristic curve: TCC)을 추정함으로써 특정 능력 수준 θ를 가진 학생이 그 시험에서 획득할 것으로 기대되는 점수(진점수)를 추정하고, 검사에서 이 기대값(진점수)과 같은 점수(관찰점수)를 획득한 학생의 능력 수준을 θ라고 간주한다(Kolen & Brennan, 2004, pp.176-181). 즉, θ 수준의 능력을 가진 학생이 얻을 것으로 기대되는 진점수가 있다면, 이와 동일한 관찰점수를 획득할 것이라는 것을 전제하고 있다. 이러한 진점수 척도화는 문항의

난이도와는 관계없이 동일한 수의 문항을 맞힌 학생은 동일한 능력 θ을 가졌다고 가정하기 때문에, 이들에게 동일한 능력점수 및 척도점수를 부여하는 방식이다([그림 1] 참조).



[그림 1] 문항반응이론에서의 문항특성곡선(ICC)과 검사특성곡선(TCC)

2. 문항반응이론 진점수에 기반한 척도화

문항반응이론에 기반한 척도점수는 두 가지 방법으로 구분할 수 있는데, 패턴 채점(pattern-scoring) 방식으로 학생의 능력점수(θ)를 추정하는 방법과 학생의 관찰점수(원점수)를 진점수로 보고 검사특성곡선(TCC)에서 학생의 능력점수를 찾는 방법으로 구분할 수 있다(김경희 외, 2003, pp.95-99). 일반적으로 원점수와 척도점수 간의 관계는 단조 증가 경향을 보여야하는데, 패턴 채점에 의한 척도는 원점수와 척도점수의 순위가 뒤바뀌기도 한다는 단점이 있다(김경희 외, 2003, pp.98-99). 따라서 본 연구에서는 학업성취도 평가에서 활용하고 있는 문항반응이론에 기반한 척도화 방법 중 진점수를 활용하는 척도화 방법에만 초점을 맞추기로 한다.

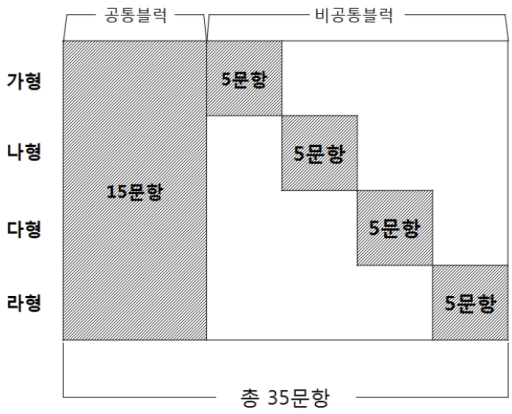
문항반응이론에서는 학생의 능력 수준(θ)이 변함에 따라 각 문항에 대한 정답을 맞힐 확률을 추정하여 문항특성곡선을 생성할 수 있다. 또한 검사에 포함되어 있는 개별 문항의 문항특성함수를 모두 합하면 검사특성곡선을 얻을 수 있는데, 검사특성곡선을 이용하여 θ의 능력을 가진 학생이 해당 검사에서 얻을 것으로 기대되는 진점수를 추정할 수 있다(Kolen & Brennan, 2004, p.176). 즉 이러한 문항반응이론 진점수 기반의 척도는 학생이 검사를 통해 실제로 얻은 원점수(관찰점수)를 검사특성곡선

(TCC)에 의해 부합하는 원점수(진점수)와 연계하여 능력점수 θ를 부여하는 방식이다. 이러한 능력 모수 추정 방법은 TCC 방법이라 불리며, 그 특성은 Kolen과 Tong(2010), Kim(2012) 등에 의해 연구되었다.

한편 문항반응이론에 기반한 또 다른 척도화 방식인 패턴 채점 방법에 의하면 동일한 개수의 문항을 맞힌 피험자여도 어려운 문항을 맞힐수록 피험자의 능력 모수는 높은 값으로 추정되므로, 그 능력 모수 추정치에 대해 높은 척도 점수가 부여되는 방식이라 논리적일 수는 있으나 원점수와 척도점수에서의 학생 순위가 뒤바뀌는 현상이 발생할 수 있어 일반인들에게 이해시키기가 힘들다는 단점이 있다(김경희 외, 2003, pp.98-99). 이러한 문제점을 고려하여 패턴 채점 방법보다는 검사특성곡선을 활용하여 각 진점수에 연계되는 학생의 능력점수(θ)를 찾아내는 방법을 사용할 수 있다. 물론 능력점수(θ)는 대체적으로 -3 ~ +3 범위의 값을 갖는 연속변수이므로 이를 그대로 사용하기보다는 최종적인 척도를 위해 특정한 평균과 표준편차를 갖도록 선형변환 절차를 거치는 경우가 많다(시기자 외, 2015, p.66).

3. 행렬표집 시행 체제

행렬표집 시행 체제에서는 학생을 집단별로 구분하여 각각 다른 검사 유형을 배분하므로 학생 개개인은 부담 없는 검사 길이로 구성된 1개 유형의 검사를 실시하되, 전체적으로 시행된 문항 수를 합산하면 학생 개인에게 배분되었던 문항 수보다 최소한 1.5~2배 정도까지의 문항 에 대한 응답 자료를 확보할 수 있다는 장점이 있다(김회경 외, 2019, p.46). 검사 길이가 대폭 축소되는 2019년 고등학교 학업성취도 평가에서도 학생의 입장에서는 제한된 시간 내에 풀기에 알맞은 문항 수로 평가를 시행하되, 평가를 설계하고 결과를 활용하고자 하는 입장에서는 교육과정 상의 교과별 성취기준 대표성을 확보할 수 있도록 문항 수를 늘릴 수 있는 행렬표집 설계를 도입할 필요성이 있다. 국외의 대규모 학업성취도 평가 사례를 살펴보면, 국제 학업성취도 비교 연구인 PISA, TIMSS, 그리고 미국의 PARCC, NAEP에서 효율적인 시행을 위해 행렬표집 체제를 적용하고 있고, 이러한 행렬표집 설계를 도입한 평가 체제에서는 일반적으로 전체 학생이 동일한 문항을 시행하지 않은 경우에 점수를 추정하기에 적합한 문항반응이론에 기반한 척도점수를 활용하고 있다(OECD, 2017, pp.141-144; Yamamoto & Kulick, 2000, pp.237-238). [그림 2]에는 행렬표집 설계의 예시를 제시하였다. 이 예시에서는 학생을 4개의 집단으로 구분하고, 각 집단은 가~다형 가운데 1개의 검사유형을 배분받게 된다. 즉 학생 개인은 공통문항(15문항), 비공통문항(5문항)을 합산하여 20문항을 풀지만, 전체적으로는 공통문항(15문항), 4가지의 비공통문항(5 × 4 = 20문항)을 합산하면 총 35문항에 대한 자료 수집이 가능하다.



