교육과정평가연구
The Journal of Curriculum and Evaluation
2025, Vol. 28, No. 2, pp.217~248
DOI: https://doi.org/10.29221/jce.2025.28.2.217

Saltus 모형을 활용한 문항 특성에 따른 잠재적 차별기능문항군 탐색[1)]

장윤선 (대구교육대학교 조교수)*

요약

이 연구는 문항 특성에 따른 문항반응이 이질적인 잠재집단을 탐색함으로써 검사에 존재하는 문항군 단위의 차별기능을 탐색하고자 하였다. 이를 위해 혼합 문항반응이론 모형의 하나인 Saltus 모형을 활용하여 PISA 2022 과학 영역 검사에 대한 한국 학생들의 응답자료를 분석하였다. 문항 특성 정보로 문항 유형, 맥락, 역량, 지식 차원을 반영하여 문항군을 정의하였다. 연구 결과에 따르면 문항 유형에 따른 잠재적 차별기능문항군은 없었고, 맥락 차원에서는 개인적 맥락과 지역적/국가적 맥락의 문항군이 모두 유리하게 기능한 잠재집단이 존재하는 것으로 나타났다. 역량 차원에서는 자료 및 증거의 과학적 해석 문항군이 차별적으로 기능하였는데 한 잠재집단에서 해당 문항군은 불리하게 기능하였음이 확인되었다. 지식 차원에 따른 잠재적 차별기능문항군은 절차적 지식 문항군에서만 확인되었다. 더불어 맥락 및 역량에 따른 잠재적 차별기능문항군이 존재하는 것으로 확인된 잠재집단 간 학생의 배경 변인과 과학 과목 학습 관련 교육맥락 변인의 유의미한 차이가 존재함이 확인되었다. 이 연구는 문항 특성에 따른 잠재적 차별기능문항군을 탐색함으로써 다양한 차별기능문항(군) 분석 방법의 경험적 근거를 제공하고자 하였다. 더불어 잠재적 차별기능문항(군) 탐색의 결과를 검사 타당도 근거뿐 아니라 교육적 지원의 정보로 활용하는 접근을 제시하였다는 점에서 의의가 있다.

주제어: 잠재적 차별기능문항군, 문항 특성, 혼합 문항반응이론, Saltus 모형, PISA 2022 과학 영역 검사

1) 이 논문은 2023년 대구교육대학교 교내학술연구비 지원을 받아 수행된 연구임(RC2023334).
* 제1저자 및 교신저자, ysj@dnue.ac.kr

I. 서 론

검사 결과에 대한 타당성과 공정성을 확보하는 것은 무엇보다 중요하다. 특히 눈에 보이지 않는 대상의 잠재 구인을 측정하기 위해 검사 도구를 사용하는 경우, 해당 검사를 통해 측정하는 잠재 구인이 일관되어야 한다는 측정의 동일성(measurement invariance)이 충족되어야 그 검사 도구를 통해 측정된 점수가 타당해질 수 있다. 측정의 동일성이 충족되었는지를 검증하는 대표적인 방법 가운데 하나는 검사 도구를 구성하는 문항 가운데 차별적으로 기능하는 문항의 존재를 탐색하는 것이다. 이와 같은 문항을 차별기능문항(differential item functioning: DIF)이라 부르며, 검사를 통해 측정하고자 하는 잠재 구인의 수준이 같음에도 불구하고 검사 대상의 특성으로 인하여 문항에 대한 반응이 달라지는 것을 의미한다.

전통적으로는 문항이 차별적으로 기능할 것이라고 예상되는 집단의 특성(예: 여학생과 남학생)을 사전에 정하고 각 집단의 문항 모수를 비교하여 차별기능문항 여부를 판정한다. 만약 차별기능문항이 존재한다면 그 원인이 무엇인지를 파악하여 문항을 수정하거나 다른 문항으로 변경하는 등의 조치를 한다. 차별기능문항을 적절하게 수정하거나 삭제 여부를 결정하기 위해서 검사 문항이 차별적으로 기능하는 원인이 무엇인지 파악하는 것이 필요하다.

