교육과정평가연구
The Journal of Curriculum and Evaluation
2024, Vol. 27, No. 1, pp. 209~243
DOI: https://doi.org/10.29221/jce.2024.27.1.209

컴퓨터 기반 평가에서 디지털 기기 친숙도에 따른 차별기능문항 분석: eTIMSS 2019 4학년 수학을 중심으로

서경숙 (이화여자대학교 교육학과 박사과정)*
안해연 (이화여자대학교 교육학과 박사수료)
최윤정 (이화여자대학교 교육학과 교수)**

요약

본 연구는 컴퓨터 기반 평가의 차별기능문항 분석을 통해 디지털 기기 활용 경험이 상대적으로 부족한 학생에게 불리하게 작용하는 문항의 유형을 탐색하고자 하였다. 이를 위해 eTIMSS 2019의 우리나라 초등학교 4학년 수학 자료를 활용하여 차별기능문항 분석을 실시하였다. 분석에는 정규 수학에 응시한 557명, PSI 수학 문항에 응시한 556명의 자료를 활용하였으며, 디지털 기기 친숙도가 낮은 집단을 초점집단(focal group), 디지털 기기 친숙도가 높은 집단을 참조집단(reference group)으로 선정하였다. 차별기능문항의 분석 방법으로는 로지스틱 회귀분석 방법, Mantel-Haenszel 방법, IRT 우도비 검정(likelihood ratio test) 방법을 사용하여 결과를 비교하였다. 분석 결과 디지털 기기 친숙도가 낮은 집단의 학생들은 컴퓨터 상호 작용이 강화된 문항, 숫자 키패드로 답을 입력하는 문항들에 불리한 것으로 나타났다. 본 연구의 결과는 향후 컴퓨터 기반 평가의 타당도를 제고하기 위한 정보로 활용될 수 있을 것으로 기대한다.

주제어: 컴퓨터 기반 평가, 차별기능문항, 디지털 기기 친숙도, eTIMSS 2019

* 제1저자, ksseo7575@gmail.com
** 교신저자, younjengchoi@ewha.ac.kr

I. 서 론

디지털 전환 시대를 맞아 미래 사회의 요구에 부응하기 위해 교육 환경이 빠르게 변화하고 있다. 학교 현장에는 초·중등 학생을 위한 디지털 기기 지원이 확대되고 있으며, 인공지능, 에듀테크 등 다양한 디지털 기술을 활용하는 수업이 활발하게 이루어지고 있다. 이러한 변화에 발맞추어 교육의 질을 점검하고 개선하기 위한 컴퓨터 기반 평가(computer based test; CBT)가 주목받고 있다. 국제적으로는 Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study(TIMSS) 등의 대규모 학업성취도 평가가 컴퓨터 기반으로 전환되었으며, 우리나라에서도 2022년부터 국가수준 학업성취도 평가인 '맞춤형 학업성취도 자율 평가'를 컴퓨터 기반으로 시행하고 있다.

컴퓨터 기반 평가는 시험 과정을 간소화하며, 표준화된 관리를 보장하고, 채점 오류를 줄일 수 있다. 또한 멀티미디어 기술의 사용으로 지필평가로는 구현할 수 없었던 역량 평가적 내용과 형식을 실현할 수 있으며, 평가 결과를 데이터베이스로 축적하여 환류 체계를 구축하도록 돕는다(백종호, 이재봉, 구자옥, 2023). 이처럼 컴퓨터화 검사는 발전된 컴퓨터 기술을 통해 평가의 개발, 시행, 채점 및 해석 등 검사과정의 전 영역에서 시험 관리자와 응시자의 편의를 도모한다는 장점이 있다(Miller & Lovler, 2016).

이러한 장점에도 불구하고 컴퓨터 기반 평가의 단점도 존재한다. 컴퓨터 불안(computer anxiety), 일반화된 해석(generalized interpretation), 시험 전략 변경의 어려움 등이 그것이다(Miller & Lovler, 2016). 지필평가를 주로 경험해 온 학생들은 컴퓨터로 시험을 치르는 것에 대해 낯설어할 수 있으며, 디지털 기기에 접근하기 어려운 소외 계층이나 컴퓨터 활용 경험이 상대적으로 부족한 어린 학생들의 경우에는 더욱 어려움을 느낄 수 있어, 컴퓨터 기반 평가의 타당도 문제, 측정 효율성 저하 등의 우려도 제기되고 있다.

이처럼 컴퓨터 기반 평가는 학생의 능력을 새롭게 측정하고 데이터에 기반하여 종합적으로 분석할 수 있도록 돕지만, 학생의 디지털 기기에 대한 경험, 친숙도, 활용 능력의 차이가 차별적으로 기능하여 정확한 평가를 어렵게 할 수도 있다. 따라서 컴퓨터 기반 평가를 통해 학생의 실질적인 능력을 측정하기 위해서는 학생이 새로운 기술에 적응이 가능한 수준과 범위 내에서 점진적으로 이루어질 필요가 있다(이재봉 외, 2020).

본 연구는 eTIMSS 2019의 초등학교 4학년 자료를 활용하여 컴퓨터 기반 평가에서 디지털 기기 친숙도에 따른 차별기능문항을 탐색하고자 하였다. eTIMSS는 컴퓨터 기반의 국제 대규모 학업 성취도 평가로서, 학생들의 정보 기기 활용 경험 차이가 컴퓨터 기반 평가에서 학생의 수행 능력에 미치는 영향을 분석하기에 타당한 자료라 할 수 있다.

현재 시행 중인 컴퓨터 기반의 맞춤형 학업성취도 평가는 2024년부터 그 대상이 초등학교 3학년까지 확대 시행될 예정이며(교육부, 2023), 학교 수업에서의 디지털 기기의 활용이 크게 확대되고 있어, 변화하는 평가 환경에 대한 대처가 요구되고 있다. 이에 본 연구는 컴퓨터 기반 평가의 차별기능문항

분석을 통해 디지털 기기 활용 경험이 상대적으로 부족한 학생에게 불리하게 작용하는 문항 유형을 탐색하고자 하였으며, 연구의 결과는 향후 컴퓨터 기반 평가의 타당도를 제고하기 위한 정보로 활용될 수 있을 것이다.

본 연구의 연구 문제는 다음과 같다.

첫째, eTIMSS 2019의 초등학교 4학년의 수학 영역의 공통 정규 문항 검사에 있어 디지털 기기 친숙도 간 차별기능문항은 어떤 특성이 있는가?

둘째, eTIMSS 2019의 초등학교 4학년의 수학 영역의 문제해결 및 탐구 과제 문항 검사에 있어 디지털 기기 친숙도 간 차별기능문항은 어떤 특성이 있는가?

II. 이론적 배경

1. 컴퓨터 기반 평가

컴퓨터 기반 평가(computer based test; CBT)는 학생의 평가 수행을 포함한 평가 관리의 전 과정이 컴퓨터 시스템을 통해 이루어지는 평가를 뜻한다(김명화 외, 2011). 교육 분야에서의 디지털 기술 도입과 CBT의 편의성으로 인해, PISA, TIMSS 등의 대규모 국제 학업성취도 평가를 비롯한 많은 평가가 컴퓨터 기반으로 시행되고 있어(Keng et al., 2008; Russo, 2002; Thelwall, 2000), CBT에 관한 연구의 필요성이 증가하고 있다.

CBT는 학습자 개인의 학업에 대한 자율적 평가, 유용한 피드백, 학습공간과 시간의 유연성, 멀티미디어의 활용 가능성 등 여러 면에서 장점을 가지고 있는 것으로 나타났다(McDowell & Sambell, 1999). 또한 미디어, 도구 조작, 정보 검색 등 다양한 컴퓨터 기능을 통해 문제해결 과정을 현실적으로 재현할 수 있어, 기존 지필평가(paper based test; PBT)의 한계를 넘어 학생들의 문제해결 역량을 효과적으로 측정할 수 있는 장점이 있다(교육부, 교육과정평가원, 2022).

그러나 PBT에서 CBT로의 전환은 학생의 학업성취도에 영향을 미칠 가능성이 있는 것으로 나타났다(Bennett et al., 2008; Jeong, 2014, Jerrim et al., 2018; Mazzeo & von Davier, 2014, Pommerich, 2004; Pomplun et al., 2002; Steven, 2006), Jeong(2014)의 연구에서는 우리나라의 초등학교 6학년 학생들의 PBT와 CBT 점수를 비교한 결과, CBT에서 더 낮은 성취를 보이는 것으로 나타났다. 이유로는 학생들이 지침을 읽기 위해 되돌아갈 수 없었던 점, 지시 사항('다음 질문으로 이동', '다음 페이지로 이동', '텍스트 밑줄', '답변 입력')이 PBT보다 복잡하여 문제해결에 방해가 될 수 있는 점을 들었다. Steven(2006)의 연구에서는 PISA, TIMSS와 같은 저부담(Low-Stakes) 컴퓨터 기반 평가에서 지문이 너무 길거나 부담스러운 경우, 응시자의 동기가 낮아져 실제 실력보다 낮게 평가된다고 보고하였다.

또한 CBT에서의 평가 결과는 개인의 특성인 성별(Volman et al., 2005; Vekiri & Chronaki, 2008; Anderson et al., 2008; Meelissen & Drent, 2008), 디지털 기기 친숙도 및 자신감(Bennett et al., 2008; Horkay et al., 2006) 등에 따라 달라질 수 있는 것으로 나타났다.

2. 디지털 기기 친숙도

본 연구에서의 디지털 기기 친숙도(familiarity with digital devices)는 컴퓨터와 태블릿에 대한 친숙도 및 자신감을 의미한다. TIMSS 2019는 많은 국가에서 컴퓨터 기반 버전으로 참여하였으나, 일부 학생들은 디지털 기기를 통한 평가 경험이 적었을 것이 예상되었으므로, eTIMSS 2019(TIMSS 2019의 컴퓨터 기반 버전)는 참여하는 학생들에게 디지털 기기 친숙도를 묻는 몇 가지 추가 설문에 답하도록 하였다.

설문의 1번 문항에서는 컴퓨터나 태블릿으로 응시한 소감을 물었으며, 2번 문항에서는 컴퓨터나 태블릿 기반의 학습 빈도를 조사하였다. 우리나라 초등학교 4학년 학생의 컴퓨터 기반의 학습 빈도는 높지 않은 것으로 나타났다. '학교 숙제하기'의 경우 '한 달에 한두 번(35.0%)'으로 응답한 학생이 가장 많았으며, '거의 없음(30.6%)', '한주에 한두 번(24.1%)', '거의 매일(9.4%)' 순으로 나타났다. 수학 수업에 컴퓨터를 활용하는 빈도는 '거의 없음(47.0%)', '한 달에 한두 번(20.9%)', '한주에 한두 번(18.5%)', '거의 매일(11.6%)' 순으로 나타나, 수학 수업에서 컴퓨터를 활용하는 경우는 상대적으로 많지 않은 것을 알 수 있다. 또한 컴퓨터를 활용한 퀴즈나 시험의 빈도는 '거의 없음(44.6%)', '한 달에 한두 번(25.9%)', '한 주에 한두 번(20.6%)', '거의 매일(10%)' 순의 결과를 보여, 응답자 중 절반에 가까운 학생들이 평가에 컴퓨터를 거의 활용하지 않는 것으로 나타났다.

3번 문항은 컴퓨터나 태블릿 사용에 대한 자기효능감을 묻는 문항이며, 컴퓨터 활용 기능에 대한 7개의 진술문에 대하여 각각 '많이 동의함', '조금 동의함', '조금 동의하지 않음', '많이 동의하지 않음'으로 동의하는 정도를 선택하도록 하였다. 각 진술문의 내용은 다음과 같다. (a) 저는 컴퓨터를 잘합니다, (b) 저는 타자를 잘 칩니다, (c) 저는 컴퓨터, 태블릿, 스마트폰의 터치스크린을 사용할 수 있습니다, (d) 저는 인터넷에서 정보를 쉽게 찾을 수 있습니다, (e) 저는 인터넷에서 단어의 뜻을 찾을 수 있습니다, (f) 저는 컴퓨터로 문장과 문단을 쓸 수 있습니다, (g) 저는 컴퓨터로 문장을 수정할 수 있습니다.

eTIMSS 2019는 해당 진술문에 대한 응답을 바탕으로 학생의 디지털 기기 친숙도를 상·중·하의 척도(ASDGSEC)와 척도 점수(ASBGSEC)로 제공하고 있다. 우리나라 초등학교 4학년 학생 중 디지털 기기 친숙도가 높은 학생들은 평균 11.2점의 척도 점수를 받았는데, 이는 7개 진술문 중 6개 이상에서 '많이 동의함', 나머지 한 진술문은 '조금 동의함'에 해당한다. 디지털 기기 친숙도가 낮은 학생들은 평균 5.7점의 척도 점수를 받았으며, 이는 학생들이 7가지 진술문에 대해 평균적으로 '조금 동의하지 않음'에 해당한다. 디지털 기기에 대한 중간 정도의 친숙도를 가지고 있는 학생은 평균 8.1점의 척도 점수를 받았으며, 7개 진술문에 대해 평균적으로 '조금 동의함'에 해당한다. 디지털 기기 친숙도 상·중·하 척도에 따른 학생 수와 비율은 아래 〈표 1〉과 같다.