[그림 2] 행렬표집 설계의 예시

III. 연구 방법

1. 모의 검사의 문항 구성

본 연구를 위해, 문항 수 축소 전의 검사 구성은 2018년 학업성취도 평가의 고등학교 3개 교과별(국어, 수학, 영어) 실제 검사 구성 자료를 기초로 하였다. 2018년 고등학교 교과별 검사 구성 자료에 나타난 문항 수, 문항 배점, 총점(원점수)은 〈표 3〉과 같다.

〈표 3〉 문항 수 축소 이전의 검사 구성(2018년 실제 검사 구성)

시간 제한	교과	선다형			서답형			총 문항 개수	총점 (원점수)	서답형 문항의 총점 대비 비율(%)
		문항 개수	문항 배점	점수 합계	문항 개수	문항 배점	점수 합계			
교과별 60분	국어	26	1	26	6	2점: 4문항 3점: 2문항	14	32	40	35.0
	수학	25	1	25	4	2점: 1문항 3점: 1문항 4점: 2문항	13	29	38	34.2
	영어	30	1	30	6	2점: 3문항 3점: 1문항 4점: 2문항	17	36	47	36.2

또한 역량평가 도입을 위해 문항 수가 축소되는 검사 구성을 위해서는 동효관 외(2018, p.124)에서 2019년 본시행 준비를 위한 예비시행을 실시함으로써, 개별 학생이 시험 시간 50분 내에 풀기에 적절

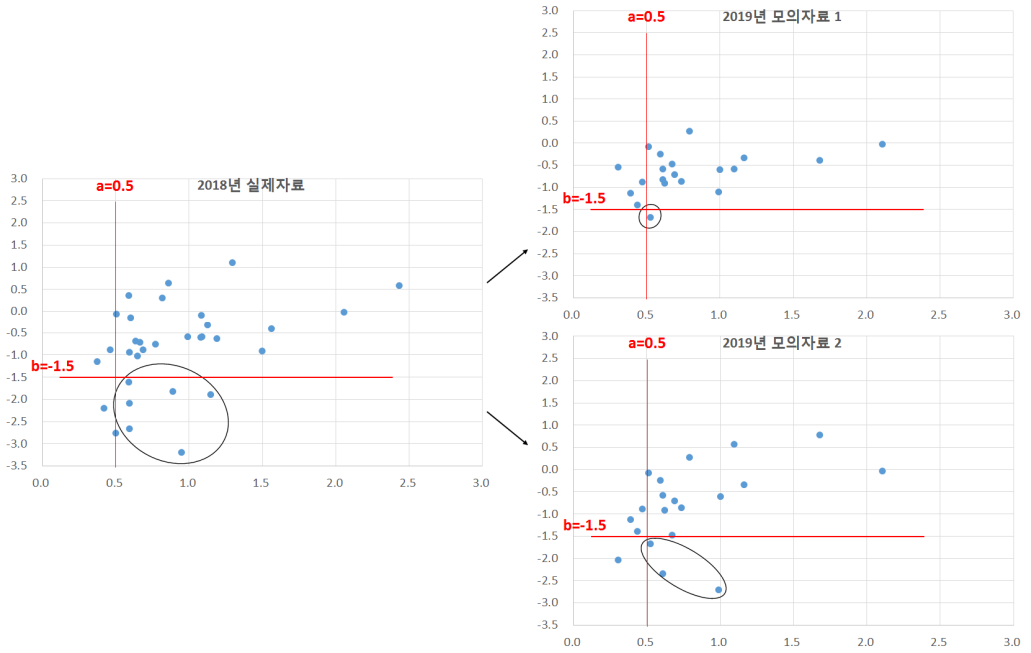
한 총 문항 수 및 서답형 문항이 차지하는 비율을 제시한 바가 있어, 이를 활용하였다. 따라서 2019년에 문항 수가 축소되는 상황을 반영하기 위한 모의 검사는 〈표 4〉와 같은 검사 구성이 되도록 설계하였다. 2015 개정 교육과정에서 중시하는 교과역량을 반영한 평가를 도입하는 상황임에도 불구하고, 학교 수업 시간에 맞춰 시험 제한시간은 오히려 기존 60분에서 50분으로 줄었고, 교과별 총 문항 수는 약 30문항에서 약 20문항으로 축소된 반면 교과역량 측정을 위해 서답형 문항의 비중이 늘어나 서답형 점수 합계가 총점에서 차지하는 비율은 34~36% 정도에서 42~64% 정도로 증가하였다.

〈표 4〉 문항 수 축소 이후의 검사 구성(2019년 예정 검사 구성)

시간 제한	교과	선다형			서답형			총 문항 개수	총점 (원점수)	서답형 문항의 총점 대비 비율(%)
		문항 개수	문항 배점	점수 합계	문항 개수	문항 배점	점수 합계			
교과별 60분	국어	12	1	12	8	2점: 5문항 3점: 2문항 5점: 1문항	21	20	33	63.6
	수학	15	1	15	5	2점: 4문항 3점: 1문항	11	20	26	42.3
	영어	16	1	16	5	2점: 3문항 3점: 1문항 5점: 1문항	14	21	30	46.7

〈표 3〉과 〈표 4〉과 같이, 고등학교 국어교과 검사를 살펴보면, 2018년 학업성취도 평가는 32개 문항으로 구성되었으나 2019년에는 총 20개 문항으로 감축하여 구성될 것으로 예정되었다. 따라서 2018년 학업성취도 평가의 고등학교 국어교과 평가 결과 자료로부터 실제로 산출된 문항모수를 활용하여 2019년 검사 구성안에 따라 두 가지 버전의 문항 특성을 갖춘 모의 검사를 구성하였다. 또한 본 연구에서는 기초학력과 기초학력 미달 학생을 구분하기에 적절한 문항은 난이도(b모수)는 충분히 낮은 반면에 변별력(a모수)은 양호한 조건을 갖추어야 한다고 보고, b모수 〈-1.5, 동시에 변별도 $a \geq 0.5$ 라는 두 가지 조건을 기준으로 삼았다. 이러한 두 가지 조건을 제시한 근거는 〈표 8〉에 제시된 바와 같이 국가수준 학업성취도 평가에서 설정된 기초학력/기초학력 미달을 구분하는 능력 모수 지점과 연관되어 있다. 교과별로 어느 정도 차이는 있으나 능력모수가 -2보다 낮은 경우에 기초학력 미달로 분류되므로 이를 고려하면 난이도 b모수가 -1.5 미만이면 충분히 능력모수 -2 지점을 포함할 것으로 판단하였다.

[그림 3]은 국어교과 2018년 실제 검사와 2019년 두 가지 모의 검사의 문항 특성을 나타낸다. '2018년 실제 검사'의 문항 특성을 설명하기 위해 2018년 국어교과 검사를 구성하였던 32개 문항의 실제 자료에서 산출된 문항반응이론 기반 문항모수(난이도, 변별도)를 산포도로 나타내었다.



