교육과정평가연구
The Journal of Curriculum and Evaluation
2022, Vol. 25, No. 1, pp. 247~267
DOI: https://doi.org/10.29221/jce.2022.25.1.247

국가수준 학업성취도 평가의 시행 매체 변화가 학업성취도에 미치는 영향[1)]

이소라 (한국교육과정평가원, 부연구위원)*
김용남 (서울대학교, 조교수)
구남욱 (한국교육과정평가원, 연구위원)**

요약

본 연구는 고등학교 2학년 학생을 대상으로 시행된 2020년 국가수준 학업성취도 평가에서 수집된 지필평가(paper-based test, 이하 PBT)와 컴퓨터 기반 평가(computer-based test, 이하 CBT) 데이터를 활용하여 시행 매체에 따른 학업성취도를 교과와 문항 유형별로 비교·분석하였다. 먼저, 신뢰도 분석 결과, 국어, 수학, 영어 교과에서 두 시행방식의 안정적인 신뢰도 지수가 산출되었다. 구인동등성 확인 결과, 국어와 영어 교과에서 매체 간의 구인동등성이 의심되는 결과가 나타났다. 매체효과 확인 시 학교수준의 특성을 통제하기 위해 PBT와 CBT가 모두 시행된 학교를 대상으로 다층모형을 적용하여 분석하였고, 모형적합도가 높은 무선절편 모형의 회귀 계수를 해석하였다. 그 결과 국어 교과와 영어 교과의 서답형 문항 유형에서 분명한 매체효과가 발견되었으며 CBT의 점수가 유의하게 더 높았다. 다음으로 성별, 지역 규모, 학교성별유형과 같은 하위집단별 매체효과의 차이를 분석하였다. 그 결과 수학과 영어 교과에서 남학생들이 CBT에서 여학생보다 높은 점수를 얻는 경향이 있었고, 여학생들은 매체에 영향을 받지 않는 것으로 나타났다.

주제어 : 국가수준 학업성취도 평가, 컴퓨터 기반 평가, 매체효과, 다층분석

1) 이 논문은 한국교육과정평가원의 보고서(RRE 2021-1)인 구남욱 외(2021)의 일부를 발췌·수정하여 발표하는 것임.

* 제1저자, slee486@kice.re.kr
** 교신저자, gu9971@kice.re.kr

I. 서 론

국가수준 학업성취도 평가(이하, 학업성취도 평가)는 학생들의 학업성취 수준을 파악하고 그 추이를 분석하여 학교교육의 성과를 국가수준에서 점검함과 동시에 교육정책 수립의 기초자료를 확보하는 것을 목적으로 한다(교육부, 2019.11.30.). 학업성취도 평가에서는 2019년 고등학교, 2020년 중학교 평가에 2015 개정 교육과정을 적용하여 교과별 내용과 역량을 종합적으로 측정하기 위한 새로운 문항을 도입하였다. 이러한 새로운 유형의 문항을 효과적으로 활용하기 위해 지필평가(paper-based tests, 이하 PBT)에서 컴퓨터 기반 평가(computer-based tests, 이하 CBT) 체제의 도입을 준비하고 있다(교육부, 2019.3.29.). 그동안 학생들의 능력을 측정하는 보편적인 방법은 PBT 방식이었으나, PISA, TIMSS, NAEP와 같은 해외 및 국제 학업성취도 평가에 CBT 방식이 점차 확산되고 있어 우리나라의 학업성취도 평가도 이러한 흐름에 발맞춰 2022년 CBT로의 전환을 계획하고 있다. CBT 시행 방식을 도입하게 되면 PBT에서는 구현하기 어려웠던 교과역량 및 학생의 문제해결력을 다양한 방식으로 측정할 수 있고, 그 과정에 대한 정보를 수집하여 개별 학생에 대한 맞춤형 피드백 제공이 가능해질 것이다.

안정적인 컴퓨터 기반 학업성취도 평가로의 전환을 위해서 우선시 되어야 하는 것은 기존의 PBT 점수와 새로운 CBT 점수가 평가 매체 변화에 따라 영향을 받는지, 즉 매체효과를 확인할 필요가 있다. 검사 상황에서 매체효과가 있다는 것의 의미는, 응답 데이터를 수집하는 방법(또는 매체)이 피험자의 응답에 변화를 일으킨다는 것이다. 이것을 알아내기 위한 매체효과 분석은 PBT와 CBT 방식이 각각 측정하는 구인이 동일한지, 즉 시행 방식의 변화가 측정하고자 하는 대상을 변화시키지 않는지, 점수의 의미가 동일하게 해석될 수 있는지 등을 점검하는 것이다. PBT 점수와 CBT 점수의 의미가 동일할 때, 두 점수는 동일한 척도 위에서 비교 가능하고 나아가 해석이 가능해 진다. 학업성취도 평가의 주요 목적 중 하나는 학업성취도의 연도 간 추이를 분석하는 데 있으므로 PBT 점수와 CBT 점수가 비교 가능한 것인지 확인하는 것은 매체 전환의 필수적 단계라고 할 수 있다. ETS, ACT와 같은 해외의 대규모 학업성취도 평가 연구기관에서는 시행방식에 따른 점수의 비교 가능성을 매체효과 분석을 통해 점검하고 있다. 각 기관의 선행 연구에서 해당 지역의 피험자 특성과 검사의 성격에 따라 다양한 결과가 도출되고 있기 때문에, 우리나라의 학업성취도 평가 상황에 맞는 매체효과 점검을 하고 그 결과에 따라 점수 연계 방안을 제시할 필요가 있다. 본 연구에서는 국가수준 학업성취도 평가 데이터를 활용하여 매체효과를 분석해 봄으로써 평가 시행방식 전환에 따른 과목별 검사 점수의 비교 가능성을 점검하고자 한다.

II. 이론적 배경

학업성취도 평가 매체효과 분석에 앞서 선행연구에서 활용한 연구 설계 및 분석 방법을 확인하기 위해 해외 평가기관의 매체효과 연구에 대해 살펴보았다.

1. 미국의 대규모 성취도평가의 매체효과 연구

먼저, 미국의 교육 관련 통계 분석 및 학업성취도 평가 기관인 NCES(National Center for Education Statistics)에서 시행하는 NAEP(national assessment of educational progress)는 PBT로 시행되어 오다가 2017년부터 디지털 기반 평가(digital based assessment, 이하 DBA)를 도입하였다. Jewsbury 외(2020)에 따르면, 2017년에 4학년과 8학년의 읽기, 수학 교과를 대상으로 매체효과를 확인하기 위해 PBT와 DBA를 모두 시행하였고, 당시에 활용한 점수연계(bridge study) 방법을 제시하였다. 우선, 표집을 위해 층화무선표집 방법을 사용하였고, 표집 크기는 PBT 약 29만 명, DBA 약 7만 명이었다. 학교 간 차이를 최소화하기 위해 각 학교에서 PBT와 DBA 응시자를 모두 표집하였다. DBA 문항의 경우, PBT와 동일한 문항 90%와 DBA를 위한 새로운 문항 10%로 구성하였다. PBT 피험자들의 능력 추정의 경우, 직전 시행인 2015년도 문항 응답 데이터와 2017년도 응답 데이터를 동시에 활용하고, 두 시점에서 공통적으로 사용된 공통문항의 모수를 동일하게 제약한 후 능력모수를 추정하였다. DBA 피험자들의 능력 추정은 DBA 응답 데이터만으로 별도로 추정한 뒤, 선형변환식[2]을 이용하여 PBT의 피험자 능력 추정치와 같은 척도에서 산출하여 비교가능 하도록 하였다. 매체효과 분석 방법은, 고전검사이론(classical test theory, 이하 CTT)과 문항반응이론(item response theory, 이하 IRT)에 기반하여 난이도, 변별도 등의 문항특성을 분석하였고, 지역, 성별, 인종, 학년 등 하위 집단별로 매체별 평균 차이인 매체 잔차(mode residual)를 산출한 뒤 t-검정을 통해 매체 간 성취도를 분석하였다. 결과를 요약하면, 문항 수준의 매체효과는 DBA가 PBT보다 유의하게 난이도가 높은 것으로 나타났으며, 선다형 보다는 서답형에서, 8학년보다는 4학년에서 매체 간 차이가 더 크게 나타났다. 이러한 매체효과는 소집단별로 유의한 차이가 나타나지 않아, 시행 방식과 하위 집단 간 상호작용은 존재하지 않았다.