〈표 1〉 eTIMSS 2019에서 우리나라 초등학교 4학년 학생의 디지털 기기 친숙도(N=4,448)

디지털 기기 친숙도	학생 수	비율(%)
상	1,376	30.9
중	2,609	58.7
하	449	10.1
결측 및 무효	14	3.1

3. 차별기능문항의 개념과 분석방법

차별기능문항(differential item functioning; DIF)은 동일한 수준의 잠재적 능력을 지닌 피험자들이 속한 집단의 문화적·경험적 배경 등으로 인해 문항 정답 확률이 달라지는 현상을 의미한다(Holland & Wainer, 2012). 이는 학업성취도 평가에서 중요한 고려사항이다. 평가는 학습자의 능력을 정확하게 측정해야 하며, 특정 집단에 대한 편향(bias) 없이 공정(fairness)하게 이루어져야 한다. 만약 평가 문항이 편향성을 가진다면, 평가 결과를 통해 피험자의 특성을 정확하게 파악할 수 없다.

차별기능문항에 관한 연구는 1960년대 미국에서 시작되었으며, 개인의 수행 능력의 차이가 평가 결과에 영향을 준다고 보고되었다(Angoff, 2012). 선행연구에서 밝혀진 차별기능문항의 원인으로는 사용 언어(Linn & Drasgow, 1987; Linn, Levine, Hastings et al., 1981; Shepard, Camilli, & Averill, 1981), 번역(Yildirim & Berberoglu, 2009), 교육과정의 차이(이태구, 손지영, 양희원, 2016; Ercikan & Koh, 2005) 등이 있으며, 정규 학교 교육 이외의 '방과 후 학습 시간'이나 '과거의 교육 경험'과 같은 학생의 문화적 배경 변수(Clauser, Nungester, & Swaminathan, 1996; Wu & Ercikan, 2006)도 원인으로 지목되었다. 국내 연구의 경우에는 오랜 시간 단일 민족 문화를 고수해 왔기 때문에, 문화적 차이보다는 성별에 따른 차별기능문항 연구가 주로 이루어졌다(김종민, 이문수, 안성훈, 2016; 노언경, 김진호, 김수진, 2010; 성태제, 1994).

차별기능문항의 분석에는 여러 방법이 사용되었다. Mantel-Haenszel 방법, SIBTEST 방법(노언경, 2007; 이영주, 2012; 조윤동, 강은주, 고호경, 2012; 진수정, 성태제, 2004), 로지스틱 회귀분석 방법(강태훈, 2018; 손원숙, 2012), 그리고 문항반응이론(item response theory; IRT)을 활용한 우도비 검정 방법(김종민, 이문수, 안성훈, 2016; 이태구, 2016) 등이 주로 활용되었다. 본 연구에서는 초등학생의 디지털 기기 친숙도에 따른 차별기능문항을 탐색하기 위해 고전검사이론에 기반한 Mantel-Haenszel 방법, 로지스틱 회귀분석(Logistic Regression) 방법과 문항반응이론에 기반한 IRT 우도비 검정(IRT likelihood ratio test) 방법을 사용한다. 차별기능문항 연구에서는 한 방법으로 차별기능문항을 선별한 후 다른 방법으로 추가 검증하는 다양한 접근법을 효과적인 전략으로 간주한다(Lai et al., 2005).

가. 로지스틱 회귀분석 방법

로지스틱 회귀분석 방법은 집단변수(초점 또는 참조집단), 문항 총점, 집단과 문항 총점의 상호 작용에 대한 회귀계수를 추정하고, 모형을 비교함으로 유의도를 검증한다(Camilli & Shepard, 1994). 로지스틱 회귀분석 방법을 사용한 차별기능문항 분석에서는 초점집단 또는 참조집단인 집단변수와 문항의 총점인 대응변수를 이용하여 문항의 정답률을 예측하게 된다(Swaminathan & Rogers, 1990).

$$P(u=1 \mid \theta, \Gamma)=\frac{e^{\tau_0+\tau_1\theta+\tau_2\Gamma+\tau_3(\theta\Gamma)}}{1+e^{\tau_0+\tau_1\theta+\tau_2\Gamma+\tau_3(\theta\Gamma)}} \qquad (1)$$

식 (1)의 모형에서 P는 정답을 맞출 확률, θ는 문항 총점 또는 두 집단의 능력 추정치, 개인의 잠재변수인 대응변수에 해당하며, Γ는 초점집단 또는 참조집단으로 코딩된 피검사자의 집단을 의미한다. 일반적으론 $\Gamma=1$은 초점집단, $\Gamma=0$은 참조집단을 나타낸다. $\theta\Gamma$는 대응변수와 집단변수 사이의 상호 작용을 의미이며, τ는 회귀계수로 τ_0는 절편(intercept), τ_1은 문항에 대한 응답(0 또는 1)과 개인의 잠재 변수 상의 위치(θ)를 나타내며, τ_2는 피검사자들 간의 능력 차이, τ_3는 집단과 능력의 상호작용을 의미한다. 이 모형에서 $\tau_2 \neq 0$ and $\tau_3=0$이면 균일(uniform) 차별기능문항이 존재함을 의미하고, $\tau_2>0$이면 참조집단, $\tau_2<0$이면 초점집단에 유리한 문항이다. τ_2의 값에 상관없이 $\tau_3 \neq 0$인 경우 비균일(nonuniform) 차별기능문항이 존재함을 의미한다. 만약 $\tau_3>0$이면 집단과 능력(총점) 간 양의 상관이란 의미로 초점집단의 총점이 참조집단보다 더 높아 초점집단에 유리하게 기능함을 의미한다. 반면 $\tau_3<0$인 문항은 참조집단의 총점이 초점집단보다 더 높아 참조집단에 유리하게 기능하는 것으로 해석한다. 0과 유의미하게 다른 값이라면 비균일 차별기능문항이 존재함을 의미한다.

로지스틱 회귀분석을 이용하여 차별기능문항을 추출할 때는 3가지 모형을 설정한다.

모형 1(완전 모형) : $z=\tau_0+\tau_1\theta+\tau_2\Gamma+\tau_3(\theta\Gamma)$
모형 2(축소된 모형) : $z=\tau_0+\tau_1\theta+\tau_2\Gamma$
모형 3(영 모형) : $z=\tau_0+\tau_1\theta$

모형 1은 완전 모형(full model)으로 위에서 설명한 모형과 같은 것이고, 모형 2는 축소된 모형(reduced model)으로 집단과 문항 총점의 상호 작용을 제외한, 집단과 문항 총점만을 독립변수로 한 모형이며, 모형 3은 영 모형(null model)으로 문항 총점만을 독립변수로 포함한 모형이다. 모형 검증은 〈표 2〉와 같이 2가지 단계로 이루어진다.

〈표 2〉 로지스틱 회귀분석의 3가지 모형과 모형 검증 방법

모형 1(완전 모형)	모형 2(축소된 모형)	모형 3(영 모형)
첫 번째로 모형 1과 모형 2 사이를 비교함. '$H_0 : \tau_3 = 0$'이 기각되면, 비균일 차별기능문항이 존재함.		
	'$H_0 : \tau_3 = 0$'이 채택되면, 모형 2와 모형 3 사이를 비교함. '$H_0 : \tau_2 = 0$'이 기각되면, 균일 차별기능문항이 존재함.	

모형1과 모형 2, 모형 2와 모형 3의 차이 비교로 이루어진다. 간단히 표현하면 다음의 G_1, G_2와 같다.

G_1 = [-2ln(모형2의 최대우도)]-[-2ln(모형1의 최대우도)]
G_1 = [-2ln(모형3의 최대우도)]-[-2ln(모형2의 최대우도)]

G_1과 G_2는 카이제곱(χ^2) 분포를 하며, G_1과 G_2의 자유도는 두 모형의 차이이므로 각각 1이다. 완전 모형과 축소된 모형을 다루는 G_1은 비균일 차별기능문항의 지수이고, 영 모형과 축소된 모형을 다루는 G_2는 균일 차별기능문항의 지수이다.

Zumbo(1999)에 의해 제안된 보완된 접근 방식은 모형의 $\triangle R^2$을 사용하여 차별기능문항의 크기를 평가한다. 로지스틱 회귀분석 방법에는 다양한 유형의 R^2 통계량이 존재하며 Zumbo의 접근 방식에는 최소한 두 가지 변형이 있다. 그중 하나는 Nagelkerke R^2이고, 다른 하나는 가중 최소 제곱 R^2이다. 본 연구에서 사용되는 효과 크기는 Nagelkerke R^2 값을 사용하며, 효과 크기에 대한 판단 기준은 Zumbo와 Thomas(1997)의 연구에서 사용한 기준을 참고하였다. 그 판단 기준은 $\triangle R^2 <$ 0.035으면 매우 작은 수준(A 수준), $0.035 \leq \triangle R^2 <$ 0.070으면 중간 수준(B 수준), $\triangle R^2 \geq 0.070$으면 큰 수준(C 수준)이다.

나. Mantel-Haenszel 방법

Mantel-Haenszel 방법은 Mantel과 Haenszel이 생물 통계학 분야를 분석하기 위하여 사용되었다(1959). Holland가 처음 적용하였으며(1985), Holland와 Thayer가 차별기능문항 추출 방법을 정립하였다(1988). Mantel-Haenszel 방법은 Scheuneman과 Camilli의 카이제곱(χ^2) 통계 방법과 문항에 차별기능이 없다는 영가설을 검증하는 방식에서 비슷하다(추정아, 성태제, 1993).

Mantel-Haenszel 방법은 초점집단과 이를 비교할 참조집단의 특성과 능력 수준을 동일하게 대응(matching)시키기 위해 원점수에 따라 j개의 점수구간을 만든다. 그 다음 각 점수구간별로 집단과 정답 여부에 따라 2x2 분할표로 정리하며 이는 〈표 3〉과 같다.

〈표 3〉 j번째 점수구간의 집단과 정답 여부에 따른 피검사자 수의 2x2 분할표

		집단		전체 집단
		참조집단	초점집단	
정답여부	1(정답)	A_j	C_j	m_{1j}
	0(오답)	B_j	D_j	m_{0j}
전체		n_{Rj}	n_{Fj}	T_j

T_j는 초점집단과 참조집단의 총 피검사자 수, n_{Rj}와 n_{Fj}는 참조집단의 피검사자 수와 초점집단의 피검사자 수, m_{1j}와 m_{0j}는 정답을 맞춘 총 피검사자 수와 정답을 맞추지 못한 총 피검사의 수이며, A_j는 참조집단의 정답을 맞춘 피검사자 수, B_j는 참조집단의 정답을 맞추지 못한 피검사자 수, C_j는 초점집단의 정답을 맞춘 피검사자 수, D_j는 초점집단의 정답을 맞추지 못한 피검사자 수이다.

차별기능문항을 식별하기 위해 Mantel-Haenszel χ^2의 식은 다음과 같다.

$$MH\chi^2 = \frac{(\left|\sum_{j=1} A_j - \sum_{j=1} E(A_j)\right| - 0.5)^2}{\sum_{j=1} Var(A_j)}$$

$$E(A_j) = \frac{n_{Rj} m_{1j}}{T_j}$$

$$Var(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)}$$

Mantel-Haenszel χ^2는 어떤 문항이 집단 간에 차별적으로 기능하지 않는다는 영가설 하에서 운영되며, 자유도가 1인 χ^2 분포를 따른다(Dorans & Holland, 1993). 그러나 이 접근법은 문항이 차별기능문항을 나타내는지를 검증할 뿐, 차별기능의 방향이나 크기는 점수 수준에서 차별적 기능의 정도를 나타내는 지수들의 값을 합산하여 추정치의 값으로 알 수 있다. 이 값은 0에서 +∞까지의 범위를 가질 수 있다. 문항이 차별적으로 기능하지 않는다는 영가설이 참인 경우, 이 추정치의 기댓값은 1이다. 이 값은 참조집단이 정답을 맞힐 확률이 초점집단이 정답을 맞힐 확률을 초과하는 정도를 나타낸다. 따라서 1보다 큰 값은 문항이 참조집단에게 더 유리하다는 것을 의미하며, 1보다 작은 값은 초점집단에게 편향되어 있다는 것을 의미한다(노언경, 2007). 이 방법은 통계적 검정력과 계산 측면에서 모두 효과적이다.

Mantel-Haenszel 방법은 유의성 검증과 함께 차별기능의 정도를 나타내는 추정치를 사용하며, 문항반응이론에 기반한 방법들이 요구하는 것보다 훨씬 작은 표본 크기로도 신뢰할 수 있는 결과를 도출할 수 있다는 장점이 있다(Swaminathan & Rogers, 1990).

차별기능문항의 여부를 판별한 후에, 방향과 효과 크기에 대한 정보를 알기 위해 공통 승산비인 α_{MH}와 $\triangle_{MH}$의 크기를 이용한다. α_{MH}는 모든 점수 수준에서 α_j를 더한 값이며, $\triangle_{MH}$는 해석을

위해 α_{MH}에 자연로그를 취한 후 상수 -2.35를 곱한 값이다(Holland & Thayer, 1985). 이 방법의 효과에 대한 판단 기준은 〈표 4〉와 같다(노언경, 김진호, 김수진, 2010).