[그림 3] 국어교과 2018년 실제 검사 및 2019년 모의 검사의 문항 특성 산포도

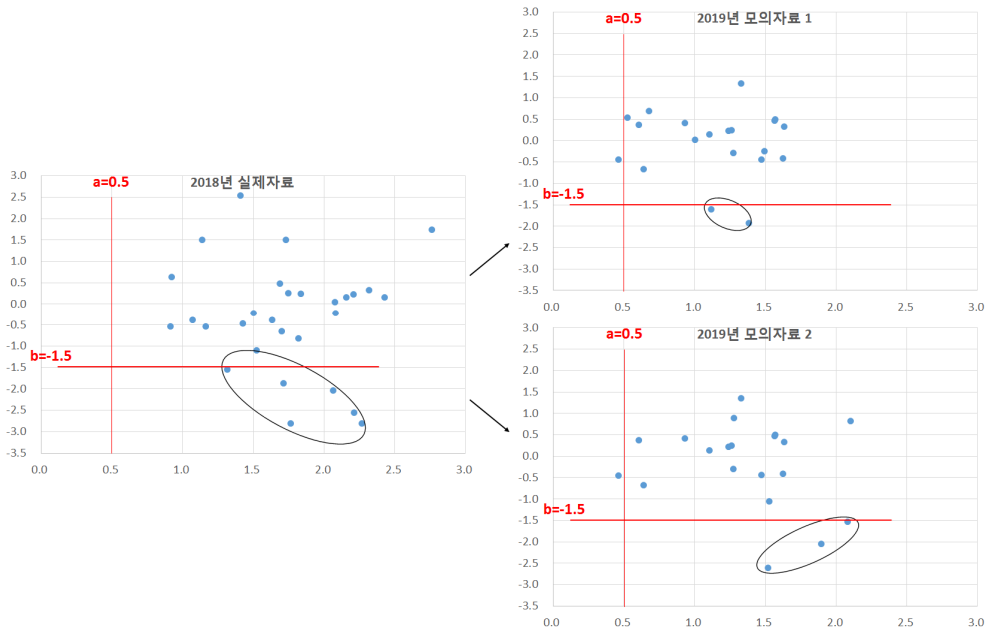
[그림 3]의 산포도를 살펴보면, 2018년 국어교과 실제 검사에는 기초학력과 기초학력 미달 학생을 구분하는 데 필요한 문항 난이도는 낮으면서(난이도 b 〈-1.5) 변별력은 양호한(변별도 a ≥ 0.5) 수준의 문항이 6개 포함되었음을 알 수 있다. 문항 수가 축소되는 상황을 반영하는 2019년 모의 검사 2가지 버전 중에서 '모의 검사 1'은 2018년 실제 검사에 포함되었던 문항에서 문항의 특성(난이도, 변별도)을 고려하지 않고 무작위로 20개의 문항을 추출하여 구성하였다. 결과적으로 '모의 검사 1'의 문항모수 산포도를 살펴보면 기초학력 미달 학생을 판별하기 위한 문항 난이도가 낮으면서(난이도 b 〈-1.5) 변별력은 충분히 양호한(변별도 a ≥ 0.5) 문항이 1개만 포함되었다. 문항 수가 축소되는 2019년 모의 검사 가운데 '모의 검사 2'는 기초학력 미달 학생 판별을 위해 문항 난이도는 낮으면서(난이도 b 〈-1.5) 변별력은 충분히 양호한(변별도 a ≥ 0.5) 문항을 의도적으로 최소한 3개 포함하도록 구성하였다. 이러한 2018년 실제 검사와 비교한 2019년 두 가지 버전의 모의 검사 특성을 〈표 5〉에 요약하였다.

〈표 5〉 국어교과 2018년 실제 검사 및 2019년 모의 검사 특성 요약

	2018년 실제 검사	2019년 모의 검사 1	2019년 모의 검사 2
총 문항 수(비율)	32 (100%)	20 (100%)	20 (100%)
양호한 문항 조건을 충족하는 문항 수* (비율)	5 (15.6%)	0 (0.0%)	3 (15.0%)

*문항 난이도 b 〈 -1.5, 문항 변별도 a ≥ 0.5 조건을 동시에 충족하는 문항 수

다음으로 수학교과의 경우, 〈표 3〉과 〈표 4〉에 제시된 바와 같이 2018년 수학교과 실제 검사의 총 문항 수 29개는 2019년에는 총 20개 문항으로 감축되었다. 국어교과와 마찬가지로 2018년 수학교과 평가 결과 자료에서 실제로 산출된 문항모수를 고려하여 2019년의 두 가지 모의 검사를 문항 수가 축소되도록 조정해보았다. [그림 4]에 수학교과에서의 2018년 실제 검사와 2019년 문항 수가 축소 조정된 두 가지 모의 검사의 특성을 나타냈다. '2018년 실제 검사'의 문항 특성 산포도에는 2018년 수학교과 실제 검사에 포함되었던 29개 문항에 대한 학생 응답 자료에서 산출된 문항반응이론 기반 문항모수(난이도, 변별도)가 제시되어 있다. 산포도에 따르면 2018년 수학교과 실제 검사에는 기초학력 미달 학생을 판별하기에 효과적인 난이도가 낮으면서(난이도 $b < -1.5$) 변별력은 어느 정도 양호한(변별도 $a \geq 0.5$) 문항이 6개 포함되어 있었던 것을 알 수 있다. 문항 수가 축소된 2019년 '모의 검사 1'은 2018년 실제 검사에서 문항 특성과 관계없이 무작위로 20개의 문항을 추출하였고, 결과적으로 산포도에 따르면 기초학력 미달 학생을 판별하기 적절하도록 난이도가 낮은 수준이면서(난이도 $b < -1.5$) 변별력은 충분히 양호한(변별도 $a \geq 0.5$) 문항이 2개 포함되어 있음을 알 수 있다. 또한 국어교과와 마찬가지로 수학교과에서도 2019년 '모의 검사 2'는 기초학력 미달 학생을 판별을 위해 효과적인 문항(난이도 $b < -1.5$, 변별도 $a \geq 0.5$) 문항을 의도적으로 최소한 3개 포함하도록 설계하였다. 이러한 2018년 실제 검사와 2019년 2가지 모의 검사의 특성을 〈표 6〉에 요약하였다.



