The Mathematical Education. Vol. 61, No. 4, 597-612
http://doi.org/10.7468/mathedu.2022.61.4.597

Print ISSN 1225-1380
Online ISSN 2287-9633

Research Article

Exploring the change in achievement by the transition of the test mode from paper to computer: Focusing on the national assessment of educational achievement of high school mathematics

Jung, Hye-Yun[1] · Song, Chang-Geun[2] · Kim, Young-Jun[2] · Lee, Kyeong-Hwa[3]*

[1]Assistant Research Fellow, Korea Institute for Curriculum and Evaluation

[2]Student, Seoul National University

[3]Professor, Seoul National University

*Corresponding Author: Kyeong-Hwa, Lee (khmath@snu.ac.kr)

ABSTRACT

Recently, large-scale mathematics assessments are shifting from traditional paper-based tests to computer-based tests, nationally and internationally. This study explored the mode effect (the difference in student achievement by the change of test mode) according to the types of test items, the technological function reflected in the items, the characteristics of students' computer use, and the computer-based test environment. To this end, we analyzed the results of the 2020 national assessment of educational achievement of high school mathematics conducted on a paper and computer basis. As a result, firstly, the mode effect induced by the mode transition was generally insignificant, but the mode effect was larger in the extended response type than other types. Secondly, there were differences in the mode effect according to the transition to test with computer mode where innovative items were added. Thirdly, the difference between mode effects was statistically significant according to the student's sense of efficacy in computer use. The results of this study suggest that innovative items should be introduced deliberately according to the targeted content and competency in evaluation, and that assessment design and environment preparation need to be carefully developed so that nonessential abilities other than students' mathematical ability or incidental situation do not distort the assessment results.

Key words: computer-based test of mathematics, large-scale assessment, mode effects, innovative item

종이에서 컴퓨터로의 매체 전환에 따른 평가 결과의 변화 탐색: 고등학교 수학 국가수준 학업성취도 평가를 중심으로

정혜윤[1] · 송창근[2] · 김영준[2] · 이경화[3]*

[1]한국교육과정평가원 부연구위원 · [2]서울대학교 대학원 학생 · [3]서울대학교 교수

*교신저자: 이경화 (khmath@snu.ac.kr)

초록

최근 국내외 대규모 평가를 중심으로, 수학 평가가 전통적인 지필 평가에서 컴퓨터 기반 평가로 전환되고 있다. 본 연구는 평가 문항의 유형, 문항에 반영된 기술공학적 기능, 학생의 컴퓨터 사용 특성 및 컴퓨터 평가 환경에 따라 평가 매체 전환에 따른 학생의 성취도 차이가 어떻게 변화하는지 탐색하였다. 이를 위해 지필 환경과 컴퓨터 환경에서 실행된 2020 고등학교 수학의 국가수준 학업성취도 평가 결과를 분석하였다. 연구결과, 첫째, 단순 매체만 전환된 경우 모드 효과는 대체로 미미하였지만, 서술형 문항의 경우 단답형이나 선택형 문항에 비해 모드 효과가 더 크게 나타났다.

Received October 21, 2022 **Revised** November 01, 2022 **Accepted** November 01, 2022
2000 Mathematics Subject Classification : 97C40, 97U70

둘째, 매체 전환과 기술공학적 기능의 반영이 함께 이루어진 경우, 문항에 반영된 기술공학적 기능에 따라 모드 효과에 차이가 있었다. 셋째, 컴퓨터 기반 평가 시행환경, 학생의 컴퓨터 활용 빈도와는 달리, 학생의 컴퓨터 활용 효능감에 따라 컴퓨터 기반 평가에서 통계적으로 유의한 점수 차이가 나타났다. 이와 같은 연구 결과는 향후 평가 내용과 역량에 맞춘 적절한 기술공학적 기능의 도입이 필요하며, 학생의 수학 학습 능력 외의 부수적인 능력이나 상황이 평가 결과에 영향을 미치지 않도록 세밀한 평가 설계와 환경 구비가 필요함을 시사한다.

주요어: 컴퓨터 기반 수학 평가, 대규모 평가, 모드 효과, 혁신 문항

서론

4차 산업혁명 시대에 교육개혁의 방향은 학생의 디지털 리터러시를 강화하는 쪽으로 나갈 것이며, 이러한 경향성이 교육 평가의 영역에서도 컴퓨터 활용의 증가를 요구할 것이라는 사실은 잘 알려져 있다(Clements, 2020; Spires, Paul, & Kerkhoff, 2019). 특히 4차 산업혁명 시대를 마주하는 시점에서 발생한 COVID-19 사태와 이로 인한 언택트 교육에 대한 논의의 증가(e.g., Heyd-Metzuyanim, Sharon, & Baram-Tsabari, 2021; Lee, Ham, Lee, & Park, 2020; Park, 2020; Parshall, Harmes, Davey, & Pashley, 2009)는 교육 평가 분야에서의 컴퓨터 활용에 대한 요구를 가속화하였다. 실제로, 국제 수준의 대규모 평가로 널리 알려진 PISA (Programme for International Student Assessment)는 2006년부터(Martin, 2008), TIMSS(Trends in International Mathematics and Science Study)는 2019년부터 컴퓨터 기반 평가(computer based test, 이하 CBT)로 이행하였으며(Martin et al., 2020), 이후 여러 나라의 국내 수준 대규모 평가가 CBT로 전환되는 등 시대의 변화와 함께 컴퓨터 평가는 널리 확장되고 있다. 구체적으로, 미국의 NAEP와 호주의 NAPLAN이 이미 CBT로 전환하였으며(Bennett et al., 2008; Rose et al., 2020), 우리나라의 국가수준 학업성취도 평가 역시 이러한 국제적 흐름에 맞추어 2021년 예비시행(2월)과 병행시행(9월)을 거쳐, 2022년 CBT로의 전면 시행을 준비 중이다.

사회와 시대의 변화를 반영한다는 측면 외에, CBT는 교육적인 맥락에서도 의미를 갖는다. 전통적인 종이 기반 평가(paper based test, 이하 PBT)가 교육과정에서 목표로 하는 역량과 실천을 측정하는 데 제약이 있다는 비판과 함께(Schoenfeld, 2017), CBT가 이를 보완할 수 있다는 주장이 제기되었다. 구체적으로, CBT는 기존 평가에서와 다른 문항 형식을 이용하여, 고등 사고, 비판적 사고, 창의성, 문제해결 등 21세기 역량 측정을 가능하게 하고(Hoogland & Tout, 2018), 현실적인 과제를 통해 학생의 다차원적 특성을 평가할 수 있게 한다(Csapó et al., 2012).

이처럼 CBT는 사회적, 교육적 측면에서 주목받고 있지만, CBT로의 전환은 신중히 이루어져야 한다. 새롭게 도입되는 평가 방식에 대한 학생의 적응이 필요함은 물론, 평가 방식의 변화에 따라 새롭게 도입되는 혁신적인 문항, 그리고 평가 결과의 변화에 영향을 미치는 요인에 대한 분석이 필요한 것이다. 이와 관련하여, CBT에 대한 연구는 앞서 언급한 국제, 국가 수준의 대규모 평가가 CBT로 전환되기 시작한 2000년대부터 주목받기 시작하였다. 먼저, Bennett 외 (2008), Drasgow와 Mattern (2006), Sireci와 Zenisky (2006) 등의 연구자는 PBT와 차별화되는 CBT의 혁신적인 문항(예를 들어, 아래로 펼치기 기능, 끌어놓기 등이 반영된 문항)을 제시한 바 있다. 이후 PBT와 CBT를 비교, 분석한 연구가 수행되었는데, 주로 PBT와 CBT 사이의 점수 차를 확인하여 모드 효과를 분석한 연구가 이루어졌다(Keng, McClarty, & Davis, 2008; Wang, Jiao, Young, Brooks, & Olson, 2007). 뿐만 아니라, 모드 효과의 원인이 되는 학생 특성 및 외부 환경 요인에 대한 탐색도 진행되었다(Ebrahimi, Toroujeni, & Shahbazi, 2019; McClelland & Cuevas, 2020; McDonald, 2002; Sandene et al., 2005).

다만, 이들 연구가 의미 있는 시사점을 제공해주긴 하였지만, 대규모 평가를 중심으로 CBT로의 전환이 확산되고 있는 지금의 상황에서, 매체 전환에 따른 점수 차를 분석하는 데에서 나아가 모드 효과의 원인이 종이에서 컴퓨터로의 단순 매체 전환에 의한 것인지, 매체 전환과 함께 도입된 기술공학적 기능이 적용된 문항에 의한 것인지 확인하는 것이 필요하다. 특히, CBT에 제시된 문항(Bennett et al., 2008)의 각 기술공학적 기능이 문항의 정답률 변화에 미치는 영향을 논의하는 연구로의 확장이 필요하다. 기술공학적 기능으로 인해 학생의 문제 해결 과정에 의도하지 않은 변화나 장애물이 발생한다면, 평가로부터 측정하고자 하는 학생의 지식을

정확히 측정하는 데 어려움이 존재할 수 있기 때문이다(Buerger, Koehler, & Goldhammer, 2019). 또한, 학생의 컴퓨터 활용 빈도와 효능감 면에서의 특성과 컴퓨터 사용 환경 요인이 평가에 어떠한 영향을 미치는지 역시 아직 국내의 대규모 평가의 환경을 대상으로는 거의 연구되지 않았기에 이에 대한 탐색이 필요하다. 이는 수학 교과에서도 마찬가지이다. 수학 교과는 특히 CBT에 대한 연구가 부족한 바, 수학 교과에서 수행된 CBT 결과를 분석하고 학생의 컴퓨터 사용 면에서의 특성과 컴퓨터 사용 환경이 평가 결과에 미치는 영향을 파악하여 CBT로의 전환이 수학 교과의 평가 결과에 미치는 영향을 파악하는 것이 필요하다. 이를 통해 수학 교과에서 측정하고자 하는 수학 학습 역량에 부합하도록 평가 환경을 개선하고 CBT 문항을 개발해 나갈 수 있을 것이다.