다음으로, 대학과 직업 준비도를 평가하는 PARCC(the partnership for assessment of readiness for college and careers)에서는 2015년에 영어 교과는 3~11학년 학생들을 대상으로, 수학 교과는 3~8학년 학생을 대상으로 PBT와 CBT를 시행하였다. Liu 외(2016)에 따르면 참여 학교가 PBT와 CBT 중 한 가지 방법을 고를 수 있도록 하였다. 시험의 영역별로 30% 이상의 문항들이 매체 간에 동일하도록 설계하였다. 분석 방법의 경우, 앞에서 언급했듯이 참여 학교들이 시행방식을 선택할 수 있었기 때문에 표집이 랜덤하지 않아 피험자 집단을 동질적으로 만들기 위해 각 주별로 학교 수준의 공변인[3]을 이용하여 경향점수매칭 기법을 활용하였다. 이 때, 공변인이 하나의 시행방식에 치중되어있지 않는지 확인하여 모형 투입에 고려하기 위해 Cohen's d 값을 활용하여 효과크기를 산출하였고, 이 값이 0.2 이하인 공변인들만 활용하였다. 이 연구에서는 모든 매칭에서 가장 최적의 거리가 나오는 최적완전매칭(optimal full matching) 기법을 활용하여 동일한 특성의 피험자를 선별하였다. 선별된 피험자를 대상으로, 검사구인 동등성 확인, 차별기능문항(differential item functioning, 이하 DIF) 확인, 문항 모수 비교, 학생 점수 비교, 검사특성 곡선 비교 방법들을 통해 매체효과를 검증

2) $\theta_{최종DBA} = A\theta_{초기DBA} + B$, $(A = \hat{\sigma}_{PBT}/\hat{\sigma}_{초기DBA},\ B = \hat{\mu}_{PBT} - A\hat{\mu}_{초기DBA})$

3) 흑인/히스패닉/백인 학생 비율, 여학생 비율, 빈곤층 학생 비율, 장애 학생 비율, 영어 외 모국어 사용 학생의 비율

하였다. 결과를 보면, 영어, 수학 영역에서 대체적으로 CBT에 비해 PBT 문항의 난이도가 쉽게 나타나며, 매체별로 문항들의 난이도 순서는 차이가 없는 것으로 나타났다. DIF 분석에서는 영어보다는 수학에서 DIF 문항이 많이 나타나며, 특히 학년이 올라갈수록 DIF 문항 수가 늘어나는 것으로 나타나 문항의 난이도가 DIF에 영향을 미칠 수 있을 것으로 유추할 수 있다. 학생들의 성취도를 살펴보면, PBT 학생들의 점수가 CBT에 비해 유의하게 높게 나타났다. PARCC의 평가에 대해 매체효과 분석을 한 또 다른 보고서인 Steedle 외(2016)에 따르면, 경향점수를 이용하여 매칭하기보다는 같은 인구특성을 가진 학생들을 추출해 내는 'CEM(coarsened exact matching)' 기법을 사용하여 문항·검사·학생수준의 매체효과를 분석하였다. 문항수준에서는 정답으로 응답한 학생들의 빈도를 토대로 χ^2-검정을 실시하고, IRT 문항 분석을 통한 문항 모수를 비교하였다. 검사수준에서는 학생들의 점수를 사용하여 영역별, 학년별 효과크기를 산출하고, 매체별 신뢰도와 수렴타당도를 산출하여 비교하였다. 학생수준에서는 매체별 피험자 적합도 지수(person-fit statistics)를 비교하였다. 연구 결과를 살펴보면, 문항수준에서는 수학에서 부분적으로 매체효과가 있었지만 대체적으로 매체효과가 나타나지 않았으며, 검사수준과 학생수준에서도 매체효과를 나타낼만한 근거는 발견되지 않았다.

미국의 대학입학시험인 SAT(scholastic aptitude test)와 예비 SAT시험인 PSAT10(preliminary scholastic aptitude test, 10학년을 대상으로 시행), PSAT8/9(8학년과 9학년을 대상으로 시행)의 매체효과 연구(Proctor et al., 2019)에서는 읽기, 쓰기와 언어, 수학 교과에 대해 CBT와 PBT 결과를 비교하였다. 연구에 지원한 학교 중 CBT 시행이 가능한 환경을 갖춘 학교를 대상으로 실시하였고, 각 학교 내에서 무선적으로 PBT와 CBT에 학생들을 배정하여 표집함으로써 학생들 간 특성이 유사하도록 설계하였다. 분석 방법을 살펴보면, 검사 수준에서는 매체별 성취도 평균 점수와 효과크기를 비교하였고, 문항수준에서는 정답률과 Mantel Haenszel(MH) DIF를 산출하였으며, 하위 집단 수준에서는 집단별 평균 점수를 비교하였다. 결과를 살펴보면, 읽기 영역에서는 CBT의 점수가 유의하게 높게 나타났으며 수학, 쓰기와 언어 교과의 경우 PSAT10에서는 PBT 점수가, SAT와 PSAT8/9에서는 CBT 점수가 더 높은 것으로 나타났으나 효과크기는 0.1 미만으로 매체 간 차이가 크지 않은 것으로 나타났다. 소집단별 비교에서 거의 모든 하위집단에서 CBT의 점수가 PBT보다 평균적으로 높게 나타났다. 문항 수준의 분석 결과도 대체로 CBT 문항의 난이도가 낮았으나, 수학 서답형 문항의 경우 PBT의 난이도가 낮게 나타났다. 매체 간 DIF 결과를 보면, 각 교과에서 1~2개 정도의 매우 적은 문항에서 DIF가 발견되었으며 어느 매체에 유리하게 작용하였는지는 혼재적인 양상을 보였다.

미국대학입학자격시험 중 하나인 ACT(American college testing)에서는 2019년 10월, 12월, 2020년 2월에 시행된 PBT와 CBT의 데이터를 분석하여 매체효과 연구(Steedle, Pashley, & Cho, 2020)를 수행하였다. 학생들은 무작위로 PBT와 CBT에 응시하였고, 선택형 문항으로만 구성된 영어, 수학, 읽기, 과학 교과[4] 시험을 치렀다. 분석방법으로는 문항 수준에서는 정답률 및 MH-DIF 검정, 점수 동등성 비교에서는 KS(Kolmogorov-Smirnov) 검정을 통해 시행방식에 따른 척도점수의 분포에 차이가 있는지, 독립표본 t-검정을 통해 평균 점수 차이가 있는지 등을 확인하였다. 또한, 각 시행방식마다 동백분위 동등화(equipercentile equating) 방법으로 원점수에서 척도점수로 변환한 후, 같은

4) 쓰기 영역은 2월에만 시행 하였으며, 문항은 서답형 유형으로 제시됨.

원점수가 두 방식에서 어떠한 척도점수로 변환되었는지 살펴보았다. 구인동등성 확인을 위해 전체 점수의 분산에 대해 각 영역의 분산이 차지하는 비율을 계산하였고, 시행방식에 따른 확인적 요인분석을 통해 모형적합도 비교도 확인하였다. 그리고 다집단 확인적 요인분석을 활용해 시행방식 간 측정 동등성도 검증하였다. 그 결과, 문항 수준에서는 CBT 문항의 정답률이 더 높게 나타났고, 영어, 읽기의 경우, 검사 후반부에 위치한 문항들에서 매체효과가 더 큰 것을 확인하였다. 무응답 비율은 PBT에서 더 높게 나타났다. 시행방식 간 DIF를 분석한 결과, 영어에서 DIF를 보이는 문항들이 많이 나타났으며 문항 내용과 DIF 결과에는 상관이 없는 것으로 나타났다. 다음으로, 성취도 결과를 살펴보면, PBT보다 CBT에서 학생들의 성취도가 높게 나타났고, 같은 원점수라면 CBT의 척도점수가 PBT보다 더 낮게 나타나 CBT의 정답률이 더 높았음을 알 수 있었다. KS 검증 결과에 따르면 영어, 읽기, 과학에서 시행방식 간 성취도 점수 분포가 다른 것으로 확인되었고 수학 교과는 대체적으로 분포가 비슷한 것으로 드러났다. 다음으로 구인 및 측정 동등성은 두 시행방식 간 충족되는 것을 확인하였고, 특히 수학의 경우 동등성이 강하게 충족되는 것으로 나타났다. 추가적으로 쓰기의 경우, CBT의 성취도가 PBT보다 높은 것으로 나타나 컴퓨터 자판을 이용한 글자 입력 방법이 글쓰기에 유리한 것으로 판단할 수 있었다.