〈표 4〉 Mantel-Haenszel 방법 차별기능문항의 효과에 대한 판단 기준

통계량	효과 크기	차별기능의 방향
α_{MH}	$0 < \alpha_{MH} < 1$	초점집단에 유리
	$\alpha_{MH} = 1$	차별기능 안 함
	$\alpha_{MH} \geq 1$	참조집단에 유리
$\triangle_{MH}$	$\lvert\triangle_{MH}\rvert \leq 1.0$	A 수준(유의하지 않은 정도)
	$1.0 < \lvert\triangle_{MH}\rvert < 1.5$	B 수준(중간 정도)
	$\lvert\triangle_{MH}\rvert \geq 1.5$	C 수준(심한 정도)

다. IRT 우도비 검정 방법

IRT 우도비 검정(IRT likelihood ratio test)은 Thissen, Steinberg와 Wainer(1988)에 의해 발전된 방법으로, 문항반응이론 내에서 주변최대우도추정(marginal maximum likelihood estimation) 알고리즘을 사용하여 모수를 추정한다. 영가설은 두 집단 간 문항모수에 차이가 없다는 것을 가정한다. 이 검정은 문항모수가 모두 동일하다고 가정하는 축소 모형(compact model)과 동일성 제약을 해제한 확장 모형(argument model)의 우도를 비교하여 차별기능문항을 판별한다. 우도비 검정 통계량은 다음과 같이 표현된다.

$$G^2 = -2\ln L_c - (-2\ln L_A) = -2\ln(L_c - L_A) \sim X^2_{df}$$

검정 통계량 G^2값이 모수에 통계적으로 유의미한 차이가 있음을 나타내면($G^2 > \chi^2_{(\alpha=.05,\, df)}$), 개별 모수에 대한 검정을 통해 차별기능문항의 유형을 파악할 수 있다(노언경, 김진호, 김수진, 2010). 우도비 검정은 균일 차별기능문항과 비균일 차별기능문항을 모두 탐지할 수 있다. 이는 초점집단과 참조집단 각각의 문항특성곡선(item characteristic curve; ICC) 일치 여부를 χ^2 차이 검정을 활용하여 우도비 검정 통계량(G^2)을 통하여 검정한다. 만약 두 집단의 문항특성곡선에 통계적으로 유의미한 차이가 없다면, 해당 문항이 차별기능문항이 아니라는 의미이다(강태훈, 2017).

III. 연구 방법

1. 연구 대상

eTIMSS 2019는 국제 교육성취도 평가 협회(International Association for the Evaluation of Educational Achievement; IEA)에서 주관하는 평가로, 58개국의 초등학생 약 33만 명, 39개국의 중학생 약 25만 명이 참여하였으며, 우리나라에서는 2018년 12월에 170개 초등학교 4학년생 5,855명과 175개 중학교 2학년생 6,246명이 참여하였다(교육부, 2020).

eTIMSS 2019는 컴퓨터 기반 평가로 진행되었으나, 초등학교 4학년생들은 중학교 2학년생들에 비해 디지털 기기 사용에 익숙하지 않으며, 타이핑(typing), 드래그 앤 드롭(drag and drop), 숫자 키패드(number keypad) 사용하기 등과 같은 컴퓨터 사용 능력이 다소 부족하였을 것으로 가정할 수 있다(서민희 외, 2021). 따라서 본 연구는 초등학교 4학년 학생을 연구 대상으로 하였으며, 수학 영역을 중심으로 분석하였다.

수학은 정규 문항으로 행렬표집 방식에 의해 14개의 검사지(booklet 1~14)와 문제해결 및 탐구 과제(problem solving and inquiry tasks; PSI) 문항으로 구성된 2개의 검사지(booklet 15, 16)로 구성되어 있기에 정규 수학 문항은 일부 문항만 제한적으로 연구에 사용하였다. 각 피험자는 하나의 검사지로 시험을 치르고, 각 검사지는 동등화를 목적으로 공통 문항을 포함하게 된다. 이러한 이유로 본 연구에서는 가장 많은 학생 수를 확보하기 위하여, 정규 수학 문항 중 1번과 14번 검사지(booklet 1, 14)와 PSI 문항 검사지(booklet 15, 16)를 연구에 사용하였다.

최종 데이터에 포함된 우리나라 초등학교 4학년 학생 총 4,448명 중에서 1번과 14번 검사지(booklet 1, 14)에 응시한 557명과 PSI 문항 검사지(booklet 15, 16)에 응시한 556명을 최종 연구 대상으로 선정하였다. 본 연구에 활용한 문항은 1번과 14번 검사지(booklet 1, 14)에 공통된 수학 22문항(ME51043A ~ ME51507B)과 PSI 수학 43문항(MQ11A01 ~ MQ11P07B)이며, 정규 수학 문항으로 연구에 사용한 22개 문항들과 PSI 수학 문항 43개 문항들은 각각 하나의 검사로 간주하였고, 각 문항은 이분 문항으로 처리하였다. 최종 연구 문항과 연구 대상 수는 〈표 5〉의 음영 부분이며 연구 대상의 인구통계학적 특성은 〈표 6〉과 같다.

〈표 5〉 검사지 문항 및 연구 대상 수

구분	booklet	제외된 문항	선택된 문항	성별		
				남	여	계
정규 수학	1	ME71219 ~ ME71204 (32 문항)	ME51043A ~ ME51507B (22 문항)	147 (53.1%)	130 (46.9%)	277 (100%)
	14	ME71024 ~ ME71205 (20 문항)		139 (49.6%)	141 (50.4%)	280 (100%)
	계			286	271	557
PSI 수학	15, 16		MQ11A01 ~ MQ11P07B (43 문항)	288	265	553

〈표 6〉 연구 대상의 인구통계학적 특성

변수	유형	정규 수학 문항 그룹		PSI 수학 문항 그룹	
		빈도	비율(%)	빈도	비율(%)
성별	남	286	51.3	288	51.9
	여	271	48.7	265	47.7
거주지역	인구밀도가 높은 도시	255	45.8	256	46.0
	교외 지역	102	18.3	96	17.3
	중규모 도시 또는 대도시	143	25.7	148	26.6
	소도시 또는 마을	57	10.2	56	10.1
가정 내 컴퓨터(태블릿) 보유 여부	보유	520	93.4	518	93.2
	미보유	36	6.5	35	6.3
	무응답	1	0.2	3	.5
디지털 기기 친숙도	높음	174	31.2	168	30.3
	보통	322	57.8	319	57.5
	낮음	61	11.0	64	11.5
계		557명		556명	

2. 측정 도구

eTIMSS는 TIMSS 2019에서 새롭게 도입되는 컴퓨터, 태블릿 기반의 평가를 지칭하는 용어로, 새로운 문항 형식인 문제해결과 탐구 과제(PSI)가 추가되었으며, 태블릿 환경에 적합한 학생들의 답변 유형 4개가 추가되었다. eTIMSS 2019의 성취도 평가틀은 평가하려는 내용과 인지 과정의 두 가지로 구성된다. 본 연구의 대상인 초등학교 4학년에 수학의 내용영역은 수(number), 측정과 기하(measurement and geometry), 자료(data)로 구분되고, 인지영역은 알기(knowing), 적용하기(applying), 추론하기(reasoning)로 구분된다(교육부, 2020).

연구 대상이 되는 공통 정규 수학 검사지와 PSI 수학 검사지에 대한 내용영역과 인지영역별 문항 수는 〈표 7〉과 같다. 내용영역으로 살펴보면, 공통 정규 수학 검사지는 수 16문항(72.7%), 측정과 기하 4문항(18.2%), 자료 2문항(9.1%) 순으로 많았으며, PSI 수학 검사지는 수 32문항(74.4%), 자료 7문항(16.3%), 측정과 기하 4문항(9.3%) 순으로 많이 나타났다. 한편 인지영역으로 살펴보면, 공통 정규 수학 검사지에서는 알기가 13문항(59.1%), 적용하기가 7문항(31.8%), 추론하기가 2문항(9.1%) 순으로 많았으며, PSI 수학 검사지에서는 적용하기가 25문항(58.1%), 추론하기 11문항(25.6%), 알기 7문항(16.3%) 순으로 많이 나타났다.

〈표 7〉 초등 4학년 수학 내용영역과 인지영역별 문항 수

구분	내용영역	인지영역			계 (%)
		알기	적용하기	추론하기	
공통 정규 수학 검사지	수	10	5	1	16 (72.7)
	측정과 기하	2	1	1	4 (18.2)
	자료	1	1	0	2 (9.1)
	계 (%)	13 (59.1)	7 (31.8)	2 (9.1)	22 (100)
PSI 수학 검사지	수	4	19	9	32 (74.4)
	측정과 기하	2	2	0	4 (9.3)
	자료	1	4	2	7 (16.3)
	계	7 (16.3)	25 (58.1)	11 (25.6)	43 (100)

3. 분석 방법

본 연구에서는 수학 과목의 성취도에 차이를 보이게 하는 이차적인 차원인 디지털 기기 활용 능력에 관심을 두고 있으므로 디지털 기기 친숙도가 낮은 집단(ASDGSEC=3)을 초점집단(focal group), 디지털 기기 친숙도가 높은 집단(ASDGSEC=1)을 참조집단(reference group)으로 선정하고자 하였다. 따라서 eTIMSS 2019의 '디지털 기기 친숙도' 척도가 낮은 학생(공통 정규 수학 61명, PSI 수학 64명)을 초점집단으로, 높은 학생(공통 정규 수학 174명, PSI 수학 168명)을 참조집단으로 선정하였다. 이 두 집단에 속한 피험자들의 능력 수준을 통제한 후, 응답에 있어 집단 간 차이가 나타나는지 확인하기 위하여 세 가지 차별기능문항 방법을 사용하여 결과를 비교하였다.

본 연구에서는 디지털 기기 친숙도에 따른 차별기능문항 추출을 위해 고전검사이론을 적용한 모수적 검정인 로지스틱 회귀분석 방법, 비모수적 검정인 Mantel-Haenszel 방법과 문항반응이론을 적용한 우도비 검정 방법을 적용하였다. 이는 다양한 조건 하에서 데이터의 차별기능문항을 탐색하기 위함이다.

첫 번째로 로지스틱 회귀분석 방법는 응답을 0과 1로 하는 이분형 로지스틱 회귀 방법으로, 문항 점수를 종속변수로 사용하였다(강태훈, 2018). 통계적 유의성과 함께 실제적인 중요성을 판단하기 위하여 차별기능문항의 효과 크기(effect size)를 고려하였다. 로지스틱 회귀분석은 R의 difR 패키지를 사용하여 분석하였다.

두 번째로 Mantel-Haenszel 방법은 추출된 차별기능문항 유무와 방향의 크기에 대해 사용 기준은 α_{MH}와 $\triangle_{MH}$를 이용하였다. α_{MH} 값은 양수이며, 차별기능문항이 아닌 경우 α_{MH}는 1이다. α_{MH}가 1보다 작으면 초점집단에 유리하고, 1보다 크면 참조집단에 유리하다고 판단하였다. $\triangle_{MH}$의 경우, 0일 때 차별기능이 없고, 0보다 작으면 참조집단에 유리하며, 0보다 크면 초점집단에 유리하다고 판단하였다. Mantel-Haenszel 방법은 R의 difR 패키지를 사용하여 분석하였다.

세 번째로 IRT 우도비 검정 방법은 문항반응이론의 문항 모수의 차이를 통해 차별기능문항을 추출한다. 본 연구에서는 IRTLRDIF 컴퓨터 프로그램(Thissen, 2001)을 사용하여 분석하였으며, 이 프로그램의 집단 번호는 초점집단의 경우 1, 참조집단의 경우 2로 설정하였다. 응답이 없는 문항은 공백으로 표시하였다. 이분적 채점 자료로 데이터가 전처리된 정규 수학 문항과 PSI 문항 자료에 대하여 3-모수 IRT모형을 적용하였다. 3-모수 모형은 모든 문항 모수가 참조집단과 초점집단 간에 동일하다는 가설에 대해 G^2 검정에서 χ^2 분포의 $\alpha = 0.05$에서 자유도가 3일 때의 임계값인 7.82를 초과하지 않으면 차별기능문항이 아니고, 7.82를 초과하면 차별기능문항이라고 하였다.

IV. 연구 결과

본 연구는 eTIMSS 2019 수학 영역에서 디지털 기기 친숙도에 따른 집단 간 문항 수준의 차별적 기능을 하는 문항을 추출하기 위하여 정답 문항은 1점, 오답이거나 무응답인 문항은 0점으로 점수화하였으며, 디지털 기기 친숙도가 높은 집단을 참조집단(reference group)으로, 디지털 기기 친숙도가 낮은 집단을 초점집단(focal group)으로 설정하여 일관적으로 분석하였다. eTIMSS 2019 수학 영역에서 집단에 따른 성취도와 기술통계는 〈표 8〉과 같다.

〈표 8〉 정규 수학과 PSI 수학의 집단별 기술통계

		공통 정규 수학(*N*=235)		PSI 수학(*N*=232)	
		참조집단 (*N*=174)	초점집단 (*N*=61)	참조집단 (*N*=168)	초점집단 (*N*=64)
수학 성취도	평균	611.30 (59.24)	571.41 (68.48)	616.27 (58.06)	596.27 (63.49)
	최댓값	745.19	681.36	760.83	709.42
	최솟값	439.42	390.31	463.31	411.27
성별(남=1, 여=0)		0.54	0.49	0.58	0.52
지역 규모		5.48 (0.75)	5.21 (1.17)	5.38 (0.99)	5.33 (1.05)
디지털 기기 친숙도		11.33 (1.24)	5.72 (0.91)	11.04 (1.20)	8.74 (1.98)

주: 성별(1=여, 2=남), 지역 규모는 지역의 인구수를 나타내며, 숫자가 클수록 대도시를 나타냄.(6 = 500,000명 이상, 5 = 100,001 ~ 500,000, 4 = 50,001 ~ 100,000, 3 = 30,001 ~ 50,000, 2 = 15,001 ~ 30,000, 1 = 3,001 ~ 15,000), 숫자는 평균(Mean)을 나타내며, 괄호( ) 안의 숫자는 표준편차(standard deviation)를 나타냄.