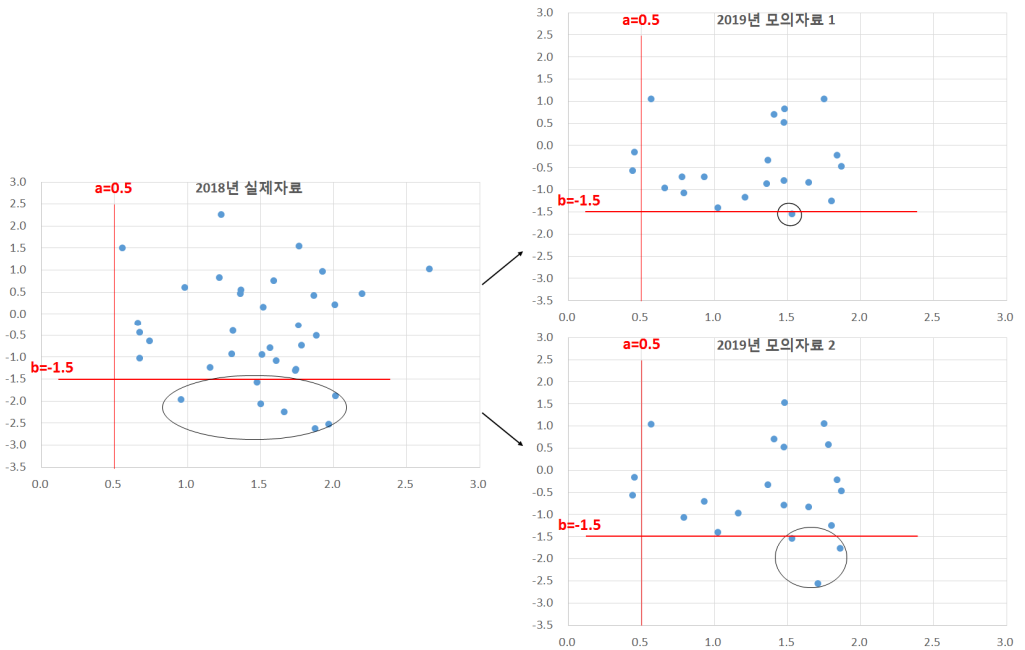
[그림 4] 수학교과 2018년 실제 검사 및 2019년 모의 검사의 문항 특성 산포도

〈표 6〉 수학교과 2018년 실제 검사 및 2019년 모의 검사 특성 요약

	2018년 실제 검사	2019년 모의 검사 1	2019년 모의 검사 2
총 문항 수(비율)	29 (100%)	20 (100%)	20 (100%)
양호한 문항 조건을 충족하는 문항 수* (비율)	6 (20.6%)	2 (10.0%)	3 (15.0%)

*문항 난이도 b 〈 −1.5, 문항 변별도 a ≥ 0.5 조건을 동시에 충족하는 문항 수

마지막으로 고등학교 영어교과의 경우, 〈표 3〉과 〈표 4〉에 제시된 바와 같이 2018년 검사는 총 문항 수 36개였으나 2019년에는 21개 문항으로 감축하도록 예정되었다. 2018년 학업성취도 평가 고등학교 영어교과 평가 결과 자료에서 실제로 산출된 문항모수를 활용하여 2019년 검사 구성안을 적용하여 두 가지 버전의 모의 검사를 설계하였다. [그림 5]에 영어교과의 2018년 실제 검사 및 2019년 모의 검사를 구성하는 문항 특성을 산포도로 제시하였다.



[그림 5] 영어교과 2018년 실제 검사 및 2019년 모의 검사의 문항 특성 산포도

문항 수가 축소되기 이전인 '2018년 실제 자료'에는 2018년 영어교과 실제 검사를 구성하였던 36개 문항에 대한 실제 문항반응이론 문항모수 특성을 산포도로 제시하였다. 2018년 영어교과 실제 검사에는 기초학력 미달 학생을 판별하기에 효과적인 위한 난이도가 낮으면서(난이도 b 〈-1.5) 변별력이 양호한(변별도 a ≥ 0.5) 문항이 7개 포함되어 있었다. 2019년에 축소되는 검사 구성을 반영하는 '모의 검사 1'은 2018년 실제 검사에서 문항 특성과 관계없이 무작위로 20개의 문항을 추출한 것이고,

결과적으로 '모의 검사 1'의 문항 특성 산포도를 살펴보면 기초학력 미달 학생 판별을 위한 난이도가 '하' 수준이면서(난이도 b 〈-1.5) 적절한 변별력을 갖춘(변별도 a ≥ 0.5) 문항이 1개만 포함된 것을 알 수 있다. 다음으로 2019년 '모의 검사 2'는 기초학력 미달 학생을 효과적으로 판별하기 위한 문항(난이도 b 〈-1.5, 변별도 a ≥ 0.5) 을 최소한 3개 갖추도록 구성하였다. 이러한 실제 검사와 2가지 버전의 모의 검사의 특성을 〈표 7〉에 요약하였다.

〈표 7〉 영어교과 2018년 실제 검사 및 2019년 모의 검사 특성 요약

	2018년 실제 검사	2019년 모의 검사 1	2019년 모의 검사 2
총 문항 수(비율)	36 (100%)	21 (100%)	21 (100%)
양호한 문항 조건을 충족하는 문항 수* (비율)	7 (19.4%)	0 (0.0%)	3 (14.3%)

*문항 난이도 b 〈 −1.5, 문항 변별도 a ≥ 0.5 조건을 동시에 충족하는 문항 수

2. 모의 응답 자료 생성 및 분석

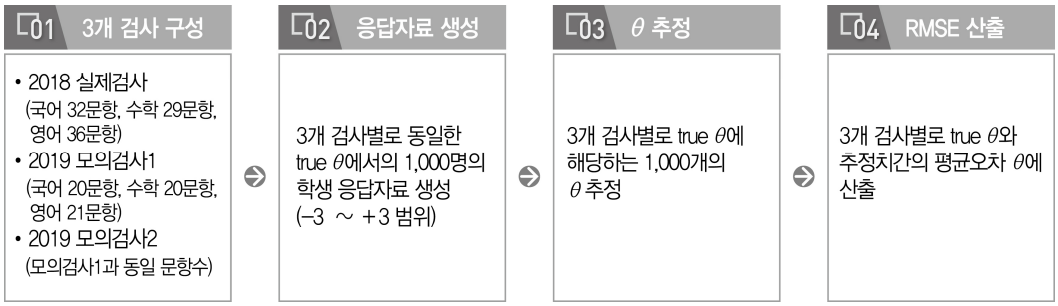
본 연구에서는 2019년에 예정된 학업성취도 평가 고등학교 검사지의 문항 수 축소가 점수척도의 안정성에 어떻게 영향을 끼치는지 점검하고자 하였다. 2019년 학업성취도 평가의 고등학교 검사에서는 2015 개정 교육과정 적용을 위한 새로운 평가틀이 도입됨에 따라 역량을 측정하기 위한 새로운 유형의 문항을 도입하고자 하였고, 따라서 2018년에 비교해 문항 수 감축이 불가피한 상황이었다. "2019년 학업성취도 평가의 고등학교 검사가 이전과 비교하여 축소된 문항 수로 구성됨에도 불구하고 척도의 점수대별 오차가 안정적인가?" 에 대한 답을 찾고자 2019년 문항 수 감축이 계획된 검사구성안을 반영하여 고등학교 교과별 모의 검사를 구성하였다 즉 2019년 계획된 검사구성안(문항 수, 배점 등)을 반영하여 문항 수가 축소된 두 가지 버전의 모의검사를 마련하였으며, '모의검사 1'은 문항의 특성(난이도, 변별도)을 고려하지 않고 2018년 실제 문항 중에서 무작위로로 문항을 추출하였고, '모의검사 2'는 기초학력 미달 학생을 안정적으로 판별하기 위해 문항의 난이도가 낮으면서(난이도 b 〈-1.5) 문항의 변별력도 어느 정도 양호한(변별도 a ≥ 0.5) 문항을 모든 교과에서 최소한 3개를 갖추도록 구성한 것이다.