2. 국제 학업성취도 평가의 매체효과 연구

TIMSS(trends in international mathematics and science study)는 CBT인 eTIMSS를 2019년에 도입하였다(Fishbein et al., 2018). 기존의 TIMSS와 eTIMSS의 문항동등성을 확인하기 위해 4학년과 8학년의 수학, 과학 교과에 대해 별도로 학생 표집을 실시하여 시행하였고, 2015년도에 사용된 동일한 문항(선택형과 서답형)을 사용하였다. 연구 설계를 살펴보면, 동일한 특성으로 구성된 8개의 검사 블록을 만들어 모든 학생들이 PBT 방식인 TIMSS의 한 블록과 CBT 방식인 eTIMSS의 한 블록에 대해 응답하여 동일한 학생들이 두 검사 방식에 모두 응시하도록 하였다. 또한 학생들의 50%는 TIMSS, 50%는 eTIMSS부터 시작함으로써 순서효과를 최소화 하였다. TIMSS의 매체효과 연구에서 특이한 사항은 내용 전문가들이 각 문항을 검토하여 매체가 바뀌더라도 동등성이 기대되는 문항 군과 동등성이 기대되지 않는 문항 군으로 분류하여 각 문항 군별로 문항 특성 분석을 실시한 점이다. 성별, 사회경제적지위, 디지털 자기효능감 각각의 수준별로 소집단을 나누고, 소집단과 시행방식 간 반복측정 분산분석으로 상호작용효과를 검증하였다. 또한 매체별 성취도 차이에 영향을 미치는 요인을 찾기 위해 시행방식 간 점수 차이를 종속변인으로 하고, 다양한 공변인을 독립변인으로 통제한 회귀분석을 실시하였다. 문항단위 결과를 보면, 동등성이 기대되는 문항군에서는 예상대로 매체효과가 나타나지 않았고, 동등성이 기대되지 않는 문항군에서는 PBT보다 CBT를 더 어렵게 느끼는 것으로 매체효과가 나타났다. 학업성취도의 경우 PBT의 평균 성적이 더 높고 특히 수학에서 차이가 더 큰 것으로 나타났다. 다만 매체별 학생들의 성취도 순위에는 변동이 없는 것으로 나타나 두 매체에서 측정된 구인이 동일한 것으로 드러났다. 소집단별 분석 결과를 보면, 대체적으로 매체효과가 크지 않은 것으로 보이나 수학에서 집에서 보유한 도서의 수와 시행 방법 간에 상호작용 효과가 통계적으로 유의하게 나왔으며, 과학에서는 성별과 시행 방법 간에 상호작용 효과가 유의한 것으로 나타났다. 그러나 상호작용 효과의 크기는 작았다.

PISA(Programme for International Student Assessment) 2015에서는 대부분의 참여 국가들이 CBT로 전환하였고, Jerrim 외(2018)의 연구에서는 PISA 2015 예비검사를 통해 매체효과를 확인하였다. 연구 설계를 살펴보면, 독일, 스웨덴, 아일랜드의 만 15세 학생들을 PBT, PBT와 동일한 문항의 CBT, 새로운 문항의 CBT 등 세 개의 검사군 중 하나의 시험을 무작위로 치르게 함으로써 각 검사군의 학생들이 동질적인 특성을 갖는다고 가정하였다. 분석 방법으로는 1-모수 IRT 모형을 이용하여 성취도를 추정하였으며, 문항 수준의 분석은 정답률을 비교하였다. 매체효과의 분석을 위해 성취도를 종속 변수로, 시행방식과 성별, 학년 등의 인구특성을 독립 변수로 설정한 회귀모형을 이용하였다. 이때 각 공변인별로 시행방식과의 상호작용 효과를 투입하여 소집단별 매체효과를 확인하였다. IRT 모형의 모형적합도를 활용해 매체효과가 존재하는지, 존재한다면 개인마다 동일하게 또는 다르게 작용하는지 분석하였다. 먼저 성취도 결과를 보면, 학생들은 PBT에서 더 높은 성취도를 나타냈고, 특히 아일랜드 학생들의 경우 읽기 영역에서 여학생들이 CBT를 PBT보다 더 어려워하는 경향이 있는 것으로 나타났다. 사분회귀분석(quantile regression)으로 성취수준별 매체효과를 각각 살펴본 결과, 상·하위권 집단별 매체효과의 차이가 관찰되지는 않았다. 문항 수준 결과를 보면, CBT의 문항들이 더 어려운 것으로 나타났으나, 문항들의 난이도 순서는 CBT와 PBT에서 동일하게 나타났다. 세 가지 IRT 모형의 정보 준거지수인 AIC, BIC를 비교한 결과 학생마다 갖는 매체효과 크기에 차이가 있었으며, 매체효과는 학생들의 사전 컴퓨터 활용 능력에 큰 영향을 받지 않는 것으로 나타났다.

이상의 매체효과 선행 연구의 연구 설계, 방법 및 결과에 관한 공통점을 〈표 1〉에 요약하였다. 본 연구에서는 이 중에서 검사 특성, 성취도 점수, 소집단 수준의 분석에 초점을 맞추었다.

〈표 1〉 매체효과 관련 선행 연구 요약

연구 단계		주요 내용
	연구 설계	• 무선집단 설계, 공통문항 설계, 혼합 설계
분석방법	문항 수준	• 문항의 정답률, 난이도, 변별도 분석 • MH-DIF 분석 • 시행 매체별 문항 난이도 순서를 비교하여 측정동일성 확인
	성취도(점수)	• 확인적 요인분석을 이용한 구인동등성 확인
	소집단 수준	• 인종, 성별, 모국어 등에 따라 소집단을 나누고 하위집단 별로 성취도 평균점수 비교 • 소집단 특성 변수와 시행방식 간 상호작용 효과를 회귀분석 또는 ANOVA 분석에 투입하여 집단별 매체효과 차이 확인
연구결과	문항 수준	• CBT보다 PBT의 정답률이 높은 것으로 나타남 (ACT 결과 예외) • 문항들의 정답률(난이도) 순서는 PBT와 CBT 각각에서 동일하게 나타남
	성취도(점수)	• 대부분의 연구에서 CBT 보다는 PBT에서 성취도가 더 높은 것으로 나타났으나, ACT와 SAT의 연구에서는 CBT의 성취도가 더 높은 것으로 나타나기도 함
	소집단 수준	• 검사 및 피험자 특성에 따라 다양한 양상이 나타남

III. 연구 방법

CBT 전환에 따른 매체효과 분석을 위해 활용한 데이터의 수집 방법 및 연구방법은 다음과 같다. 데이터 수집의 경우, 2020년 학업성취도 평가는 COVID-19에 의한 등교 중지 상황 및 향후 컴퓨터 기반 평가로의 전환에 대비하기 위해 인터넷 기반 검사 서비스를 준비하여 CBT를 부분적으로 시행[5]하였다. 고등학교 2학년의 경우 결시자 및 미응답자를 제외하고 476명의 응답 자료가 수집되었다. CBT는 PBT 표집학교를 대상으로 추가 학급을 요청하여 시행되었으며, 학급 단위로 PBT와 CBT 중 한 곳에 무작위로 배치되어 1회 평가를 실시하였다.

1. PBT-CBT 시행 방식

2020년 국가수준 학업성취도 평가의 시행은 3% 표집 학생을 대상으로 하는 PBT 방식이 원칙이었으나, CBT 방식의 온라인 평가도 부분적으로 병행 시행하였다. CBT 평가 대상은 (1) 평가 시행일에 등교 중지가 발생한 표집학교 중, 평가를 희망하는 학급의 학생들이 가정에서 온라인으로 참여하였고, (2) 표집학교에 추가적으로 CBT 시행을 요청하고 이를 수락한 학교의 CBT에 배정된 학급이 학교에서 참여하였다. CBT도 PBT와 동일한 날짜의 유사한 시간에 시행되었으며, 시험 시간은 PBT와 동일하게 50분으로 제한[6]하였다. 〈표 2〉는 PBT와 CBT 시행 규모를 나타낸다.