공통 정규 수학과 PSI 수학 그룹 모두, 초점집단의 학생들이 디지털 기기 친숙도와 수학 성취도에서 더 낮은 점수를 보이는 것을 확인하였다. 성별, 연령은 차이를 보이지 않았으며, 거주지역에서는 초점집단 학생 거주지의 지역 규모가 다소 작은 것을 확인할 수 있다.

본 연구의 결과에서는 공통 정규 수학 문항과 문제해결 및 탐구 과제(PSI) 수학 문항에 대해 로지스틱 회귀분석 방법, Mantel-Haenszel 방법, IRT 우도비 검정 방법을 사용하여 추출한 차별기능문항 결과를 제시하였다.

1. 공통 정규 수학 문항에 대한 차별기능문항 추출 결과

공통 정규 수학 문항에서 디지털 기기 친숙도에 따른 비균일 차별기능문항을 추출하기 위한 방법으로, 첫 번째로 로지스틱 회귀분석 방법으로 추출한 결과는 〈표 9〉와 같다.

로지스틱 회귀분석 기법을 사용하여 분석한 결과, 1번과 10번 문항에서 비균일 차별기능문항이 추출되었으며, 3번 문항에서 균일 차별기능문항이 추출되었다. 1번과 10번 문항을 좀 더 살펴보면, 두 문항 모두 G1 값이 $\alpha = 0.05$ 수준에서 자유도 1인 카이제곱 분포의 임계치인 3.84를 넘어 $\tau_3 = 0$이라는 영가설을 기각하므로 비균일적 차별기능문항이 된다. 3번 문항은 G2 값이 임계치인 3.84를 넘고 G1 값은 임계치를 넘지 못하므로 $\tau_2 \neq 0$, $\tau_3 = 0$이 되어 균일적 차별기능문항으로 디지털 기기 친숙도가 낮은 집단에게 유리하였다. 1번 문항에 대한 차별기능문항 효과의 크기는 'A 수준'으로 매우 작은 크기로 나타나며, 3번과 10번 문항은 'B 수준'으로 효과의 크기가 중간 크기로 나타났다.

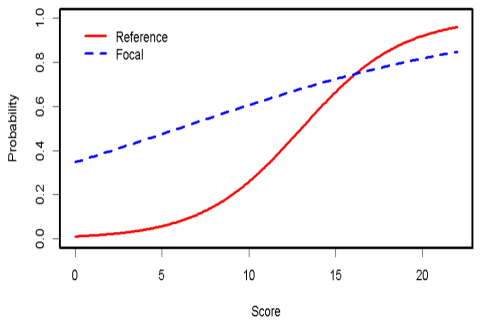
〈표 9〉 공통 정규 수학 문항의 로지스틱 회귀분석 방법 차별기능문항

문항	로지스틱 회귀분석(비균일)			로지스틱 회귀분석(균일)			DIF 유형	유리	
	G1	p-value	$\triangle R^2$	G2	p-value	$\triangle R^2$		초점집단	참조집단
I1	5.64	0.02*	A	2.46	0.12	A	비균일 DIF		
I2	2.30	0.13	A	0.00	0.96	A			
I3	0.06	0.81	A	6.89	0.01**	B	균일 DIF	○	
I4	0.00	0.97	A	0.44	0.51	A			
I5	0.02	0.88	A	0.24	0.62	A			
I6	3.81	0.05	B	3.84	0.05	B			
I7	0.01	0.92	A	0.21	0.65	A			
I8	1.24	0.27	A	1.29	0.26	A			
I9	1.02	0.31	A	0.00	0.96	A			
I10	4.32	0.04*	B	0.06	0.81	A	비균일 DIF		
I11	1.95	0.16	A	1.78	0.18	A			
I12	2.47	0.12	A	2.48	0.12	A			

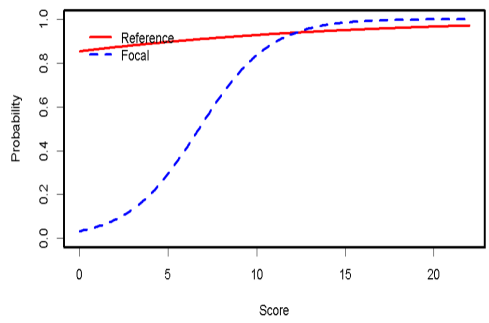
문항	로지스틱 회귀분석(비균일)			로지스틱 회귀분석(균일)			DIF 유형	유리	
	G1	p-value	$\triangle R^2$	G2	p-value	$\triangle R^2$		초점집단	참조집단
I13	0.21	0.65	A	0.59	0.44	A			
I14	0.10	0.76	A	0.43	0.51	A			
I15	3.31	0.07	A	1.43	0.23	A			
I16	3.31	0.07	A	1.43	0.23	A			
I17	0.03	0.87	A	0.46	0.50	A			
I18	2.40	0.12	A	0.10	0.75	A			
I19	0.01	0.93	A	0.14	0.71	A			
I20	1.22	0.27	A	0.88	0.35	A			
I21	0.04	0.85	A	0.02	0.89	A			
I22	0.53	0.47	A	0.69	0.41	A			

주 1: ** $p < .01$, * $p < .05$
주 2: 디지털 기기 친숙도가 낮은 집단이 초점집단, 디지털 기기 친숙도가 높은 집단이 참조집단.

[그림 1]에 제시되어 있는 차별기능곡선은 모두 비균일 차별기능문항을 보이는 문항이다. (a) 1번 문항의 경우, 동일한 능력을 가지고 있다 하더라도 특히, 상위능력집단에서는 참조집단인 디지털 기기 친숙도가 높은 집단이 유리하게 나타났다. 반면에 (b) 10번 문항의 경우, 디지털 기기 친숙도가 높은 집단은 전반적인 능력집단에서 .8이상의 높은 정답률을 나타내고, 하위능력집단에서 디지털 기기 친숙도가 낮은 집단이 불리하게 나타났다.



(a) 1번 문항



(b) 10번 문항

[그림 1] 공통 정규 수학 문항의 비균일 차별기능곡선

두 번째로 균일 차별기능문항을 추출하기 위한 방법으로 Mantel-Haenszel 방법을 이용하였으며, 〈표 10〉에서는 Mantel-Haenszel 방법으로 추출한 차별기능문항을 제시하였다.

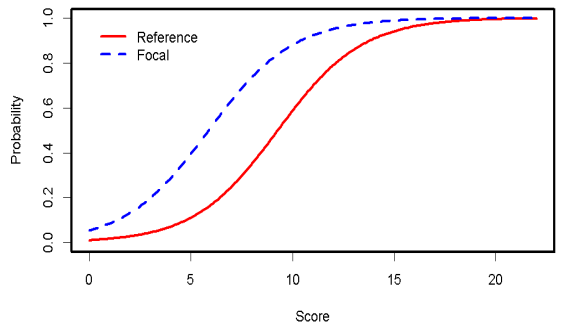
Mantel-Haenszel 방법을 사용하여 분석한 결과, 3번 문항만이 유의하게 균일적 차별기능문항으로 추출되었다. 3번 문항의 차별기능문항 통계량은 $\alpha_{MH} = 0.27$, $\triangle_{MH} = 3.10$으로 나타났으며, 이

는 $0 < \alpha_{MH} < 1$이고 $\triangle_{MH} > 0$이므로 초점집단인 디지털 기기 친숙도가 낮은 집단에 유리한 문항임을 알 수 있다. 또한 $\triangle_{MH}$의 차별기능문항 기준에 따르면 'C 수준'으로 효과의 크기가 심한 정도의 차별기능문항의 수준으로 나타났다.

〈표 10〉 공통 정규 수학 문항의 Mantel-Haenszel 방법 균일 차별기능문항

문항	MHx^2	p-value	α_{MH}	$\triangle_{MH}$		문항	MHx^2	p-value	α_{MH}	$\triangle_{MH}$	
I1	2.73	0.10	0.57	1.34	B	I12	1.98	0.16	1.76	-1.33	B
I2	0.01	0.91	1.04	-0.09	A	I13	0.55	0.46	1.36	-0.73	A
I3	4.02	0.04	0.27	3.10	C	I14	0.71	0.40	1.37	-0.74	A
I4	1.83	0.18	0.59	1.24	B	I15	1.75	0.19	1.67	-1.20	B
I5	0.01	0.93	0.97	0.07	A	I16	1.75	0.19	1.67	-1.20	B
I6	3.65	0.06	0.20	3.82	C	I17	0.57	0.45	1.48	-0.92	A
I7	0.89	0.35	0.70	0.82	A	I18	0.13	0.72	1.24	-0.51	A
I8	0.89	0.35	0.47	1.78	C	I19	0.10	0.75	1.13	-0.29	A
I9	0.03	0.86	0.93	0.16	A	I20	1.79	0.18	0.60	1.20	B
I10	0.35	0.55	1.47	-0.90	A	I21	0.04	0.84	0.94	0.15	A
I11	1.68	0.19	1.62	-1.14	B	I22	0.45	0.50	1.37	-0.74	A

[그림 2]에 제시되어 있는 차별기능곡선은 균일 차별기능문항을 보이는 3번 문항이다. 전체적으로 디지털 기기 친숙도가 낮은 집단(초점집단)의 곡선이 디지털 기기 친숙도가 높은 집단(참조집단)보다 위쪽에 놓여있는 것으로 보아 디지털 기기 친숙도가 낮은 집단(초점집단)이 유리한 문항임을 알 수 있다.



[그림 2] 공통 정규 수학 3번 문항 균일 차별기능곡선

세 번째로 공통 정규 수학 문항의 IRT 우도비 검정 방법은 〈표 11〉에서 추출한 차별기능문항을 제시하였다. 본 연구에서는 3-모수 모형을 적용하였고, G^2 값이 $\alpha = 0.05$인 수준에서 자유도 3인 카이제곱 분포의 임계치가 7.82 이상인 경우가 없으므로 차별기능문항은 추출되지 않았다.

〈표 11〉 정규 수학 문항의 우도비 검정 방법 차별기능문항

문항	G2	참조집단			초점집단				
		a	b	g	a	b	g	*M*	*SE*
1	2.3	48.65	0.49	0.41	41.23	0.5	0.52	−0.05	0.7
2	0.4	0.68	−0.14	0.28	0.41	−0.29	0.25	−0.05	0.7
3	5.2	1.49	−1.03	0.21	3.38	−1.26	0.25	−0.07	0.72
4	0	77.27	0.38	0.45	77.01	0.39	0.48	−0.05	0.7
5	0.4	50.15	0.47	0.42	42.68	0.5	0.4	−0.04	0.7
6	1	32.58	−1.26	0.23	25.48	−1.49	0.25	−0.06	0.71
7	0	104.57	0.41	0.4	104.79	0.42	0.43	−0.05	0.7
8	0.7	37.84	−1.31	0.24	34.27	−1.25	0.24	−0.04	0.68
9	0	50.97	0.72	0.18	51.06	0.7	0.2	−0.05	0.69
10	5.2	0.1	−15.9	0.25	2.13	−1.19	0.26	−0.05	0.7
11	3.6	0.54	1.1	0.19	1.46	0.89	0.18	−0.05	0.69
12	5.1	0.43	−1.52	0.25	1.2	−0.21	0.27	−0.05	0.69
13	1.1	0.52	−1.3	0.25	0.49	−0.85	0.25	−0.05	0.7
14	0	617.96	−0.02	0.21	618.22	−0.02	0.2	−0.05	0.7
15	0.4	83.62	−0.11	0.19	84.28	−0.11	0.14	−0.05	0.7
16	0.4	83.62	−0.11	0.19	84.28	−0.11	0.14	−0.05	0.7
17	2.1	0.72	−1.76	0.25	1.51	−0.8	0.28	−0.05	0.69
18	1.8	0.6	−2.6	0.25	−0.01	127.41	0.25	−0.05	0.7
19	1	0.54	−1.26	0.26	0.92	−0.7	0.25	−0.05	0.69
20	1.1	9.01	1.97	0.18	790.29	2.2	0.24	−0.05	0.7
21	0.1	0.3	0.5	0.24	0.25	0.81	0.25	−0.05	0.7
22	0.6	0.96	1.92	0.19	7.54	1.35	0.17	−0.05	0.7

2. PSI 수학 문항에 대한 차별기능문항 추출 결과

PSI 수학 문항에서 디지털 기기 친숙도에 따른 차별기능문항을 추출하기 위한 방법으로 첫 번째로 로지스틱 회귀분석 방법을 이용하였으며, 추출한 결과는 〈표 12〉와 같다.

