교육평가연구
Journal of Educational Evaluation
2021, 제34권 제3호, 587-614
http://dx.doi.org/10.31158/JEEV.2021.34.3.587

프로세스 데이터를 활용한 응답속도와 정확도의 관계 분석: PISA 2015 미국과 한국의 사례를 중심으로

신 효 정*
Educational Testing Service

최근 PISA를 비롯한 국제학업성취도 평가가 컴퓨터 기반 검사로 전환되면서 단순 응답 결과뿐 아니라, 피험자가 문항에 응답하는 과정을 기록한 프로세스 데이터의 수집 및 활용이 가능해졌다. 프로세스 데이터 중에서도 응답시간은 가장 주목받고 있으며, 문항반응이론에 근거하여 응답속도와 정확도의 관계를 분석하는 다양한 심리측정모형이 개발되고 있다. 본 연구에서는 실험심리학에서 수립된 speed-accuracy tradeoff 개념과 van der Linden의 위계적 모형(2007)을 소개하고, 이를 응답속도가 문항 유형(구성형 또는 선다형)을 반영할 수 있도록 확장시켰다. 두 가지 심리측정 모형을 통해 한국과 미국의 PISA 2015 과학 영역 자료를 분석한 결과, 한국과 미국 모두 정확도가 높은 학생들이 더 천천히 응답하는 경향이 있는 것으로 나타났으며, 이러한 경향은 미국보다 한국에서 더욱 강하게 나타났다. 또한 선다형 문항보다 구성형 문항에서 천천히 응답할수록 정확도가 높아지는 경향성이 더 강하게 나타났다.

주제어 : 응답속도, 정확도, 프로세스 데이터, 다차원 문항반응이론 모형, PISA 2015, 문항 유형

* 교신저자: 신효정, Senior Measurement Scientist at Educational Testing Service, hj1424@gmail.com

I. 서 론

국제학업성취도 검사(international large-scale assessments)는 한국을 비롯한 전세계에서 증거 기반 교육 정책 수립의 주요 자료로 많은 관심을 받고 있다. 특히 PISA(Programme for International Student Assessment)는 2000년에 처음 OECD(Organisation for Economic Co-operation and Development)에 의해 30 여개국에서 시행된 이래, 가장 최근인 2018년에는 참여대상국이 80 여개국으로 늘어날 만큼 많은 주목을 받고 있다. PISA는 3년마다 참여국의 만 15세 학생들의 읽기(reading), 수학(mathematics), 과학(science) 세 영역에서의 역량을 측정하는 대표적인 국제 역량 평가이다. 주요 세 영역(core domains)은 매 주기마다 실시되며, 혁신적인 평가 영역이 매주기에 새롭게 한 영역씩 추가된다. 예를 들어 PISA 2015에는 협력적 문제해결 능력(collaborative problem solving), PISA 2018에는 글로벌 역량(global competence) 영역을 새롭게 시행하여 국제비교연구가 다양한 분야에서 이루어질 수 있도록 하였다. 이 밖에도 PISA는 성별과 과목에 대한 흥미 같은 학생들의 성취에 영향을 미칠 수 있는 교육맥락변인 특성, 즉 비인지적인(non-cognitive) 요소도 설문 문항으로 포함한다. 이를 바탕으로 참여국 학생들의 인지적 성취 및 성취에 영향을 미치는 다양한 변인에 대한 데이터베이스를 공공데이터(Public Use File; PUF)의 형태로 제공한다.

한국은 PISA가 처음으로 시행되던 2000년부터 참여하며 상위권의 성취도를 보여주고 있다. 우리나라의 PISA를 주관하는 한국교육과정평가원에서는 주기마다 PISA와 관련된 시행보고서 및 연구보고서를 발간하고 있으며, PISA 데이터를 활용한 연구물들도 최근 증가하고 있다. 양길석(2019)의 키워드 네트워크 분석을 통한 <교육평가연구>의 최근 연구 경향 분석에서도 드러나듯이 한국 연구자들 사이에서도 PISA 데이터를 활용한 분석은 많은 관심을 받고 있다. 우리나라 학생들의 학업성취에 미치는 다양한 요인들에 대한 분석(김혜숙, 함은혜, 2019; 손윤희, 박현정, 박민호, 2020; 조성민, 구남욱, 2020a; 한정아, 2020), 우리나라 학생들의 주기 간 학업성취 변화를 통해 교육정책 및 교수법에 대한 평가(김준형, 2019; 김현정, 구남욱, 2019), 우리나라 학생들의 다양한 교육맥락변인을 국제수준에서 비교하는 연구(김갑수, 2020; 김혜숙, 김한성, 김진숙, 신안나, 2017; 이빛나, 손원숙, 2019; 조성민, 구남욱, 2020b; 김난옥, 손원숙, 2019; 이희숙, 2019) 등 많은 연구들이 국내 교육 성취의 객관적인 평가와 함께 증거에 기반한 향후 정책 수립에 시사점을 제공해 왔다.

PISA에서 비교적 최근에 일어난 중요한 변화는 2015년부터 검사가 지필 방식에서 컴퓨터 기반 방식으로 전환된 것이다. 이러한 전환은 학생들의 응답률을 높이고 지필검사 방식으로는 측정이 불가능했던 영역을 테크놀로지에 기반한(technology-enhanced) 문항들을 통해 측정할 수 있도록 하였다(OECD, 2017). 이는 단순한 검사방식의 전환에 그치지 않고, 심리측정학

적으로도 다양한 진전을 가져왔다. 예를 들어, PISA 2015에서는 기존의 라쉬 모형(Rasch model) 뿐만 아니라 문항에 따라 2모수 로지스틱 모형과 혼용으로 사용하는 하이브리드(hybrid) 모형을 측정모형(measurement model)으로 사용함으로써 비교성(comparability)을 높이고 측정오차(measurement error)를 줄이고자 하였다. 더욱이 해당 주기, 특정 나라의 자료 분석에만 의존하지 않고 기존의 2000년부터 2012년까지 모든 참여국의 자료를 다집단 문항반응모형(김성훈, 2013; Bock & Zimowski, 1997)을 통해 분석하고 각 영역을 새롭게 척도화 하여, PISA의 가장 큰 목적이라고 할 수 있는 국가 간 그리고 주기 간 비교성을 높였다(von Daiver, Yamamoto, Shin et al., 2019). 또 다른 괄목할 만한 변화는 컴퓨터 기반 검사를 통해 학생들이 문항에 응답한 결과뿐만 아니라 그 과정에 대해서도 기록할 수 있게 되었다는 점이다. 이렇게 학생들이 문항에 응답하는 과정에 대한 기록은 흔히 프로세스 데이터(process data)라고 불리며, 대표적인 교육분야의 빅데이터로서 로그 파일(log file) 등의 형태로 저장된다. 이는 기존의 지필 검사로는 알 수 없었던(예: 문항에 대한 응답시간, 수학 영역에서 계산기의 사용 등), 학생들의 검사 참여와 관련한 귀중한 자료를 제공하여, 최근 전세계적으로 연구자들의 많은 관심을 받고 있다(von Davier, Khorramdel, He et al., 2019).

응답시간(response time)은 프로세스 데이터 중에서도 가장 주목받고 있는데, 이는 응답시간을 분석함으로써 검사 시행이 얼마나 올바르게 이루어졌는지 확인할 수 있을 뿐만 아니라(Yamamoto & Lennon, 2018), 학생들이 검사에 참여하는 수준(engagement)을 유추하거나(Goldhammer, Martens, Christoph et al., 2016), 다양하게 문제에 접근하고 해결하는 방식 등을 깊이 있게 이해하는 데 유용한 정보를 제공하기 때문이다(He, von Davier, Han, 2018). 특히 흥미로운 지점은 예외적으로 빠른 한국 학생들의 응답속도이다. 국제 수준의 PISA 데이터를 분석한 기술보고서(technical report Chapter 9)에 따르면, 한국은 모든 PISA 참여국 중에서도 가장 빠른 응답시간을 보이며, 이는 2015년과 2018년에서 공통적으로 나타나는 현상이다(OECD, 2017, 2020a). 2015년의 예를 들면, 주요 영역으로서 과학 영역 문제를 풀기 위해 60분의 시간이 주어지는데, 우리나라 학생들만 평균적으로 30분도 안되는 짧은 시간 안에 과학 영역 검사에 응답한 것으로 나타났다([그림 9.6], OECD, 2017). 이러한 짧은 검사 시간은 모든 인지적 영역에서 공통적으로 확인되었으며, 이는 성취 기준 상위권 나라들 중에서도 예외적으로 무척 빠른 편이다.

그러나 한국 학생들의 예외적으로 빠른 응답속도에 대한 깊이 있는 연구는 아직 부족한 실정이다. 예외적으로 김화경 등(2020)이 PISA 2018 자료에서 나타난 응답시간과 동작(action)을 중심으로 수학 기초학력 미달 학생의 특징을 분석하였고, 구자옥 등(2017)은 PISA 2015 자료에 나타난 한국의 빠른 응답속도는 여학생보다 남학생에게서 더욱 두드러진 것을 지적하며, 이를 남학생 성취 하락의 원인으로 지적하였다. 이러한 연구에는, 짧은 응답시간이 낮

은 검사 참여도를 의미하는 것으로, 학생들이 문제해결에 노력을 덜 들이고 결과적으로 낮은 학업 성취로 이어졌을 것이라는 가정이 깔려있다. 하지만 실제 PISA 검사에서 나타난 우리나라 학생들의 응답시간과 학업 성취의 관계에 대한 본격적인 분석은 이루어지지 않고 있다.

본 연구에서는 PISA 2015에 나타난 한국의 응답속도와 정확도의 관계를 실증적으로 분석하고자 한다. 응답속도와 정확도의 관계에서 가장 널리 알려진 이론적 토대는 실험심리학에서 정립된 speed-accuracy tradeoff이며, 교육측정 분야에서는 van der Linden(2007)이 제안한 위계적 모형(hierarchical model)이 응답속도와 정확도의 관계를 탐구하는 모형으로서 가장 널리 활용되어 왔다. 이하에서는 응답속도와 정확도의 관계를 이해하는 이론적 배경으로서 speed-accuracy tradeoff 및 van der Linden의 위계적 모형을 소개하고, 이를 문항 유형(구성형 또는 선다형)을 고려할 수 있는 모형으로 확장시킨다. 다음으로, 두 가지 심리측정 모형을 이용하여 PISA 2015 과학 영역의 한국과 미국 자료를 분석하고 그 결과를 비교한다. 미국은 보통 PISA에서 평균 수준의 학업 성취도를 보이며, 검사도구 및 문항 개발에서 밑그림(master version)을 제공할 때 면밀히 검토되는 국제수준 영어권 자료의 큰 부분을 차지한다. 또한 PISA를 비롯한 국제비교 연구에서 주로 분석의 대상이 되기 때문에 평균적인 비교 집단으로 적절하다고 판단되었다. 이를 통해 본 연구에서 답하고자 하는 연구 질문은 다음과 같다.