우리나라의 경우, 교육부가 2019년에 국가수준 학업성취도 평가에서의 CBT 도입 3단계를 제시하고, 2020년 예비시행, 2021년 병행시행을 거쳐 2022년 전면 시행을 할 것을 발표하였으며, 이에 국가수준 학업성취도 평가를 시행하고 있는 한국교육과정평가원에서 2021년 예비시행[1]과 병행시행을 거쳐, 2022년 전면 시행을 위한 컴퓨터 시스템의 구비 및 문항 개발을 추진 중에 있다(Lee et al., 2021). 이러한 상황 속에서, 본 연구는 2021년 2월에 시행된 컴퓨터 기반 국가수준 학업성취도 평가 예비시행 결과를 분석하여, 향후 수학 교육 평가에서 CBT가 안정적으로 도입되기 위해 필요한 시사점을 과제 유형 및 기술공학적 기능이 반영된 문항의 특성과 학생 특성, 외적 환경 요인의 측면에서 모색하고자 한다. 이를 위해, 먼저 종이에서 컴퓨터로의 매체 전환에 따른 평가 결과의 차이를 살펴보고, 기술공학적 기능이 반영된 문항의 특성과 학생의 특성 및 외적 환경이 평가 결과에 미치는 영향을 살펴보고자 한다. 구체적인 연구 질문은 다음과 같다.

연구 문제1. 종이에서 컴퓨터로의 단순 매체 전환에 따른 문항별 정답률의 차이는 문항의 유형에 따라 어떻게 나타나는가?

연구 문제2. 종이에서 컴퓨터로의 매체 전환과 더불어 기존의 문항에 기술공학적 기능이 반영된 경우, 문항별 정답률의 차이는 어떻게 나타나는가?

연구 문제3. 학생의 특성(컴퓨터 활용 빈도, 컴퓨터 활용 효능감)과 외적 환경(컴퓨터 시행환경)에 따라 CBT 평가 결과의 차이는 어떻게 나타나는가?

선행연구 분석

CBT에 관한 선행연구 분석

교육 분야에서 평가는 오랜 기간 주로 PBT 형식으로 이루어져 왔다. 하지만 최근 전통적인 PBT의 한계가 여러 연구자로부터 지적되고(e.g. Hoogland & Tout, 2018; Schoenfeld, 2017) 그에 대한 대안으로 CBT가 주목을 받으면서(Csapó et al., 2012; Poggio, Glasnapp, Yang, & Poggio, 2005), 국가 또는 국제 수준의 대규모 평가를 중심으로, 평가 매체가 종이에서 컴퓨터로 전환되어 가고 있다(Bennett et al., 2008; Martin, 2008; Martin et al., 2020; Rose et al., 2020).

CBT로의 전환은 단순히 평가 매체가 바뀌는 것을 넘어, 새로운 기술공학적 기능을 반영한 문항을 평가에 활용할 수 있음을 의미한다. 이러한 새로운 문항을 선행연구에서는 '혁신적인 문항(innovative item)'(Drasgow & Mattern, 2006)이라 부른다. 이와 관련하여, Sireci와 Zenisky (2006)는 문항의 성분을 과제의 내용이 제시되는 자료와 학생의 답이 기록되는 응답으로 구분한 바 있다. 이에 따르면 혁신적인 문항은 문항의 자료 성분에 PBT에는 추가될 수 없는 기능(e.g. 미디어, 비디오, 소리, 애니메이션)이 더해지거나, 응답의 형식이나 응답에 반응하는 행위가 전통적인 시험 방식과 다를 수 있다. 예를 들어, 컴퓨터 환경에서는 객관식 문항의 응답 부분을 아래로 펼치기, 끌어놓기, 체크박스 기능 등 다양한 기술공학적 기능(Figure 1, 2 참고)을 활용하여 새로운 포맷으로 제시할 수 있다. 반영된 기술공학적 기능에 따라 혁신적인 문항이 성취도에 미치는 영향은 달라질 수 있지만(Buerger et al., 2019; Hu et al., 2021), 아직 국내에서는 그 영향을 분석한 연구가 거의 이루어지지 않았다.

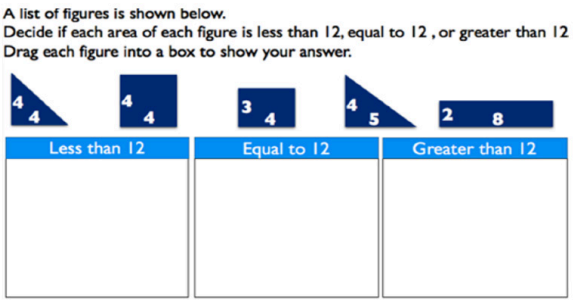

Figure 1. An example of drag and drop item (Arslan et al., 2020)

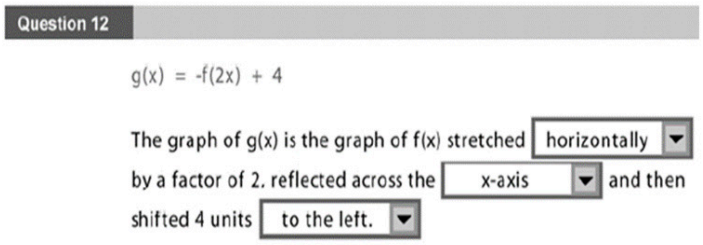

Figure 2. An example of pull-down item (Scalise & Gifford, 2006)

CBT에 대한 연구는 주로 PBT가 CBT로 전환됨에 따른 모드 효과 분석을 중심으로 이루어졌다. 평가에서의 모드 효과가 존재한다는 것은 평가 매체의 변화가 피험자의 응답에 영향을 주어 응답에 변화가 발생하였음을 의미하는 것으로, 여러 선행연구에서는 평가 매체의 변환에 따른 기존의 점수와 새로운 점수 사이의 차이를 통해 모드 효과가 발생하였음을 판단한다(e.g., Ebrahimi, Toroujeni, & Shahbazi, 2019; Fishbein, Martin, Mullis, & Foy, 2018). 예컨대, Fishbein 외 (2018)는 2019 eTIMSS와 paperTIMSS의 평균 정답률을 비교한 뒤 paperTIMSS에서의 평균 정답률이 높았다는 점을 근거로 모드 효과가 존재함을 주장하였다. 한편, Fishbein 외 (2018)가 모드 효과가 있음을 주장한 것과 달리, 다수의 연구(Ebrahimi, Toroujeni, & Shahbazi, 2019; Eid, 2005; Karay, Schauber, Stosch, & Schüttpelz-Brauns, 2015; Poggio et al., 2005; Wang et al., 2007)에서는 CBT와 PBT의 결과 비교를 통해 통계적으로 유의한 모드 효과가 발견되지 않았음을 주장한다. 다만, Fishbein 외 (2018) 이외에도 모드 효과가 존재했다고 밝히는 연구 결과(Backes & Cowan, 2019; Bennett et al., 2008)가 제시되는 등 모드 효과에 대한 연구 결과는 연구마다 엇갈리고 있다(Wang et al., 2007). 특히 모드 효과의 존재성을 밝히는 연구의 대부분은 PBT에서의 학생 성적이 CBT에서보다 높았음을 밝히고 있다.

이처럼 모드 효과에 대한 연구 결과가 엇갈림에 따라, 연구의 초점은 정답률 차이를 단순 비교하는 것에서 나아가 두 평가 간의 정답률에 차이가 나는 문항의 특성을 분석하는 것으로 확장되어 갔다(Keng et al., 2008; Noyes & Garland, 2008; Pommerich, 2004). 몇몇 선행연구(Horkay et al., 2006; Russell, 1999; Russell & Tao, 2004)에서는 학생이 주어진 응답 중 적절한 것을 고르는 선택적 응답에 비해, 본인의 응답을 생성해야 하는 구성적 응답의 경우에 모드 효과가 더 강하다는 것을 확인한 바 있다.