〈표 12〉 PSI 수학 문항의 로지스틱 회귀분석 비균일과 균일 차별기능문항

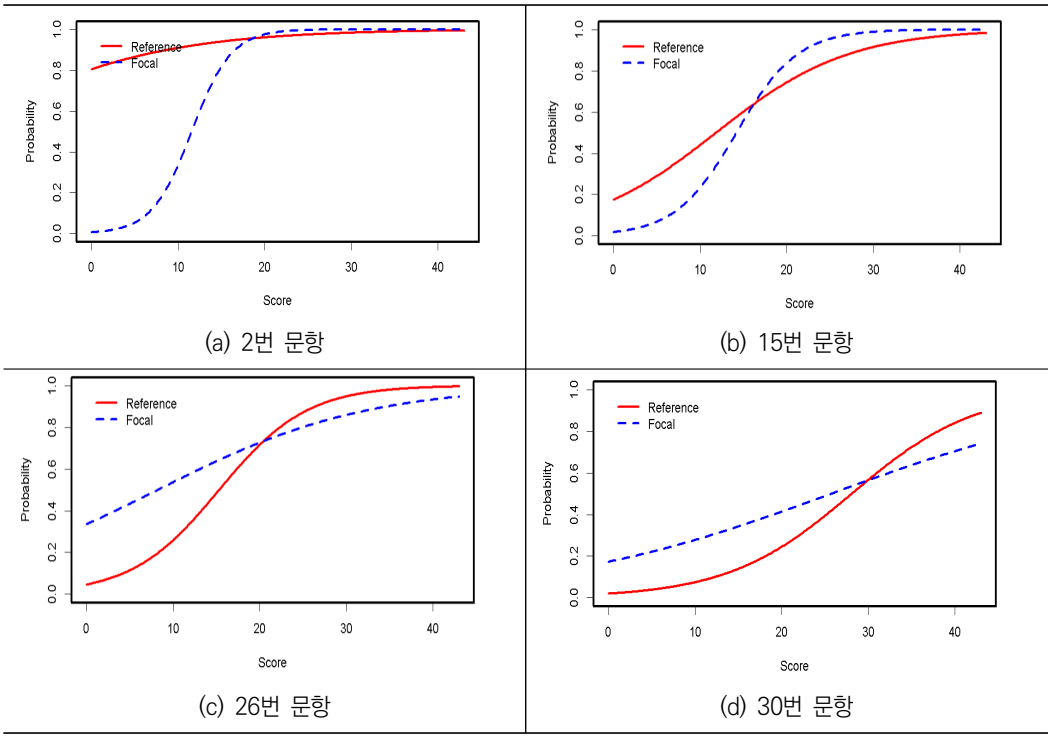
문항	비균일		균일		유리		문항	비균일		균일		유리	
	G1	$\triangle R^2$	G2	$\triangle R^2$	초점집단	참조집단		G1	$\triangle R^2$	G2	$\triangle R^2$	초점집단	참조집단
I1	1.04	A	6.63	B	○		I23	1.34	A	0.34	A		
I2	8.57	B	5.44	B			I24	1.88	A	6.59	A		○
I3	0.18	A	0.01	A			I25	0.03	A	1.66	A		
I4	1.44	A	0.67	A			I26	5.01	A	0.37	A		
I5	0.42	A	0.04	A			I27	2.88	A	7.53	B	○	
I6	2.81	A	0.01	A			I28	0.01	A	0.05	A		
I7	1.13	A	0.00	A			I29	0.08	A	0.01	A		
I8	0.01	A	0.30	A			I30	4.35	A	2.31	A		
I9	0.34	A	0.63	A			I31	0.12	A	0.01	A		
I10	0.08	A	0.45	A			I32	0.90	A	0.00	A		
I11	0.19	A	0.91	A			I33	0.18	A	2.32	A		
I12	0.74	A	6.14	A	○		I34	0.69	A	0.14	A		
I13	3.53	A	0.81	A			I35	1.97	A	0.01	A		
I14	1.66	A	0.01	A			I36	0.55	A	0.00	A		
I15	4.72	A	0.03	A			I37	0.05	A	0.04	A		
I16	0.03	A	0.58	A			I38	0.00	A	0.32	A		
I17	0.32	A	0.27	A			I39	0.94	A	0.17	A		
I18	0.01	A	0.00	A			I40	0.57	A	0.01	A		
I19	0.41	A	0.26	A			I41	1.47	A	0.06	A		
I20	0.04	A	0.50	A			I42	2.39	A	2.48	A		
I21	3.06	A	1.05	A			I43	0.37	A	1.65	A		
I22	2.30	A	0.16	A									

주: 디지털 기기 친숙도가 낮은 집단은 초점집단, 디지털 기기 친숙도가 높은 집단은 참조집단.

로지스틱 회귀분석을 사용하여 분석한 결과, 2번, 15번, 26번, 30번 문항에서 비균일 차별기능문항이 추출되었으며, 1번, 12번, 24번, 27번 문항에서 균일 차별기능문항이 추출되었다. 차별기능문항의 효과의 크기를 살펴보면, 12번, 15번, 24번, 26번, 30번 문항은 'A 수준'으로 매우 작은 크기로 나타나며, 1번, 2번, 27번 문항은 'B 수준'으로 효과의 크기가 중간 크기로 나타났다.

[그림 3]에 제시되어 있는 차별기능곡선은 모두 비균일 차별기능문항을 보이는 문항이다. (a) 2번 문항의 경우, 동일한 능력을 가지고 있다 하더라도 특히, 하위능력집단에서는 참조집단인 디지털 기기 친숙도가 높은 집단이 유리하게 나타났다. (b) 15번 문항의 경우, 하위능력집단과 상위능력집단의 유리한 집단이 서로 반대로 나타나고 있는데 하위능력집단은 디지털 기기 친숙도가 높은 집단이 유리하고 상위능력집단에서는 디지털 기기 친숙도가 낮은 집단이 유리하게 나타났다. (c) 26번 문항과 (d) 30번 문항은 하위능력집단에서는 디지털 기기 친숙도가 낮은 집단이 유리하고 상위능력집단에서는

디지털 기기 친숙도가 높은 집단이 유리하게 나타났다.



[그림 3] PSI 수학 문항의 비균일 차별기능곡선

두 번째로 균일 차별기능문항을 추출하기 위한 방법으로 Mantel-Haenszel 방법을 이용하였으며, 〈표 13〉에서는 Mantel-Haenszel 방법으로 추출한 결과를 제시하였다.

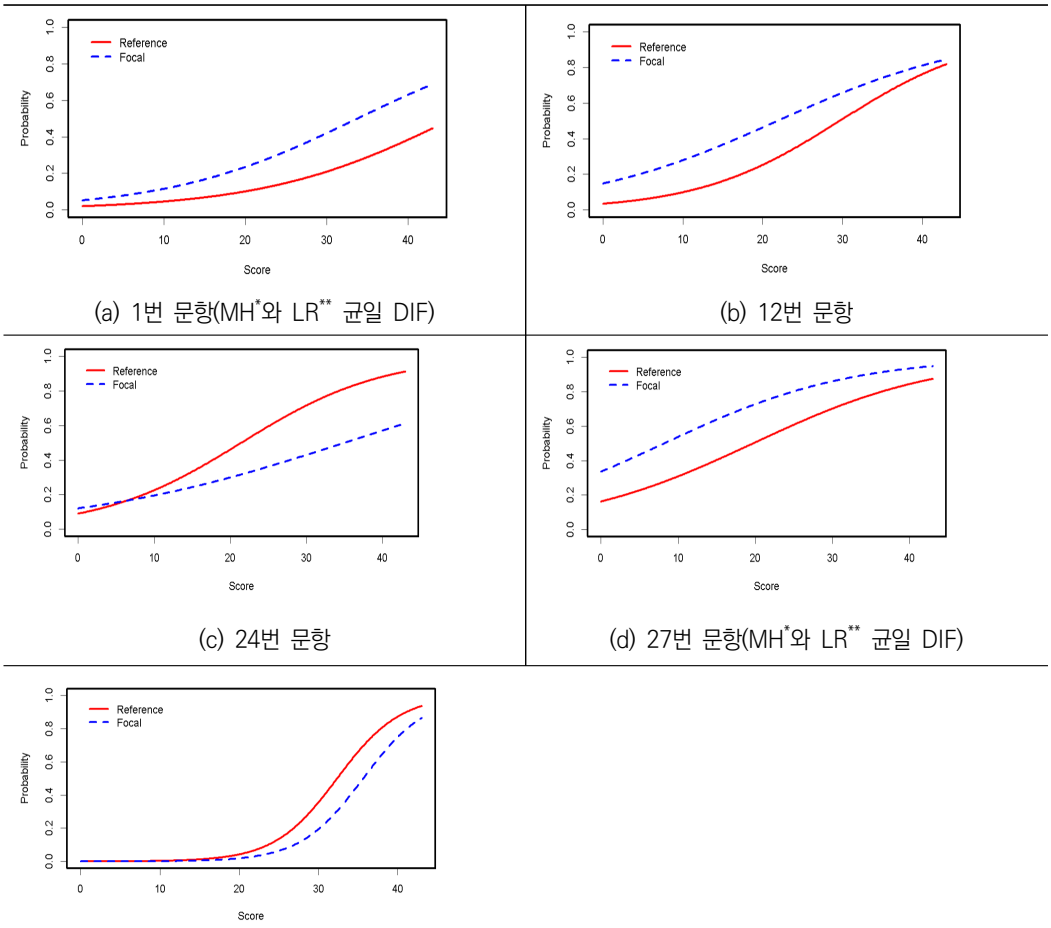
〈표 13〉 PSI 수학 문항의 Mantel-Haenszel 방법 차별기능문항

문항	MHx^2	p-value	α_{MH}	$\triangle_{MH}$		문항	MHx^2	p-value	α_{MH}	$\triangle_{MH}$	
I1	7.32	0.01	0.30	2.86	C	I23	0.02		1.19	-0.42	A
I2	2.22	0.14	3.25	-2.77	C	I24	2.98	0.08	2.00	-1.63	C
I3	0.38	0.54	0.70	0.82	A	I25	0.89	0.35	1.54	-1.01	B
I4	0.07	0.79	1.25	-0.52	A	I26	0.04	0.85	0.99	0.02	A
I5	0.00	0.96	0.92	0.19	A	I27	4.47	0.03	0.43	1.97	C
I6	0.00	0.98	0.90	0.25	A	I28	0.16	0.69	1.24	-0.51	A
I7	0.00	0.96	0.86	0.36	A	I29	0.00	0.97	1.14	-0.31	A
I8	0.00	0.97	1.10	-0.23	A	I30	1.26	0.26	0.65	1.03	B

문항	MHx^2	p-value	α_{MH}	$\triangle_{MH}$		문항	MHx^2	p-value	α_{MH}	$\triangle_{MH}$	
I9	0.10	0.75	1.21	-0.45	A	I31	0.05	0.82	1.19	-0.40	A
I10	0.01	0.94	1.11	-0.24	A	I32	0.00	0.97	0.92	0.19	A
I11	0.43	0.51	0.70	0.84	A	I33	0.55	0.46	1.36	-0.73	A
I12	3.48	0.06	0.45	1.86	C	I34	0.09	0.77	0.75	0.68	A
I13	0.11	0.74	0.70	0.85	A	I35	0.02	0.89	0.83	0.45	A
I14	0.08	0.78	1.40	-0.79	A	I36	0.00	0.97	0.87	0.32	A
I15	0.06	0.80	1.00	0.00	A	I37	0.00	0.95	1.04	-0.08	A
I16	0.01	0.91	0.96	0.09	A	I38	0.50	0.48	1.57	-1.06	B
I17	0.68	0.41	0.61	1.16	B	I39	0.05	0.82	1.20	-0.43	A
I18	0.00	0.97	0.95	0.12	A	I40	0.01	0.92	0.85	0.37	A
I19	0.18	0.67	1.27	-0.57	A	I41	0.28	0.6	0.74	0.71	A
I20	0.18	0.68	1.28	-0.58	A	I42	4.57	0.03	5.57	-4.04	C
I21	0.53	0.47	1.61	-1.12	B	I43	0.55	0.46	1.82	-1.41	B
I22	0.01	0.91	1.05	-0.12	A						

Mantel-Haenszel 방법 분석 결과, 1번, 27번, 42번 문항이 유의하게 균일적 차별기능문항으로 추출되었다. 1번 문항의 차별기능문항 통계량은 $\alpha_{MH} = 0.30$, $\triangle_{MH} = 7.32$로 나타났으며, 이는 균일 차별기능문항으로서 초점집단인 디지털 기기 친숙도가 낮은 집단(초점집단)에 유리한 문항임을 알 수 있다. 또한 $\triangle_{MH}$의 차별기능문항 기준에 따르면 'C 수준'으로 효과의 크기가 심한 정도의 차별기능문항의 수준으로 나타났다. 그리고 27번 문항의 차별기능문항 통계량은 $\alpha_{MH} = 0.43$, $\triangle_{MH} = 1.97$으로 나타났으며, 균일 차별기능문항으로 디지털 기기 친숙도가 낮은 집단(초점집단)에 유리한 문항임을 알 수 있다. 42번 문항의 차별기능문항 통계량은 $\alpha_{MH} = 5.57$, $\triangle_{MH} = -4.04$로 나타났으며, 이는 균일 차별기능문항으로 참조집단인 디지털 기기 친숙도가 높은 집단에 유리한 문항임을 알 수 있다. 로지스틱 회귀분석의 균일 차별기능문항과 Mantel-Haenszel 방법으로 추출된 균일 차별기능문항에서 공통으로 나온 문항은 1번과 27번 문항이다.