더불어 차별기능문항의 원인에 대한 정보를 교육적으로 활용할 수도 있다. 이를테면 검사 개발자가 의도한 대로 교육과정에 부합하는 내용 및 인지 과정을 정확하게 측정하는 문항이 차별기능문항으로 판정될 수 있을 것이다. 이런 경우, 해당 문항을 검사에서 제외하거나 수정하는 것 보다 이들 문항이 불리하게 기능하는 집단을 위해 추가로 제공해야 할 교육적 지원의 내용을 결정하는 하나의 정보로 활용하는 더 적절할 것이다(Bundsgaard, 2019; Gierl et al., 2001). 즉, 차별기능문항 연구는 통계적 방법을 활용하여 검사의 차별기능문항을 탐색하는 것뿐 아니라 차별기능의 원인을 파악하고 이를 검사의 타당성을 근거를 확보하거나 교육적 지원을 통한 검사의 공정성 문제 해결을 위해 필요한 유의미한 정보를 제공하는 중요한 역할을 한다.

성별 또는 국가 등 차별기능문항이 존재할 것으로 예측되는 특정 관찰집단(manifest class)을 연구자가 사전에 선정하고, 집단에 따른 차별기능문항을 탐색하는 전통적인 방법은 차별기능의 원인에 대한 정보를 제공하는 데 한계가 있다. 왜냐하면 차별기능문항을 유발하는 원인은 매우 다양할 수 있으며 집단의 특성이 성별이나 국가 등 관찰이 가능한 변인이 아닌 응답 경향 또는 문제해결전략 등 잠재적 변인에 기인하여 문항이 차별적으로 기능할 수 있기 때문이다(Embreston & Reise, 2000; Kang & Cohen, 2006;

Li, Cohen, & Ibarra, 2004). 또한, 때에 따라서는 여러 특성이 복합적으로 작용하여 차별기능문항이 유발될 수도 있을 것이다.

이러한 문제를 보완하기 위하여 문항 반응이 다른 잠재집단(latent class)을 구분하고, 잠재집단별로 추정된 문항 모수의 차이를 평가하여 차별기능문항을 판정하는 탐색적 접근을 활용할 수 있다. 탐색적 접근의 대표적인 분석 방법은 문항반응이론(item response theory: IRT)과 잠재집단분석(latent class analysis: LCA)을 결합한 혼합 문항반응이론(mixture IRT) 모형을 활용하는 것이다(De Boeck, Cho, & Wilson, 2011; Kang & Cohen, 2006; Oliveri et al., 2016; Oliveri, Ercikan, & Zumbo, 2013). 혼합 문항반응이론 모형은 수집된 문항반응 자료에 기반하여 잠재집단을 추출하고 집단별 문항 모수를 다르게 추정하기 때문에 연구자가 예측할 수 없었던 잠재적 차별기능문항의 존재를 파악할 수 있는 장점이 있다.

잠재적 차별기능문항은 차별기능문항의 개념을 유연하게 확장한다는 측면에서 장점이 있으나 차별기능이 발생한 잠재집단의 특성이 무엇인지 명확하게 알 수 없다는 측면에서 한계가 존재한다. 일반적으로 잠재적 차별기능문항의 원인을 규명하기 위하여 차별기능문항이 나타나는 잠재집단의 질적 특성을 파악하는 사후분석을 수행하며, 주로 집단 간 차이가 큰 문항모수에 기반하여 잠재집단을 정의한다. 하지만 잠재집단 간 문항모수의 차이만으로 잠재집단의 질적 특성을 명확하게 정의하는 것이 어려운 경우가 많다. 이런 경우, 문항이나 검사 대상에 대한 추가적인 정보를 잠재집단 추출을 위한 통계적 모형에 직접적으로 반영할 수 있다면, 잠재집단 특성 정의 및 차별기능문항 원인 파악이 더욱 쉬워질 수 있을 것이다.