외부 변인에 대한 선행연구 분석

CBT에서는 학생이 컴퓨터를 이용해 문항에 대한 답을 제시해야 한다는 점에서 컴퓨터와 관련된 학생 변인이나 외부 환경 요인들이 학생의 성적에 영향을 줄 수 있으므로, 이에 관련된 연구들이 진행되었다(McDonald, 2002). 기존 연구에서는 컴퓨터와 관련된 학생 변인으로 컴퓨터에 대한 친숙도를 주로 다루었으며, 이를 학생들의 컴퓨터에 대한 경험, 접근성, 태도, 활용 능력 등으로 정의한 뒤 CBT 결과와의 상관관계를 분석하였다(e.g. Eid, 2005; Sandene et al., 2005). 더불어 모니터와 네트워크의 질 등의 CBT 시행환경 또한 성적에 영향을 미칠 수 있는 외부 환경 요인으로 분석된 바 있다(Kingston, 2008).

한편, 선행연구에서는 컴퓨터에 대한 학생의 경험이나 태도가 CBT의 성취도에 영향을 줄 수 있다는 논의가 제시되었지만(e.g. Russell, 1999; Sandene et al., 2005), 대부분의 최근 연구에서는 이러한 요인들이 CBT의 점수에 유의한 영향을 주지 않는다는 결론이 제시되고 있다(Ebrahimi et al., 2019; McClelland & Cuevas, 2020). 이는 시간이 지남에 따라 점차 일반 학생들의 컴퓨터에 대한 경험과 접근성이 높아진 데에 따른 결과라고 해석할 수 있으며(McDonald, 2002; Zilles et al., 2018), 지금의 학생들이 CBT 성적에 부정적인 영향을 미치지 않을 만큼의 기본적인 컴퓨터 친숙도를 갖추고 있음을 시사한다(McClelland & Cuevas, 2020). 또한 컴퓨터 경험이 부족하더라도, CBT 상황에 대한 적절한 연습 과정이 제공된다면 컴퓨터 경험 차이에 의한 영향은 충분히 줄일 수 있음을 의미한다(Kingston, 2008; Taylor, Jamieson, Eignor, & Kirsch, 1998). 실제로 연구 참여 학생들의 인터넷 사용 경험이 거의 없었던 Eid (2005)의 연구에서도 학생들은 연습 과정을 거치는 동안 쉽게 CBT에 익숙해졌으며, 결과적으로 인터넷 경험 부족은 평가 결과에 영향을 미치지 않았다. 이외에도 컴퓨터에 관한 경험과는 별개로 여러 차례의 CBT를 겪으면서 CBT 자체에 대한 친숙도가 증가한다면 성취도의 상승이 있을 수 있다는 분석이 있다(Backes & Cowan, 2019).

Bennet 외 (2008)는 컴퓨터 경험 외에 추가적으로 입력 정확도 및 입력 속도 또한 컴퓨터 친숙도의 요소로 보고 각각 성취도와의 상관 정도를 분석하였다. 그 결과 CBT의 성적과 유의미한 상관관계가 없었던 컴퓨터 경험 항목과는 달리, 입력 정확도 및 입력 속도와 같은 컴퓨터 활용 능력이 높을수록 성취도도 높아짐을 확인하였다. CBT의 환경에서 학생들은 PBT 환경에 비해 잘못된 입력을 하기 더 쉽고(Shacham, 1998) 시험 시간의 최적화를 더 어려워하기 때문에(Zilles et al., 2018), 입력의 정확도와 입력 속도같은 컴퓨터 활용 능력이 학생들의 성적에 상대적으로 더 영향을 미칠 수 있다. 또한 Russell (1999)은 특히 구성적 응답이 필요한 CBT 문항의 경우에 학생의 타이핑 능력이 결과에 영향을 준다는 점을 지적하였다. 이처럼 CBT에서는 시험이 평가하고자 하는 능력 외의 다른 컴퓨터 관련 활용 능력이 함께 평가됨을 지적하는 연구들이 존재한다(McDonald, 2002).

컴퓨터 시행환경과 관련하여, 구체적인 수치를 기반으로 CBT 결과를 비교 분석한 최근 연구는 찾아보기 어렵다. 다만, 몇몇 연구(e.g., Bennet et al., 2008; Kingston, 2008)에서 시험이 시행되는 곳의 네트워크 제약이나 컴퓨터 상황에 따른 시험의 중단 등의 기술적인 문제가 학생의 성취도에 부정적인 영향을 줄 수 있다는 일화를 제시한 바 있다. 더불어 Zilles 외 (2018)는 학생들이 CBT의 과정에서 느낀 불만 사항들이 대부분 이러한 기술적인 문제 때문이었음을 언급하였다.

연구방법

연구 참여자

본 연구의 참여자는 PBT로 이루어진 2020 국가수준 학업성취도 평가, 2020 PBT를 컴퓨터로 모드로 전환한 평가(internet-based test, 이하 IBT), 2020 PBT에서 일부 문항이 혁신적인 문항(예를 들어, 아래로 펼치기, 끌어놓기 문항 등)으로 변화된 CBT에 참여한 학생들이다. 연구 참여자는 모두 고등학교 2학년이다. PBT와 IBT는 동시에 진행되었으며, 평가 참여 학생은 2단계 층화 표집 절차를 거쳐 각각 10,526명, 473명이 선발되었다. 2단계 층화 표집은 고등학교의 3%를 표집학교로 선발한 뒤, 표집학교에 속한 일반 학급 중에서 임의로 두 학급을 선정하는 과정을 거쳐 이루어졌으며, 선정된 학급 중에서 COVID-19로 인해 원격 수업을 진행하고 있는 학급에 속한 학생이 IBT에 참여하였다. CBT는 PBT에 제시된 문항 중 일부 문항을 혁신적인 문항으로 변환하는 과정을 거친 뒤, 약 2개월 뒤에 시행되었다. 표집 학생의 수학 학습능력이나 컴퓨터 시행환경 등 문항별 정답률 차이에 영향을 미칠 수 있는 요인을 최소화하기 위하여 학교의 학업성취도와 컴퓨터 시행환경 등을 고려하여 대도시와 중소도시에서 각각 1개의 학교가 표집되었으며, 두 학교의 2학년 전체 학생 총 373명이 평가에 참여하였다.

본 연구에서는 연구 참여자의 컴퓨터 활용 능력 등의 학생 변인과 환경 변인이 CBT 결과에 영향을 미칠 수도 있다는 선행연구를 토대로, CBT 참여 학생의 학생 변인과 환경 변인에 따른 CBT 결과를 비교하고자 하였다. 학생 변인과 환경 변인으로 CBT 참여 학생의 컴퓨터 활용 빈도(한글, ppt 등의 활용 빈도)와 컴퓨터 활용 효능감(컴퓨터를 활용한 자료 검색, 수집 및 제작에 대한 자신감) 및 시험 응시 환경(컴퓨터 음향, 화질, 속도 등)을 확인하였다. 해당 정보는 CBT에 참여하는 학생과 교사의 진술을 통해 확인하였으며,

CBT 평가에 참여한 전체 학생 373명 중 진술이 확보된 359명의 학생을 대상으로 외부 변인에 따른 CBT 결과를 비교, 분석하였다 (Table 1 참고). CBT 시행 당시, 학생들은 COVID-19로 인해 약 일 년간 원격 수업에 참여한 경험이 있었다. 이로 인해 컴퓨터 활용 빈도, 컴퓨터 활용 효능감이 대부분 높은 상태였으며, 학교의 시행환경 역시 양호한 편이었다.

Table 1. The frequency in the computer use, the sense of efficacy in computer use, and the environment for CBT

The frequency in the computer use			The sense of efficacy in computer use				The environment for CBT		
Rarely used	At least once a week	Sum	Low	Medium	High	Sum	Inconvenience	Convenience	Sum
41 (11.4)	318 (88.6)	359 (100.0)	41 (11.4)	177 (49.3)	141 (39.3)	359 (100.0)	47 (13.1)	312 (86.9)	359 (100.0)

평가 도구

모든 유형의 시험은 총 20문항으로 구성되었으며, 고등학교 1학년 학생이 공통으로 배우는 '수학' 과목의 전 범위를 반영하였다. PBT, IBT, CBT를 구성하는 문항의 구체적인 특징은 Table 2와 같다. IBT 문항의 경우, PBT 문항을 컴퓨터로 매체만 전환한 형태이므로 선다형과 서답형 등의 문항 특징에 변화가 없었다[2]. CBT 문항의 경우, 첫째, PBT 문항을 그대로 사용하면서 응답 방식이나 자료 탐색 측면에 기술공학적 기능을 반영하거나, 둘째, PBT 문항과 동일한 성취기준과 유사한 난이도를 갖는 동형의 문항으로 변형한 뒤 기술공학적 기능을 반영[3]하였으며, 그 결과 기술공학적 기능이 반영된 문항 8개가 최종적으로 제시되었다. 기술공학적 기능이 반영되기 전의 동형 문항의 경우, 기존 문항과 동일한 성취기준 및 유사한 난이도를 가질 수 있도록 2020 PBT 개발에 참여한 수학 교사 4명과 수학교육 전문가 1명, 수학교육 박사 1명으로 구성된 공동체를 별도로 구성한 뒤, 공동체 내에서 개발 및 검토, 수정하는 과정을 거쳤다. 기술공학적 기능이 반영된 문항 중 자료탐색 측면에서 동영상, 시뮬레이션 등의 기능이 총 4개의 문항에 반영되었으며, 학생 응답 측면에서 아래로 펼치기, 체크박스, 끌어놓기 등의 기능이 총 4개의 문항에 반영되었다. CBT 시험에 익숙하지 않음을 고려하여, 자료탐색과 학생응답 측면에서의 기술공학적 기능을 동시에 반영하지 않았으며, 수식입력기 역시 익숙치 않음을 고려하여 서술형 문항의 개수를 줄이고 단답형을 늘렸다.