[그림 4]에 제시되어 있는 차별기능곡선은 균일 차별기능문항을 보이는 1번, 12번, 24번, 27번 문항이다. (a), (b), (d) 문항은 디지털 기기 친숙도이 낮은 집단에 유리한 문항이며 (c), (e) 문항은 참조집단이 디지털 기기 친숙도가 높은 집단이 유리하게 나타났다.



(a) 1번 문항(MH*와 LR** 균일 DIF)

(b) 12번 문항

(c) 24번 문항

(d) 27번 문항(MH*와 LR** 균일 DIF)

(e) 42번 문항

*: Mantel-Haenszel 방법
**: 로지스틱 회귀분석 방법

[그림 4] PSI 수학 문항의 균일 차별기능곡선

세 번째로 우도비 검정 방법을 사용하였으며, 추출한 결과는 〈표 14〉와 같다.

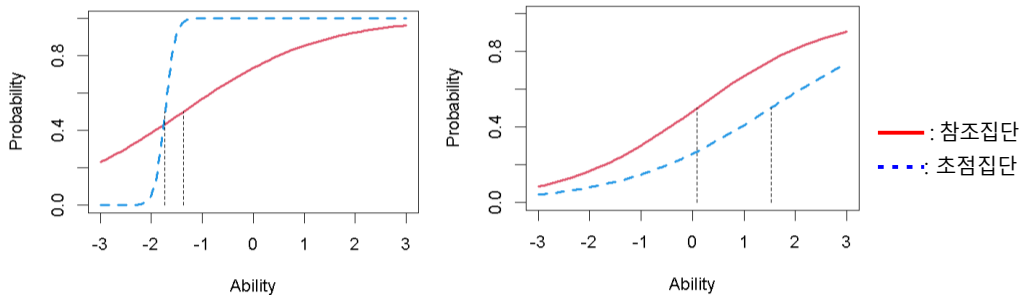
〈표 14〉 PSI 수학 문항의 우도비 검정 방법 차별기능문항

문항	G^2	참조집단			초점집단				
		a	b	c	a	b	c	*M*	*SE*
1	5	1.77	1.69	0.14	0.76	1.67	0.21	-1.14	1.65
2	6.9	0.35	-5.71	0.25	1.85	-2.02	0.22	-1.08	1.58
3	0	0.42	-1.07	0.23	0.44	-1.14	0.24	-1.14	1.65
4	0.5	0.33	-2.98	0.25	0.46	-2.08	0.25	-1.13	1.65
5	0.1	0.56	1.55	0.22	0.36	2.2	0.21	-1.14	1.66

6	4.4	0.81	−1.69	0.24	0.25	−3.44	0.25	−1.15	1.68
7	0	2.02	0.1	0.13	2.02	0.02	0.12	−1.14	1.65
8	0	1	−0.03	0.22	0.71	−0.07	0.16	−1.14	1.66
9	0.2	0.91	0.17	0.24	0.64	0.19	0.16	−1.14	1.66
10	0	1.1	0.49	0.21	0.69	0.63	0.15	−1.14	1.66
11	1.2	0.59	−0.18	0.24	0.58	−0.81	0.21	−1.14	1.66
12	3.8	0.99	0.93	0.23	0.5	0.32	0.25	−1.14	1.65
13	**8.1**	**0.74**	**−1.37**	**0.22**	**11.04**	**−1.73**	**0.25**	**−1.16**	**1.66**
14	1.1	0.94	−1.31	0.23	1.25	−1.63	0.19	−1.15	1.66
15	3.1	0.7	−1.32	0.22	1.33	−1.63	0.18	−1.14	1.65
16	0.4	0.98	−0.85	0.26	0.83	−0.98	0.2	−1.13	1.66
17	0	1.16	0.51	0.14	1.23	0.3	0.14	−1.14	1.65
18	0.4	0.69	−0.63	0.28	0.5	−0.99	0.24	−1.14	1.66
19	0	0.55	−0.78	0.24	0.56	−0.77	0.22	−1.13	1.65
20	0	0.82	−0.58	0.22	0.85	−0.39	0.22	−1.13	1.65
21	1.8	0.7	−1.61	0.24	1.52	−0.93	0.28	−1.12	1.65
22	0.6	2.11	−0.05	0.18	1.89	0.09	0.23	−1.13	1.66
23	0	1.92	0.03	0.14	2.04	0.02	0.12	−1.13	1.65
24	**8.5**	**0.77**	**0.09**	**0.25**	**0.69**	**1.53**	**0.23**	**−1.13**	**1.66**
25	1.4	0.67	−1.26	0.24	0.49	−1	0.22	−1.13	1.66
26	1.2	1.15	−0.78	0.3	12.07	0.01	0.47	−1.13	1.66
27	6.5	0.79	0.27	0.31	0.11	−3.45	0.25	−1.14	1.66
28	1.6	0.64	−0.22	0.27	1.72	0.1	0.3	−1.13	1.66
29	0.1	0.32	−2.96	0.25	0.27	−3.54	0.25	−1.14	1.66
30	4.4	0.88	0.63	0.18	0.27	1.83	0.25	−1.14	1.67
31	0	1.02	0.95	0.14	0.93	0.99	0.13	−1.13	1.66
32	0	1.04	−0.76	0.27	1.23	−0.6	0.33	−1.13	1.65
33	1.7	0.41	−0.1	0.25	0.54	0.38	0.21	−1.13	1.65
34	0	99.63	0.39	0.12	119.56	0.43	0.13	−1.14	1.65
35	1.3	68.25	0.35	0.16	88.7	0.26	0.13	−1.14	1.65
36	0	262.65	0.46	0.1	300.84	0.47	0.11	−1.14	1.65
37	0.3	0.66	0.73	0.33	0.43	0.56	0.24	−1.13	1.65
38	0	0.52	−1.78	0.25	0.5	−1.79	0.22	−1.13	1.65
39	0.4	1.55	0.71	0.16	0.96	1.13	0.16	−1.13	1.66
40	0	1.46	0.89	0.14	1.12	1.1	0.15	−1.13	1.66
41	1.3	1.51	0.56	0.21	1.2	0.27	0.14	−1.14	1.65
42	1.2	1.46	0.86	0.11	19.79	0.77	0.1	−1.13	1.66
43	1.8	151.91	1.22	0.15	3.82	1.11	0.1	−1.13	1.65

주: 디지털 기기 친숙도가 낮은 집단은 초점집단, 디지털 기기 친숙도가 높은 집단은 참조집단.

분석 결과는 G^2 값이 $\alpha = 0.05$인 수준에서 자유도 3인 카이제곱 분포의 임계치가 7.82 이상인 문항은 13번, 24번 문항이었고, 이 문항들에 대해 각각의 모수의 차이가 있는지를 분석하기 위해 문항특성곡선(ICC)을 그려본 결과, 13번 문항은 비균일 차별기능문항으로 나타났으며, 24번 문항은 균일 차별기능문항으로 문항특성곡선은 [그림 5]와 같다.



(a) 13번 문항 (b) 24번 문항

[그림 5] 정규 수학 13번, 24번 문항의 차별기능문항곡선

비균일 차별기능문항인 13번 문항은 능력이 낮은 경우는 참조집단인 디지털 기기 친숙도가 높은 집단이 유리하다가 능력이 높아지면 초점집단인 디지털 기기 친숙도가 낮은 집단이 유리한 문항으로 나타났다.

균일 차별기능문항인 24번 문항은 참조집단인 디지털 기기 친숙도가 높은 집단이 초점집단인 디지털 기기 친숙도가 낮은 집단보다 유리한 문항으로 나타났다.

3. 차별기능문항 추출 결과 분석

가. 공통 정규 수학 문항 차별기능문항 추출 결과 분석

공통 정규 수학 문항에 대하여 디지털 기기 친숙도에 따른 차별기능을 탐색하기 위하여 로지스틱 회귀분석, MH 방법, 우도비 검정 방법의 분석 결과를 종합하면 〈표 15〉와 같다.

〈표 15〉 공통 정규 수학 문항 차별기능문항 분석 결과

문항	로지스틱 회귀	MH	우도비 검정	초점집단			참조집단		
				유리	*M*	*SD*	유리	*M*	*SD*
1	비균일 DIF				.70	.46		.65	.48
3	균일 DIF	균일 DIF		○	.95	.22		.86	.35
10	비균일 DIF				.93	.25		.95	.22

주 1: 디지털 기기 친숙도가 낮은 집단이 초점집단, 디지털 기기 친숙도가 높은 집단이 참조집단.
주 2: 우도비 검정은 차별기능문항이 존재하지 않음.

총 22문항 중 19개 문항은 균일적 또는 비균일적 차별기능이 전혀 추출되지 않았다. 로지스틱 회귀분석과 Mantel-Haenszel 방법에서 공통으로 균일적 차별기능문항으로 추출된 문항은 3번 문항이다. 정규 수학 문항에서 추출된 차별기능문항의 실제 문항에 대한 정보는 〈표 16〉과 같다. 수학 내용영역에 따른 문항유형은 선다형(multiple choice items)과 구성형(constructed response items)으로 나뉜다. 구성형 문항의 답변의 경우 eTIMSS에서 채점한 후 정답 여부를 재코딩해서 제공한다. 추출된 문항은 모두 비공개 문항으로 구체적인 내용과 화면 구성을 파악하기는 어렵지만, 문항 내용을 통해 유추해 보면, 1번과 3번 문항은 약수의 개념을 알고 있는지를 확인하는 선다형 문항이다. 10번 문항은 분수의 개념을 이용하여 분수를 표현하는 문항이다. 우리나라 수학 교육과정에서는 5~6학년군에서 약수와 배수의 개념을 다루고 있어(서민희 외, 2021), 1번, 3번, 10번 문항은 초등학교 4학년 학생들의 교육과정에서 벗어난 문항에 해당한다.

〈표 16〉 공통 정규 수학 문항의 차별기능문항 정보

내용영역	문항 번호		문항 내용	인지영역	문항유형
수	1	ME51043A	6을 약수로 갖는 수 찾기 - 3	알기	선다형
수	3	ME51043C	6을 약수로 갖는 수 찾기 - 9	알기	선다형
수	10	ME51040	전체의 3/4을 색칠한 그림 찾기	알기	선다형

'디지털 기기 친숙도' 문항은 TIMSS 평가에서 학생들의 디지털 기술에 대한 숙련도 또는 경험을 평가하기 위한 문항들로 이루어져 있다. 이 문항들은 학생들이 디지털 기기를 이용하여 교육 자료에 접근하고, 학습 자료와 상호 작용하며, 기술 기반 과제를 수행하는 방식에 영향을 미칠 수 있다. 하지만, 차별기능문항으로 추출된 문항들(1, 3, 10번)은 단순 선택형 선다형 문항으로 키보드나 마우스 조작에 익숙하지 않은 학생도 특별한 어려움 없이 풀 수 있는 문항에 해당하므로 디지털 기기 친숙도가 아닌 학업성취도, 선행학습 여부, 문화적 배경 등에 의해 발생한 것으로 보아야 할 것이다. 특히 3번 문항의 경우, 초점집단에서 맞출 확률은 정답 학생 58명, 오답 학생 3명으로 95%이며, 참조집단에서 맞출 확률은 정답 학생 150명, 오답 학생 24명으로 84%로, 초점집단에서 맞춘 학생의 비율이 더 크게 나타났다. 이는 집단의 샘플 수가 작아 나타나는 현상으로 보인다.

나. PSI 수학 문항 차별기능문항 추출 결과 분석

PSI 수학 검사에 대하여 디지털 기기 친숙도에 따른 차별기능을 탐색하기 위하여 로지스틱 회귀분석, Mantel-Haenszel 방법, 우도비 검정 방법의 분석 결과를 종합하면 〈표 17〉과 같다.

〈표 17〉 PSI 문항 차별기능문항 분석 결과

문항	로지스틱 회귀	MH	우도비 검정	유리	
				초점집단	참조집단
1	균일 DIF	균일 DIF		○	
2	비균일 DIF				
12	균일 DIF			○	
13			비균일 DIF		
15	비균일 DIF				
24	균일 DIF		균일 DIF		○
26	비균일 DIF				
27	균일 DIF	균일 DIF		○	
30	비균일 DIF				
42		균일 DIF			○

주: 디지털 기기 친숙도가 낮은 집단은 초점집단, 디지털 기기 친숙도가 높은 집단은 참조집단.

총 43문항 중 33개 문항은 균일적 또는 비균일적 차별기능이 전혀 추출되지 않았다. 균일적 차별기능문항으로 추출된 문항은 1번, 12번, 24번, 27번, 30번, 42번 문항이고, 비균일적 차별기능문항으로 추출된 문항은 2번, 13번, 15번, 26번, 30번 문항이다. PSI 수학 문항 중 동시에 비균일적 차별기능문항과 균일적 차별기능문항으로 추출된 문항은 없다. 로지스틱 회귀분석과 Mantel-Haenszel 방법에 공통으로 추출된 문항은 1번, 27번 문항이고 로지스틱 회귀분석과 우도비 검정에 공통으로 추출된 문항은 24번 문항이다. 로지스틱 회귀분석, Mantel-Haenszel 방법, 우도비 검정 방법 모두에 공통으로 추출된 문항은 없다.