측정학적으로도 문항반응이론 모형에 검사 대상의 특성 정보(예: 성별, 가정의 사회경제적지위, 학습 동기 수준 등)를 부가적인 변수로 투입했을 때 차별기능문항 판별의 정확도가 향상된다는 연구 결과가 보고되었다(Cohen & bolt, 2005; Dai, 2013; Tay, Newman, & Vermunt, 2011; Tay, Vermunt, & Wang, 2013). 최근에는 기계학습 기법을 적용하여 검사 대상에 대한 다양한 변수를 종합적으로 활용한 차별기능문항 탐색 방법이 주목받고 있다. 대표적으로 재귀적 자료 분할 기법인 Rasch-tree(Strobl, Kopf, & Zeileis, 2015)를 적용하여 한 번의 분석으로 활용할 수 있는 모든 변수의 조합을 고려하여 차별기능문항을 분석하였다(장윤선, 이주연, 2023). 그러나 검사 대상에 대한 정보가 매우 제한적으로 수집되는 경우가 일반적이기 때문에 이를 부가적인 변수로 활용하거나 기계학습 기법에 기반한 분석을 적용하는 것에 한계가 있다.

한편, 검사를 구성하는 문항의 특성은 모든 검사 상황에서 기본적으로 파악할 수 있는 정보이므로 차별기능문항 탐색에 이를 활용하는 것도 고려해 볼 수 있다. 실제로 많은 연구에서 차별기능문항으로 판별된 문항의 유형이나 문항이 측정하는 내용 영역 등

문항 특성에 대한 정보를 바탕으로 차별기능의 원인을 파악한다(Bundsgaard, 2019, Li, Cohen, & Ibarra, 2004). 문항 특성 정보를 사후분석에서 활용하는 대신, 차별기능문항 분석에 직접적으로 반영하는 것을 고려해 볼 수 있다. 이와 같은 접근으로 검사 개발의 기본 틀(framework)을 바탕으로 문항이 사용한 내용 요소, 과제가 요구하는 인지 과정, 문항 유형 등에 따라 정의된 문항군 수준의 차별적 기능을 판별하는 차별기능문항군(differential bundle functioning, DBF) 연구가 수행되었다(손원숙, 2012; Finch, 2012; Gierl et al., 2001; Latifi et al., 2016; Oshima et al., 1998).

차별기능문항군 분석은 개별 문항 수준이 아닌 공통의 특성을 가진 여러 문항으로 구성된 문항군 단위의 차별기능을 탐색함으로써 차별기능의 원인을 명확하게 파악하고, 이러한 정보를 보다 유용하게 활용할 수 있는 이점이 있다(손원숙, 2012; Gierl et al., 2001). 특히 교육적 정보로 활용하는 측면에서 차별기능문항군 분석 결과가 유용할 수 있을 것이다. 아울러 개별 문항 수준에서 작은 수행의 차이가 결합되어 문항군 수준에서는 차별기능에 따른 수행의 차이가 증폭될 수 있는 상황에서 차별기능문항보다 차별기능문항군 분석이 더 적합한 방법이라 할 수 있다(손원숙, 2012; Gierl et al., 2001; Nandakumar, 1993).

기존에 수행된 차별기능문항군 분석 연구에서는 대부분 문항 특성 정보를 사용하여 문항군을 정의하고, 전통적인 차별기능문항 연구와 같이 성별, 국가 등의 관찰집단에 따른 차별기능문항군 탐색에 중점을 둠에 따라 차별기능의 잠재적 기저를 탐색하는 것에 한계가 여전히 존재한다. 이에 이 연구에서는 검사의 문항에 대한 특성 정보를 하나의 변인으로 활용한 혼합 문항반응이론 모형을 설정하고, 문항 특성에 따른 문항반응이 이질적인 잠재집단을 탐색함으로써 검사에 존재하는 문항군 단위의 차별기능을 탐색하고자 하였다. 즉, 문항 특성에 따른 잠재적 차별기능문항군을 탐색하고, 차별 기능이 나타난 문항 특성과 잠재 집단의 특성을 바탕으로 차별기능군에 대한 해석을 교육적 지원 차원으로 제안해 보고자 하였다.

Ⅱ. 이론적 배경

1. 문항 특성으로 인한 차별적 기능

차별기능문항 연구는 주로 성별, 국가 등 집단 간 문항 반응의 차이를 밝히는 것에 중점을 두어왔다. 하지만 문항 특성 역시 문항 반응에 영향을 주며, 이러한 문항 특성의

영향은 집단에 따라 다를 수 있다. 일반적으로 차별기능문항 또는 차별기능문항군 분석에서는 문항이 측정하는 인지 요소(이해, 분석, 평가 등), 문항의 내용 영역(수학 검사의 대수, 기하 등), 문항의 유형(선다형 또는 구성형), 문항 위치를 문항 특성으로 고려한다(Bolsinova et al., 2024; Oshima et al., 1998).