첫째, PISA 2015에 나타난 응답속도와 정확도의 관계는 한국과 미국에서 어떻게 나타나는가?

둘째, 응답속도와 정확도의 관계는 문항 유형에 따라 다르게 나타나는가?

Ⅱ. 응답속도와 정확도의 관계

1. Speed-accuracy tradeoff

우선 “응답속도”(response speed)와 “응답시간”(response time)의 용어 차이에 주목할 필요가 있다. 정확도(accuracy 또는 ability)가 문항에 대한 정답 여부로 측정되는 것과 같이, 속도는 문항에 대한 응답시간을 토대로 측정된다. 그래서 거의 모든 심리측정 연구에서는 응답속도와 정확도는 피험자의 특성을 나타내는 잠재변수(latent variable)로, 응답시간과 정답 여부는 피험자의 각 문항에 대한 관찰변수 (observed variable)로 설정한다. 특히 응답속도가 어떤 방향성을 지니고 있는지에 따라 해석에 주의를 기울일 필요가 있다. 응답시간이 오래 걸린다는 것은 응답속도가 느리다고 해석될 수 있으므로, 응답“시간”을 통해 응답“속도”를 명세화

할 때, 잠재변수의 방향성에 따라 응답속도는 빠르게(-) 또는 느리게(+) 해석될 수 있다 (자세한 설명은 III.2. 분석방법 참고).

응답속도와 정확도의 관계를 이해하는 관점 중 가장 널리 알려진 것은 실험심리학을 통해 확립된 speed-accuracy tradeoff이다(Luce, 1986; Heitz, 2014). 이는 tradeoff라는 표현에서도 알 수 있듯이, 피험자가 어떤 문항에 응답할 때 빠르게 응답할수록 정확성이 떨어지며 천천히 응답할수록 정확성이 높아진다는 것이다. 따라서 speed-accuracy tradeoff는 기본적으로 응답"속도"와 정확도 간의 부적 상관관계 또는 응답"시간"과 정확도 간의 정적 상관관계로 도식화된다(van der Linden(2007)의 [그림 1] 또는 Heitz(2014)의 [그림 2] 참고). 중요한 점은 실험심리학에서 이야기하는 speed-accuracy tradeoff는 한 개인이 특정 문항에 응답하는 상황을 가정할 때 적용된다는 것이다. 이러한 기본적인 인식을 바탕으로 PISA 데이터에 나타난 응답속도와 정확도의 관계를 이해할 때 두 가지 측면을 고려할 필요가 있다.

우선, speed-accuracy tradeoff는 한 피험자 개인의 상황(within-person)을 가정할 때 성립하는 것이며, 이를 어떤 집단(population)의 수준에서 이해할 때에는 전혀 다른 관계가 나타날 수 있다. 특히 집단 수준에서 응답속도와 정확도는 매우 복잡한 관계를 보이는데, 실제로 여러 경험적 연구들에서 오답 응답자들의 응답시간이 더 긴 경우가 관찰되기도 하였다(Swanson, Case, Ripkey, Clauser, & Holtman, 2001). 이는 정답을 알지 못하는 경우에 생각하는 시간이 더 오래 걸리는 상황에 해당하며, 관찰점수를 이용해 정확도를 계산한 경우에도 흔히 나타날 수 있다. 즉, 피험자 특정 개인을 가정한 상황(개인의 능력치를 고정한 상황)에는 speed-accuracy tradeoff에 따라 거의 모든 문항에서 응답속도와 정확도 간에 부적 상관관계가 나타날 것이라고 예상할 수 있지만, 이를 피험자 집단(능력치가 다양한 상황)으로 확장시키거나, 정확도를 측정하는 데 관찰점수를 이용하면 정적 상관관계나 거의 0에 가까운 상관관계를 보일 수도 있다는 것이다. 다음으로 주의해야 할 점은, speed-accuracy tradeoff가 특정 문항에 대한 응답 상황을 가정한 것이며, 다양한 난이도와 특성이 다른 여러 문항으로 구성된 검사 수준에서는 적용되지 않을 수 있다는 것이다. 집단 수준에서 speed-accuracy 관계가 다양한 피험자 능력 수준에 따라 여러 면모를 보일 수 있듯이, 여러 문항으로 구성된 검사에서는 문항난이도나 어떤 영역을 측정하느냐에 따라 다양한 관계가 성립될 수 있기 때문이다. 예를 들어, 지문을 읽고 추론하여 응답해야 하는 다소 복잡하고 어려운 문항의 경우에는 응답시간이 길어질수록 정확도가 높아질 수 있지만, 컴퓨터의 기본적인 작동만 할 줄 알면 되는 아주 쉬운 문항의 경우에는 응답시간이 짧다고 해서 정확도가 낮아지거나 그 반대 상황이 적용되지는 않는다(Dodonova and Dodonova, 2013; Goldhammer et al., 2014).

이를 토대로 보면, PISA와 같이 여러 문항으로 이루어진 대규모 피험자집단 검사 상황에서의 응답속도와 정확도의 관계는 기존의 speed-accuracy tradeoff에서 예측되는 것과 다른 패

턴이 나타날 수 있으며, 따라서, 이에 대한 경험적인 분석이 필요하다. 참고로 PISA 2015 기술보고서 [그림 9.6]의 경우는 학업성취와 응답시간의 자료를 참여국, 검사 수준으로 갈무리한 것이기 때문에 기술 통계치로서의 효용성을 가질 뿐, speed-accuracy tradeoff의 관점으로 이해해서는 안 된다. 즉, 한국의 빠른 응답속도가 낮은 정확도를 초래하는 것이므로, 국가 수준에서 응답시간이 길어진다고 하여 학업 성취가 올라갈 것이라고 기대하기는 어렵다는 것이다. 따라서 한국에서의 응답속도와 정확도의 관계가 어떻게 나타나는지 경험적 분석을 통해 살펴보고, 이를 다른 나라에서 보이는 양상과 비교하는 것은 한국 학생들이 PISA 검사 상황에서 보이는 빠른 응답속도를 이해하는 데 도움이 될 것이다. 이하에서는 본 연구에서 적용하고자 하는 수리적 모형을 설명한다.

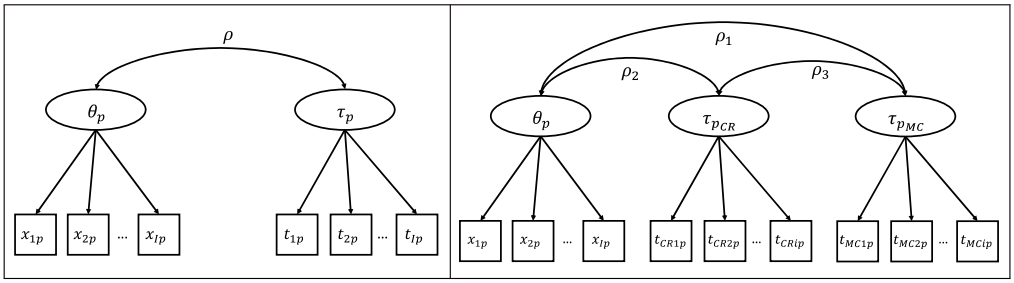
2. 응답속도와 정확도에 대한 수리적 모형

1) van der Linden의 위계적 모형

응답속도를 종속 변인 또는 설명 변인으로 활용한 심리측정 모형은 다양하게 제안되어 왔으며, De Boeck & Jeon(2019)에 잘 정리되어 있다. 이 중에서도 심리측정 분야에서는 응답속도와 정확도의 관계를 공동으로 모형화하는 방법으로 van der Linden의 위계적 모형(2007)이 가장 널리 활용되고 있다. van der Linden의 위계적 모형은 문항들에 대한 응답"시간"과 정답 여부를 통해 각각 응답"속도"와 정확도라는 두 개의 서로 상관된(correlated) 잠재변수를 모형화하는 다차원 문항반응이론 모형(multidimensional item response theory model)이라고 할 수 있다. van der Linden의 위계적 모형은 크게 두 부분으로 이루어져 있는데, 응답속도와 정확도에 대한 각각의 측정모형이 상위 레벨에서 상관관계를 통해 연결되어 있다. 이를 도식화하면 [그림 1]의 왼쪽과 같다.

우선 흔히 잠재능력(latent ability)로 일컬어지는 정확도(θ_p)는 I개의 문항으로 이루어진 검사에 정답 또는 오답(1 또는 0)으로 응답함으로써 측정된다. 여기에서 학생 p의 문항 i에 대한 응답 점수는 x_{ip}로 표시되었다. 이때 잠재변수인 정확도와 관찰변수인 개별 문항에 대한 응답을 연결하는 측정모형으로 van der Linden(2007)에서는 3모수 정규 오자이브 모형(3-parameter normal ogive model)이 사용되었지만, 라쉬모형, 2모수 로지스틱 모형 (2-parameter logistic model; 2PL) 등이 다양하게 사용될 수 있다. 다음으로 학생 p의 응답속도(τ_p)는 개별 문항 i에 소요된 시간(t_{ip})을 통해 측정된다. 이때 잠재변수인 응답"속도"와 관찰변수인 응답"시간"의 관계에는 주로 로그노멀 모형(lognormal model)이 사용된다. 응답속도의 측정모형에서는 시간난이도(time intensity)와 시간변별도(time discrimination)의 모수가 추정된다(두 모수에 대한 자세한 설명은 수식과 함께 제시된 아래 III. 2 분석 방법 참고). 응답시간의 분포를 다

른 형태(Weibull)로 가정하는 연구들도 있지만(Loeys, Rosseel, Baten, 2011), 응답시간을 로그 변환(log transformation)하여 정규분포를 따른다고 가정하는 경우가 일반적이다.