PSI 수학 문항에서 추출된 차별기능문항의 실제 문항에 대한 정보는 〈표 18〉과 같다. 문제해결 및 탐구 능력을 평가하는 문항인 PSI 문항은 TIMSS 2019에서 새롭게 추가된 문항으로 과제 중심의 스토리텔링형 문항으로 특정 상황을 중심으로 서로 연결된 문항으로 구성되어 있다.

〈표 18〉 PSI 수학 문항의 차별기능문항 정보

내용영역	문항 번호		문항 내용	인지영역	문항 유형
수	12	MQ11A05A	먹이를 식으로 표현하기	적용하기	선다형
수	13	MQ11A05BA	Food-pictograph-bony fish	적용하기	선다형
수	24	MQ11R04	로봇이 좋아하는 수	추론하기	구성형
수	26	MQ11P01BA*	티켓 판매 수익 구하기 400 + (6.50 + 6.00)	적용하기	선다형
수	27	MQ11P01BB*	티켓 판매 수익 구하기 (400 X 6.50) - (400 X 6.00)400 X (6.50 + 6.00)	적용하기	선다형

내용영역	문항 번호		문항 내용	인지영역	문항 유형
수	30	MQ11P01BE*	티켓 판매 수익 구하기 400 X (6.50 – 6.00)	적용하기	선다형
측정	1	MQ11A01	사진들	적용하기	선다형
측정	2	MQ11A02A	황제펭귄의 키 구하기	알기	구성형
자료	15	MQ11A05B	먹이를 그림그래프로 나타내기	적용하기	구성형
자료	42	MQ11P07A*	티켓 판매 수익 나타내기	알기	구성형

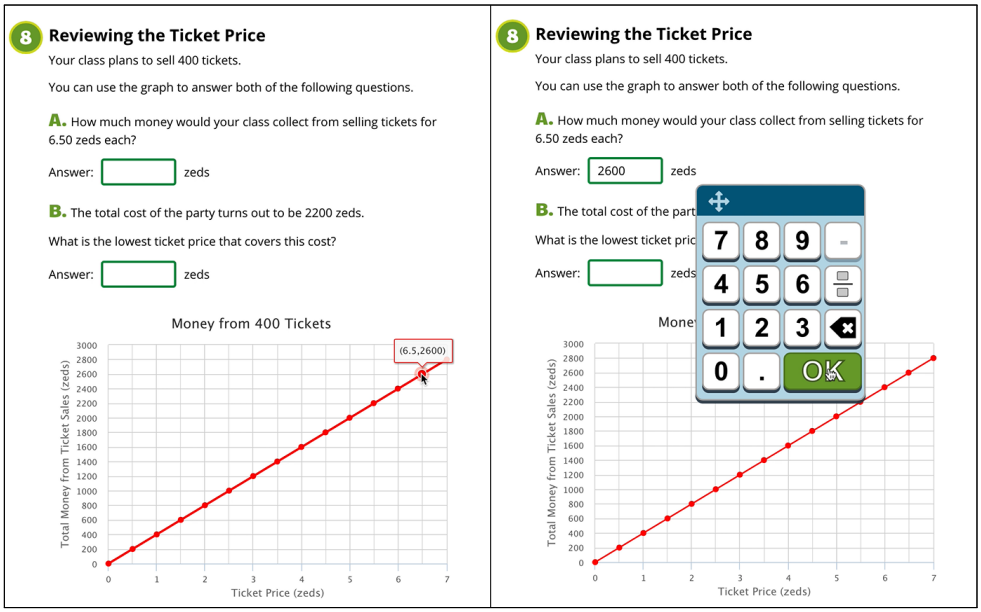
주: 문항 번호의 *표는 공개 문항임.

다. 차별기능문항 원인 분석 결과

분석을 통해 추출된 문항 중 TIMSS 2019에서 출판물과 홈페이지를 통해 공개하고 있는 문항을 중심으로 구체적인 원인을 추론하고자 하였다. 다음의 [그림 6]에 나타난 8-A번 문항은 PSI 검사지 중 디지털 기기 친숙도가 낮은 학생에게 불리한 것으로 나타난 MQ11P07A 문항이다. 문항의 답인 '티켓 가격이 6.50 zed일 때 400장을 판매해서 얻은 총수익'을 구하기 위해서는 그래프 위에 마우스를 올려 값을 확인하고, 이를 키패드로 입력하는 과정을 거쳐야 한다. 디지털 기기 친숙도가 낮은 학생들은 화면을 통해 주어진 문제상황을 이해하고 사용자-컴퓨터 간 상호 작용을 통하여 필요한 정보를 찾고, 답을 숫자 키패드로 입력하는 과정에서 어려움이 있었을 것으로 예상된다.

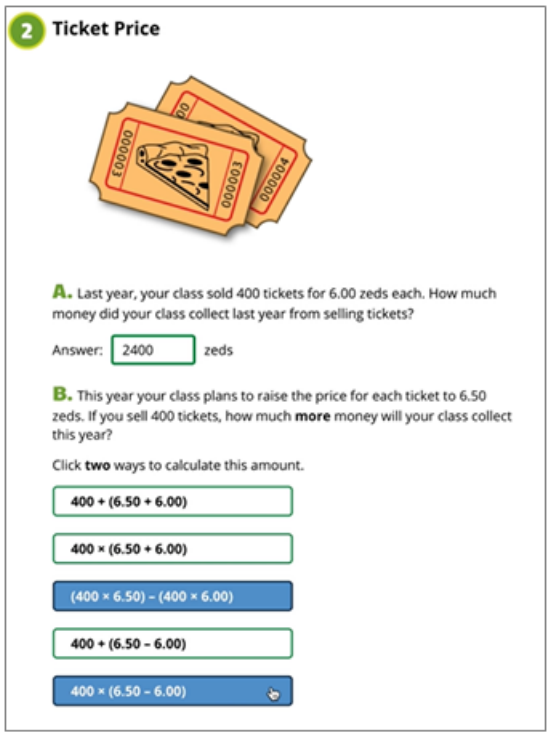
우리나라 수학과 교육과정은 수학 수업에서 계산기를 사용하도록 권장하고 있지만, 실제 초등학교 수학 수업에서 계산기는 거의 사용하고 있지 않으며(최지선, 2019), 디지털 기기 친숙도 설문 결과에서도 수학 수업에서 컴퓨터(태블릿) 사용 빈도가 낮은 것(한 달에 한두 번(35.%), 거의 없음(30.6%))으로 나타났으므로, 학생들은 숫자 키패드 사용 과정에서 어려움을 겪었을 가능성이 있다.

MQ11P07A 문항은 기존의 지필평가와는 다른 상호 작용이 강화된 문항이다. 문제해결을 위해서는 화면 아래에 주어진 선 그래프를 읽어 필요한 정보를 확인해야 하며, 선 그래프를 읽기 위해서는 그래프 위에 마우스를 올려 해당 지점의 값을 확인해야 한다. 따라서 컴퓨터(태블릿) 기반의 학습 경험이 부족한 학생들은 문항의 풀이 방법을 파악하지 못했을 가능성이 있다. 해당 문항에서 우리나라 학생의 정답률은 23%로 참가국 중 6위를 차지하였으며, 수학 영역의 전체 평균 점수에서 3위를 차지한 것에 비해 상대적으로 낮은 성취도를 보였다(Mullis et al., 2021).



[그림 6] MQ11P07A 문항의 화면(Mullis et al., 2021)

다음 [그림 7]에 나타난 2-B번 문항은 PSI 검사지 중 디지털 기기 친숙도가 낮은 집단에게 유리한 것으로 나타난 MQ11P01BB 문항이다. 이 문항에서 학생들은 괄호가 포함된 혼합계산식을 머릿속으로 구상하여 정답 선지를 고르면 되었으므로, 컴퓨터 조작의 미숙으로 인한 어려움을 예상하긴 어렵다. 자연수의 혼합계산은 2009 개정 수학과 교육과정에서는 4학년에서 다루어졌으나, 2015 개정 교육과정부터 5학년으로 이동한 내용으로(박경미 외 2015), 교육과정을 벗어난 문항에 해당한다. 이를 고려할 때 MQ11P01BB 문항이 차별기능문항으로 추출된 원인은 집단의 또 다른 특성인 학업성취도, 선행학습 여부, 문화적 배경 등이 영향을 미쳤을 가능성과 보조 수단으로 종이와 필기구 등을 활용한 학생이 유리했을 가능성 등이 존재한다.



[그림 7] MQ11P01BB 문항의 화면(Mullis et al., 2021)

이러한 결과를 종합할 때, 컴퓨터 기반 평가에서 디지털 기기 친숙도가 낮은 집단이 불리한 문항은 MQ11P07A 문항과 같이 화면 구성이 다소 복잡하고, 상호 작용이 강화된 문항이라 할 수 있다. 이는 멀티미디어와 상호 작용이 강화된 컴퓨터(태블릿) 기반의 학습 경험의 부족, 디지털 기기에 대한 미숙함 등에서 비롯되었다고 볼 수 있다.

V. 요약 및 시사점

우리나라는 컴퓨터 기반 평가인 eTIMSS 2019 수학 영역에서 참여국 중 2위라는 높은 성취를 보였지만, 평가 학생을 대상으로 디지털 기기 친숙도를 조사한 추가 문항에서는 컴퓨터나 태블릿을 이용한 학습의 빈도가 높지 않은 것으로 나타났다. 수학 수업에 컴퓨터를 활용하는 빈도를 묻는 설문에는 '거의 없다'라고 응답한 학생이 47.0%로 나타났으며, 컴퓨터를 활용한 퀴즈나 시험의 빈도를 묻는 설문에서도 '거의 없음'이 44.6%로 나타나 응답자 중 절반에 가까운 학생들이 수학 수업과 평가에 컴퓨터를 거의 활용하지 않는 것으로 나타났다. 이러한 설문 결과를 볼 때 우리나라 초등학교 4학년 학생의 컴퓨터 기반 평가에 대한 준비도가 다소 부족한 것으로 판단되어, 디지털 기기 친숙도에 따른 차별기능문항을 분석하였다.

본 연구에서 추출된 차별기능문항을 살펴보면 다음과 같다.

내용영역으로 살펴본 결과에서는, 공통 정규 수학 문항에서 추출된 차별기능문항 3문항 모두 수 영역의 연산 문항이었다. 공통 정규 수학 문항에서 수 영역이 차지하는 비율은 약 72.7% 내외를 차지하였으나, 차별기능문항 중에서는 수 영역이 차지하는 비율이 100%에 달하는 것으로 나타났으며, 수 영역 중 2개의 문항은 약수의 개념을 확인하는 문항으로 나타났다. 컴퓨터 기반 평가의 요소가 더 많이 반영된 PSI 수학 검사지에서는 총 10개의 차별기능문항이 나타났다. 수 영역이 총 6문항으로 가장 많았으며, 측정과 기하 영역이 2문항, 자료 영역이 2문항이었다. 수 영역의 6개 문항 모두 연산 영역의 문항으로 나타나 우리나라의 학생들은 컴퓨터 기반 평가에서 자료를 통해 식을 세우는 문항에서 정답률의 차이를 보이고 있음을 확인할 수 있다. 자료 영역에서는 주어진 자료를 그래프로 나타내거나 그래프를 해석하는 문항이 추출되었다.

문항 유형으로 살펴본 결과에서는, 공통 정규 수학 문항 중 선다형 문항의 비율은 50%에 달했으나, 차별기능문항 중에서는 75%의 비율로 나타났다. PSI 수학 문항 중 선다형 문항은 16.3%에 불과했으나, 차별기능문항에서는 63.6%를 차지하여 더 높은 비율을 나타냈다. 선다형 비율이 높게 나온 결과를 살펴볼 때, 문항의 유형보다는 문항의 내용영역 또는 형식 등이 정답률에 더 영향을 미치고 있다고 가정할 수 있으나 선다형 문항이라 해도 풀이 과정에서 컴퓨터 활용 능력이 요구되는 경우가 있으므로, 해당 문항에 대한 구체적이고 질적인 분석을 통해서 해석할 필요가 있다.

마지막으로 인지 영역으로 살펴본 결과에서는, 공통 정규 수학의 차별기능문항으로 '알기' 영역 3문항, '추론하기' 영역 1문항으로, '알기' 영역의 비율이 높게 나타났다. PSI 수학의 차별기능문항 중

에서는 '적용하기' 영역이 7문항으로 63.6%의 비율로 나타났다. PSI 문항 전체에서는 '적용하기' 문항이 58.1%를 차지한 것에 비해 차별기능문항 중에서는 비율이 약간 상승한 것을 확인할 수 있었다. 다음으로는 '알기' 2문항, '추론하기' 1문항이 추출되었다.

추출된 문항 중 공개 문항을 통해 구체적인 원인을 추론한 결과, 디지털 기기 친숙도가 낮은 학생들은 화면을 통해 주어진 문제상황을 이해하고 상호 작용을 통하여 정답을 찾고, 이를 숫자 키패드 등으로 입력하는 문항에 어려움이 있었으며, 이는 멀티미디어 기반의 화면 구성, 상호 작용이 강화된 컴퓨터(태블릿) 기반의 학습 경험의 부족에서 비롯된 것으로 추론할 수 있었다.