〈표 2〉 PBT와 CBT의 시행 규모

PBT			CBT		
학교 수	학생 수	시행 장소	학교 수	학생 수	시행 장소
215	11,324	학교 (215개교)	10	476	가정(3개교) 학교(7개교)

평가 영역은 국어, 수학, 영어의 가형[7]이고, CBT와 PBT의 문항은 매체만 변경하여 100% 동일하게 구성하였다.

2020년 수집된 CBT 데이터는 무선표집 방법에 의해 학교를 선표집하지 않았기 때문에 PBT 성적과 CBT 성적은 매체 선정, 즉 처치 변수와 독립적인 관계로 볼 수 없는 상황이다. 선택편의가 발생하는 경우이므로 통계적인 방법을 통해 CBT와 PBT 집단을 최대한 동질적으로 구성하여 비교할 필요가

5) 2020년 국가수준 학업성취도 평가는 COVID-19 상황에 대비하여, 가정에서 응시할 수 있도록 IBT(internet based tests) 서비스를 제공하였으며, 실제 등교가 불가능한 일부 학교에서 IBT로 응시하였음. 본 연구에서는 PBT와의 대칭적인 의미로 서술하기 위해 IBT를 CBT로 통일하여 서술하였음.

6) 온라인 시행 시스템에서 시간을 제한하였고, 부정행위를 방지하기 위해 시행을 잠시 멈추거나 화면을 내리고 다른 화면을 여는 행위 등은 제한됨.

7) 학업성취도 평가는 '가, 나, 다, 라' 유형으로 구성되어 있어, PBT 방식의 경우 네 가지 유형으로 시행되었으나, CBT의 경우 '가' 유형만 시행됨.

있다. 집단별 동질성이 담보되지 않은 상황에서 매체에 따른 성취도를 비교하려면 주어진 개인 및 학교수준 특성을 이용하여 비슷한 특성을 갖는 표본끼리의 매칭(matching)을 하거나, 개인 및 학교수준 특성들의 영향력을 통제한 후 매체효과 분석을 수행할 필요가 있다.

본 분석에서는 일부 학교에서 PBT와 CBT를 모두 응시한 점을 이용하여, 학교 특성을 완벽히 통제(혹은 매칭)한 뒤, 매체효과를 추정하고자 하였고 7개 고등학교가 이에 해당되었다. 이 때 같은 학교 내에서 CBT와 PBT에 학급이 무선적으로 배정되었는지를 확인하여, 이에 부합하지 않는 1개교를 제외하고 6개 학교를 분석대상으로 선정하였다. 결론적으로 해당 6개 학교는 등교 중지 대상 학교가 아니었기 때문에 가정에서 시행한 경우는 없었으며, PBT 표집 학교에 CBT 추가 시행을 요청한 경우만 포함되었다.

학업성취도 평가의 학생 및 학교장 설문조사를 통해 수집된 〈표 3〉의 개인 및 학교 특성 변인을 활용하여 조건들을 최대한 통제하였다.

〈표 3〉 PBT와 CBT에서 수집된 공통 설문 요소

설문 대상	변인 이름
학생수준	• 검사 시행방식(1=CBT, 0=PBT) • 성별(1=남학생, 0=여학생)
학교수준 (학교장 설문)	• 학교 정보: 지역규모(대도시, 중소도시, 읍면지역), 설립유형(공립, 사립), 학교성별(남녀공학, 남학교, 여학교), 특목고유형(일반고, 과학고, 예술고) • 학교 설문: 학교장 성별/경력, 학교 규모(학생 수), 교육비 지원 대상 학생 수, 다문화 가정 학생 수, 직무연수 정규 시간을 채운 교사 수, 교원학습공동체 참여 교사 수, 교사 특성(수업 질 개선 노력, 생활지도 노력, 학생과의 관계, 교직 만족도 등), 재학생 특성(학업 열의, 타인과의 조화, 교우 관계, 교사와의 관계 등)

2. 매체효과 분석 방법

성취도 점수에 대한 매체의 영향력을 분석하기 위하여 활용된 기본 회귀 모형은 식 (1)과 같다.

$$Y_i = B_0 + B_1 X_i + \sum_{k=2}^{n} B_k (\text{공변인})_{ik} + \epsilon_i \quad (1)$$

i는 학생, Y_i는 학생 i의 성취도 원점수, X_i는 시행방식 더미 변인(0=PBT, 1=CBT)이며, B_1은 매체효과, B_k는 공변인 k의 효과이다. 다음으로, 학교별 평균 차이를 이용하여 자료의 중층구조를 반영하기 위한 개인-학교수준의 2수준 다층모형은 식 (2)와 같다. 학생들이 학교에 소속되어 있어 학교수준 특성에 영향을 받기 때문에, 개인뿐만 아니라 학교 간 차이를 통제하여 매체효과의 크기가 적절하

게 추정될 수 있도록 모형을 선정하였다.

$$\begin{aligned} &1\text{수준: } Y_{ij} = B_{0j} + B_{1j}(\text{시행방식}) + B_{2j}(\text{성별}) + \cdots + \epsilon_{ij} \\ &2\text{수준: } B_{0j} = \gamma_{00} + u_{0j} \\ &\qquad\quad B_{1j} = \gamma_{10} + u_{1j} \end{aligned} \tag{2}$$

Y_{ij}는 j 학교에 다니는 i 학생의 성취도 원점수, B_{0j}는 절편(j 학교의 평균 성취도 원점수), B_{1j}는 학생수준의 시행방식이 성취도에 미치는 영향이며, γ_{00}는 전체 학교의 평균 성취도, u_{0j}는 성취도의 학교 간 차이, γ_{10}은 전체 학교의 평균 매체효과, u_{1j}는 매체효과의 학교 간 차이로 해석할 수 있다. 식 (2)에 대해서 각 학교에서 매체효과가 모두 동일할 것이라고 가정하는 무선절편(random intercept) 모형과 학교마다 매체효과의 차이를 고려한 무선기울기(random slope) 모형을 사용하였으며, 모형 비교를 통해 적합도가 높은 모형의 회귀계수를 중점적으로 해석하였다. 1수준 모형에 투입한 공변인은 두 시행방식에 대해 공통적으로 수집[8]된 변인들인 〈표 3〉과 동일하다.

IV. 연구 결과

1. 검사 특성

먼저, 전체 데이터를 대상으로 매체별 평균 점수를 산출하였다. 〈표 4〉의 결과를 보면, 세 과목 모두에서 PBT의 평균 점수가 약간 높은 것으로 나타났다.

〈표 4〉 2020 학업성취도 평가 점수의 기술통계

교과	CBT					PBT				
	사례 수	평균	표준편차	최솟값	최댓값	사례 수	평균	표준편차	최솟값	최댓값
국어	473	22.16	8.11	0	33	11324	22.17	7.48	0	33
수학	473	13.02	7.29	0	27	11324	13.97	7.73	0	27
영어	471	17.32	9.21	1	32	11324	17.97	9.03	0	32

8) CBT로 응시한 학생들에게는 학생 수준의 설문조사를 실시하지 못하였고, 학교 수준의 설문조사만 실시되어 학생 수준의 변인 활용에 한계가 있음.

각 매체의 검사 신뢰도를 확인하기 위해 각 검사의 Cronbach's alpha 값을 살펴보았다. 〈표 5〉의 결과를 보면 PBT와 CBT 모두 높은 신뢰도를 보였으며 수학은 PBT, 국어와 영어는 CBT의 신뢰도가 더 높은 것으로 확인되었다.