〔그림 1〕 van der Linden의 위계적 모형(M1; 왼쪽)과 문항 유형을 고려하도록 발전시킨 모형(M2; 오른쪽)

이 모형의 특징은 각각의 응답속도 모형과 정확도 모형이 상위 레벨에서 연결된다는 것이다. 특히 피험자 모수인 정확도(θ_p)와 응답속도(τ_p)는 다변량 정규분포(multivariate normal distribution)를 따르는 것으로 가정하는 경우가 일반적이며, 여기에서 두 잠재변인 간의 상관관계(latent correlation; ρ)가 관심 모수이다. 피험자 모수가 상관관계를 지닐 수 있도록 한 것처럼 기존의 van der Linden(2007) 모형에서는 문항 모수들 역시 다변량 정규분포를 따른다고 가정함으로써 상관관계를 지닐 수 있도록 하였다. 즉, 정확도에서 추정된 문항난이도와 응답속도에서 추정된 시간난이도가 서로 상관될 수 있도록 하여, 어려운 문항에서 평균적으로 더 오랜 시간이 소요되는지 살펴볼 수 있다.

2) 문항 유형을 고려하여 발전시킨 모형

[그림 1]의 왼쪽에서 볼 수 있는 것과 같이 응답속도와 정확도에 관한 기존의 위계적 모형은 두 개의 잠재변인들을 각각 일차원적으로 규정하였다. 즉, 측정하고자 하는 구인(정확도)이 일차원적이며, 검사 상황에서 보이는 응답속도 또한 일차원적이라고 가정하는 것이다. 이는 지금까지 van der Linden의 모형이 사용되고 적용된 경우가 대부분 선다형 문항(multiple choice items; MC)으로 구성된 검사에서 사용되었기 때문으로 보이며, 선행연구 중 구성형 문항(constructed-response items; CR)과 선다형 문항이 혼재하는 검사(mixed- format tests)에 적용된 경우는 거의 없다. 다만, PISA 데이터의 응답시간만을 이용하여 척도화한 Shin 등(2020)의 연구에서 선다형 문항과 구성형 문항의 응답속도는 일차원적이라기보다는 문항 유형에 따라 2차원으로 보는 것이 더 적합함을 밝혔다. 따라서 PISA를 비롯해 선다형 문항과 구성형 문항이 골고루 섞여 있는 검사 상황에서의 적용 가능성을 고려할 때, van der Linden의 위계적 모

형(2007)을 문항 유형을 고려할 수 있는 형태로 확장시킨 모형이 필요하다.

그중 한 가지 방법은 응답속도 잠재변수를 문항 유형에 따라 선다형 문항에서의 응답속도($\tau_{p_{MC}}$)와 구성형 문항에서의 응답속도($\tau_{p_{CR}}$)로 구별하여 반영하는 것이다. 이를 도식화하여 표현하면, [그림 1]의 오른쪽과 같으며, 기존의 2차원 모형이 3차원으로 확장된 것을 알 수 있다. 이 모형의 가장 큰 장점은 피험자의 응답속도를 문항 유형에 따라 다르게 모형화하여, 선다형 문항에서는 빠르게 응답하고 구성형 문항에는 천천히 응답하는 경우, 혹은 그 반대 상황 등을 고려할 수 있게 된다는 것이다. 또한 각 문항 유형에서의 응답속도와 정확도의 관계가 다르게 추정할 수 있어서 구성형 문항에서의 응답속도와 정확도의 관계가 선다형 문항에서의 관계보다 더 강한 상관관계를 보이는지 알 수 있다. 이와 같이 모형이 확장될 경우, 한 개의 관심 모수였던 응답속도와 정확도간의 잠재상관관계(ρ)가 세 개로 늘어난다. 구체적으로, 정확도와 선다형 문항에서의 응답속도(ρ_1), 정확도와 구성형 문항에서의 응답속도(ρ_2), 선다형 문항에서의 응답속도와 구성형 문항에서의 응답속도(ρ_3)로 나누어진다. 이때 응답속도가 문항 유형을 반영한 것처럼 정확도 역시 문항 유형에 따라 잠재변수를 나누는 것도 고려할 수 있을 것이다. 그러나 Shin, von Davier, Yamamoto (2019)은 PISA 데이터의 채점자 효과를 분석하기 위해 정확도를 문항 유형에 따라 나누어 분석한 결과, 정확도는 여전히 매우 일차원적임(구성형 문항 점수와 선다형 문항 점수의 상관관계가 0.95 이상)을 보고하였다. 따라서 본 연구에서는 문항 유형이 영향을 미치는 잠재변수는 응답속도로 제한하기로 한다.

Ⅲ. 연구 방법

1. 연구 대상

1) PISA 2015 프로세스 데이터

미국의 대표적인 국가수준 학업성취도 평가인 NAEP(National Assessment of Educational Progress)의 경우에는 프로세스 데이터를 연구자가 요청하는 경우에 한하여 제공하고 있으나, PISA의 경우에는 2015년 이후 프로세스 데이터 관련 변인을 공공데이터에 제공하여 연구자들의 활용도가 매우 높다. <표 1>은 컴퓨터 기반 검사가 시작된 2015년 이후 공공데이터에 제공되는 문항 수준의 변수들을 정리한 것이다. 이중에서 음영 처리된 부분이 흔히 프로세스 데이터로 분류되며 2015년 이후부터 제공되기 시작한 변수들로, 특히 응답시간에 대한 관심이 높아 연구가 활발하게 이루어지고 있다.

〈표 1〉 PISA 2015와 2018 공공데이터에 제공되는 문항 수준 변수 목록

문항 수준 변인		PISA 2015	PISA 2018
실제 응답 (“R”로 끝나는 변인)*	학생들이 고른 선택지 (e.g., A, B, C, D)	√	√
채점자에 의해 부여된 점수 (“C”로 끝나는 변인) 또는 기계 채점된 점수 (“S”로 끝나는 변인)	0 : 오답 1: 정답(이분문항) 또는 부분 정답(다분문항) 2: 정답 (다분문항)	√	√
응답시간1 (“T”로 끝나는 변인)	학생이 “마지막으로” 해당 문항을 보게 되었을 때 응답하는 데 소요된 시간(milliseconds 단위로 기록)	√	√
동작 횟수 (“A”로 끝나는 변인)	마우스 클릭, 마우스 더블 클릭, 키 프레스, 드래그 앤 드랍 등 학생이 문항에 응답하며 보인 동작의 횟수	√	√
첫 동작까지의 시간 (“F”로 끝나는 변인)	학생이 첫 동작을 하기까지 걸린 시간 (milliseconds 단위로 기록)		√
응답시간2 (“TT”로 끝나는 변인)	학생이 해당 문항을 보고 응답하는 데 소요된 시간의 총합 (milliseconds 단위로 기록)	√	√
방문 횟수 (“V”로 끝나는 변인)	해당 문항을 방문한 횟수	√	√

* 선다형 문항에서 학생들이 고른 선택지는 알 수 있으나, 구성형 문항에 대한 학생들의 응답은 공공데이터에 제공되지 않는다.

PISA 컴퓨터 기반 검사에서는 한 화면에 한 문항만 제시되도록 하는 경우가 거의 대부분이다. 이에 따라 PISA 2018 기술보고서의 부록(Annex K)에 기술된 바와 같이, 공공데이터에 제공된 문항의 응답시간은 ‘학생이 화면에서 해당 문항을 보기 시작하고 다음 문항으로 넘어가기 이전까지의 시간’으로 정의된다(OECD, 2020a). 따라서 한 화면에 두 개 이상의 문항이 제시된 경우, 개별 문항에 대한 응답시간으로 분리하는 것은 거의 불가능하다. 주의할 점은 PISA 2015와 2018의 경우, 문항 수준 변인으로 두 종류의 응답시간이 제공되고 있다는 것인데, “T”로 끝나는 응답시간1의 경우는 학생이 해당 문항을 “마지막으로” 보았을 때 그 화면에서 머무른 시간을 계산한 것이다. 즉, 인지적 영역 검사에 응답하면서 이전 문항으로 돌아갔다가 다시 해당 문항으로 돌아왔을 경우(multiple visits), 응답시간1은 마지막 방문에서의 화면 시간만 계산하기 때문에 해당 문항에 들인 시간의 일부만 포함된다. 반면, 비교적 최근에 제공된 “TT”로 끝나는 응답시간2의 경우에는 해당 문항을 방문한 횟수가 2회 이상

일 경우, 각 방문에서 해당 문항을 본 시간을 합산하여 계산하기 때문에, 응답속도와 정확도의 관계를 파악하는 데 더 적합하다. 따라서 본 연구에서는 PISA 2015와 연계된 별도의 파일에 제공된 “TT”로 끝나는 응답시간2를 PISA 2015 공공데이터의 응답 점수와 결합하여 분석에 사용하였다. 이 응답시간2 변인은 방문횟수 변인과 함께 최근에 추가로 제공된 프로세스 데이터 변인들에 속하며, 기존의 공공데이터가 아닌 별도의 파일로 PISA 웹사이트를 통해 다운로드받을 수 있다(링크는 참고문헌에 제시). PISA 2018의 별도 파일은 2020년 9월에 제공되었으며 (OECD, 2020c), PISA 2015의 별도 파일은 2020년 11월에 제공되었다(OECD, 2020b).

2) 분석 자료

본 연구에서는 PISA 2015 본검사에 참여한 한국과 미국의 자료를 분석에 활용하였다. 2015년에는 과학 영역이 주영역으로, 참여한 모든 학생들은 전체 2시간의 인지적 영역 검사 중 1시간을 과학 영역 검사에 임하였다. 과학 영역은 총 184개의 문항으로 구성되었는데, 각 문항은 12개 중 하나의 문항그룹(cluster)으로 배치되었다. 이때 6개의 문항그룹(S01-S06; trend clusters)은 이전 검사에서도 실시되어 주기 간에 공통척도를 구성할 수 있도록 기능하는 “척도연계 문항”(linking 또는 trend item)들로 구성되며, 나머지 6개의 문항 그룹(S07-S12; new clusters)은 새로 확장된 평가 영역 구조에 따라 2015년에 새로 개발되고 실시된 문항들로 구성된다. 척도연계 문항들은 지필 방식으로 시행되었던 문항들이 컴퓨터 기반으로 시행되는 것이며, 새롭게 개발된 문항들은 지필 방식으로는 불가능했던 시뮬레이션이나 시나리오에 근거한 테크놀로지 기반 문항들이 대부분이다. 학생들이 각 문항 그룹에 30분 이내에 응답할 것으로 가정하여, 총 12개의 문항그룹 중 2개의 문항그룹이 위치 효과(position effects)를 고려하여 각 검사도구(form)에 골고루(balanced) 배치된다. 이러한 검사 설계는 NAEP과 PISA과 같은 집단수준 점수보고 검사(large-scale assessments for group-level reporting)에서 전통적으로 널리 쓰인 방법으로 balanced incomplete block 검사 설계라고 불린다(Messick, Beaton, & Lord, 1983). 이에 따라 과학 영역이 포함되어 있는 검사도구가 전체 396개로 구성되었다(금융 소양 영역 제외). 또한 2015년도에 새롭게 실시된 문항에 대한 모수 추정의 정확성을 높이기 위해 약 두 배수의 학생들이 새로운 문항 그룹(S07-S12)에 응답하도록 안배되었다. 한국과 미국에서도 PISA 검사설계(assessment design)에 따라 표집된 학생들에게 396개의 검사도구 중 한 개가 무선적으로 할당되었다. 이를 학생 개인의 입장에서 보자면 과학 영역 184개 전체 문항 중 무선 할당된 2개의 문항그룹(평균적으로 31개 문항)에 응답한 셈이 된다.