연구 결과를 바탕으로 컴퓨터 기반 평가의 타당도 제고를 위한 시사점을 제시하면 다음과 같다. 첫째, 차별기능문항이 나타난 문항 유형에 대한 심층분석을 통해 향후 개선사항에 대한 제안이 필요하다. 연구 결과 차별기능문항은 스토리텔링 형식과 상호 작용이 강화된 문항으로 구성된 PSI 수학 검사지에서 더 많이 추출되었다. PSI 수학 문항은 학생들의 참여와 동기를 향상할 수 있으며, 문제해결 과정을 현실적으로 재현할 수 있어 학생들의 문제해결 역량을 효과적으로 측정할 수 있는 장점이 있는 것으로 알려져 있다. 그러나 PSI 수학 문항에서 차별기능문항이 더 많이 추출되었으며, PSI 문항의 완료율이 66%로 공통 정규 수학 문항의 완료율이 92%인 것에 비해 현저하게 낮은 것을 볼 때, 학생들이 문제를 푸는 과정에서 어려움이 있었음을 예상할 수 있다. 따라서 향후 초등학교 저학년 학생에게 컴퓨터 기반 평가를 적용하기 위해서는, 학생들이 문항에 대한 응답 방법을 알고 어려움 없이 조작할 수 있는지 확인할 필요가 있다. 또한 본 연구의 결과에서 확인하였듯, 복잡한 상호작용을 포함한 다소 도전적인 과제를 대규모 학업성취도 평가에 적용하는 것은 더욱 신중하게 검토될 필요가 있다.

둘째, 공교육 현장에서 활용이 가능한 디지털 기기 기반의 학습 및 평가도구를 개발·보급할 필요가 있다. 우리나라 대부분의 가정에서는 컴퓨터 또는 태블릿을 보유하고 있으며 인터넷이 연결되어 있고, 최근 초·중·고 학교에는 스마트 디바이스 보급 사업이 활발하게 이루어지고 있어 디지털 기반 학습 환경이 비교적 잘 구축되어 있다. 그러나 eTIMSS 2019의 디지털 기기 친숙도 설문 결과에서는 절반에 가까운 학생들이 수학 수업과 평가에서 컴퓨터와 태블릿을 거의 활용하지 않는 것으로 나타났다. 이러한 상황에서 컴퓨터 기반 평가를 통해 학업성취도를 평가하는 것은 학습과 평가에 디지털 기기를 거의 사용하지 않는 학생들에게 불리하게 작용할 수 있다. 컴퓨터 기반 평가의 타당도를 높이기 위해서는 실제 수학 수업과 평가에서 디지털 기기의 활용이 우선되어야 하며, 이를 위해서는 실제 수업에서 활용이 가능한 컴퓨터 기반 학습 도구와 평가도구가 더 많이 보급될 필요가 있다.

본 연구는 우리나라 초등학교 4학년 학생의 디지털 기기 친숙도에 따른 차별기능문항을 검출하고, 디지털 기기 친숙도가 낮은 학생에게 불리하게 작용할 수 있는 문항의 유형을 분석하여, 향후 컴퓨터 기반 평가의 개발과 적용에 대한 시사점을 제시하였다는 점에서 연구의 의의를 지닌다.

컴퓨터 기반 평가로의 전환은 디지털 전환 시대를 맞아 더 이상 거스를 수 없는 변화이다. 컴퓨터 기반 평가는 지필평가의 단점을 보완하고 개별 학생의 성취에 대한 정밀한 측정과 즉각적인 피드백을 가능하게 할 수 있다. 그러나 컴퓨터 기반 평가의 타당도와 신뢰도를 높이기 위해서는 학생의 발달 수준과 디지털 기기 친숙도에 맞춘 문항의 유형과 화면 구성에 관한 연구가 필요하다. 학생의 발달 수준과 디지털 기기 친숙도에 맞는 컴퓨터 기반 평가를 통해 학생의 학습 동기를 높이고, 개별 맞춤형 교육이 실현되기를 기대한다.

참고문헌

강태훈(2018). K-CASA 글로벌역량 검사에 대한 차별적 기능문항 분석. **핵심역량교육연구, 2**(2), 1-13.

교육부(2020). 수학·과학 성취도 추이변화 국제비교 연구(TIMSS) 2019 결과 발표.

교육부(2023). 맞춤형 학업성취도 평가로 기초학력 끌어올린다. (2023.12.13. 보도자료). https://www.moe.go.kr.

교육부, 한국교육과정평가원(2022). 2022년 맞춤형 학업성취도 자율평가 시행안.

김명화, 박은아, 최혁준, 김경성(2011). 컴퓨터 기반 문제해결능력 평가 모형 개발. 한국교육과정평가원 연구보고 PRE 2011-5.

김종민, 이문수, 안성훈(2016). 2015년도 국가수준 초·중학생 ICT 리터러시 검사의 성별에 따른 차별기능문항 분석. **교육평가연구, 29**(2), 301-324.

노언경(2007). 중학생 적성검사 중 공간능력 영역에서의 성별에 따른 차별기능문항 추출. 이화여자대학교 석사학위논문.

노언경, 김진호, 김수진(2010). PISA 2003 문제해결력 영역에 대한 성별 차별기능문항 추출. **교육방법연구, 22**(4), 165-194.

박경미, 박선화, 권점례, 윤상혁, 강현영, 이경진, 최지선, 강은주, 김민정, 이광상, 김재영, 이광연, 한준철, 김선희, 방정숙, 이경은, 도종훈, 이문호, 황선미, 임해미, 이화영, 조혜정, 박정숙, 이승훈, 박문화, 김성여, 임미인, 권영기, 서보억, 이은정, 김완일, 장혜원, 이만근, 권오남, 인현정, 이지윤, 강성권, 강태석, 김화경, 신동관, 오택근, 전인태(2015). 2015 개정 수학과 교육과정 시안 개발 연구 II. 교육부 & 한국과학창의재단 연구보고서 BD15120005.

백종호, 이재봉, 구자옥(2023) 컴퓨터 기반 평가와 지필평가 간 학생 응답 특성 탐색 -컴퓨터 기반 국가수준 학업성취도 평가 병행 시행 결과를 중심으로. **한국과학교육학회지, 43**(1), 17-28.

서민희, 김경희, 이재원, 전성균, 김슬비, 민여준(2021). TIMSS 2019 결과 및 변화 추이 심층 분석. 한국교육과정평가원 연구보고 RRE 2021-5.

성태제(1994). 1994학년도 제1차 대학수학능력시험의 성별에 따른 차별기능문항 추출. **교육평가연구, 7**(2), 87-101.

손원숙(2010). 국제학업성취도 평가의 DIF 원인 탐색을 위한 로지스틱 회귀분석의 적용. **교육평가연구, 23**(2), 371-390.

손원숙(2012). 차별기능의 원인파악을 위한 차별기능문항군 기법의 적용. **중등교육연구, 60**(4), 917-935.

이영주(2013). 차별기능문항분석을 통해 살펴본 초등학교 다문화가정 학생의 영어 학업성취 특성.

초등영어교육, 19(1), 169-184.

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 PRE 2020-5.

이태구, 손지영, 양희원(2016). 문항반응이론의 차별기능문항 분석을 활용한 단위학교 차별적 교육과정 탐색. **체육과학연구, 27**(1), 153-168.

조윤동, 강은주, 고효경(2013). 수학과 국가수준 학업성취도 평가 결과를 통한 다문화·탈북 가정 학생 차별기능문항 분석. **수학교육학연구, 23**(2), 75-94.

진수정, 성태제(2004). MH 방법과 SIBTEST 방법을 이용한 문항 유형에 따른 차별기능문항의 탐색. **교육평가연구, 17**(2), 215-235.

최지선(2019). 초등학교 수학 수업에서 계산기 사용에 대한 국가별 비교. **초등수학교육, 22**(1), 83-94.

추정아, 성태제(1993). Mantel-Haenszel 방법과 Raju 방법에 의한 제4차, 제5차 대학수학능력시험 실험평가의 성별에 따른 차별기능문항 추출. **교육평가연구, 6**(2), 259-286.

Anderson, N., Lankshear, C., Timms, C., & Courtney, L. (2008). 'Because it's boring, irrelevant and I don't like computers': Why high school girls avoid professionally-oriented ICT subjects. *Computers & Education, 50*(4), 1304-1318.

Angoff, W. H. (2012). Perspectives on differential item functioning methodology. In *Differential item functioning* (pp. 3-23). Routledge.

Bennett, S., Maton, K., & Kervin, L. (2008). The 'digital natives' debate: A critical review of the evidence. *British Journal of Educational Technology, 39*(5), 775–786.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Sage.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement, 33*(4), 453-464.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization 1, 2. *ETS Research Report Series, 1992*(1), i-40.

Dorans, N. J., & Potenza, M. T. (1994). Equity assessment for polytomously scored items: A taxonomy of procedures for assessing differential item functioning 1. *ETS Research Report Series, 1994*(2), i-33.

Ercikan, K., & Koh, K. (2005). Examining the Construct Comparability of the English and

French Versions of TIMSS. *International Journal of Testing, 5*(1), 23-35.

Holland, P. W. (1985). On the study of differential item performance without IRT. *Proceedings of the military testing association, 1*, 282-287.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). *Lawrence Erlbaum*.

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 5*(2), n2.

Jerrim, J., Micklewright, J., Heine, J. H., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it?. *Oxford Review of Education, 44*(4), 476-493.

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour and Information Technology, 33*(4), 410-422.

Keng, L., McClarty, K. L., & Davis, L. L. (2008). Item-level comparative analysis of online and paper administrations of the Texas assessment of knowledge and skill. *Applied Measurement in Education, 21*(3), 207-226.

Lai, J.-S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning for small sample sizes. *Evaluation and the Health Professions, 28*(3), 283-294.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological measurement, 5*(2), 159-173.

Linn, R. L., & Drasgow, F. (1987). Implications of the Gorden Rule settlement for test construction. *Educational Measurement: Issues and Practice, 6*(2), 13-17.

Mazzeo, J., & von Davier, M. (2014). Linking scales in international large-scale assessments. *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, 229-258.

McDowell, L., & Sambell, K. (1999). The experience of innovative assessment: student perspectives. *Assessment matters in higher education: Choosing and using diverse approaches*, 71-82.

Meelissen, M., and Drent, M. (2008). Gender differences in computer attitudes: does the

school matter?. *Computers in Human Behavior, 24*, 969-985.

Miller, L. A., & Lovler, R. L. (2016). *Foundations of psychological testing: A practical approach.* Sage publications.

Mullis, I. V., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). TIMSS 2019 international results in mathematics and science. *Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls. bc. edu/timss2019/international-results*.

Pommerich, M. (2004). Developing computerized versions of paper tests: Mode effects for passage-based tests. *The Journal of Technology, Learning, and Assessment, 2*(6), 1-44.

Pomplun, M., Frey, S., & Becker, D.F. (2002). The score equivalence of paper and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*(2), 337-354.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items:a framework for classification and evaluation. *Applied Psychological Measurement, 19*(1), 23-37.

Russo, A. (2002). Mixing technology and testing. *The School Administrator, 59*(4), 6-12.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability *criteria. Journal of Educational Statistics, 6*(4), 317.

Steven, L. W. (2006). An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test, *Applied Measurement in Education, 19*(2), 95-114.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Thelwall, M. (2000). Computer-based assessment: a versatile educational tool. *Computers & Education, 34*(1), 37-49.

Thissen, D. (2001). Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. *University of North Carolina at Chapel Hill*.

Vekiri, I., & Chronaki, A. (2008). Gender issues in technology use: perceived social

support, computer self-efficacy and value beliefs, and computer use beyond school. *Computers & Education, 51*(3), 1392-1404.

Volman, M., et al. (2005). New technologies, new differences. Gender and ethnic differences in pupils' use of ICT in primary and secondary education. *Computers & Education, 45*(1), 35-55.

Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing, 6*(3), 287-300.

Yildirim, H. H., & Berberoglu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing, 9*(2), 108-121.

Zumbo, B. (1999). A handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type item scores. *Directorate of Human Resource Research and Evaluation, National Defense*.

Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. *University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science*.

· 논문접수 : 2024.1.5. / 수정본접수 : 2024.1.31. / 게재승인 : 2024.2.7.

ABSTRACT

Analysis of Differential item functioning according to digital device familiarity in computer-based test: focusing on eTIMSS 2019 4th grade mathematics

Seo, Kyungsook
Doctoral Student, Department of Education, Ewha Womans University
Ahn, Haeyeon
Doctoral Candidate, Department of Education, Ewha Womans University
Choi, Younjeng
Associate Professor, Department of Education, Ewha Womans University

This study attempted to explore the types of items that adversely affect students who have relatively insufficient experience in using digital devices through the analysis of differential functional questions in computer-based evaluation. To this end, an analysis of differential functional questions was conducted using data from the 4th grade mathematics of eTIMS 2019. For the analysis, data from 557 people who took regular mathematics and 556 people who took PSI questions were used, and the group with low digital device familiarity was selected as the focul group and the group with high digital device familiarity as the reference group. As a method of analyzing differential functional questions, the results were compared using the logistic regression analysis method, the Mantel-Haenszel method, and the likelihood ratio test method. As a result of the analysis, it was found that students in the group with low digital device familiarity were disadvantageous in terms of items with enhanced computer interaction and items for entering answers with a numeric keypad. The results of this study are expected to be used as information to improve the validity of computer-based evaluation in the future.

Key Words: *Computer based test, differential item functioning, familiarity with digital devices, eTIMSS 2019*