교과별 모의 검사 구성이 완료된 후, 3가지의 검사 구성(2018년 실제검사, 2019년 모의검사 1, 2019년 모의검사 2)에 대한 학생 응답 자료를 생산하였는데, 문항반응이론을 적용하여 동일한 능력수준(θ)별로 1,000명의 응답 자료를 생산하였다. 구체적으로, 김성훈 외(2010)에서와 같이, 각 검사에 대해, θ=-3에서 θ=+3까지 0.5 간격의 13개 능력 지점 각각에서 1,000명의 문항반응 자료를 모의 생성하고, TCC 방법을 통해 각 피험자의 능력점수를 추정하였다. 능력모수 지점을 이렇게 선정한 것은 학생의 능력점수는 일반적으로 -3에서 +3의 구간에 분포하고, 0.5 간격의 지점을 사용하더라도 추정오차의 패턴은 무리 없이 파악할 수 있는 것으로 나타났기 때문이다(김성훈 외, 2010). 특정 능력수준 θ(true θ)에서 r번째 추정된 능력점수를 $\hat{\theta}_r$이라고 하였을 때, 추정의 오차는 다음과 같이 평균제곱오

차의 제곱근(root of mean squared error: RMSE)으로 계산하였다.

$$\mathrm{RMSE}(\hat{\theta}\,|\,\theta) = \sqrt{\sum_{r=1}^{1000}(\hat{\theta}_r - \theta)^2/1000} \quad \text{(식 1)}$$

모의실험 결과를 적절히 해석하기 위하여, 2018년 실제 검사에서 산출된 RMSE와 2019년 모의 검사에서 산출된 RMSE를 비교하였다. [그림 6]은 이러한 모의 검사 구성, 응답 자료 생성, 분석 방법 과정을 요약하여 제시한 것이다.



[그림 6] 모의 실험 절차

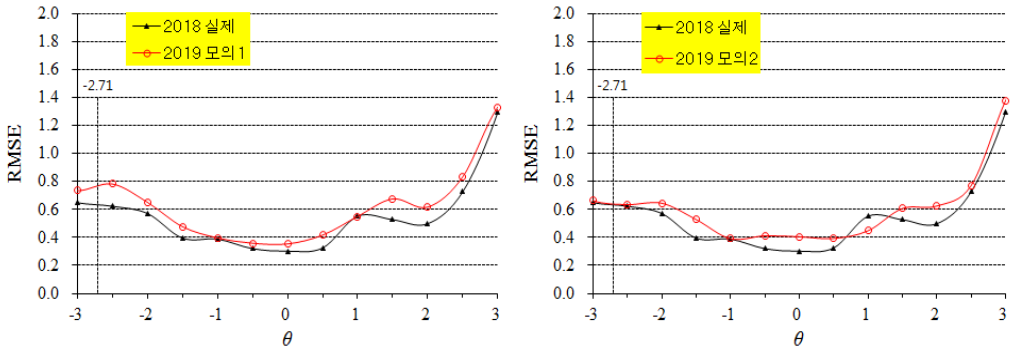
IV. 연구 결과

학업성취도 평가에서는 학생의 성취도를 4단계 수준(우수/보통/기초/기초학력 미달)으로 구분하여 각 성취수준에 해당하는 학생 비율에 대한 연도간 추이 결과를 보고하는데, 특히 기초학력 미달 학생의 비율에 가장 큰 사회적인 관심이 집중된다. 따라서 학업성취도 평가 점수척도의 안정성은 매년 '기초학력 미달' 학생 판별에 대한 안정성을 담보할 수 있는지 여부에 중요한 무게를 둔다고 볼 수 있다. 〈표 8〉은 마지막으로 학업성취도 평가의 점수척도를 개선했던 2015년에 설정된 능력점수(θ) 상에서 교과별로 기초학력 미달을 판별하는 분할점수이다.

〈표 8〉 2015년 설정된 학업성취도 평가 고등학교 교과별 기초학력/기초학력 미달 분할점수(θ점수)

	기초학력과 기초학력 미달 경계선의 분할 점수		
	국어	수학	영어
능력점수(θ)	-2.7097	-2.0521	-2.3814

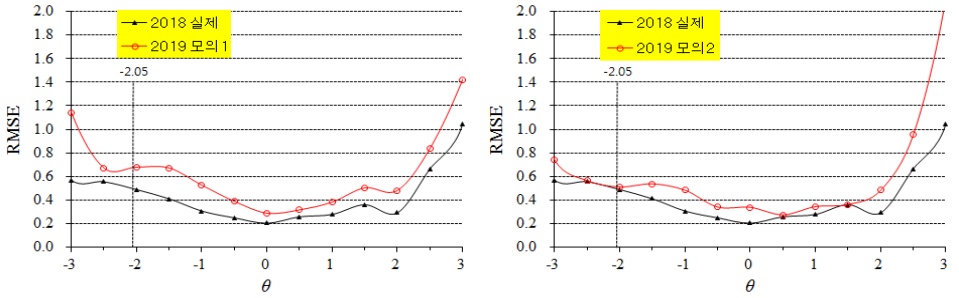
[그림 7]은 고등학교 국어교과의 세 가지 검사(실제 검사, 2가지 모의 검사)에 대한 θ 추정치의 RMSE를 나타낸 것이다(이하에서는 편의상 RMSE를 간단히 "오차"로 부른다). 왼쪽 그래프에는 '2018년 실제 검사'와 '2019년 모의 검사 1'의 RMSE 패턴을 비교하여 제시하였다. 또한 2015년 설정되었던 국어교과에서 기초학력 미달을 판별하는 θ점수 = -2.71을 기준선으로 표시하였다. 이는 특히 기초학력 미달을 판별하는 분할점수 부근에서의 오차 크기 변화를 점검하기 위함이다. '2019년 모의 검사 1'은 문항의 특성(난이도, 변별도)과 관계없이 무작위로 문항을 추출하여 검사를 구성한 경우를 나타낸다. '2019년 모의 검사 1'에서 기초학력 미달을 판별하는 분할점수(θ = -2.71)에서 나타나는 오차는 '2018년 실제 검사'의 오차와 비교하였을 때, 약 0.1 이상 크다.