이와 같은 검사 설계에 대한 이해를 바탕으로, 본 연구에서는 분석에 필요한 충분한 사례수(sample size)를 확보하고 결측치를 효과적으로 통제하기 위해, 12개의 문항그룹을 각각의

분석단위(analysis unit)로 보았다. 즉, 396개의 검사도구 중 해당 문항그룹을 포함한 검사도구에 응답한 학생들을 한 집단으로 취급하여 분석에 사용하였다. 예를 들어, 문항그룹(S08)에 응답한 학생들은 첫 번째 검사 시간(first hour)에 응답하였거나, 두 번째 검사 시간(second hour)에 응답하였거나, 같은 시간 내에 해당 문항그룹을 먼저 응답하였거나(S08-S07) 나중에 응답하였거나(S07-S08), 동일한 분석단위로 포함하였다. 이는 응답속도와 정확도의 관계를 위한 모형을 추정하는 데 어느 문항그룹을 먼저 검사했는지 위치 효과가 없었다고 가정하는 것으로, 한국과 미국의 거의 모든 학생들이 1시간 이내에 검사를 종료한 것을 보면 합리적인 가정이라고 볼 수 있다.

이와 같은 자료 준비과정을 거쳐 한국과 미국에서 각각 12개의 고유한 분석단위를 구성하였으며, 동일한 분석을 각 분석단위에서 시행하였다. 12개의 분석단위에서 각 단위에 포함된 학생 수와 문항 수의 분포는 <표 2>에 정리되어 있다.

〈표 2〉 문항그룹에 따른 12개의 분석단위에서 각 단위에 포함된 학생 수와 문항 수의 분포

	문항 그룹	전체		한국 (개별 분석단위)				미국(개별 분석단위)			
		한국	미국	최소	평균	최대	표준편차	최소	평균	최대	표준편차
학생 수	S01-S06	5,564	5,597	610	615	619	3.2	611	627	660	17.6
	S07-S12			1,215	1,241	1,263	19.3	1,218	1,251	1,306	33.9
문항 수	S01-S06	181	183	9	14	18	3.9	9	14.2	18	4.1
	S07-S12			16	16.2	17	0.4	16	16.3	17	0.5

공공데이터에는 한국 학생 총 5,581명, 미국 학생 총 5,677명의 자료가 제공되었으나, 본 연구에서는 해당 분석단위에서의 응답 점수와 응답시간 기록이 없는 경우를 제외하여 5,564명의 한국 학생, 5,597명의 미국 학생의 자료를 분석에 포함하였다. 전체 자료를 문항그룹에 따른 12개의 분석단위로 나누었을 때 한국의 경우, 척도연계 문항그룹(S01-S06)에는 610~619(평균 615)명의 학생들이 포함되었으며 새롭게 개발된 문항들로 구성된 문항그룹(S07-S12)에는 약 두 배인 1,215~1,263(평균 1,241)명의 학생들이 포함되었다. 미국의 경우에도 분석단위 당 포함된 학생들의 수가 한국과 비슷하나, 표준편차는 한국보다 약간 큰 경향을 보였다. 다음으로 문항의 분포를 보면, 과학 영역 전체 184개 문항 중 동일한 화면에 제시된 문항 1개를 제외하고 183개의 문항을 분석 대상으로 삼았고, 한국의 경우에는 국제수준의 데이터 분석에서 추가적으로 2개의 문항이 적합하지 않은 것으로 나타나(PISA2015 기술보고서 Annex A; OECD, 2017), 본 연구의 모든 분석에서도 제외하였다. 분석단위 당 문항

의 분포를 보면 최빈치(mode)가 16으로 대부분의 문항그룹이 16~18개의 문항을 포함하고 있었으며, 척도연계 문항그룹 중 S04(10개), S05(13개), S06(9개)가 이례적으로 적은 수의 문항을 포함하고 있었다. 이때 분석단위와 관련하여 해석에 주의해야 할 점은 전체 문항들은 상호배반적으로(mutually exclusive) 분석단위에 나누어졌지만, 학생들의 경우에는 동일한 학생이 두 개의 분석단위에 포함되었다는 점이다. 이는 앞서 설명한대로 학생 개인이 2개의 문항그룹에 응답하도록 설계된 검사 설계를 반영한 결과이며, 학생 개인의 응답속도가 검사 중에 일정하게 유지된다는 가정을 필요로 한다. van der Linden(2007)의 기존 모형 역시 같은 가정에 기반하고 있다.

2. 분석 방법

본 연구에서는 응답속도와 정확도의 관계를 규명하기 위하여 두 가지 심리측정 모형을 분석에 이용한다. 구체적으로, 앞서 [그림 1]에 소개된 van der Linden의 위계적 모형(왼쪽; 모형 1)과 이를 문항 유형을 반영하도록 발전시킨 모형(오른쪽; 모형 2)을 분석한다. 본 연구에서 분석하고자 하는 두 가지 심리측정 모형을 수식으로 표현하면 아래와 같다.

정확도에 관한 모형(모형1, 모형2): $P(x_{ip}=1|\theta_p,\ a_i,\ b_i)=\Phi(a_i\theta_p-b_i)$ (1)

응답속도에 관한 모형(모형1, 모형2): $\ln(t_{ip}\mid\tau_p)=\gamma_i\tau_p+\beta_i+\epsilon_{ip},\quad \epsilon_{ip}\sim\ N(0,\ \alpha_i)$ (2)

피험자 모수에 대한 통계적 가정(모형1): $(\theta,\ \tau)\sim MVN\left((\mu_\theta,\ \mu_\tau),\ \begin{pmatrix}\sigma_\theta^2 & \sigma_{\theta\tau}\\ \sigma_{\theta\tau} & \sigma_\tau^2\end{pmatrix}\right)$ (3)

피험자 모수에 대한 통계적 가정(모형2):

$$(\theta,\ \tau_{MC},\ \tau_{CR},)\sim MVN\left((\mu_\theta,\ \mu_{\tau_{MC}},\ \mu_{\tau_{CR}}),\ \begin{pmatrix}\sigma_\theta^2 & \sigma_{\theta\tau_{MC}} & \sigma_{\theta\tau_{CR}}\\ \sigma_{\theta\tau_{MC}} & \sigma_{\tau_{MC}}^2 & \sigma_{\tau_{MC}\tau_{CR}}\\ \sigma_{\theta\tau_{CR}} & \sigma_{\tau_{MC}\tau_{CR}} & \sigma_{\tau_{CR}}^2\end{pmatrix}\right) \quad (4)$$

이 연구에서는 본래 van der Linden(2007)에서 제안된 위계적 모형에서 약간 변형된 형태를 사용하였다. 구체적으로, 수식 (1)에 표현된 바와 같이 정확도는 프로빗 링크(probit link)를 통한 2모수 노멀 오자이브 모형을 이용해 모형화했으며, 이에 따라 각 문항의 문항변별도(a_i)

와 문항난이도(b_i)를 추정하였다. 위의 수식은 이분문항(binary)의 경우를 예로 들어 표현하고 있으며, 3개의 점수(0,1,2)가 가능한 다분문항(polytomous)의 경우에는 한 개의 문항 모수가 추가적으로 추정되었다. 다음으로 응답속도(수식 2)는 van der Linden의 모형을 따라 로그변환된 문항별 응답시간의 정규분포를 가정하는 로그노멀 모형을 사용하였다. 주의해야 할 점은 본 모형의 수식에서 문항에서의 응답시간(t_{ip})과 응답속도(τ_p)의 관계가 가법으로(+) 표현되었다는 것이다. 이에 따라 학생의 응답시간이 길어질수록 응답속도(slowness)도 높은 값을 가지는 것으로 해석되었다. 또한 정확도가 문항별로 두 개의 문항 모수를 추정하는 것처럼, 응답속도 모형에서도 문항별로 두 개의 문항 모수를 추정하게 된다. β_i는 시간난이도로서, 추정치가 높을수록 피험자들이 해당 문항에 평균적으로 소요하는 시간이 길어진다는 것을 의미한다. γ_i는 시간변별도로서 응답속도의 문항변별도에 해당하는 개념으로, 추정치가 높을수록 해당 문항이 피험자의 응답속도에 미치는 영향이 크다는 것을 의미한다. 여기에서 ϵ_{ip}는 로그변환된 응답시간에서 응답속도, 시간난이도, 시간변별도에 따라 예측하고 남은 잔차(residual)를 의미하며, 잔차는 개별 문항에 따라 평균 0, 분산 α_i인 정규분포를 따를 것으로 가정된다.

마지막으로 응답속도 모형에서 유의할 점은 원래 van der Linden의 모형에서는 라쉬모형과 비슷하게 시간변별도가 모두 1로 고정된 형태가 분석되었지만, Molenaar, Tuerlinckx와 van der Maas(2015)에서처럼 개별 문항의 시간변별도가 다르게 추정될 수 있도록 일반화되었다는 것이다. 즉, 본 연구에서는 정확도에 2 모수 모형을 이용한 것과 같이 응답속도에서도 시간난이도와 시간변별도를 모두 추정하는 일반화된 형태의 모형을 사용하였다. 다만 본래 van der Linden 연구에서는 문항 모수들이 다변량 정규분포를 따르며 문항 모수 간의 상관관계가 고려되었지만, 본 연구에서는 피험자 모수 간의 상관관계(수식 3 또는 4)만을 고려하였다. 즉, 본 연구에서 피험자 모수는 다변량 정규분포를 따르는 무선 효과(random effects)로, 문항 모수는 특정 분포를 고려하지 않은 고정효과(fixed effects)로 설정되었다.