Table 2. The item types presented in PBT, IBT, and CBT

Item type		PBT	IBT	CBT
Selected response item		15	15	13
Constructed response item	Short answer item	3	3	6
	Extended response item	2	2	1
Sum		20	20	20
Innovative item		0	0	8

자료 수집 및 분석

본 연구에서는 분석 자료로 2020 국가수준 학업성취도 평가의 PBT, IBT, CBT 자료를 사용하였다. 해당 자료에는 시험 유형별 각 문항의 평균 정답률, CBT에 참여 학생 대상 인터뷰 자료가 포함된다. 인터뷰는 CBT에 반영된 기술공학적 기능이 문제해결에 미친 영향에 대한 질문으로 구성되었으며, 학생이 자유롭게 이야기할 수 있도록 편안한 환경에서 반구조화된 방식으로 진행되었다.

자료 분석은 연구 질문에 맞추어 다음의 세 단계로 진행되었다. 첫째, PBT와 IBT의 정답률을 비교하여 동일 문항의 컴퓨터 전환에 따른 정답률 차이를 확인하였다. 둘째, PBT와 CBT의 정답률을 비교하여 매체 전환과 함께 기술공학적 기능이 반영된 경우 발생하는 정답률의 변화를 확인하였다. 셋째, 학생의 컴퓨터 활용 빈도, 컴퓨터 활용 효능감, 시험 시행환경에 따른 CBT 정답률의 차이를 확인하였다. CBT에는 아래로 펼치기, 끌어놓기, 시뮬레이션 등의 기술공학적 기능이 반영되었기 때문에, PBT와 달리 학생의 수학적 지식 외에 컴퓨터 활용 능력 등이 점수에 영향을 미칠 수 있다. 이에, SPSS ver.26을 이용하여 이들 세 가지 변수와 점수 사이의

상관분석을 시행하였다. 컴퓨터 활용 빈도와 시험 시행환경 변인의 경우 각각 거의 사용 안 함과 주 1회 이상, 불편과 양호의 두 가지 기준에 따른 결과를 분석해야 하므로 t 검정을 수행하였으며, 컴퓨터 활용 효능감 변인의 경우 저, 중, 고의 세 가지 기준으로 결과를 분석해야 하므로 ANOVA 검정을 수행하였다(Yoo, 2013).

연구결과

종이에서 컴퓨터로의 단순 매체 전환에 따른 문항의 정답률 차이

Figure 3은 문항별 PBT에서의 정답률(축)과 IBT에서의 정답률(축)의 차이를 나타낸다. 직선(y=x)은 PBT와 IBT에서의 정답률의 차이가 없는 경우를 의미하며, 직선에서 멀어질수록 두 모드에서 해당 문항의 점수차가 크다는 것을 의미한다. 또한, 직선 아래쪽에 위치한 번호는 PBT에서의 정답률이 더 높음을, 직선 위쪽에 위치한 번호는 IBT에서의 정답률이 더 높음을 의미한다.

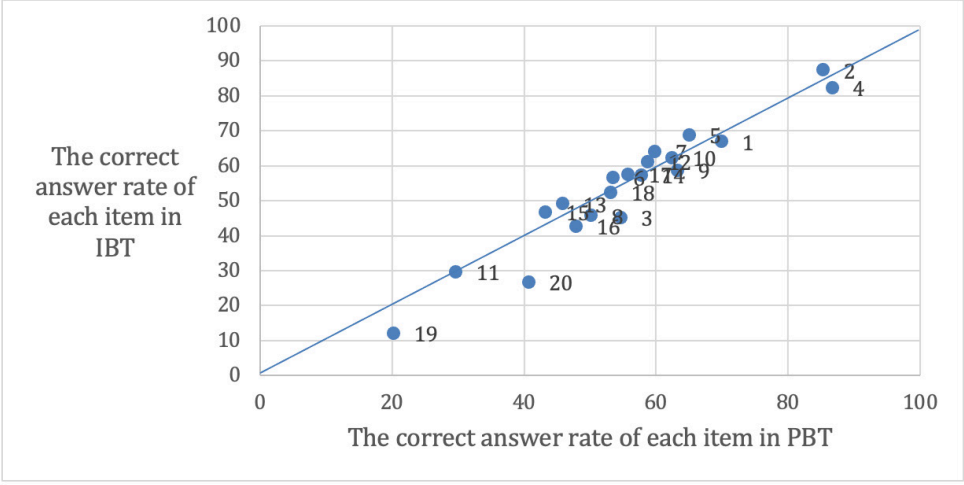

Figure 3. The correct answer rate of each item presented in PBT and IBT (unit: %)

Figure 3을 살펴보면, 대부분의 문항이 직선 주변에 위치하고 있어, 대부분 문항에서의 PBT와 IBT의 정답률 차이가 크지 않음을 확인할 수 있다. 실제로, Pearson r=0.96으로, 문항별 PBT의 정답률과 IBT의 정답률 사이에는 매우 강한 양의 상관관계가 존재한다. 예외적으로 직선에서 떨어진 문항은 3번, 19번, 20번 문항으로, 다른 문항들에 비해 PBT에서의 정답률과 IBT에서의 정답률의 차이가 크다는 사실을 알 수 있다. 특기할 만한 점은, Figure 3을 살펴보면, 서술형 문항에 해당하는 19번과 20번의 경우 PBT와 IBT에서의 정답률이 모두 가장 낮은 편에 속한다는 점이다. 문항의 정답률이 낮음은 난이도가 높음을 의미한다는 점에 비추어, 난이도가 높은 문항은 매체에 상관없이 학생들이 어려움을 느낀다는 것을 알 수 있다.

Figure 3에서 PBT와 IBT의 정답률 사이에 강한 상관관계가 있음을 확인하였지만, 문항별 정답률 차이는 조금씩 다르게 나타났다. 실제로, 전체 20개의 문항 중 12개 문항의 정답률이 PBT에서 더 높게 나타났으며, 8개 문항의 정답률이 IBT에서 더 높게 나타났다. 이와 관련하여, Table 3은 각 문항의 PBT 정답률과 IBT 정답률에 대한 구체적인 수치를 보여주며, Figure 4는 문항 유형별 PBT와 IBT에서의 정답률의 평균을 및 문항 유형별 PBT와 IBT에서의 평균 정답률의 차이를 보여준다. 아래에서는 Table 3과 Figure 3, Figure 4를 통해 확인할 수 있는 매체 전환에 따른 문항 유형별 정답률 차이를 살펴본다.

Table 3. Item type and the correct answer rate for PBT and IBT of each item

Number	Item type	PBT(%)	IBT(%)	Number	Item type	PBT(%)	IBT(%)
1	Selected response item	69.97	66.81	11	Selected response item	29.73	29.60
2		85.34	87.32	12		58.77	61.10
3		54.68	45.03	13		45.94	49.05
4		86.73	82.24	14		57.81	57.29
5		65.04	68.71	15		43.29	46.72
6		53.49	56.66	16	Short answer item	47.93	42.71
7		59.89	64.06	17		55.84	57.40
8		50.21	45.67	18		53.22	52.22
9		63.34	58.56	19	Extended response item	20.33	12.05
10		62.47	62.16	20		40.72	26.71

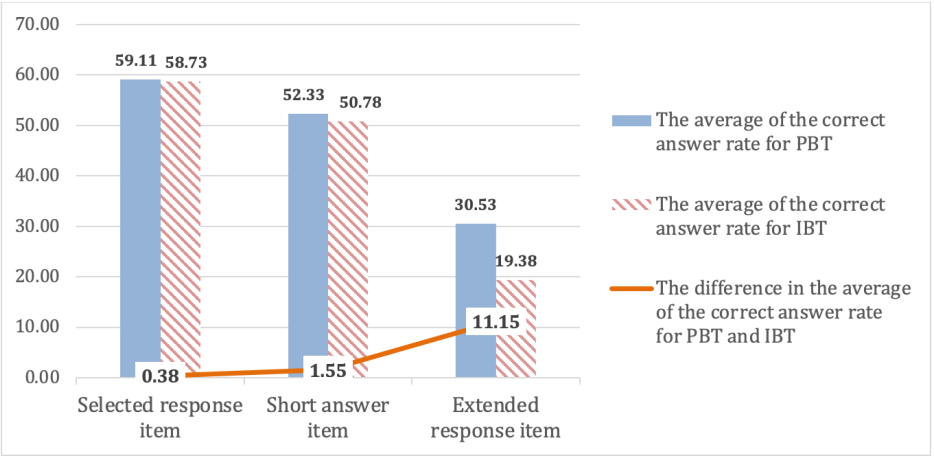

Figure 4. The average of the correct answer rate for PBT and IBT (unit: %) and the difference in the average of the correct answer rate for PBT and IBT (unit: %p) by the item type

Figure 4에서 알 수 있듯이, 선다형 문항의 경우 PBT에서의 평균 정답률은 59.11%, IBT에서의 평균 정답률은 58.73%로, 정답률 차이가 거의 나타나지 않았다. 단답형 문항의 경우 PBT에서의 평균 정답률은 52.33%, IBT에서의 평균 정답률은 50.78%, 이들 평균 정답률의 차이는 1.55%p로, 단순 매체 전환에 따른 평균 정답률의 차이가 선다형 문항보다 조금 더 크게 나타났다. 서술형 문항의 경우 PBT에서의 평균 정답률은 30.53%, IBT에서의 평균 정답률은 19.38%, 이들 평균 정답률의 차이는 11.15%p로, 단순 매체 전환에 따른 평균 정답률의 차이가 다른 문항 유형에 비해 가장 크게 나타났다. 또한, 선택형과 단답형이 2%p 이내의 차이를 보인 것에 반해, 서술형의 경우 10%p가 넘는 차이를 보였다. 이는 학생들이 다른 유형의 문항에 비해 서술형 문항에서 PBT로의 전환에 따른 어려움을 더 많이 경험하였음을 의미한다.