〈표 5〉 2020 학업성취도 평가 매체별 신뢰도(Cronbach's alpha)

교과	CBT	PBT
국어	0.869	0.827
수학	0.899	0.901
영어	0.878	0.873

또한 CBT와 PBT의 구인동등성(construct equivalency)과 차원성(dimensionality)을 검증하기 위해 탐색적 요인분석을 수행하였다. 일요인 모형과 이요인 모형을 비교하여 매체 간 내재적 구조의 일관성 및 동일한 구인을 측정하고 있는지 살펴보았다. 분석 시 선다형 문항만을 사용하여 문항 유형에 따른 차원성 간섭을 최소화 하였다. 고유치(eigenvalue)를 이용해 스크리 도표(scree plot)를 산출한 결과, 요인의 개수가 최대 2개가 적합하여, 〈표 6〉에 매체별 일요인 및 이요인 모형의 적합도를 제시하였으며, 두 적합도의 차이도 제시하였다.

〈표 6〉 일요인 모형과 이요인 모형의 적합도

		CBT			PBT		
		일요인	이요인	차이값	일요인	이요인	차이값
국어	CFI	0.895	0.970	−0.075	0.795	0.961	−0.166
	TLI	0.871	0.954	−0.083	0.749	0.940	−0.191
	RMSEA	0.060	0.036	0.024	0.088	0.043	0.045
	SRMR	0.050	0.035	0.015	0.080	0.030	0.050
수학	CFI	0.917	0.974	−0.057	0.927	0.963	−0.036
	TLI	0.904	0.964	−0.060	0.914	0.948	−0.034
	RMSEA	0.057	0.035	0.022	0.051	0.039	0.012
	SRMR	0.049	0.033	0.016	0.037	0.025	0.012
영어	CFI	0.910	0.979	−0.069	0.852	0.940	−0.088
	TLI	0.897	0.971	−0.074	0.830	0.919	−0.089
	RMSEA	0.067	0.035	0.032	0.080	0.055	0.025
	SRMR	0.054	0.030	0.024	0.060	0.037	0.023

※ 음영 표시: 모형 적합도가 낮은 경우

일차원성 검증 결과를 살펴보면, 수학은 일요인 모형의 모형 적합도가 적합한 수준으로 나타나 검사의 일차원성을 만족하는 것으로 나타났으나, 국어와 영어교과의 경우 PBT에서 일요인 모형으로는 모형이 적합하지 않은 것(〈표 6〉 음영 표시 부분)으로 나타났다. 다음으로, 매체 간 구인동등성 확인을 위

해 일요인과 이요인 모형의 적합도 지수의 차이를 PBT와 CBT 간에 비교해 보면, 국어에서 매체 간 차이가 가장 크게 나타났다. 따라서 국어 교과에서 매체 간 차원성과 구인 동등성이 가장 다를 것으로 추정된다.

2. 매체효과

학교 특성을 통제하기 위해 같은 학교 내에서 PBT와 CBT를 동시에 시행하면서도 학급 표집에 특별한 의도가 없었던 6개교의 2학년 학생들이 대상으로, 전체 학생 수는 1,151명이었으며 1,009명은 PBT, 142명은 CBT에 응시하였다. 매체효과는 교과 및 문항 유형별로 분석하였다. 동일한 학교에서 데이터가 수집되었으므로 수집된 측정변인 뿐만 아니라 잠재적인 특성들도 같은 값을 가지는 것으로 가정할 수 있기 때문에 학교수준에서 매칭 된 경우라고 해석할 수 있다. 학교 내에서 PBT와 CBT 학급을 무작위로 선정하였기 때문에 학생수준의 선택 편의도 거의 존재하지 않는다고 볼 수 있다. 그러므로 본 결과에서 도출된 매체효과는 타당하게 해석할 수 있을 것이다. 〈표 7〉은 각 교과의 성취도 원점수의 기술통계치를 보여준다.

〈표 7〉 국어, 수학, 영어 교과의 원점수 기술통계치

		국어			수학			영어		
		평균	표준편차	범위	평균	표준편차	범위	평균	표준편차	범위
선다형	CBT	8.51	1.91	2-12	9.47	3.45	2-15	11.72	3.31	2-16
	PBT	8.40	2.23	0-12	8.50	3.66	0-15	11.02	3.80	0-16
서답형	CBT	16.68	3.36	1-21	4.57	3.18	0-12	7.49	4.42	0-16
	PBT	13.71	5.02	0-21	4.58	3.81		6.36	4.92	

성별에 따른 학생 수(비율)는 여학생 769명(66.8%), 남학생 382명(33.2%)이고, 지역규모에 따른 학생 수(비율)은 대도시 166명(14.4%), 읍면지역 554명(48.1%), 중소도시 431명(37.4%)이며, 학교 성별에 따른 학생 수(비율)는 남녀공학 400명(34.8%), 남학교 167명(14.5%), 여학교 584명(50.7%), 학교설립유형에 따라 국공립 720명(62.6%), 사립 431명(37.4%)로 나타났다. 6개교 모두 목적유형에 따른 분류에서는 일반 고등학교에 해당되었다.

각 교과에서 종속 변수는 선다형과 서답형 문항 유형 각각에 대한 원점수를 사용하였다. 학생 및 학교 수준 모두 공변인을 투입하지 않은 기본 다층 모형을 분석하여 국어, 수학, 영어 성취도의 전체 분산 중 학교 간 분산이 차지하는 비율을 확인하기 위해 ICC(intraclass correlation coefficient)를 산출하였다. 국어 선다형의 경우 0.012(1.2%), 국어 서답형의 경우 0.013(1.3%)로 나타났다. 수학 선다형은 0.139(13.9%), 수학 서답형은 0.098(9.8%)이며, 영어 선다형은 0.022(2.2%), 영어 서답형은 0.026(2.6%)으로 나타났다. 즉, 국어, 영어성취도의 차이는 수학에 비해 상대적으로 학교 간 차이가 작은 것을 확인했으나, 본 연구의 목적이 혼란 변수(confounding variable)를 최대한 통제한 후 매체효과를 찾는데 있으므로 학교 간 차이를 통제하기 위해 회귀모형과 더불어 다층회귀모형도 함께 분

석하였다.

먼저, 국어 선다형 점수에서 위계 관계에 있는 두 다층모형의 적합도를 비교하기 위한 χ^2-차이 검정 결과, 무선기울기 모형보다 단순한 무선절편 모형이 더 적합한 것으로 나타났다($\Delta\chi^2$=4.373, Δdf=2, p=.112). 이 때 〈표 8〉에 나타난 매체효과 추정치는 0.136으로 CBT 점수가 PBT에 비해 약간 높았으나 유의미한 차이는 아니었다. 국어 서답형 점수에서도 무선절편 모형의 적합도가 좋았으며($\Delta\chi^2$=4.592, Δdf=2, p=.101), 매체효과 추정치는 2.952로 CBT 점수가 PBT에 비해 유의미하게 높게 나타났다. 분석에 활용한 모든 모형에서 국어 서답형의 유의미한 결과가 나타나, CBT 방식이 학생들에게 유리한 경향이 있는 것으로 분명하게 드러났다. 이는 ACT의 선행연구에서 쓰기 교과의 CBT 점수가 PBT 보다 유의하게 높은 것으로 나온 결과와 일치하고, NAEP의 선다형보다 서답형에서 매체효과가 더 크게 나타난 결과와도 일치한다. 서답형 쓰기 수행에 있어 컴퓨터를 활용한 응답 입력 방식이 유리할 수 있음을 시사한다.