분석에는 Mplus 8.0(Muthén & Muthén, 2017)이 사용되었다. 위 두 가지 모형으로 한국과 미국 각 나라에서 구성된 12개의 분석단위를 반복적으로 분석하기 위해 R의 MplusAutomation 패키지를 사용하였다(Hallquist & Wiley, 2018). 또한 모든 분석은 표집 가중치(sampling weights)를 반영하여 수행되었다. 표집 가중치는 표집 설계(sampling design)와 미응답 비율(non-response rates)을 반영하는 것으로(Rust & Johnson, 1992), 표본 조사로부터 모집단에 대한 추론을 가능하게 한다. 마지막으로 추정에는 평균 및 분산조정 가중 최소제곱법(mean and variance adjusted weighted least squares; WLSMV)이 사용되었으며(김수영, 2016; Asparouhov & Muthén, 2007), 반복 횟수 및 수렴 기준은 초기값(default)을 따랐다. 이 추정법

은 표준오차 추정의 편향이 거의 없는 매우 강건한 방법으로, Mplus의 범주형 내생변인 분석에도 기본 추정방식으로 선택되어 있다(Muthén & Muthén, 2017). 특히 본 연구에서와 같이 범주형(categorical) 변수인 응답점수와 연속형(continuous) 변수인 응답시간의 상관관계 추정에 가장 적합한 방식으로 알려져 있다(Asparouhov & Muthén, 2007). 사용된 Mplus 코드는 부록에 제시되었으며, R 코드는 저자에게 요청하는 경우 제공될 수 있다.

Ⅳ. 연구 결과

1. PISA 2015에 나타난 응답속도와 정확도의 관계

첫 번째 연구 질문인 "PISA 2015에 나타난 응답속도와 정확도의 관계는 한국과 미국에서 어떻게 나타나는가?"에 대해서는 van der Linden(2007)의 위계적 모형에 상응하는 모형 1의 분석을 통해 탐구하였다. 앞서 설명한 대로 각 문항그룹에 따라 구성된 한국과 미국의 12개의 분석단위마다 분석이 시행되었지만, 예외적으로 척도연계 문항그룹 중 하나인 S06은 9개의 문항(구성형 문항 1개, 선다형 문항 8개)만으로 이루어져 있어 한국과 미국 자료에서 모두 수렴된 결과를 얻을 수 없었다. 따라서 이하에서는 그 외 11개 분석단위에서 나타난 결과를 요약하며, 비교를 용이하기 위해 정확도와 응답속도 두 잠재변인의 평균이 각각 0으로 고정되고 분산도 1로 표준화된 경우의 추정치를 보고하고 결과를 해석하였다.

가장 관심이 되는 모수는 정확도와 응답속도의 관계를 상위 수준인 피험자 모수에서 추정하는 잠재 상관관계(ρ)이다. <표 3>은 한국과 미국의 11개의 분석단위에서 분석된 잠재 상관관계 추정치를 정리한 것이다. 이때 수식 (2)에 표현된 것과 같이 응답시간과 응답속도의 관계는 가법으로 모형화되어 응답속도는 slowness로 해석되어야 한다. 이에 따라 두 나라의 모든 분석단위에서 잠재 상관계수는 모두 정적 상관관계를 보여, 정확도가 높은 학생일수록 응답하는 데 더 오랜 시간이 걸리는 경향이 있다고 해석할 수 있다. 이는 결국 speed-accuracy tradeoff의 응답속도(speed)와 정확도의 부적 상관관계와 같은 방향성을 보이는 것으로, PISA와 같은 저부담 검사 상황에서도 비슷한 양상이 나타났음을 경험적 분석을 통해 보여주는 사례이다. 또한 흥미롭게도 모든 분석단위에 걸쳐 한국에서의 상관관계가 정적으로 더 강한 경향성을 띄는 것으로 확인되었다. 상관계수의 분포가 한국에서는 0.49~0.70(평균 0.63)이었던 데 반해 미국에서는 0.28~0.50(평균 0.41)로 나타났다. 즉, 정확도가 높은 학생들이 더 천천히 응답 하는 경향성이 미국보다 한국에서 더욱 강하게 나타났다.

〈표 3〉 한국과 미국의 11개 분석단위에서 정확도와 응답속도의 잠재 상관관계 추정치

		한국		미국	
		상관계수(ρ)	표준오차	상관계수(ρ)	표준오차
이전 주기에 실시된 척도연계 문항 (지필 검사와 컴퓨터 기반 검사에서 동일한 문항)	문항그룹 1	0.638	0.036	0.457	0.038
	문항그룹 2	0.655	0.028	0.451	0.038
	문항그룹 3	0.612	0.027	0.405	0.038
	문항그룹 4	0.488	0.038	0.321	0.045
	문항그룹 5	0.626	0.040	0.278	0.044
2015년 주기에 새롭게 개발되어 실시된 문항 (시나리오, 시뮬레이션 등 테크놀로지 기반 검사 문항이 대다수)	문항그룹 7	0.675	0.018	0.453	0.026
	문항그룹 8	0.702	0.018	0.499	0.026
	문항그룹 9	0.589	0.022	0.407	0.030
	문항그룹 10	0.622	0.023	0.363	0.028
	문항그룹 11	0.658	0.020	0.440	0.027
	문항그룹 12	0.630	0.021	0.435	0.026

2. 문항 유형에 따른 응답속도와 정확도의 관계

두 번째 연구 문제인 “응답속도와 정확도의 관계는 문항 유형에 따라 다르게 나타나는가?”를 탐구하기 위해 van der Linden의 위계적 모형을 문항 유형을 반영하여 확장시킨 모형 2를 분석하였다. 모형 2는 정확도는 일차원임을 가정하되 구성형 문항과 선다형 문항에서의 응답속도가 달리 반영될 수 있도록 문항 유형에 따라 응답속도를 이차원으로 발전시킨 것이다. 만약 문항 유형에 따라 응답속도와 정확도의 관계가 유의하게 다르게 나타난다면, 이는 모형 적합도가 개선되는 결과로 나타날 것이며, 선다형 문항과 구성형 문항에서의 응답속도와 정확도의 관계가 흥미로운 차이를 보여줄 것이다. 한국 자료에 나타난 각 분석단위에서 모형 적합도의 비교는 <표 4>와 같이 정리할 수 있다.

Chen(2007)은 시뮬레이션 연구를 통해 사례 수가 300명 이상일 때 측정 불변성(measurement invariance)을 검증하는 기준으로 CFI(comparative fit index)의 차이가 0.01 보다 크면서 RMSEA (root mean square error of approximation)가 0.015 이상 차이를 보여야 한다고 권고한 바 있다. 이에 따르면 모든 분석단위에서 CFI 기준 모형 적합도의 증가는 유의하지만, RMSEA 기준에는 미치지 못한다. <부록 표>에서도 볼 수 있듯이 미국 자료의 분석 결과도 비슷한 양상을 보인다. 미국의 경우에는 한 분석단위를 제외하고는 CFI 기준 유의한 모형 적합도의 증가를 보였지만 여전히 모든 분석단위에서 RMSEA의 기준에는 미치지 못하였다. 종합하자면, 응답

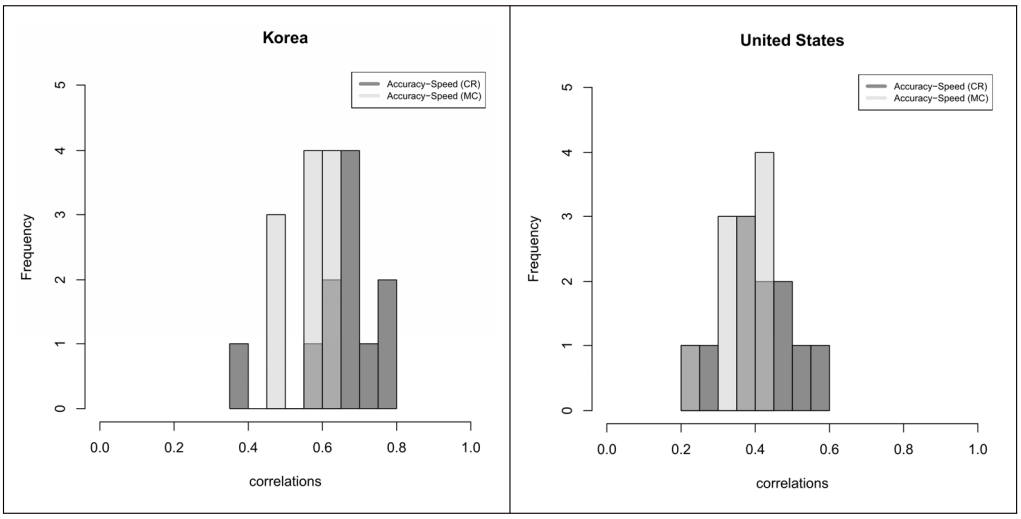
속도와 정확도의 관계를 탐구하는 van der Linden의 위계적 모형을 응답속도가 문항 유형을 반영하도록 발전시켰을 때 모형 적합도의 개선이 어느 정도 나타나지만, Chen(2007)의 기준에 의하면 뚜렷하게 개선되었다고 보기는 어렵다.

〈표 4〉 한국 자료에 나타난 모형 1과 모형 2의 모형 적합도의 비교

문항 그룹	모형 1			모형 2			모형 1 - 모형 2	
	CFI	TLI*	RMSEA	CFI	TLI	RMSEA	|△CFI|	△RMSEA
1	0.898	0.892	0.053	0.918	0.913	0.047	0.020	0.006
2	0.876	0.868	0.058	0.899	0.892	0.052	0.023	0.006
3	0.932	0.927	0.046	0.946	0.942	0.042	0.014	0.004
4	0.910	0.899	0.056	0.925	0.915	0.052	0.015	0.004
5	0.876	0.865	0.064	0.911	0.902	0.055	0.035	0.009
7	0.882	0.874	0.058	0.905	0.898	0.053	0.023	0.005
8	0.823	0.810	0.072	0.852	0.841	0.066	0.029	0.006
9	0.810	0.796	0.080	0.849	0.837	0.072	0.039	0.008
10	0.902	0.895	0.053	0.922	0.916	0.047	0.020	0.006
11	0.895	0.887	0.060	0.913	0.906	0.054	0.018	0.006
12	0.859	0.849	0.064	0.869	0.859	0.062	0.010	0.002

* TLI(Tucker-Lewis Index)

다음으로 각 문항 유형에서의 응답속도와 정확도의 관계가 어떻게 나타나는지 살펴보았다. 모형 2에서는 세 개의 잠재 상관관계가 추정되는데 이 중 하나는 구성형 문항과 선다형 문항의 응답속도 간의 관계이며, 나머지 둘은 문항 유형에 따른 응답속도와 정확도의 관계이다. 선다형 문항에서의 응답속도와 구성형 문항에서의 응답속도 간의 상관계수(ρ_3)는 한국에서는 0.73~0.87(평균 0.83), 미국에서는 0.59~0.84(평균 0.76)으로 추정되었으며, 이러한 결과는 선다형과 구성형 문항으로 이루어진 PISA 검사 상황에서 응답속도의 일차원성 가정이 적합하지 않을 수 있음을 보여준다. Shin et al. (2020)의 연구에서도 PISA 2015 본검사에 컴퓨터 기반 방식으로 참여한 모든 참여국의 응답시간을 분석한 결과, 선다형 문항과 구성형 문항의 응답속도 간의 잠재 상관관계가 평균 0.5~0.6 정도에 머물러 응답속도 연구에서 문항 유형을 고려하는 것이 필요함을 보여주었다. 특히 한국이 미국에 비해 전반적으로 더 강한 정적 상관관계를 나타냈으며, 이는 선다형 문항에 빨리 응답할수록 구성형 문항에도 빨리 응답하는 경향성이 한국에서 더 강했다는 것으로 해석할 수 있다.