[그림 7] 2018 실제 검사 대비 2019년 모의 검사에 대한 능력 추정치의 RMSE(국어)

또한 [그림 7]에서 '2019년 모의 검사 2'는 검사 구성 시 기초학력 미달 판별을 위해 난이도가 낮으면서(난이도 b <-1.5), 변별력은 어느 정도 양호한(변별도 a ≥ 0.5) 문항을 최소한 3개 포함한 경우를 예시한다. 결과적으로 [그림 7]의 오른쪽에 '2018년 실제 검사'와 '2019년 모의 검사 2'에 대한 θ 추정치의 RMSE를 나타냈다. '2019년 모의 검사 2'에서는 기초학력 미달 학생을 판별하는 분할점수(θ = -2.71) 부근에서 2018년 실제 검사에 대한 오차와 거의 유사한 크기를 보였다(차이는 약 0.01).

[그림 8]은 고등학교 수학교과의 '2018년 실제 검사'와 '2019년 모의 검사'에서 산출된 θ 추정치의 RMSE 패턴을 비교하여 제시한 것이다. [그림 8]의 왼쪽에는 '2018년 실제 검사'와 '2019년 모의 검사 1'의 RMSE를 비교하여 제시하였다. '모의 검사 1'은 문항 특성(난이도, 변별도)와 관계없이 무작위로 문항을 추출하여 검사를 구성한 경우를 나타낸다. '모의자료 1'에서 수학교과의 기초학력 미달을 구분하기 위한 분할점수(θ = -2.05) 근처의 오차는 '2018년 실제 검사'의 오차와 비교하였을 때, 약 0.1 이상 크게 나타났다.

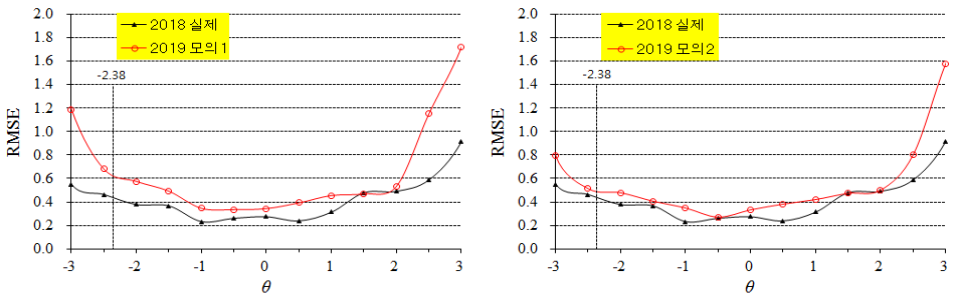


[그림 8] 2018 실제 검사 대비 2019년 모의 검사에 대한 능력 추정치의 RMSE(수학)

또한 [그림 8]의 오른쪽 그래프에서 '모의 검사 2'는 검사 구성 시 기초학력 미달 판별을 위해 난이도가 낮으면서(난이도 b 〈-1.5) 변별력은 충분히 갖춘(변별도 a ≥ 0.5) 문항을 최소한 3개 포함한 경우를 나타낸다. 그 결과로 '2018년 실제 검사'와 '2019년 모의 검사 2'의 RMSE를 비교하여 제시하였다. 국어교과에서와 마찬가지로 수학교과 모의실험에서도 '모의 검사 2'에서는 기초학력 미달을 결정하는 분할점수(θ = -2.05)에서 2018년 실제 검사의 오차와 거의 유사한 크기로 나타났다(차이는 약 0.02).

[그림 9]는 고등학교 영어교과의 '2018년 실제 검사'와 '2019년 모의 검사'에서 산출된 θ 추정치의 RMSE 패턴을 비교하여 제시한 것이다. [그림 9]의 왼쪽에는 '2018년 실제 검사' 대비 '2019년 모의 검사 1'의 RMSE를 제시하였다. '모의 검사 1'은 문항 특성(난이도, 변별도)에 관계없이 무작위로 검사를 구성한 경우를 의미한다. '모의 검사 1'에서 영어교과의 기초학력 미달을 판별하는 분할점수(θ = -2.38) 부근의 오차는 2018년의 오차와 비교하였을 때, 약 0.2 이상 크게 나타나 국어교과와 수학교과에 비해서도 비교적 큰 RMSE가 산출되었다.

또한 '모의 검사 2'는 검사 구성 시 기초학력 미달 판별을 위해 난이도는 낮고(난이도 b 〈-1.5)이면서 변별력은 양호한(변별도 a ≥ 0.5) 문항을 최소한 3개 포함한 경우를 예시한다. 그 결과, [그림 9]의 오른쪽에는 '2018년 실제 검사'와 '2019년 모의 검사 2'에 대한 θ 추정치의 RMSE를 비교하여 제시하였다. 국어교과 및 수학교과에서와 마찬가지로 영어교과에서도 '모의 검사 2'에서는 영어교과의 기초학력 미달을 결정하는 분할점수(θ = -2.38) 지점에서 '2018년 실제 검사'의 오차와 비교하면 약 0.05 정도 크게 나타났지만 '모의검사 1'에 비해 오차가 1/4 수준으로 줄어든 수준이었다.



[그림 9] 2018 실제 검사 대비 2019 모의 검사에 대한 능력 추정치의 RMSE(영어)

이와 같이, 대규모 평가에서 문항 수 감축이 불가피한 상황에서 학생의 기초학력 미달 여부 판별을 기존과 유사한 수준으로 안정적으로 수행할 수 있는지를 점검한 결과, 문항 특성을 고려하여 검사 구성을 설계할 필요가 있음을 알 수 있었다. 즉 문항의 난이도는 '하' 수준이면서 변별력을 충분히 갖춘 문항을 3개 이상 확보하는 경우에 기존 검사 길이에서와 근접한 수준의 오차를 보이는 것을 확인하였다.

V. 결론 및 논의

대규모 평가의 검사 구성 체제가 변화하게 되는 원인은 다양하다. 2015 개정 교육과정을 처음으로 적용하게 된 2019년 학업성취도 평가는 교과역량에 대한 학생의 성취수준을 타당하게 측정하기 위해 고차원적인 문항을 도입하게 된 점이 검사 구성 체제 변화의 핵심 원인이라 할 수 있다. 이에 단순한 지식을 넘어 문제해결과 같은 복잡한 능력을 측정하는 문항이 새로 도입된 교과별 검사를 개별 학생이 주어진 시간제한(교과별 50분) 내에 완료할 수 있도록, 문항 수 감축이 피할 수 없는 현안이 되었다. 즉 2019년 학업성취도 평가의 고등학교 검사를 구성하는 데 있어 주된 변화는 교과역량 측정을 위한 문항 도입에 따른 검사 길이 축소라고 할 수 있다. 이에 따라 개별 학생에게는 시간제한 내에 풀 수 있는 적절한 검사 길이를 갖추도록 조정하되, 전체 학생에게 노출되는 문항 수를 총합하면 교육과정 상의 교과별 성취기준 대표성을 보장할 수 있도록 행렬표집 설계를 도입할 필요성이 있음을 앞에서 설명하였다. 이러한 교과별 검사도구의 문항 수 감축이 점수척도(변환점수 척도)의 안정성에 어떠한 영향을 끼치는지 점검할 필요성이 있다. 본 연구에서는 "2019년 학업성취도 평가의 고등학교 검사(국어, 수학, 영어 검사)가 이전과 비교하여 축소된 문항 수로 구성됨에도 불구하고 척도의 점수대별 오차를 안정적으로 유지하고, 기초학력 미달 학생을 판별하는 데 문제가 없는가?"에 대한 답을 찾고자 하였다.