〈표 8〉 고등학교 국어 점수 분석 결과

			기본 회귀 모형		무선절편 모형	
			회귀계수	**표준오차**	**회귀계수**	**표준오차**
선다형	고정효과(Fixed Effects)					
	학생 수준	절편	9.287***	0.289	9.263***	0.355
		CBT(시행)	0.129	0.194	0.136	0.193
		남학생(성별)	−1.265***	0.216	−1.257***	0.215
		읍면지역(지역규모)	−0.136	0.242	−0.136	0.304
		중소도시(지역규모)	0.425*	0.214	0.425	0.283
		남학교(성별유형)	−0.278	0.312	−0.263	0.397
		여학교(성별유형)	−1.029***	0.235	−1.006***	0.285
	무선효과(Random Effects)					
			분산성분	**표준편차**	**분산성분**	**표준편차**
	1수준 잔차				4.602	2.145
	2수준 절편(u_{0j})				0.017	0.132
			회귀계수	**표준오차**	**회귀계수**	**표준오차**
서답형	고정효과(Fixed Effects)					
	학생 수준	절편	15.287***	0.631	15.229***	0.800
		CBT(시행)	2.942***	0.423	2.952***	0.421
		남학생(성별)	−3.802***	0.472	−3.784***	0.470
		읍면지역(지역규모)	0.602	0.528	0.601	0.687
		중소도시(지역규모)	0.922*	0.467	0.922	0.642
		남학교(성별유형)	1.990**	0.682	2.029*	0.897
		여학교(성별유형)	−2.438***	0.514	−2.380***	0.640
	무선효과(Random Effects)					
			분산성분	**표준편차**	**분산성분**	**표준편차**
	1수준 잔차				21.953	4.685
	2수준 절편(u_{0j})				0.098	0.314

*p<.05 **p<.01 ***p<.001

다음으로, 수학의 선다형($\Delta\chi^2$=4.828, Δdf=2, p=.089)과 서답형($\Delta\chi^2$=4.968, Δdf=2, p=.083) 모두에서 무선절편 모형의 적합도가 높았으며, 결과를 〈표 9〉에서 살펴보면, 선다형 매체효과 추정치는 0.963으로 CBT 점수가 PBT에 비해 유의미하게 높은 것으로 나타났다. 그러나 이러한 결과는 무선기울기 모형에서는 유의하지 않은 것으로 나타나, 모형의 종류나 변수 투입 방식에 따라 결과가 달라질 가능성이 있다. 수학 서답형의 경우, 모든 모형에서 유의미한 매체효과가 없었다.

국제 학업성취도 선행연구에서도 수학 교과의 뚜렷한 매체효과가 나타나기는 했으나, PBT의 점수가 더 높은 것으로 나타났다. 학업성취도 평가의 분석 결과는 반대의 양상이 관찰되었기 때문에 우리나라 학생들이 국제 평균과 비교하여 컴퓨터 활용에 더 익숙할 수 있음을 시사한다.

〈표 9〉 고등학교 수학 점수 분석 결과

			기본 회귀 모형		무선절편 모형	
			회귀계수	표준오차	회귀계수	표준오차
	고정효과(Fixed Effects)					
선다형	학생 수준	절편	9.058***	0.465	8.889***	1.136
		CBT(시행)	0.932**	0.312	0.963**	0.305
		남학생(성별)	-0.309	0.348	-0.256	0.341
		읍면지역(지역규모)	0.362	0.389	0.360	1.005
		중소도시(지역규모)	0.277	0.344	0.277	0.990
		남학교(성별유형)	1.239*	0.503	1.351	1.326
		여학교(성별유형)	-1.798***	0.379	-1.632	0.887
	무선효과(Random Effects)					
			분산성분	표준편차	분산성분	표준편차
		1수준 잔차			11.546	3.398
		2수준 절편(u_{0j})			0.433	0.658
			회귀계수	표준오차	회귀계수	표준오차
	고정효과(Fixed Effects)					
서답형	학생 수준	절편	5.956***	0.481	5.873***	0.756
		CBT(시행)	-0.021	0.322	-0.005	0.320
		남학생(성별)	-0.567	0.360	-0.541	0.357
		읍면지역(지역규모)	-0.321	0.402	-0.322	0.658
		중소도시(지역규모)	0.162	0.356	0.162	0.631
		남학교(성별유형)	0.649	0.520	0.705	0.863
		여학교(성별유형)	-2.340***	0.392	-2.258***	0.598
	무선효과(Random Effects)					
			분산성분	표준편차	분산성분	표준편차
		1수준 잔차			12.683	3.561
		2수준 절편(u_{0j})			0.137	0.370

*p<.05 **p<.01 ***p<.001

다음으로 영어 교과의 선다형 점수에서 무선절편 모형의 적합도가 높았으며($\Delta\chi^2$=4.698, Δdf=2, p=.095), 이 때 결과를 〈표 10〉에서 살펴보면, 매체효과 추정치는 0.69로 CBT 점수가 PBT에 비해 유의미하게 높은 것으로 나타났다. 그러나 이 결과는 무선기울기 모형에서는 유의하지 않은 것으로 나타나 모형의 종류나 변수 투입 방법에 따라 결과가 달라질 수 있을 것으로 추정된다. 영어 서답형에서도 무선절편 모형이 더 적합한 것으로 나타났으며($\Delta\chi^2$=4.829, Δdf=2, p=.089), 매체효과 추정치는 1.133으로서 CBT에서 유의하게 높은 점수를 받는 것으로 나타났다.

〈표 10〉 고등학교 영어 점수 분석 결과

			기본 회귀 모형		무선절편 모형	
			회귀계수	표준오차	회귀계수	표준오차
선다형	고정효과(Fixed Effects)					
	학생 수준	절편	12.273***	0.491	12.212***	0.681
		CBT(시행)	0.678*	0.329	0.690*	0.327
		남학생(성별)	−2.316***	0.367	−2.299***	0.365
		읍면지역(지역규모)	0.341	0.410	0.341	0.588
		중소도시(지역규모)	0.629	0.363	0.629	0.557
		남학교(성별유형)	1.002	0.530	1.043	0.770
		여학교(성별유형)	−2.024***	0.400	−1.963***	0.542
	무선효과(Random Effects)					
			분산성분	표준편차	분산성분	표준편차
	1수준 잔차				13.244	3.639
	2수준 절편(u_{0j})				0.090	0.300
			회귀계수	표준오차	회귀계수	표준오차
서답형	고정효과(Fixed Effects)					
	학생 수준	절편	7.465***	0.644	7.363***	0.975
		CBT(시행)	1.114**	0.432	1.133**	0.428
		남학생(성별)	−1.880***	0.482	−1.848***	0.478
		읍면지역(지역규모)	0.266	0.539	0.265	0.848
		중소도시(지역규모)	0.628	0.476	0.628	0.810
		남학교(성별유형)	1.114	0.696	1.182	1.111
		여학교(성별유형)	−1.977***	0.525	−1.877*	0.772
	무선효과(Random Effects)					
			분산성분	표준편차	분산성분	표준편차
	1수준 잔차				22.752	4.770
	2수준 절편(u_{0j})				0.217	0.465

*p<.05 **p<.01 ***p<.001

다층분석 결과의 검증을 위해 학교수준의 회귀분석을 추가적으로 수행하였다. 앞서 제시된 결과들은 분석 단위가 학생이었으나, 학교수준 분석[9]은 학교별 집계 데이터(aggregated data)를 이용하여 학교별 평균 점수를 구하고, 이 값을 종속 변수로 활용하여 CBT와 PBT 시행 학교 간에 차이가 있는지 분석하는 방법이다. 이러한 분석 방법을 활용하는 것의 장점은 앞선 다층 모형에서 활용될 수 없었던 학교가 추가적으로 분석에 포함될 수 있다는 것이다. 다층분석에서는 CBT와 PBT를 모두 시행한 6개교만을 분석 대상으로 함으로써, CBT만을 응시한 학교가 제외되었다. 즉, 3개교가 CBT를 시행하였음에도 불구하고 다층분석에서 제외되었으나 학교 단위 분석을 위해 포함[10]하였다. 이 분석에서 학생특성 변수는 활용되지 않았으며 회귀 모형에서 학교 특성만 통제 변수로 활용하였다. 절편과 매체효과 회귀계수만 각 교과와 문항 유형별로 정리하면 〈표 11〉과 같다.

〈표 11〉 고등학교 국어, 수학, 영어 교과의 학교수준 매체효과 분석 결과

종속 변수		국어		수학		영어	
		회귀계수	표준오차	회귀계수	표준오차	회귀계수	표준오차
선다형	절편	7.172***	0.748	6.294***	1.448	9.555***	1.427
	CBT(시행)	0.193	0.323	1.200	0.625	1.016	0.615
서답형	절편	10.751***	1.899	1.915***	1.573	4.897**	1.717
	CBT(시행)	2.866***	0.819	0.531	0.678	1.183	0.741

*p<.05 **p<.01 ***p<.001

학교수준 분석에서는 국어 서답형의 매체효과만 유의하였고, 앞선 결과와 마찬가지로 CBT에서 더 높은 점수를 받는 것으로 드러났다. 수학 선다형과 영어 서답형의 경우, 통계적으로 유의한 결과가 나타나지는 않았으나 회귀계수가 큰 편으로 나타나 앞선 결과와 일치하는 경향성을 확인하였다. 이러한 일관적인 경향을 통해 앞선 다층모형 분석 결과의 신뢰성을 확인하였다.