〔그림 2〕 문항 유형에 따른 정확도와 응답속도 간의 관계 분포 (한국: 왼쪽, 미국: 오른쪽)

[그림 2]는 문항 유형에 따른 응답속도와 정확도의 관계를 히스토그램으로 표현한 것이다. 히스토그램의 총 빈도는 문항 그룹 11개와 문항 유형 두 종류가 고려된 총 22개(11X2)이며, 히스토그램에서 겹치는 구간은 중간 회색으로 표현되었다. 우선 선다형 문항에서의 응답속도와 정확도의 관계(ρ_1; 옅은 회색)는 한국에서 0.48～0.62(평균 0.57), 미국에서는 0.22～0.45(평균 0.37)로 나타난 데 반해, 구성형 문항에서의 관계(ρ_2; 진한 회색)는 한국에서 0.40～0.79(평균 0.65), 미국에서는 0.21～0.55(평균 0.41)로 나타났다. 즉, 대체적으로 한국과 미국 모두 천천히 응답할수록 정확도가 높아지는 경향성이 선다형 문항보다 구성형 문항에서 약간 더 강하게 나타나는 것을 볼 수 있다. 이는 단순히 문항에서 주어진 옵션 중에 선택을 하는 것보다 실제 응답을 작성해야 하는 구성형 문항의 특성을 고려하면 일견 당연한 결과처럼 보이기도 한다. 그러나 이러한 결과가 분석단위에 따라 일관되지 않게 나타났다는 것은 흥미롭다. 특히 새로 개발된 테크놀로지 기반 문항이 대다수인 문항그룹(S07-S12)의 경우에는 정확도와 구성형 문항에서의 응답속도의 상관관계가 언제나 선다형의 경우보다 높았던 반면($\rho_2 > \rho_1$), 이전 주기에 시행된 척도연계 문항그룹(S01-S06)의 경우에는 혼재된 양상이 나타났다. 이는 향후 응답속도 연구에서 통상적인 선다형 또는 구성형의 문항 유형 분류뿐만 아니라, 동일한 지필 검사 문항이 컴퓨터 기반 검사 상황에서 시행된 경우(척도연계 문항)나 테크놀로지를 적극 활용한 문항에서 일어나는 고유한 응답과정을 살펴볼 필요가 있음을 시사한다.

Ⅴ. 결론 및 논의

본 연구에서는 PISA를 비롯한 국제학업성취도 평가들이 컴퓨터 기반 검사로 전환됨에 따라 새롭게 주목받고 있는 프로세스 데이터 중, 응답시간에 주목하여 PISA 검사 상황에서 나타난 우리나라 학생들의 응답속도와 정확도의 관계를 규명하고자 하였다. 특히 미국과의 비교분석을 통해 PISA 2015와 PISA 2018에 일관적으로 나타나고 있는 한국 학생들의 매우 빠른 응답속도를 이해하고자 하였다. 이를 위해 실험심리학 분야에서 확립된 speed-accuracy tradeoff의 개념을 소개하고, 이를 PISA와 같이 피험자 집단을 대상으로 여러 문항으로 이루어진 검사에 적용시킬 경우 유의해야 할 점을 서술하였다. 다음으로 응답속도와 정확도의 관계를 규명하는 데 가장 널리 활용되는 van der Linden(2007)의 위계적 모형을 소개하고, 이를 PISA 검사 상황에 맞추어 문항 유형(구성형 또는 선다형)에 따라 응답속도와 정확도의 관계가 달라질 수 있는 모형으로 확장시켰다. 그리고 PISA 2015의 검사 설계를 고려하여 한국과 미국 각 나라에서 12개의 분석단위를 구성한 다음, 두 가지의 심리측정 모형을 통해 자료를 분석하고, 수렴된 결과를 바탕으로 11개 분석단위의 결과를 보고하였다.

첫 번째 연구 질문인 “PISA 2015 에 나타난 응답속도와 정확도의 관계는 한국과 미국에서 어떻게 나타나는가?”에 대한 분석 결과, 한국과 미국 모두 11개의 모든 분석단위에서 정확도가 높은 학생일수록 더 오랜 시간에 걸쳐 응답하는 경향이 나타났다. 이는 앞서 소개된 speed-accuracy tradeoff와 같은 방향성을 보이는 것으로, PISA와 같은 저부담 검사 상황에서 우리나라 학생들이 그와 비슷한 양상을 보인다고 할 수 있다. 흥미롭게도 PISA 평균 집단(영어를 사용하는 평균 학업성취 집단)으로서 비교에 포함된 미국에 비해 한국에서 이러한 경향성이 더욱 강하게 나타나고 있다(한국 상관계수 평균 0.63 vs. 미국 상관계수 평균 0.41). 즉, 미국을 평균 집단으로 생각하여 비교하면, 우리나라 학생들의 매우 빠른 응답속도가 정확도와 강한 연관성이 있는 것으로 해석된다. 다만 이를 통해 한국 학생들의 응답시간을 늘리면 정확도가 높아질 것이라는 제안을 하기는 어렵다. 본 연구에서 사용된 심리측정 모형은 관찰자료를 기반으로 응답속도와 정확도의 상관관계를 추정한 것으로, 인과적인 관계를 추론할 수는 없기 때문이다. 우리나라 학생들의 PISA 학업성취 수준을 높이기 위해서 응답시간을 늘려야 하는가에 답하기 위해서는 학생들의 검사참여도에 대한 후속 연구와 저부담 검사 상황에서 우리나라 학생들의 행동 양식을 살펴보는 실험연구 등이 요구될 것이다.

두 번째 연구 질문인 “응답속도와 정확도의 관계는 문항 유형에 따라 다르게 나타나는가?”에 대한 연구 결과, 대체로 선다형 문항보다 구성형 문항에서의 정확도와 응답속도의 관계가 더 강한 정적 상관관계를 보였다. 즉, 한국과 미국 모두 천천히 응답할수록 정확도가 높아지는 경향성이 선다형 문항보다 구성형 문항에서 더 강하게 나타났다. 이와 같이 문항

유형을 고려한 심리측정 모형을 이용하여 분석했을 때, 한국과 미국 모든 분석단위에서 모형 적합도가 개선된 것으로 나타났다. 다만, 모형 적합도의 개선이 얼마나 유의한가와 관련하여 Chen(2007)에서 제안된 기준을 모두 충족시키지는 못하였으므로, 모형 적합도가 뚜렷하게 개선되었다고 검증하기에는 한계가 있었다.

이와 같은 연구 결과를 토대로 해석에 유의해야 할 점과 앞으로의 추후 연구를 제안하며 본 논문을 마무리짓고자 한다. 먼저 PISA에서 제공되는 프로세스 데이터를 분석할 때 PISA 검사 설계에 대한 이해가 선행되어야 한다. 2015년에 컴퓨터 기반 검사로의 전환이 이루어진 이후 2018년 이후부터는 주요 영역마다 순차적으로 단계적 컴퓨터화 적응 검사(multistage adaptive testing design)로의 전환이 이루어지고 있으며(Shin, Yamamoto, Khorramdel, Robin, 2021; Yamamoto, Shin, Khorramdel, 2019), 2025년부터는 모든 주요 영역에서 컴퓨터화 적응 검사가 시행될 것으로 보인다. 이와 관련하여 프로세스 데이터의 수집과 효용성은 검사 설계와 별도로 논의할 수 없다는 점을 인식하고 연구자들은 분석 방법과 결과 해석에 주의를 기울여야 할 것이다. 구체적으로, 본 연구에서는 2015년 balanced incomplete block 검사 설계에 의해 자료 수집이 이루어졌기 때문에 개별 문항에 응답한 학생들이 표집 집단 중에서는 무선 표집되었다는 가정 하에 분석할 수 있었다. 그러나 2018년 이후부터는 컴퓨터화 적응 검사를 통해 자료가 수집되기 때문에 초기 능력치에 근거하여 선별된 학생들의 응답점수 및 프로세스 데이터가 수집되는 경우가 대부분이다. 따라서 PISA 2015년 자료에서는 유효하였던 통계적 검증 과정이라도 PISA 2018년 이후 자료부터는 좀 더 복잡성을 띄게 될 수 있으며 이를 통계적 절차에 적합하게 반영해야 할 필요가 있다.