연구목적과 관련하여 가장 먼저 학업성취도 평가의 척도점수가 변화해온 과정과 배경을 검토하였는데, 학업성취도 평가에서는 2003년에 처음으로 점수척도를 개발하였고, 이후 평가틀 변화를 겪을 때마다 재척도화를 거쳐왔으며, PISA, TIMSS, 미국 대규모 표준화 평가인 NAEP, PARCC에서도 적용하고 있는 문항반응이론 기반의 점수척도를 가장 최근인 2015년에 도입하였음을 확인하였다(박정 외, 2006, pp.82-88; 김경희 외, 2011, pp.49-73; 박인용 외, 2017, pp.71-75). 본 연구에서는 2015년에 학업성취도 평가에 도입되어 지금껏 적용되고 있는 문항반응이론 진점수를 바탕으로 한 척도화 방법을 유지하되, 문항 수 감축 상황에서도 기존의 평가와 유사한 척도 안정성을 확보하고, 특히 사회의 가장 큰 관심인 기초학력 미달 학생을 판별하는 데 유효하도록 검사구성안을 제시하는 것을 목적으로 하였다.

연구방법으로는 2019년 검사 길이를 축소하도록 계획된 검사구성안을 고려하여 고등학교 교과별(국어, 수학, 영어) 모의 검사를 두 가지 버전으로 설계하였다. 또한 구성된 모의 검사에 대해 문항반응이론에 기반하여 동일한 능력(θ)을 가진 학생별로 1,000명의 응답 자료를 생산하였다. 학업성취도 평

가에서는 매년 평가 결과로 산출되는 기초학력 미달 학생 비율에 대한 연도간 변화에 가장 큰 사회적인 관심이 집중되므로 점수척도의 안정성 여부는 '기초학력'과 '기초학력 미달'을 구분하는 지점에서의 학생의 능력을 추정 오차 크기가 핵심이라고 할 수 있다. 본 연구에서 수행한 모의실험 결과를 요약하면, 기초학력 미달 학생을 판별의 정밀성을 유지하기 위해 난이도는 '하' 수준(난이도 $b < -1.5$)이면서 변별력은 어느 정도 양호한(변별도 $a \geq 0.5$) 문항을 교과별(국어, 수학, 영어) 검사에서 공통적으로 최소한 3개 이상 유지하는 기준을 지키는 경우, 기초학력 미달 분할점수 지점에서의 학생의 능력점수(θ) 추정의 평균오차(RMSE)가 최대 0.05를 초과하지 않는 것을 확인하였다.

본 연구의 결과를 바탕으로 세 가지 시사점을 정리할 수 있다. 첫째, 학업성취도 평가 결과의 가장 큰 역할은 우리나라 학생들의 연도별 성취수준 비율 추이를 모니터링하는 것이며, 이때 특히 기초학력 미달 학생의 비율의 증가 또는 감소 여부에 사회적 관심이 크게 집중된다. 본 연구에서 수행한 모의실험 결과에 의하면, 기존의 문항 수를 줄이는 것이 불가피한 상황에서 문항의 특성을 고려하여 검사 구성을 설계하는 방법이 기존의 안정성을 유지하는 데 유효하였다. 구체적으로 난이도는 기초학력과 기초학력 미달을 구분하기에 적절하도록 낮은 수준(난이도 $b < -1.5$)이면서 변별력은 어느 정도 양호한(변별도 $a \geq 0.5$) 문항을 교과별 검사에서 공통적으로 3개 이상 확보한다면 기초학력 미달을 판별하는 분할점수 근처의 오차가 0.05 이하로 유지되므로 성취수준 모니터링의 안정성이 확보된다는 것을 확인하였다. 이러한 연구 결과는 향후 교과별 검사의 평가틀을 설정하고 출제하는 과정에서 교과역량을 반영하기 위한 문항 내용에 대한 관심뿐 아니라 문항의 특성(난이도, 변별도)을 정밀하게 조절할 수 있는 지침을 갖출 필요가 있음을 시사한다. 향후 역량평가의 필요성이 증가함에 따라 학업성취도 평가 외에 다른 대규모 학생평가 상황에서도 검사의 길이가 축소되어야 할 필요성이 제기될 수 있다. 본 연구에서와 같이 문항 수가 축소되는 경우, 기존 평가와 같은 척도의 안정성을 확보하기 위해 교과별 평가틀 설정 및 출제 시 고려해야할 지침으로 안내할 수 있을 것이다.

둘째, 점차 사회적 요구에 따라 학업성취도 평가를 비롯한 학생을 대상으로 하는 평가는 점차 역량평가로의 전환을 이뤄야할 것이 예상된다. 따라서 개별 학생이 치르는 검사 길이를 더욱 더 축소하더라도 문항 내용의 복잡성이 증가될 가능성이 있다. 따라서 짧은 검사 길이로도 학생의 성취도를 안정적으로 모니터링하는 기능은 유지할 수 있도록 효율적으로 행렬표집 설계 및 시행의 개선을 지속적으로 고민하고 추진해야 할 것이다. 지필평가 시행 체제에서는 행렬표집 설계가 지나치게 정교해지면 검사지 인쇄 및 배송 상황에서의 사고 발생 우려가 커지는 한계점이 있었으나 컴퓨터 기반 평가 체제 전환이 가속화됨에 따라 더욱 효율적이면서 정교하고 미래지향적인 시행 체제로 개선하기 위한 다양한 후속 연구가 필요할 것이다.

셋째, 지금까지 학업성취도 평가에서는 학생들에게 평가 결과로서 4수준(우수학력, 보통학력, 기초학력, 기초학력 미달)에 대한 정보만을 제공해 주었을 뿐, 척도점수에 관련된 정보는 학생이나 학교에 통보하지 않았다. 이러한 제한된 평가 결과 제공은 물론 경쟁을 유발하지 않고 준거참조평가를 통해 학교교육의 목표 달성을 확인할 수 있다는 장점이 있다. 그러나 학업성취도 평가가 전수평가에서 표집평가로 바뀌면서 어차피 개인의 서열 정보가 산출되지 않는 상황이므로 우리나라 학교교육의 질을 관리하기 위한 목적으로 보다 풍부하고 의미 있는 결과 활용이 가능한 점수척도를 개발할 것을 요청받고 있다. 추후에는 평가 결과로서 어떤 척도점수를 제공하느냐가 이전보다 더욱 중요한 문제로 인식될 것

이라 예상된다. 대규모 평가 자료를 다양하게 분석하여 데이터에 기반한 정책이나 교육적 의사결정이 효과적으로 이루어지기 위해서는 학생들의 성취도 평가 점수를 정확하게 추정하는 점수척도를 갖추는 것이 중요하다. 특히 대규모 평가 결과를 다양한 교육 환경, 학생들의 정의적 특성, 학생의 학습 과정이나 경험 관련 데이터를 수집할 수 있는 설문자료와 연계하여 심층분석를 수행함으로써 학생의 학업성취도에 영향을 미치는 주요 요인을 원인을 진단하고자 한다면, 학생들의 교과별 성취를 과학적인 방법을 통하여 정확하게 측정할 수 있도록 평가틀 및 점수척도를 보완하고 개선하는 방안이 지속적으로 탐색되어야 할 것이다.