3. 하위집단별 매체효과

소집단별로 매체효과에 차이가 있는지 확인하기 위해 회귀모형에 시행방식 변수와 하위 집단 변수 간 상호작용 효과를 투입하여 분석하였다. 즉, 식 (1)에 상호작용 효과인 '$cbt \times$ 성별', '$cbt \times$ 지역규모', '$cbt \times$ 학교성별' 항을 각각 추가하여 각 교과의 선다형 및 서답형 유형에 대해 분석하였고, 그 중 유의

9) 학교수준 분석은 학교별 표본 수를 반영하지 못하는 등 불확실성이 있기 때문에 일반적으로 잘 활용되지 않는 방법임. 주로 학생, 학교수준 특성을 모두 반영하는 다층분석을 실시하지만, 준비된 데이터를 최대한 활용하여 매체효과를 확인하기 위해 추가적으로 분석함.

10) 학교별로 PBT 또는 CBT 응시학교 둘 중 하나로 구분해야 하므로, 한 학교에서 일부라도 CBT를 시행한 경우는 CBT 시행으로 분류하여 분석함. 즉, 두 매체 모두 시행한 학교에서 PBT를 시행한 경우는 학교수준 분석에서 제외되었음.

미한 상호작용 효과가 나타난 결과만을 제시하였다.

먼저, 수학 교과의 선다형과 서답형 모두에서 학교성별 유형의 하위집단에 따라 매체효과에 차이가 있는 것으로 나타났다. 구체적으로 선다형 문항 점수의 경우, 남녀공학(β_{cbt}=0.420, p<0.001)과 남학교(β_{cbt}=0.806, p<0.01) 집단에서는 CBT 점수가 PBT보다 유의하게 높은 것으로 나타나지만, 여학교(β_{cbt}=0.445, p=0.715)에서는 매체효과가 없었다. 이를 통해 여학생들은 CBT 전환에 따른 수학 선다형 점수의 변화가 없는 것으로 확인되었다. 수학 서답형 점수의 경우, 남녀공학(β_{cbt}=1.201, p<0.01)에서 CBT 점수가 PBT보다 유의하게 높은 것으로 나타나지만, 남학교(β_{cbt}=0.348, p=0.676)에서는 CBT 점수가 높지만 유의하지는 않았으며, 여학교(β_{cbt}=-0.790 p=0.100)에서는 유의미한 매체효과는 없었지만 PBT 점수가 더 높은 것으로 나타났다. 즉, 여학생들은 수학 서답형 문항에 대해 다른 두 학교 유형과는 반대로 CBT에서 더 어렵게 느끼는 경향이 있는 것으로 보인다.

다음으로 영어 교과의 선다형과 서답형 문항군의 점수에서 모두 성별의 하위집단에 따라 매체효과가 다르게 나타나는 것으로 나타났다. 구체적으로 선다형 점수의 경우, 남학생(β_{cbt}=1.041, p<0.05)은 CBT 점수가 PBT보다 유의하게 높은 것으로 나타나지만, 여학생(β_{cbt}=0.489, p=0.170)은 매체효과가 없는 것으로 나타났다. 즉, 여학생들은 CBT 시행에 따른 영어 선다형 점수의 변화가 없는 것으로 확인되었다. 영어 서답형 점수의 경우, 남학생(β_{cbt}=1.858, p<0.01)은 CBT 점수가 PBT보다 유의하게 높은 것으로 나타났으나, 여학생(β_{cbt}=0.586, p=0.220)은 매체효과가 없는 것으로 확인되었다. 즉, 여학생들은 CBT 시행에 따른 영어 서답형 점수의 변화가 없는 것으로 확인되었다. 수학, 영어 교과의 결과를 통해 남학생들이 CBT 전환에 따라 점수 획득에 유리한 측면이 있을 것으로 추정된다.

V. 결론 및 논의

본 연구는 2020년 11월에 시행된 국가수준 학업성취도 평가를 통해 수집된 고등학교 2학년 학생들의 PBT와 CBT 응답데이터를 활용하여 학업성취도 평가의 매체효과를 분석하였다. 실험 설계에 의한 무선표집 데이터가 아닌 PBT 표집대상 학교 중에서 추가로 CBT 시행을 요청하여 수락한 학교에서 수집된 데이터이므로 단순 평균 점수 및 정답률 비교 방법은 적합하지 않아 회귀모형과 다층회귀모형으로 분석하였다. 주요 결과와 그에 따른 학업성취도 평가의 시행 방향을 제안하면 다음과 같다.

결과에서 뚜렷한 매체효과가 관찰되는 것은 국어와 영어 교과의 서답형 유형으로, CBT 점수가 높은 것으로 관찰되었다. 특히 국어 교과의 매체효과가 두드러진 것으로 나타났는데, 국어 서답형 문항에서 매체효과를 보인 결과에 대해서 세 가지 이유를 추정해 보았다. 첫째, 국어는 다른 교과들에 비해 서답형 문항의 수가 많고 배점 비율이 높아서 더 많은 변량이 제공되었을 수 있다. 둘째로 국어 서답형 마지막 문항은 여러 개의 문장으로 문단을 완성해야 하는 축소된 논술형에 가깝기 때문에, 동일 시간 내에 필기하는 것보다 컴퓨터로 타이핑(typing)하는 것이 학생들에게 더 많은 문장을 기술할 수 있도

록 하고 수정과 재구조화가 가능하게 함으로써 인지부하를 낮춰 답안 작성에 용이하게 작용했을 가능성이 있다. CESE(2021)의 연구에서는 쓰기 활동에서 효과적인 결과를 내기 위해서는 학생들의 키보드 입력 방식 연습을 통해 쓰는 행위 자체를 자동화 하여 인지 부하를 줄일 필요가 있다고 주장하였다. 셋째로 평정자 효과(rater effects)가 개입했을 가능성이 있다. 학업성취도 평가의 서답형 문항은 교사 2명이 팀을 이루어 수동으로 채점하기 때문에 수기로 글씨를 써서 제출한 답안에서는 글씨체 특성에 따라 평정자가 응답자의 의도를 정확하게 파악하기 어려웠거나, 반대로 보다 깔끔하게 타이핑된 답안에 대해 더 후한 점수를 부여했을 가능성이 존재한다(Canz, Hoffmann, & Kania, 2020). 국어나 영어와 같은 언어 영역 서답형 유형에서 매체효과가 나타나는 명확한 원인에 대해서는 후속 연구를 통해서 확인할 필요가 있다.