마지막으로 본 연구에서 기본 모형으로 채택하고 확장시킨 van der Linden(2007)의 모형이 여러 가지 통계적 가정에 기초하고 있음을 인식하고, 그러한 가정의 타당성을 검증해 보는 연구들이 필요할 것이다. 특히 Bolsinova, Tijmstra, Molenaar, De Boeck(2017)이 제안한 것과 같이 응답속도와 정확도의 관계에 대한 연구에서 조건적 독립성(conditional independence) 가정이 위배되는 것은 통계적 추론을 방해하는 것이 아니라, 실제 검사 상황에서 응답하는 학생들의 응답과정을 보다 깊이 있게 알 수 있는 유효한 정보를 제공하기도 한다. 일례로, 본 연구에서는 분석 방법에 소개된 것과 같이 학생 입장에서 일정한 응답속도를 가정하고 분석을 시행하고 결과를 해석하였으나, 향후에는 이러한 일정한 응답속도가 타당한 가정인지 경험적으로 검증할 수 있을 것이다. 즉, 우리나라 학생들이 PISA 검사에 응답하면서 본인의 응답속도를 바꾸는 경우가 나타나는지, 나타난다면 어떤 교육맥락변인과 연관되는지에 대한 연구를 통해 한국 학생들의 빠른 응답속도에 대한 깊이 있는 이해를 도울 수 있을 것이다. 이에 더하여 본 연구에서는 한국과 PISA 평균 집단으로서 미국의 자료를 토대로 비교 연구를 시도하였으나, 한국과 같은 아시아권에 있는 나라들(예를 들어, 중국, 일본, 싱가포르)과

의 비교 연구도 흥미로울 것이다. 특히 한국과 미국의 경우에는 문항 유형에 따른 모형 적합도의 개선이 뚜렷하게 나타나지는 않았으나, 한자 기반의 언어를 컴퓨터 입력 시스템에 사용하는 나라들이 다른 나라들과는 뚜렷하게 다른 응답시간 패턴을 보인 것을 감안할 때(Ercikan, Guo, He(2020), 그러한 나라들에서는 문항 유형에 따라 뚜렷한 문항 적합도의 개선이 나타날 것으로 기대할 수 있다. 또한 문항 유형에 의한 차이뿐만 아니라, 문항의 내용영역이나 인지영역, 문항 제시 방식 등과 같은 다양한 문항 특성들이 응답속도와 정확도의 관계에 미치는 영향을 검증하는 것도 우리나라 학생들의 응답속도를 이해하는 데 도움이 될 것이다.

프로세스 데이터는 현재 교육 및 심리측정 분야에서 가장 활발히 연구되는 분야 중 하나이다. 최근에는 PISA에 참여하는 나라들의 순위를 매길 때, 전통적인 학업성취뿐만 아니라 프로세스 데이터에 나타난 응답속도의 관점도 함께 고려하자는 제언도 나오고 있다(Pohl, Ulitzsch, von Davier, 2021). 비슷한 맥락으로, 학업성취 점수(plausible value)를 산출하는 통계적 모형에 어떻게 프로세스 데이터를 반영할 것인가에 대한 연구가 진행 중이며(Shin, von Davier, Yamamoto, in press), 실제 PISA 2018에 보고된 학업성취 점수에는 응답시간 정보가 반영되기도 하였다(OECD, 2020a). 국내에서도 프로세스 데이터 전반에 대한 관심과 활용 방법에 대한 관심이 높아지고 있으나, 응답시간을 비롯한 프로세스 데이터 연구는 아직 찾아보기 어렵다. 대표적인 국제학업성취도 평가인 PISA 2015에서 공공데이터로 제공된 응답시간을 이용하여 한국과 미국에서의 응답속도와 정확도의 관계를 경험적으로 분석한 본 연구가 향후 프로세스 데이터를 활용한 국내 연구 확산과 국가수준 학업성취도 평가에서 프로세스 데이터의 수집과 활용에 도움이 될 수 있기를 기대한다.

참고문헌

구자옥, 김성숙, 이혜원, 조성민, 박혜영(2017). **OECD 국제 학업성취도 평가 연구: PISA 2015 결과 심층 분석 보고서(연구보고서 RRE 2017-9)**. 한국교육과정평가원.

김갑수(2020). PISA 2018년 데이터를 기반으로 한국 학생들의 ICT 접근성과 교과 활용도 분석. **정보교육학회논문지**, 24(1), 39-48.

김난옥, 손원숙(2019). 우리나라 학생평가 실태 연구: PISA 2015 참여국과의 국제비교를 중심으로. **교육과정평가연구**, 22(3), 173-198.

김성훈(2013). 유한혼합분포에 대한 EM 알고리듬을 사용한 다집단 IRT 추정 방법의 원리와 기능. **교육평가연구**, 26(4), 801-822.

김수영(2016). **구조방정식 모형의 기본과 확장: Mplus 예제와 함께**. 서울: 학지사.

김준형(2019). PISA자료 분석을 통한 평준화 정책의 재평가. **교육사회학연구**, 29(2), 93-129.

김현정, 구남욱(2019). PISA 설문 결과에 나타난 우리나라 과학 수업의 실태 분석. **교육과정평가연구**, 22(4), 85-104.

김혜숙, 김한성, 김진숙, 신안나(2017). OECD PISA 자료를 활용한 우리나라 학생들의 ICT 접근 및 활용 수준 추이 분석. **정보화정책**, 24(4), 17-43.

김혜숙, 함은혜(2019). PISA 2015 협력적 문제해결력에 영향을 미치는 학생 및 학교수준 특성 분석-과학성취도와의 비교를 중심으로-. **교육과정평가연구**, 22(3), 199-224.

김화경, 송민호, 최인용, 정인우(2020). 컴퓨터 기반 평가에서 수학 기초학력 미달 학생의 특징 분석: PISA 2018 문항 응답시간과 동작을 중심으로. **학교수학**, 22(3), 791-809.

손윤희, 박현정, 박민호(2020). 랜덤 포레스트를활용한 읽기소양 수준에 따른 집단 결정요인 분석: PISA 2018 자료를 중심으로. **아시아교육연구**, 21(1), 191-215.

유희영, 오윤정(2019). PISA2015 참여국의 과학성취도와 과학의 즐거움에 대한 인식비교. **교과교육학연구**, 23(4), 346-360.

이강호, 박주호(2019). 학교장 변혁적 리더십과 교사협력활동 관계에서 학부모 교육적 관여 의조절효과. **아시아교육연구**, 20(2), 405-430.

이빛나, 손원숙(2019). 과학 실험수업에서 형성평가의 역할 탐색: PISA 2015 한국, 싱가포르, 캐나다의 국제비교. **교육평가연구**, 32(4), 649-670.

이희숙(2019). 초·중등 사교육 실태와 관련 요인 분석 연구- 한국, 중국, 일본, 싱가포르를 중심으로 -. **청소년학연구**, 26(5), 469-488.

조성민, 구남욱(2020a). PISA 2018 결과에 나타난 우리나라 학생들의 수학 성취 및학교 풍토와 학생 웰빙 관련 결과 탐색. **수학교육학연구**, 30(3), 465-486.

조성민, 구남욱(2020b). PISA 2015 상위국과의 비교를 통한 우리나라 학생들의 수학영역 성취 특성 분석. **교과교육학연구**, 24(3), 262-272.

한정아(2020). 과학 관련 정의적 특성과 교수방법이 학생들의 과학 성취에 미치는 영향. **교육과정평가연구**, 23(1), 31-56.

Asparouhov, T., & Muthen, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. In proceedings of the 2007 jsm meeting insalt lake city, utah, section on statistics in epidemiology(pp. 2531 – 2535)

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), ***Handbook of Modern Item Response Theory*** (pp. 433-448). New York, NY: Springer-Verlag.

Bolsinova, M., Tijmstra, J., Molenaar, D., & De Boeck, P. (2017). Conditional dependence between response time and accuracy: an overview of its possible sources and directions for distinguishing between them. ***Frontiers in psychology,*** *8*, 202.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. ***Structural Equation Modeling: a multidisciplinary journal,*** *14*(3), 464 - 504.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. ***Frontiers in psychology,*** *10*, 102.

Dodonova, Y. A., & Dodonov, Y. S. (2013). Faster on easy items, more accurate on difficult ones: Cognitive ability and performance on a task of varying difficulty. ***Intelligence,*** *41*, 1-10. doi:10.1016/j.intell.2012.10 .003

Ercikan, K., Guo, H. & He, Q. (2020) Use of Response Process Data to Inform Group Comparisons and Fairness Research, ***Educational Assessment,*** *25*(3), 179-197, DOI: 10.1080/10627197.2020.1804353

Goldhammer, F. Martens, T., Christoph, G., & Lüdtke, O. (2016). **Test-taking engagement in PIAAC, OECD Education Working Papers** 133. Paris, France: OECD Publishing. doi:10.1787/5jlzfl6fhxs2-en

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. ***Journal of Educational Psychology,*** *106*(3), 608-626.

Hallquist, M. N. & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. ***Structural Equation Modeling: a multidisciplinary journal,*** *25*(4), 621-638. doi: 10.1080/10705511.2017.1402334.

He, Q., von Davier, M., & Han, Z. (2018). Exploring process data in problem-solving items in computer-based large-scale assessments. In H. Jiao, R. W. Lissitz, & A. Van Wie (Eds.), ***Data Analytics and Psychometrics: Informing Assessment Practices*** (pp. 53-76). Information Age Publishing.

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. ***Frontiers in neuroscience,*** *8*, 150.

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. ***Large-Scale Assessments in Education,*** *2*(8). doi:10.1186/s40536-014-0008-1

Loeys, T., Rosseel, Y., & Baten, K. (2011). A joint modeling approach for reaction time and accuracy

in psycholinguistic experiments. ***Psychometrika,*** *76*(3), 487-503.

Luce, R. D. (1986). ***Response times: Their role in inferring elementary mental organization (No. 8)***. Oxford University Press on Demand.

Messick, S. J., Beaton, A. E., & Lord, F. M. (1983). ***NAEP reconsidered: A new design for a new era (NAEP Report No. 83-1)***. Princeton, NJ: Educational Testing Service.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015). A bi-variate generalized linear item response theory modeling framework to the analysis of responses and response times. ***Multivariate Behavioral Research,*** *50*(1), 56-74. doi:10.1080/00273171.2014.962684

Muthén, B., & Muthén, L. (2017). **Mplus Users Guide**. Chapman and Hall/CRC.

Organisation for Economic Co-operation and Development (2017). ***PISA 2015 Technical Report***. OECD Publishing. http://www.oecd.org/pisa/data/2015-technical-report/

Organisation for Economic Co-operation and Development (2020a). ***PISA 2018 Technical Report***. OECD Publishing. http://www.oecd.org/pisa/data/pisa2018technicalreport/

Organisation for Economic Co-operation and Development (2020b). PISA 2015 Database. http://www.oecd.org/pisa/data/2015database/

Organisation for Economic Co-operation and Development (2020c). PISA 2018 Database. http://www.oecd.org/pisa/data/2018database/

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. ***Science,*** *372*(6540), 338-340.

Rust, K. F., & Johnson, E. G. (1992). Chapter 2: Sampling and weighting in the national assessment. ***Journal of Educational Statistics,*** *17*(2), 111-129.