넷째, 본 연구에서는 “2019년 학업성취도 평가의 고등학교 검사(국어, 수학, 영어 검사)가 이전과 비교하여 축소된 문항 수로 구성됨에도 불구하고 척도의 점수대별 오차를 안정적으로 유지하고, 기초학력 미달 학생을 판별하는 데 문제가 없는가?” 라는 가장 주목받는 이슈에 초점을 맞추었으나, 보통학력과 기초학력의 구분, 우수학력과 보통학력의 구분에 관련된 추가적인 분석에 대한 필요성도 존재한다. 사회적 관심이 기초학력 미달 학생 비율에 치중하므로, 본 연구에서는 기초학력 미달 학생 판별의 안정성에 초점을 맞추었으나 후속연구에서는 다른 성취수준 분류에 대한 안정성 분석도 수행할 필요가 있다. 또한 학업성취도 평가는 2019년에 지필평가 형태를 유지하면서 검사 길이를 20문항정도로 감축되었으나, 이후 2022년에 컴퓨터 기반 평가 (Computer Based Test)를 도입하고, 2024년 이후에는 학생의 성취수준을 고려하여 개별 맞춤형 문항을 제공하는 컴퓨터 적응형 평가 (Computerized Adaptive Test)로의 전환을 추진할 계획이다(교육부, 2021. 6. 1.). 따라서 향후에는 학업성취도 평가의 검사 길이가 보다 더 축소될 가능성이 있으므로, 어느 정도의 검사 길이부터 측정 정확도에 문제가 발생하는지를 분석하는 후속연구도 필요할 것이다.

참고문헌

교육부(2021. 6. 1.). **2020년 국가수준 학업성취도 평가 결과 및 학습 지원 강화를 위한 대응 전략 발표**. 교육부 보도자료.

김경희, 정구향, 채선희, 김재철, 남명호, 반재천, 손원숙, 남현우, 이규민. (2003). **국가수준 교육성취도 평가 점수체제 개발 연구**. 한국교육과정평가원. 연구보고 RRE 2003-17.

김성훈, 김희경, 김성숙. (2010). 혼합형 검사를 사용한 피험자 능력모수 추정에서 가중우도 방법과 사후기대 방법의 기능 비교. **교육평가연구, 23**(3), 665-685.

김희경, 김완수, 김수진, 정혜경, 김미림, 김성훈. (2019). **국가수준 학업성취도 평가 점수 체제 개선 및 결과 활용도 제고 방안**. 한국교육과정평가원. 연구보고 RRE 2019-3.

동효관, 김경주, 강민경, 장의선, 성경희, 임해미, 김성경, 이재봉, 배주경, 김소연, 최병택, 최원호, 김용진, 이기영 (2017). **2017년 국가수준 학업성취도 평가 출제 연구**. 한국교육과정평가원. 연구보고 2017-2.

동효관, 김경주, 강민경, 장의선, 성경희, 양성현, 김성경, 이재봉, 구자옥, 박상복, 김소연, 최원호, 김용진, 이기영. (2018). **2015 개정 교육과정에 따른 국가수준 학업성취도 평가 출제 방안 연구**. 한국교육과정평가원. 연구보고 RRE 2018-4.

박인용, 김완수, 정혜경, 서민희, 한정아. (2017). **2017년 국가수준 학업성취도 평가 기술보고서.** 한국교육과정평가원. 연구보고 RRE 2017-10.

박정, 김경희, 김수진, 손원숙, 송미영, 조지민. (2006). **국가수준 학업성취도 평가 기술보고서**. 한국교육과정평가원. 연구보고 RRO 2006-4.

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

이재봉, 김준식, 박지선, 성경희, 이광상, 이소라, 정혜윤, 최소영, 김감영, 안유민, 하민수 (2020). **컴퓨터 기반 국가수준 학업성취도 평가(eNAEA) 도입을 위한 출제 방안 연구.** 한국교육과정평가원. 연구보고 2020-5.

Kim, S. (2012). A follow-up study of psychometric properties of IRT proficiency estimates. *Journal of Educational Evaluation, 25*(4), 829-849.

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

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates.

Educational Measurement: Issues and Practice, 29(3), 8-14.

OECD. (2017). *PISA 2015 technical report*. Retrieved from https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf (2023. 2. 14. 검색)

Yamamoto, K., & Kulick, E. (2000). Scaling methodology and procedures for the TIMSS mathematics and science scales. In M. O. Martin, K. D. Gregory, & S. E. Stemler (Eds.), *TIMSS 1999 technical report*, pp.235-264. Leicester: International Study Center. Retrieved from https://timss.bc.edu/timss1999i/pdf/T99_TR_Chap14.pdf (2023. 2. 14. 검색)

· 논문접수 : 2023.01.05. / 수정본접수 : 2023.02.08. / 게재승인 : 2023.02.17.

ABSTRACT

Analysis of Scale Stability of NAEA in the Presence of Test Length Change

HeeKyoung Kim
Senior Research Fellow, KICE
Seong-hoon Kim
Professor, Han Yang University

The 2019 NAEA(National Assessment of Educational Achievement) for high school is planed to implement a new type of item to measure student subject-specific competency to apply the 2015 Revised National Curriculum. Since the new type of items are implemented to measure complex abilities rather than simple knowledge, the issue of reducing the test length was considered to solve higher-level items. Namely, the major change in the 2019 NAEA is the reduction in the number of items caused by the introduction of the new type of items. This study inspected how the test length change of the 2019 NAEA for high school had an effect on the scale stability. In the NAEA, the media attention is focused on the year trend of the proportion of below basic students, so the scale stability of the NAEA which categorizes students into basic and below basic, is main point. According to the results of this simulation study, in order to precisely categorize basic level students from below basic level students with less than 0.05 estimated RMSE at the cut-off score: at least three items should be included for each subject test with the 'low' difficulty level (item difficulty parameter $b < -1.5$) and the 'moderate' discrimination level (item discrimination parameter $a \geq 0.5$).

Key Words: *Large-scale test, National Assessment of Educational Achievement(NAEA), Test Length, Scale stability, Score system*