한편, 선다형 문항의 경우 15개 문항 중 총 8개 문항의 정답률이 PBT에서 높게 나타나고 7개 문항의 정답률이 IBT에서 높게 나타났으며, 서답형 문항의 경우 5개 문항 중 4개 문항의 정답률이 PBT에서 높게 나타났다. 서답형 문항별 정답률의 변화를 다시 단답형과 서술형 문항으로 나누어 살펴보면, 단답형 문항의 경우 세 개의 문항 중 두 개의 문항은 PBT에서, 다른 한 개의 문항은 IBT에서 정답률이 높게 나타났으며, 서술형 문항의 정답률은 모두 PBT에서 높게 나타났다. 이는 수학교과의 경우 대체로 PBT에서 정답률이 더 높게 나타날 수 있음을 보여준다.

Table 3을 토대로 매체 전환에 따른 정답률 차이 정도를 2%p 단위로 구분하여 빈도를 확인하면, 0~2%p 구간에 해당하는 문항의 수가 6개, 2~4%p 구간에 해당하는 문항의 수가 6개, 4~6%p 구간에 해당하는 문항의 수가 5개, 8~10%p 구간에 해당하는 문항의 수가 2개, 10%p 이상 구간에 해당하는 문항의 수가 1개이다. 4%p 이상의 구간에 해당하는 문항의 수(20개 중 8개 문항)보다 4%p 미만의 구간에 해당하는 문항의 수(20개 중 12개 문항)가 더 많았다. 또한, 선다형 문항은 대부분 4%p 미만의 구간에 해당하였지만(15개 중 10개 문항), 서답형 문항은 대부분 4%p 이상의 구간에 해당하였다(5개 중 3개). 이와 같은 수치 역시 선다형 문항보다 서답형 문항에서 매체 전환으로 인한 정답률 차이가 더 크게 나타남을 보여준다.

매체 전환과 더불어 기술공학적 기능의 반영에 따른 문항의 정답률 차이

Figure 5는 문항별 PBT에서의 정답률(축)과 CBT에서의 정답률(축)의 차이를 나타낸다. 직선(y=x)은 PBT와 CBT에서의 정답률의 차이가 없는 경우를 의미하며, 직선에서 멀어질수록 두 모드에서 해당 문항의 점수차가 크다는 것을 의미한다. 또한, 직선 아래쪽에 위치한 번호는 PBT에서의 정답률이 더 높음을, 직선 위쪽에 위치한 번호는 CBT에서의 정답률이 더 높음을 의미한다.

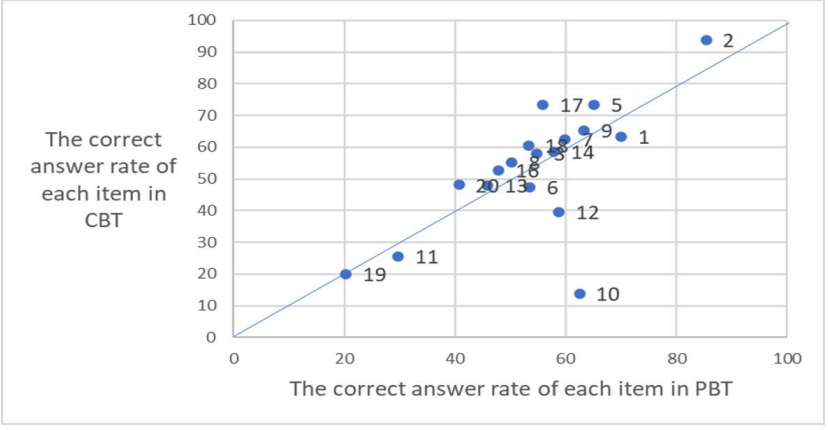

Figure 5. The correct answer rate of each item presented in PBT and CBT (unit: %)

Figure 5를 살펴보면, 대부분의 문항이 직선 주변에 위치하고 있어, 대부분 문항에서의 PBT와 CBT의 정답률 차이가 크지 않음을 확인할 수 있다. 실제로, Pearson r=0.69로, 문항별 PBT의 정답률과 CBT의 정답률 사이에는 강한 양의 상관관계가 존재한다. 예외적으로 직선에서 떨어진 문항은 10번과 12번[4] 문항으로, 다른 문항에 비해 PBT에서의 정답률과 CBT에서의 정답률의 차이가 크고 PBT에서의 정답률이 더 높다는 사실을 알 수 있다. 특기할 만한 점은, 10번과 12번의 경우 PBT에서의 정답률이 낮은 편에 속하지 않는다는 점이다. Figure 3에서 PBT와 IBT에서의 정답률 차이가 컸던 문항의 경우 PBT에서의 정답률이 낮은 문항이었던 것과 달리, Figure 5에서 PBT와 CBT에서의 정답률 차이가 큰 10번과 12번 문항은 PBT에서의 정답률이 오히려 높은 편이었다. 이와 관련하여, Table 4는 각 문항의 PBT 정답률과 CBT 정답률에 대한 구체적인 수치 및 CBT 문항에 반영된 기술공학적 기능을 보여준다[5].

정답률 차이가 가장 크게 나타난 10번 문항의 경우 PBT와 CBT에서의 정답률 차이가 48.68%p로, 두 번째로 정답률 차이가 크게 나타난 12번 문항의 경우 PBT와 CBT에서의 정답률 차이가 19.30%p로 나타났다. 이때, 10번과 12번 문항이 CBT로 전환되면서 반영된 기술공학적 기능은 각각 끌어놓기와 체크박스이다. 10번과 12번의 경우, PBT에서 IBT로 전환될 때 정답률 차이는 각각 0.31%p, -2.33%p로, PBT에서 CBT로 전환될 때의 정답률 차이보다 작음을 알 수 있다. 즉, PBT에서 CBT로의 전환에 따라 발생한 정답률의 큰 하락의 원인은 매체의 전환이 아닌 기술공학적 기능의 반영임을 추측할 수 있다.

Table 4. The correct answer rate for PBT and CBT of each item and technological function reflected in CBT

Number		Technological function reflected in CBT	The correct answer rate		Number		Technological function reflected in CBT	The correct answer rate	
PBT	CBT		PBT(%)	CBT(%)	PBT	CBT		PBT(%)	CBT(%)
1	1		69.97	63.40	11	9		29.73	25.71
2	2		85.34	93.87	12	10	Check box	58.77	39.47
3	3	Pull-down	54.68	58.07	13	11	Simulation	45.94	48.06
5	4		65.04	73.30	14	12		57.81	58.64
6	5		53.49	47.35	16	13		47.93	52.72
7	6		59.89	62.50	17	14		55.84	73.29
8	7		50.21	55.14	18	15		53.22	60.51
9	8	Video	63.34	65.33	19	20	Simulation	20.33	19.86
10	19	Drag and drop	62.47	13.79	20	17		40.72	48.09

한편, Table 4에 따르면, 기술공학적 기능이 반영된 CBT 문항은 10번과 12번 외에도 3번, 9번, 13번, 19번이 있다. 다만, 이들 문항의 경우, 10번, 12번과 달리, PBT와 CBT에서의 정답률 차이가 크지 않다. 특히, 기술공학적 기능의 반영으로 인해 정답률이 하락했던 10번, 12번 문항과는 달리, 3번, 9번, 13번 문항의 경우 기술공학적 기능이 반영된 CBT에서 정답률이 상승하였다. 주목해야 할 점은, Table 3의 결과와 종합했을 때, 종이에서 컴퓨터로의 단순 매체 전환 시에는 이들 문항의 정답률이 하락하였지만 기술공학적 기능이 반영되자 오히려 정답률이 상승했다는 점이다. 19번의 경우에도, 컴퓨터로의 단순 매체 전환 시에는 정답률이 8.28%p 하락했지만, 기술공학적 기능의 적용시에는 정답률 하락이 0.47%p로 줄어들었다. 위와 같은 정답률 변화를 종합하면, 컴퓨터로의 매체 전환 시 문항에 반영된 기술공학적 기능에 따라 정답률의 변화가 다르게 나타남을 알 수 있다.