Shin, H. J., von Davier M., Yamamoto K. (2019) Investigating Rater Effects in International Large-Scale Assessments. In B. P. Veldkamp and C. Sluijter (Eds.), ***Theoretical and Practical Advances in Computer-based Educational Measurement*** (pp. 249-268), Methodology of Educational Measurement and Assessment (book series), Springer.

Shin, H. J., von Davier, M., & Yamamoto, K. (in press). Incorporating timing data into the PISA population modeling. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), ***Innovative Computer-based International Large-Scale Assessments - Foundations, Methodologies and Quality Assurance Procedures***. Springer.

Shin, H. J., Lubaway, E., Joo, S., Robin, F., & Yamamoto, K. (2020). Comparability of response time scales in PISA. ***Psychological Test and Assessment and Modeling***, *62*(1), 107-135.

Shin, H. J., Yamamoto, K., Khorramdel, L., Robin, F. (2021). Increasing measurement precision of

PISA through multistage adaptive testing. In Wiberg, M., González, J., & Molenaar, D., Böckenholt, U., & Kim, S-J. (Eds.). ***Quantitative Psychology***. 85th Annual Meeting of the Psychometric Society, Virtual, 2020, New York: Springer.

Swanson, D.B., Case, S.E., Ripkey, D.R., Clauser, B.E., & Holtman, M.C. (2001). Relationships among item characteristics, examinee characteristics, and response times on USMLE, Step 1. ***Academic Medicine,*** *76*, 114-116

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. ***Psychometrika,*** *72*(3), 287-308.

von Davier, M., Yamamoto, K., Shin, H., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000 – 2012. ***Assessment in Education: Principles, Policy & Practice,*** *26*(4), 466-488. https://doi.org/10.1080/0969594X.2019.1586642

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. ***Journal of Educational and Behavioral Statistics,*** *44*(6), 671-705.

Yamamoto, K. & Lennon, M.L. (2018). Understanding and detecting data fabrication in large-scale assessments. ***Quality Assurance in Education,*** *26*(2), 196-212. https://doi.org/10.1108/QAE-07-2017-0038

Yamamoto, K., Shin, H., & Khorramdel, L. (2019). Introduction of multistage adaptive testing design in PISA 2018. ***OECD Education Working Papers No. 209***, OECD Publishing, Paris. https://dx.doi.org/10.1787/b9435d4b-en

◎ 논문접수: 2021. 08. 06 / 수정본 접수: 2021. 09. 09 / 게재승인: 2021. 09. 18

저 자 소 개

· 신효정 : University of California at Berkeley에서 교육 측정평가 전공으로 박사학위를 취득하고 현재 미국의 Educational Testing Service에서 Senior Measurement Scientist로 재직 중임. 관심분야는 교육분야 빅데이터, 국제학업성취도평가, 머신러닝 기법, 잠재변인 모형, 평정자 효과 분석, 교육과정 평가 등임. hj1424@gmail.com

〈ABSTRACT〉

Psychometric modeling of speed and accuracy: Analysis of PISA 2015 data from Korea and the United States

Hyo Jeong Shin

Educational Testing Service

The transition to computer-based assessment in international large-scale assessments, including PISA, enabled the collection of response processes, in addition to response outcomes. Recently, advanced psychometric models have been proposed to jointly model the response outcomes and processes. Response time, as one of the popular variables among process data, is receiving increased attention from researchers due to its accessibility. To deepen our understanding of Koreans' extremely fast responses in PISA, I introduced and fitted van der Linden's hierarchical model (van der Linden, 2007) and its extended version that accounts for the item types (multiple choice vs. constructed response items) to the PISA 2015 science data, collected in Korea and the United States. Negative accuracy-speed latent correlations were estimated between accuracy and speed dimensions, implying faster responses were associated with lower accuracy in both countries. Compared to the United States, Korea showed stronger negative accuracy-speed latent correlations. In addition, when the item types were considered, the model fit improved to some extent and showed stronger relationships between accuracy and speed on constructed-response items. Such a pattern was consistently observed on the technology-enhanced items, which were newly developed and administered in PISA 2015 both in Korea and the United States.

Keywords : response time, response accuracy, multidimensional item response theory model, process data, PISA 2015, item type, speed-accuracy tradeoff

[부록 표] 미국 자료에 나타난 모형 1과 모형 2의 모형 적합도의 비교

분석 단위	모형 1			모형 2			모형 1 - 모형 2	
	CFI	TLI	RMSEA	CFI	TLI	RMSEA	|△CFI|	△RMSEA
1	0.763	0.749	0.068	0.813	0.801	0.06	0.050	0.008
2	0.877	0.869	0.046	0.888	0.881	0.044	0.011	0.002
3	0.837	0.826	0.063	0.865	0.855	0.057	0.028	0.006
4	0.910	0.899	0.049	0.94	0.932	0.041	0.030	0.008
5	0.897	0.888	0.055	0.924	0.917	0.048	0.027	0.007
7	0.831	0.82	0.064	0.857	0.847	0.059	0.026	0.005
8	0.725	0.705	0.069	0.733	0.712	0.068	0.008	0.001
9	0.822	0.81	0.063	0.844	0.832	0.059	0.022	0.004
10	0.821	0.809	0.061	0.843	0.832	0.057	0.022	0.004
11	0.859	0.848	0.06	0.872	0.862	0.058	0.013	0.002
12	0.796	0.781	0.072	0.806	0.792	0.071	0.010	0.001

[부록] 본 연구에서 분석에 사용된 Mplus 코드의 예시 (모형 1)

```
TITLE: PISA 2015 RA-RT model for KOREA
 DATA:
            FILE IS PISA2015_RART_KOR_CL5.dat; !example of cluster 5
            TYPE = INDIVIDUAL;
            LISTWISE = OFF;
            VARIANCES = CHECK;
 VARIABLE:          !items were re-labeled considering their positions
            NAMES = CNTSTUID SENWT
             CS05X1 DS05X2 CS05X3 DS05X4 CS05X5 CS05X6 CS05X7 DS05X8 DS05X9 CS05X10 DS05X11
           DS05X12 DS05X13
             CS05T1 DS05T2 CS05T3 DS05T4 CS05T5 CS05T6 CS05T7 DS05T8 DS05T9 CS05T10 DS05T11
           DS05T12 DS05T13 ;
 USEVARIABLES =
 !cognitive items
             CS05X1 DS05X2 CS05X3 DS05X4 CS05X5 CS05X6 CS05X7 DS05X8 DS05X9 CS05X10 DS05X11
           DS05X12 DS05X13
```

```
!log response times (log transformed in the data preparation)
        CS05T1 DS05T2 CS05T3 DS05T4 CS05T5 CS05T6 CS05T7 DS05T8 DS05T9 CS05T10
        DS05T11 DS05T12 DS05T13 ;
CATEGORICAL =
        CS05X1 DS05X2 CS05X3 DS05X4 CS05X5 CS05X6 CS05X7 DS05X8 DS05X9 CS05X10
        DS05X11 DS05X12 DS05X13 ;
MISSING = ALL(9);
WEIGHT = SENWT;                                  ! weight is specified
ANALYSIS:
        TYPE = GENERAL MISSING;
        ESTIMATOR = WLSMV;
        COVERAGE = 0;
MODEL:                  ! van der Linden's model
Cognitive by            ! specify variables for ability dimension
        CS05X1 DS05X2 CS05X3 DS05X4 CS05X5 CS05X6 CS05X7 DS05X8 DS05X9 CS05X10 DS05X11
        DS05X12 DS05X13 ;
LogSpeed by             ! specify variables for speed dimension
        CS05T1 DS05T2 CS05T3 DS05T4 CS05T5 CS05T6 CS05T7 DS05T8 DS05T9 CS05T10 DS05T11
        DS05T12 DS05T13 ;
Cognitive with Logspeed ;       ! ability-speed correlation
OUTPUT:
        STANDARDIZED;
        MODINDICES (30);
```

[부록] 본 연구에서 분석에 사용된 Mplus 코드의 예시 (모형 2)

```
TITLE: PISA 2015 RA-RT by item types model for KOREA
 DATA:
        FILE IS PISA2015_RART_KOR_CL5.dat ;
        TYPE = INDIVIDUAL;
        LISTWISE = OFF;
        VARIANCES = CHECK;
 VARIABLE:
        NAMES = CNTSTUID SENWT
        CS05X1 DS05X2 CS05X3 DS05X4 CS05X5 CS05X6 CS05X7 DS05X8 DS05X9 CS05X10 DS05X11
DS05X12 DS05X13
        CS05T1 DS05T2 CS05T3 DS05T4 CS05T5 CS05T6 CS05T7 DS05T8 DS05T9 CS05T10 DS05T11
```

```
DS05T12 DS05T13 ;
 USEVARIABLES =
 !cognitive items
            CS05X1 DS05X2 CS05X3 DS05X4 CS05X5 CS05X6 CS05X7 DS05X8 DS05X9 CS05X10 DS05X11
DS05X12 DS05X13
 !log response times
            CS05T1 DS05T2 CS05T3 DS05T4 CS05T5 CS05T6 CS05T7 DS05T8 DS05T9 CS05T10 DS05T11
DS05T12 DS05T13 ;
 CATEGORICAL =
            CS05X1 DS05X2 CS05X3 DS05X4 CS05X5 CS05X6 CS05X7 DS05X8
           DS05X9 CS05X10 DS05X11 DS05X12 DS05X13 ;
 MISSING = ALL(9);
 WEIGHT = SENWT;
 ANALYSIS:
            TYPE = GENERAL MISSING;
            ESTIMATOR = WLSMV;
            COVERAGE = 0;
 MODEL:                    ! extension of van der Linden's model by item types
 Cognitive by              ! specify variables for ability dimension
            CS05X1 DS05X2 CS05X3 DS05X4 CS05X5 CS05X6 CS05X7 DS05X8
           DS05X9 CS05X10 DS05X11 DS05X12 DS05X13 ;
 HC_LogSpeed by            ! specify variables for speed on CR items
            DS05T2 DS05T4 DS05T8 DS05T9 DS05T11 DS05T12 DS05T13 ;
 MC_LogSpeed by            ! specify variables for speed on MC items
            CS05T1 CS05T3 CS05T5 CS05T6 CS05T7 CS05T10 ;
 Cognitive with HC_Logspeed ;
! correlation between ability and CR-speed
 Cognitive with MC_Logspeed ;
! correlation between ability and MC-speed
 MC_LogSpeed with HC_Logspeed ;
! correlation between speed on CR items and speed on MC items
 OUTPUT:
            STANDARDIZED;
            MODINDICES (30);
```