서답형 문항 결과와 관련하여 향후 학업성취도 평가 시행에서 고려해야 할 점을 제시하면 다음과 같다. 첫째, 적어도 PBT에서 CBT로 전환되는 시점, 즉 매체 간 점수 연계가 필요한 시점에서는 국어나 영어의 서답형 문항의 개수, 배점, 유형 및 답안의 분량 등 검사 조건을 비슷하게 유지할 필요가 있다. 서답형 점수가 CBT에서 높아지는 원인이 무엇이든 간에 두 점수를 연계하는데는 방해 요인으로 작용할 수 있기 때문에 과도기 시점 PBT와 CBT의 검사 타당도를 위해 시행 방식을 비슷하도록 제한해야 한다. 학업성취도 평가가 PBT에서 CBT로 전환되는 궁극적인 목표가 학습의 과정으로서의 평가를 확대하고, 학생들의 문제해결력 및 역량을 다양하고 효율적으로 측정하기 위한 것임에는 이견이 없을 것이다. 이를 실현하기 위해 서답형(또는 구성형) 유형 문항의 비율이나 배점이 확대되는 방향으로 평가가 변화되고 있다. 그러나 서답형 문항의 매체효과가 분명하게 드러난 이상, 과도기에 점진적인 변화를 통해 점수 연계가 안정적으로 될 수 있도록 지침을 유지해야 할 것이다. 둘째, 서답형 문항 유형에서 CBT 점수가 향상된 것이 학생들의 글쓰기 능력에 자유도가 높아졌기 때문인지, 교사들의 점수가 후해졌기 때문인지 현재로서는 알 수 없으나, 향후 학생들의 글쓰기 능력을 평가하기 위해서는 지필보다는 컴퓨터를 이용한 방식이 정착되어야 할 것으로 보인다. 현재 초·중·고등학생들은 Z세대[11]로서 디지털 또는 모바일 네이티브(digital/mobile native)로 어려서부터 인터넷 환경에서 활용 가능한 디지털 기기(스마트폰, 태블릿 PC 등)를 통해 즉각적인 정보와 소셜미디어, 가상현실, AI를 접하며 성장한 특징이 있다. 가정환경에 따라 이러한 디지털 접근성에 차이가 나지 않도록 우리나라는 꾸준한 교육정보화 정책을 통해 디지털 리터러시의 격차를 줄이기 위해 노력하였다. 그 결과 우리나라 학생들의 컴퓨팅 사고력은 ICILS 2018 참여국 중 가장 높게 나타났으며, 컴퓨터·정보소양 능력도 2위로 나타나(박상욱 외, 2019) 디지털 기기를 활용한 학습 및 사고에 익숙한 것으로 확인되고 있다. 이처럼 디지털 기술의 진보와 혁신을 자연스럽게 경험하여 익힌 세대의 학생들이 디지털 환경에서 최적화된 유연한 사고를 통해 최대한의 실력을 발휘할 수 있도록 평가 방식을 전환해야 할 것이다.

다음으로 하위집단별로 분석한 결과를 보면, 수학과 영어 교과에서 각각 학교성별유형과 성별에 따라 매체효과에 차이가 있었다. 즉, 여학생은 시행 매체가 바뀌더라도 일관적인 점수가 산출되나 남학생의 경우 CBT로 검사를 수행했을 때 점수가 더 높아지는 경향이 있어 남학생들이 컴퓨터를 활용한 평가에 더 유리한 것으로 드러났다. 이 결과는 국제학업성취도평가 결과에서 보인 우리나라 학생들의

11) 국가마다 다르긴 하나, 대체로 1997년생에서 2012년 사이에 태어난 세대를 칭함.

특성(OECD, 2009)과도 일치하는 결과이다. 학업성취도 평가가 CBT로의 전면 전환을 앞두고 있는 만큼 학생들이 매체 변화에 적응할 수 있도록 상세한 시행안내서 및 연습 프로그램 등을 보급하여 여학생들이 실력과 상관없이 불리한 상황에 몰리지 않도록 지원해야 할 것이다.

본 연구의 제한점은 다음과 같다. 본 연구에서 활용된 데이터는 IRT 문항 분석을 위한 충분한 표집 크기가 확보되지 못했기 때문에 문항 단위의 분석에 한계[12)]가 있었다. 따라서 개별 문항에 대한 특성 분석이나 세부 영역에 대한 해석을 포함하지 못하였다. 향후 후속연구에 활용될 2021년 국가수준 학업성취도 평가는 PBT와 CBT 간 공통문항을 활용한 무선 표집 설계에 의해 IRT 분석에 충분한 규모의 데이터를 수집하였다. 체계적인 데이터를 활용한 후속연구에서는 매체 간 차별기능 문항 분석을 비롯한 문항 단위의 분석 결과를 활용하여 문항 및 영역에 대한 좀 더 구체적인 해석이 가능할 것이며, 동시에 향후 PBT와 CBT 점수 연계를 위한 동등화 과정에서 활용될 문항들을 선별하는 기준으로 그 결과를 활용할 수 있을 것이다.

12) 본 연구에서는 데이터 표집 방법과 크기의 한계로 인해 IRT 기반 문항 분석과 문항 수준의 매체효과 분석 결과를 보고하지 못했으나, 각 문항별 정답률, 변별도 등 고전검사이론 기반의 문항 특성 비교 분석 및 매체 간 차이가 큰 문항들의 내용 특성 탐색은 각 교과 전문가에 의해 질적 분석 연구로 진행되었음. 본 연구와는 개별적으로 연구 결과가 보고될 예정임.

참고문헌

교육부(2019.3.29). **한 아이도 놓치지 않고 기초학력 책임진다.** 교육부 보도자료.

교육부(2019.11.30.). **2019년 국가수준 학업성취도 평가 결과 발표.** 교육부 보도자료.

구남욱, 김미림, 이소라, 곽민호, 한경택, 김동호, 김지혜(2021). **국가수준 학업성취도 평가의 컴퓨터 기반 평가 체제 구축 방안.** 한국교육과정평가원. 연구보고 RRE 2021-1.

박상욱, 김현경, 상경아, 전성균, 최인선(2019). **국제 컴퓨터·정보 소양 연구: ICILS 2018 결과 분석.** 한국교육과정평가원. 연구보고 RRE 2019-9.

Canz, T., Hoffmann, L., & Kania, R. (2020). Presentation-mode effects in large-scale writing assessments. *Assessing Writing, 45,* 100470.

Centre for Education Statistics and Evaluation (2021), *Are writing scores from online writing tests for primary students comparable to those from paper tests?* NSW Department of Education, education.nsw.gov.au/cese.

Fishbein B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education, 6*(11). Retrieved from https://largescaleassessmentsineducation.

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.

Jewsbury, P., Finnegan, R., Xi, N., Jia, Y., Rust, K., Burg, S., Donahue, P., Mazzeo, J., Cramer, E. B., Lin, A., & Weil, N. (2020). *2017 NAEP Transition to Digitally Based Assessments in Mathematics and Reading at Grades 4 and 8: Mode Evaluation Study.* Retrieved from https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional_whitepaper.pdf (검색일: 2021.5.26)

Liu, J., Brown, T., Chen, J., Ali, U., Hou, L., & Costanzo, K. (2016). *Mode Comparability Study Based on Spring 2015 Operational Test Data.* Retrieved from https://files.eric.ed.gov/fulltext/ED599049.pdf (검색일: 2021.1.5.)

Proctor, T. P., Chuah, S. C., Montgomery, M., & Way, W. D. (2019). *Comparability of performance on the SAT Suite of Assessments across pencil-and-paper and*

computer-based modes of administration. College Board.

Steedle, J., McBride, M., Johnson, M., & Keng, L. (2016). *Spring 2015 Digital Devices Comparability Research Study.* Retrieved from https://files.eric.ed.gov/fulltext/ED599032.pdf (검색일: 2021.1.5.)

Steedle, J., Pashley, P., & Cho, Y. (2020). *Three Studies of Comparability between Paper-Based and Computer-Based Testing for the ACT.* ACT Research & Policy. Research Report. ACT, Inc.

Organisation for Economic Co-operation and Development. (2009). *Equally Prepared for Life?: How 15-Year-Old Boys and Girls Perform in School.* Brussels, Belgium: OECD.

· 논문접수 : 2022.01.05. / 수정본접수 : 2022.01.28. / 게재승인 : 2022.02.09.

ABSTRACT

A study on mode effects of National Assessment of Educational Achievement(NAEA)

Sora Lee
Associate Researcher, Korea Institute for Curriculum and Evaluation
Yongnam Kim
Assistant Professor, Seoul National University
Namwook Koo
Researcher, Korea Institute for Curriculum and Evaluation

This study was conducted using paper-based test (PBT) and computer-based test (CBT) data collected from the 2020 national level academic achievement evaluation for high school sophomores. First, according to the reliability analysis results, stable reliability indices were calculated for the two implementation methods in Korean, math, and English subjects. As a result of examining construct equivalence, it was found that the construct equivalence between the two modes in Korean and English subjects was suspected. In order to control the school-level characteristics, a multi-level model was applied to the schools where both PBT and CBT were implemented, and the regression coefficients of the random intercept model showing greater model-fit were interpreted from the analysis results. As a result, a clear mode effect was found in the type of construct item in Korean and English subjects, and the CBT score was significantly higher. Next, differences in mode effects for each subgroup such as gender, regional size, and school gender were analyzed. As a result, it was found that male students tend to score favorable on CBT in mathematics and English subjects, and female students' achievement was not affected by the change in testing mode.

Key Words: *National assessment of educational achievement(NAEA), computer-based testing(CBT), mode effects, multi-level analysis*