학생의 특성과 외적 환경에 따른 CBT 평가 결과의 차이

이 절에서는 학생 변인이라고 할 수 있는 학생의 컴퓨터 활용 빈도와 컴퓨터 활용 효능감, 그리고 외적 환경 변인이라고 할 수 있는 CBT 시행환경에 따른 CBT 평가 결과의 차이를 살펴본다. Table 5는 학생의 컴퓨터 활용 빈도에 따른 CBT 점수의 차이를 보여준다. 컴퓨터를 일주일에 한 번도 사용하지 않는 학생의 점수의 평균은 11.10, 표준편차는 6.82이며, 주 1회 이상 사용하는 학생의 점수의 평균은 12.20, 표준편차는 6.38이다. 컴퓨터를 주 1회 이상 사용하는 학생의 점수의 평균이 1.10 높게 나타났으나, 유의수준 0.05에서 검증하였을 때 유의미하지 않은 것으로 확인되었다. 결론적으로, 학생의 컴퓨터 활용 빈도에 따른 유의미한 점수 차이는 나타나지 않았다.

Table 5. The difference in the achievement according to the frequency of the computer use

Frequency	N	M	SD	t	*p*
Rarely used	41	11.10	6.82	-1.038	0.300
At least once a week	318	12.20	6.38		

Table 6은 학생이 컴퓨터 시행환경에 따른 CBT 점수의 차이를 보여준다. CBT에 참여했을 당시의 시행환경에 불편함이 있었던 학생의 점수의 평균은 11.98, 표준편차는 7.13이며, 시행환경이 양호했던 학생의 점수의 평균은 12.09, 표준편차는 6.39이다. 컴퓨터 시행환경이 양호했던 학생의 점수의 평균이 0.11 높게 나타났으나, 유의수준 0.05에서 검증하였을 때 유의미하지 않은 것으로 확인되었다. 결론적으로, 컴퓨터 시행환경에 따른 유의미한 점수 차이는 나타나지 않았다.

Table 6. The difference in the achievement according to the environment for CBT

Environment	N	M	SD	t	*p*
Inconvenience	47	11.98	7.13	-0.113	0.910
Convenience	312	12.09	6.39		

Table 7은 학생의 컴퓨터 활용 효능감에 따른 CBT 점수의 차이를 보여준다. 학생의 컴퓨터 활용 효능감은 저, 중, 고 수준으로 나누어진다. 저 수준 학생의 점수의 평균은 10.20, 표준편차는 5.81이고, 중 수준 학생의 점수의 평균은 11.06, 표준편차는 6.22이며, 고 수준 학생의 점수의 평균은 13.90, 표준편차는 6.46이다. 저 수준과 중 수준 학생의 점수의 평균은 0.86 차이가 났으나 유의수준 0.05에서 검증하였을 때 유의미하지 않은 것으로 확인되었다. 저 수준과 고 수준 학생, 중 수준과 고 수준 학생의 점수의 평균은 각각 3.7, 2.84 차이가 났으며, 모두 유의수준 0.05에서 유의미한 차이가 있었다. 결론적으로, 컴퓨터 활용 효능감이 높은 학생이 낮은 학생과 중 수준 학생에 비해 더 높은 점수를 받을 수 있음을 알 수 있다. 학생의 컴퓨터 활용능력에 따라 평가 결과가 달라질 수 있음을 의미하는 것으로, 인터뷰에 참여한 학생의 "단답형, 서답형처럼 키보드를 사용하여 답안지를 작성해야 하는 경우가 있었는데, 제가 컴퓨터를 잘 못해서, 특히 타이핑 실력이 부족하기 때문에, 저에게는 조금 불리한 시험이었어요."와 같은 응답 역시 이를 뒷받침한다.

Table 7. The difference in the achievement according to the sense of efficacy in computer use

The sense of efficacy	N	M	SD	F*	Tukey**
Low	41	10.20	5.81		
Medium	177	11.06	6.22	10.127	Low<high, Medium<high
High	141	13.90	6.46		

*p<0.05

** α=0.05 수준에서 통계적으로 유의한 집단 간 평균 차이가 있음. 저-중의 유의확률 0.705, 저-고의 유의확률 0.003, 중-고의 유의확률 0.000

논의 및 결론

4차 산업혁명 시대의 도래와 함께 디지털 공학 기술에 기반한 수학교육이 확장되어 가고 있으며, 이러한 경향성은 평가 분야에서도 예외가 될 순 없다. 이에, CBT의 도입이 점차 확산되어 가고 있으며(Martin et al., 2020; Rose et al., 2020), 이와 관련한 연구의 초점 역시 매체 변환에 따른 점수 차이의 분석으로부터 CBT의 성공적인 도입 방안으로 확장되고 있다(e.g., Buerger et al., 2019; Ebrahimi et al., 2019). 이와 같은 흐름에 맞추어, 본 연구에서도 매체 변환에 따른 평가 결과의 분석을 매체와 CBT 문항에 반영되는 기술공학적 기능의 측면, 학생의 수학학습능력 외의 변인과 환경 변인의 측면에서 다각도로 확인하고자 하였다. 그 결과, 첫째, 문항별 PBT와 IBT의 점수의 차이가 대체로 미비하였지만, 일부 서술형 문항에서는 모드 효과가 예외적으로 크게 나타났다. 둘째, 문항에 반영된 기술공학적 기능에 따라 PBT와 CBT의 결과에 차이가 있음을 확인하였다. 셋째, CBT에 참여하는 학생 변인이 CBT 결과에 영향을 미칠 수 있음을 확인하였다. 이하에서는 연구 질문 각각에 대한 주요 결론과 이에 대한 논의를 제시한다.

첫째, 본 연구에서는 문항 유형에 따라 모드 효과에 차이가 발생할 수 있음을 확인하였다. 본 연구 결과, 서답형 문항의 경우 선택형 문항에 비해 대체로 모드 효과가 크게 나타났는데, 이는 선행연구(Horkay et al., 2006; Russell, 1999)를 지지하는 결과이기도 하다. 이때, 모드 효과 차이의 원인 중 하나는 Russell (1999)에서 주장한 바와 같이 타이핑 실력일 수 있지만, 수학 서술형 문항의 경우, 단순히 타이핑 속도나 정확성에 따라 모드 효과의 크기가 달라지는 것이 아닐 수 있다는 점에도 주목할 필요가 있다. 수학 문제를 해결하기 위해서는 추상적인 언어 또는 상징적인 언어를 이용하여 사고하는 과정이 필요한데(Schleppegrell, 2007), 이러한 과정이나 그 결과를 텍스트로 변환하여 타이핑하기 위해서는 타이핑 속도나 정확도로 포괄되지 않는 추가적인 능력이 필요하기 때문이다. 일상 수업이나 개별학습에서 이와 같은 사고의 과정과 자신의 사고를 텍스트로 변환하는 경험을 하지 않은 학생들에게는 이 점이 상당한 부담을 주어 모드 효과의 크기를 높일 가능성이 있다. 다만, 본 연구에서는 문항 유형에 따라 모드 효과가 어떻게 나타나는지를 확인하는 것을 목표로 진행된 연구이기에, 서로 다른 문항 유형이 정확히 어떻게 모드 효과를 일으키는지에 대해서는 후속 연구를 통해 밝혀져야 할 것이다. 나아가, 동일한 문항 유형일지라도 문항에 따라 정답률 차이가 서로 다르게 나타났는데, 이는 문항 유형 외에도 문항의 난이도와 복잡성 등이 정답률 차이의 또다른 요인이 될 수 있음을 보여준다. 이와 관련하여 후속 연구를 통해 모드 효과에 영향을 미치는 다양한 문항 특성이 밝혀져야 할 것이다.

둘째, 본 연구에서는 기술공학적 기능에 따라서 모드 효과가 다르게 나타날 가능성을 확인하였다. 본 연구에 따르면 끌어놓기와 체크박스 응답 형식이 추가된 문항의 경우 PBT의 객관식 문항에 비해 정답률이 낮았고, 반면 아래로 펼치기, 동영상, 시뮬레이션이 추가된 문항의 경우 정답률이 높았다. 이와 같은 결과는 각 기술공학적 기능이 학생의 문제해결 과정에 서로 다른 영향을 주었음을 시사한다. 먼저 주목할 만한 결과는 아래로 펼치기 기능이 추가되며 정답률이 상승한 점인데, 이는 아래로 펼치기 기능을 추가하면 그것이 없는 문항에 비해 근소하게 문항의 정답률이 내려갈 수 있다고 보고한 Buerger 외 (2019)의 결과와 상반된다. 따라서 아래로 펼치기 기능이 학생의 문제해결 과정에 미치는 영향은 후속 연구를 통해 보다 정확히 분석되어야 할 것이다. 동영상 자료가 포함됨으로써 정답률이 상승한 것은 Hu 외 (2021)의 주장과 같이 문제해결의 이해 과정에서 멀티미디어 효과가 나타난 것으로 추측할 수 있다. 기술공학적 기능의 반영에 따라 정답률이 하락한 문항 중 주목할 것은 체크박스와 끌어놓기 기능이 포함된 문항이다. 체크 박스 혹은 끌어놓기 기능이 추가된 문항의 경우, PBT와는 달리 의도하지 않은 교정 피드백(Haladyna et al., 2002)이 주어지지 않으며 PBT에서는 가능하던 거꾸로 풀기(Bridgeman, 1992) 등의 문제해결 전략이 사용될 수 없다는 특징이 있다. 이러한 특징적인 점이 학생의 문제해결 전략과 인지적 과정에 정확히 어떤 영향을 미쳤는지는 후속 연구를 통해 더욱 정확히 검증되어야 할 것이다. 한편, 기술공학적 기능의 반영에 따른 정답률의 변화는 기술공학적 기능의 반영이 문항의 난이도에 영향을 미칠 수 있음을 시사한다. 앞서 언급하였듯이, 동영상 자료는 문제를 더 쉽게 이해할 수 있도록 하고 체크 박스는 거꾸로 풀기 전략의 사용을 차단하는 등 학생의 문제해결 과정에 PBT와는 다른 도움 혹은 장애물을 제공한다. 이는 문항의 난이도에 영향을 미칠 수 있는 부분으로, 향후 후속 연구를 통해 더욱 정확히 검증되어야 할 것이다.

셋째, 컴퓨터 활용 빈도와 CBT 시행환경은 모드 효과에 큰 영향을 미치지 않지만, 학생의 컴퓨터 활용 효능감에 따라 모드 효과의 크기가 달라질 수 있다. 본 연구 결과, 컴퓨터 활용 효능감이 저수준과 고수준, 중수준과 고수준 학생 사이에서는 유의미한 점수 차이가 확인되었다. 이는 학생의 컴퓨터 활용 능력이 CBT 성취도에 영향을 미칠 수 있으므로 이를 고려해야 한다는 Bennett 외 (2008)의 주장을 뒷받침하는 결과이다. 이와 같은 본 연구 결과는 디지털-수학 역량(Geraniou & Jankvist, 2019)의 개념화 문제와 관련된 연구의 필요성 역시 보여준다. 기술공학적 기능을 활용하는 수학 평가에서는 타이핑, 탐구 소프트웨어 활용, 계산기 활용, 검색 등 다양한 컴퓨터 활용 능력이 필요할 수 있다. 컴퓨터 활용 효능감 또는 컴퓨터 활용 능력과 모드 효과 사이의 관계를 정확히 설명하기 위해서는 컴퓨터 활용에 요구되는 디지털 역량의 요소를 더욱 구체화하고 이들 요소와 수학 역량 사이의 관계를 연결지어 디지털 역량과 모드 효과 사이의 관계를 세밀하게 이해하기 위한 이론화가 진행될 필요가 있다. 한편, 학생의 컴퓨터 활용 빈도에 따라 유의미한 점수 차이가 나타나지 않았다는 결과는, 컴퓨터 경험(Ebrahimi et al., 2019; Eid, 2005; McClelland & Cuevas, 2020) 및 친숙도(Backes & Cowan, 2019; Bennett et al., 2008)가 CBT 성취도에 큰 영향을 주지 않는다는 선행연구를 지지한다. 이와는 달리 CBT 시행환경에 따른 정답률 차이가 유의미하지 않았다는 결과는 CBT 시행환경상의 문제가 성취도에 부정적인 영향을 줄 수 있다는 선행연구(Bennett et al., 2008; Kingston, 2008)와 대비되는데, 이와 같은 연구 결과는 우리나라와 같은 인프라 환경에서는 CBT 시행 환경이 CBT 결과에 큰 영향을 미치지 않는다는 것을 보여준다.

본 연구 결과로부터 다음과 같은 몇 가지 실천적인 시사점을 제안할 수 있다. 이와 같은 시사점은 대규모 평가뿐만 아니라 학교 수학에서의 안정적인 CBT 도입에도 기여할 수 있을 것이다. 먼저, 본 연구에서 확인된 기술공학적 기능의 반영 여부에 따른 성취도 결과의 차이는 기술공학적 기능의 도입이 점수에 영향을 미칠 수 있음을 보여주며, 향후 평가 내용과 역량에 맞춘 적절한 기술공학적 기능의 도입이 필요함을 보여준다. 기술공학적 기능을 도입함으로써, CBT를 시행하였을 때 학생들의 수학적 사고력을 보다 타당하게 측정할 수 있는 문항을 개발하는 것이 필요한 것이다(Hoogland & Tout, 2018). 한편, 교육과정 점검을 목표로 하는 대규모 평가와 달리 학교수학에서의 평가는 다양한 목표를 지니는 바, 목표에 대응하는 문항을 개발하는 것이 필요하다. 예컨대, 학습으로서의 평가를 목표로 하여 학생의 문제해결 참여를 고취하고자 할 경우 멀티미디어가 반영된 문항의 개발이 유용할 수 있으며, 정확한 수학적 지식의 측정을 위해 학생의 임의적인 추측을 방지하고자 할 경우 체크박스나 끌어놓기가 반영된 문항의 개발이 필요할 것이다. 나아가, 학교수학의 경우 대규모 평가에 비해 학생 개개인에 대한 평가를 좀 더 세밀하게 진행할 수 있으므로, 개별 학생의 역량과 수준에 적합한 문항 유형과 기술공학적 기능이 반영된 문항을 개발하는 것이 필요하다. 다음으로, 본 연구에서 제시한 학생 변인에 따른 성취도 결과의 차이는 수학 학습 능력 외의 부수적인 능력이나 상황이 학생의 성취도에 영향을 미칠 수 있음을 보여주며,

향후 이러한 부수적인 능력이 성취도 결과에 영향을 미치지 않도록 세밀한 시험 평가와 환경 구비가 필요함을 보여준다. 특히, 대규모 평가와 달리 단위 학교의 경우 학생의 컴퓨터 활용 효능감 등에 대한 선제적 확인과 조치가 가능하므로, 이를 고려한 문제 출제와 함께 CBT 시행을 위한 환경의 구비가 필요하다. 또한, 학교수학에서 CBT 시행을 위해서는 기술공학적 기능이 반영된 문항에 대한 교사의 이해와 개발이 요구되는 바, 학생 변인과 외부 환경 변인에 대한 통제에서 나아가 CBT와 그 문항에 대한 교사교육이 함께 이루어져야 할 것이다(Jung, 2022).

본 연구는 일부 한계점을 가지고 있다. 본 연구에서 분석한 국가수준 학업성취도 평가는 인원의 선발과 시행 일정이 교육부의 결정에 따르게 된다. 이로 인해 PBT, IBT, CBT의 참여 인원수와 시행 시기 및 시행환경에 다소 차이가 존재할 수밖에 없었으며, 본 연구의 한계점으로 남게 되었다. 향후 PBT와 IBT 및 CBT의 시행이 동시에 진행되고 이에 대한 결과가 제시된다면, 후속 연구를 진행함으로써 본 연구 결과를 정교화해 나갈 수 있을 것이다. 본 연구에서 평가 참여자는 고등학교 2학년 학생에 한정되었다. Jewsbury 외 (2020, p. 63)에 따르면 저학년에서 CBT 효과는 다르게 나타나므로, 고등학교 1학년 이하의 학생들이 참여하는 CBT에 대한 추가 연구를 통해 학교급과 학년별로 평가에 반영되는 기술공학적 기능 효과의 차이점을 확인하는 등 CBT의 효과를 정교화해 나갈 필요가 있다.

이와 같은 한계점에도 불구하고, 본 연구는 PBT, IBT, CBT의 비교를 통해 모드 효과에 대한 이해를 높이고 수학 문항에 기술공학적 기능이 반영됨에 따른 문항 특성과 난이도 변화의 가능성 등에 대한 분석을 제시하였다는 의의를 갖는다. 이는 기존 선행연구에서 찾아보기 힘든 것으로, PBT에서 CBT로의 전환 과정에서 나타날 수 있는 다양한 변인에 대한 고찰의 기회를 제공하는 한편, 수학 교과에서 CBT의 도입에 대한 긍정적인 가능성을 제공하였다. 또한, 평가 결과의 전반적인 경향성을 파악하는 것을 넘어(e.g. Jeong, 2014), 정답률과 문항 특성, 개인적, 환경적 요인을 연관지어 분석하여, 모드 효과의 원인에 대한 더 상세한 설명을 할 수 있었다. 본 연구에서의 통계적 분석 결과가 향후 수학 교과에서 수행되는 CBT의 정교화에 기여할 수 있기를 기대한다.

Endnote

[1]예비시행은 2020년 시행 예정이었으나, COVID-19 문제로 인하여 2021년 2월로 연기하여 시행되었다.

[2]IBT와 CBT의 경우 서답형 문항에 답하기 위해 서술형 답안 입력기와 수식 입력기가 반영되었는데, 이에 대해서는 기술공학적 기능이 반영된 것으로 보지 않았다.

[3]PBT 문항 중 CBT로의 전환 과정에서 문항 유형이 선택형에서 서답형으로 변경된 문항과 내용 영역이 기하 영역에서 함수 영역으로 변경된 문항이 각각 한 문항씩 존재하였다. 본 연구에서는 문항 유형의 변화와 내용 영역의 변화가 정답률 변화의 원인으로 작용할 수 있다고 판단하여, 이들 두 문항을 분석 대상에서 제외하였다.

[4]Table 4에서 확인할 수 있듯이, 이들은 CBT에서 각각 19번과 10번으로 제시되었으나, 본 연구에서는 서술과 이해의 편의성을 위해, 이들 문항을 PBT에 부여된 10번과 12번으로 칭하도록 한다.

[5]앞서 언급하였듯이, 응답 방식이 변경된 문항(4번)과 내용 영역이 변경된 문항(15번)은 분석 대상에서 제외하였다.

Acknowledgements

This paper is a revised and supplemented part of the contents of the ‘Advancing the test tools for Electronic National Assessment of Educational Achievement (eNAEA) (Lee et al., 2021)’ conducted by the Korea Institute for Curriculum and Evaluation.

References

Arslan, B., Jiang, Y., Keehner, M., Gong, T., Katz, I. R., & Yan, F. (2020). The Effect of Drag‐and‐Drop Item Features on Test‐Taker Performance and Response Strategies. *Educational Measurement: Issues and Practice, 39*(2), 96-106. https://doi.org/10.1111/emip.12326

Backes, B. & Cowan, J. (2019). Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Economics of Education Review, 68*(1), 89-103. https://doi.org/10.1016/j.econedurev.2018.12.007

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. *Journal of Technology, Learning, and Assessment, 6*(9), 1–38.

Bridgeman, B. (1992). A Comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253–271.

Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation, 62*(3), 1–9. http://dx.doi.org/10.1016/j.stueduc.2019.04.005

Clements, D. H. (2020). From exercises and tasks to problems and projects: Unique contributions of computers to innovative mathematics education. *The Journal of Mathematical Behavior, 19*(1), 9-47. http://dx.doi.org/10.1016/S0732-3123(00)00036-5

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.). *Assessment and teaching of 21st century skills*. Springer.

Drasgow, F., & Mattern, K. (2006). New tests and new Items: Opportunities and issues. In D. Bartram, & R. K. Hambleton (Eds.), *Computer-based Testing and the Internet: Issues and Advances* (pp. 59-76). John Wiley and Sons Ltd.

Ebrahimi, M.R., Toroujeni, S.M., & Shahbazi, V. (2019). Score equivalence, gender difference, and testing mode preference in a comparative study between computer-based testing and paper-based testing. *International Journal of Emerging Technologies in Learning, 14*(7), 128-143. https://doi.org/10.3991/ijet.v14i07.10175

Eid, G. K. (2005). An investigation into the effects and factors influencing computer-based online math problem-solving in primary schools. *Journal of Educational Technology Systems, 33*(3), 223-240. https://doi.org/10.2190/J3Q5-BAA5-2L62-AEY3

Fishbein, B., Martin, M. O., Mullis, I. V., & 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*(1), 1-23. https://doi.org/10.1186/s40536-018-0064-z

Geraniou, E., & Jankvist, U. T. (2019). Towards a definition of "mathematical digital competency". *Educational Studies in Mathematics, 102*(1), 29-45. https://doi.org/10.1007/s10649-019-09893-8

Haladyna, T., Downing, S., & Rodriguez, M. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5

Heyd-Metzuyanim, E., Sharon, A. J., & Baram-Tsabari, A. (2021). Mathematical media literacy in the COVID-19 pandemic and its relation to school mathematics education. *Educational Studies in Mathematics, 108*(1), 201-225. https://doi.org/10.1007/s10649-021-10075-8

Hoogland, K., & Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: pressures and tensions. *ZDM, 50*(4), 675-686. https://doi.org/10.1007/S11858-018-0944-2

Horkay, N., Bennett, R., 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. *The Journal of Technology, Learning, and Assessment, 5*(2). Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1641

Hu, L., Chen, G., Li, P., & Huang, J. (2021). Multimedia effect in problem solving: A meta-analysis. *Educational Psychology Review, 33*(4), 1717-1747. https://psycnet.apa.org/doi/10.1007/s10648-021-09610-z

Jeong, H. (2014). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology, 33*(4), 410-422. https://doi.org/10.1080/0144929X.2012.710647

Jewsbury, P., Finnegan, R., Xi, N., Jia, Y., Rust, K., Burg, S., Donahue, P., Mazzeo, J., Cramer, E. B., & Lin, A. (2017). *NAEP transition to digitally based assessments in mathematics and reading at grades 4 and 8: Mode evaluation study*. National Center for Education Statistics.

Jung, H. Y. (2022). Perceptions of students and teachers towards computer based test in National Assessment of Educational Achievement : Focused on high school mathematics test. *School Mathematics, 24*(1), 119-145. http://doi.org/10.29275/sm.2022.3.24.1.119

Karay, Y., Schauber, S., Stosch, C., & Schüttpelz-Brauns, K. (2015). Computer versus paper: Does it make any difference in test performance? *Teaching and Learning in Medicine, 27*(1), 57-62. https://doi.org/10.1080/10401334.2014.979175

Keng, L., McClarty, K., & Davis, L. (2008). Item-level comparative analysis of online and paper administrations of the texas assessment of knowledge and skills. *Applied Measurement in Education, 21*(3), 207-226. https://doi.org/10.1080/08957340802161774

Kingston, N. (2008). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22*(1), 22-37. https://doi.org/10.1080/08957340802558326

Lee, J. B., Kim, Y. H., Kim, J. S., Nam, M. W., Park, J. S., Park, J. H., Baek, J. H., Sung, K. H., Lee, S. R., Jang, E. S., & Jung, H. Y. (2021). *Advancing the test tools for Electronic National Assessment of Educational Achievement (eNAEA).* KICE.

Lee, S.-G., Ham, Y., Lee, J. H., & Park, K.-E. (2020). A case study on college mathematics education in untact DT era. *Communications of Mathematical Education, 34*(3), 201–214. https://doi.org/10.7468/JKSMEE.2020.34.3.201

Martin, M. O., von Davier, M., & Mullis, I. V. S. (2020*) Methods and Procedures: TIMSS 2019 Technical Report*. TIMSS & PIRLS International Study Center.

Martin, R. (2008). New possibilities and challenges for assessment through theuse of technology. In F. Scheuermann, & G. Pereira (Eds.). *Towards a research agenda on computer-based assessment* (pp. 6-9). Office for Official Publications of the European Communities.

Mcclelland, T., & Cuevas, J. (2020). A comparison of computer based testing and paper and pencil testing in mathematics assessment. *The Online Journal of New Horizons in Education, 10*(2), 78–89.

McDonald, A. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers and Education, 39*(3), 299-312. https://doi.org/10.1016/S0360-1315(02)00032-5

Noyes, J., & Garland, K. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*(9), 1352-1375. https://doi.org/10.1080/00140130802170387

Park, M. (2020). Applications and possibilities of artificial intelligence in mathematics education. *Communications of Mathematical Education, 34*(4), 545–561. https://doi.org/10.7468/JKSMEE.2020.34.4.545

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2009). Innovative items for computerized testing. In W. J. van der Linden (Ed.). *Elements of adaptive testing* (pp. 215-230). Springer. https://doi.org/10.1007/978-0-387-85461-8_11

Poggio, J., Glasnapp, D., Yang, X., & Poggio, A. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *The Journal of Technology, Learning, and Assessment, 3*(6).

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

Rose, J., Low-Choy, S., Singh, P., & Vasco, D. (2020). NAPLAN discourses: A systematic review after the first decade. *Discourse: Studies in the Cultural Politics of Education, 41*(6), 871-886. https://doi.org/10.1080/01596306.2018.1557111

Russell, M. (1999). *Testing on computers: A follow-up study comparing performance on computer and on paper*. Boston College.

Russell, M., & Tao, W. (2004). The influence of computer-print on rater scores. *Practical Assessment, Research, and Evaluation, 9*(10), 1-13.

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project, research and development series*. National Center for Education Statistics.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in E-Learning: A framework for constructing "Internet Constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning, and Assessment, 4*(6). Retrieved from https://www.learntechlib.org/p/103254/.

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading & Writing Quarterly, 23*(2), 139-159. https://doi.org/10.1080/10573560601158461

Schoenfeld, A. H. (2017). On learning and assessment. *Assessment in Education: Principles, Policy & Practice, 24*(3), 369-378. https://doi.org/10.1080/0969594X.2017.1336986

Shacham, M. (1998). Computer-based exams in undergraduate engineering courses. *Computer Applications in Engineering Education, 6*(3), 201-209.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing, &T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Lawrence Erlbaum Associates Publishers.

Spires, H. A., Paul, C. M., & Kerkhoff, S. N. (2019). Digital literacy for the 21st century. In D. B. A. Khosrow-Pou (Ed.). *Advanced methodologies and technologies in library science, information management, and scholarly inquiry* (pp. 12-21). IGI Global.

Taylor, C., Jamieson, J., Eignor, D. & Kirsch, I. (1998). *The Relationship between computer familiarity and performance on computer-based TOEFL test tasks*. Educational Testing Service.

Yoo, S. M. (2013*). SPSS statistical analysis for writing a thesis*. Slow& Steady.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67*(2), 219–238. https://doi.org/10.1177/0013164406288166

Zilles, C., West, M., Mussulman, D., & Bretl, T. (2018). Making testing less trying: Lessons learned from operating a Computer-Based Testing Facility. In *2018 IEEE Frontiers in Education (FIE) Conference*, IEEE.