문항이 요구하는 인지 요소 또는 과정에 따른 차별기능 및 문항 반응의 차이를 살펴보면, 다음과 같다. 수학 교과에서 나타나는 차별기능 문항의 특성을 분석해 보면, 남학생에게는 기하학 및 공간적 사고가 필요한 문항이 유리하고, 여학생에게는 대수 및 정확한 계산과 절차적 문제 해결이 요구되는 문항이 유리한 경향이 있다(Li, Cohen, & Ibarra, 2004; Liu & Wilson, 2009). 또한, 윤지영, 이윤선(2013)의 연구에서는 TIMSS 2007 수학 검사 가운데 추론을 요구하는 문항은 한국보다 싱가포르 학생에게 유리하고, 알기 관련 문항은 한국 학생에게 더 유리하다는 결과가 보고되어 인지 요소에 따른 문항 반응의 차이가 국가 간에서도 발생할 수 있음을 보여주었다.

문항 내용 및 맥락과 관련된 차별기능을 분석한 Xie & Wilson(2008)의 연구에 따르면, PISA 2003 수학 영역 검사의 문항 가운데 공간과 도형 관련 내용의 문항은 일본 학생에게 유리하였고, 불확실성 관련 내용의 문항은 캐나다 학생에게 유리하게 기능하였음이 확인되었다. 더불어, 문항의 맥락에 따른 차별기능문항 분석 결과 개인적 상황을 포함한 문항이 일본 학생에게 유리한 것으로 나타났다(Xie & Wilson, 2008). 국가 간 차별기능문항은 문항 유형에 의해서도 발생하였다. PISA 2015 과학 영역 검사의 문항 유형에 따른 국가 간 차별기능문항 분석 결과, 서술형 문항의 경우 45개 국가에서 일관된 차별기능문항이 존재하였다(Bolsinova et al., 2024).

차별기능문항군 관련 선행연구를 통해서도 인지 과정 및 내용 요소에 따른 다양한 차별기능이 존재함을 확인할 수 있다. 1997년에 실시된 캐나다 알버타 지역 6학년 학생 대상 과학 성취도 검사의 자료를 사용하여 성별에 따른 검사의 내용 요소 및 인지 수준별 차별기능문항군 판별 결과, 특정 내용 영역(관찰과 추론, 공기역학)과 인지 수준(지식, 기술) 문항군에서 차별기능이 존재한다는 결과가 보고되었다(Gierl et al., 2001). Latifi 외(2016)의 연구에 따르면, 국가 중등교육 자격 시험의 영어, 수학, 물리 과목 검사에 대한 파키스탄 10학년 학생의 응답 자료를 분석한 결과 수학과 물리 검사의 차별기능문항군은 존재하지 않았으나 영어 검사의 듣기 관련 문항군이 여학생들에게 더 유리하게 기능하였다. 또한, 읽기 검사에서는 문항의 인지 수준이나 문항이 요구하는 과업보다 읽기 지문으로 인한 강한 차별기능이 존재한다는 결과가 보고되기도 하였다(Oshima et al., 1998).

문항 특성에 따른 차별기능은 아니나 응답 반응의 차이를 분석한 연구도 있다. Kubinger(2009)에 따르면 같은 문항을 다른 응답 형식으로 제시하였을 때 문항의 난이

도가 크게 달라졌는데, 단일 정답이 있는 선다형 문항으로 제시했을 때의 난이도는 자유형으로 제시했을 때보다 낮지만, 정답 수를 알려주지 않은 선다형 문항의 난이도는 자유형보다도 더 높은 것으로 나타났다. 또한, 검사를 구성하는 문항의 위치 정보를 사용하여 문항 모수를 문항 내용에 의한 모수와 문항의 위치에 의한 모수로 나누어 추정한 결과, 문항 순서에 따른 응답 양상이 달라지는 문항 위치 효과가 유의미하였다(Kubinger, 2009). 이같이, 문항의 위치 정보를 비롯하여 문항 유형, 지문 길이 등 다양한 문항 특성에 따른 문항 반응의 차이를 탐색하고 이를 검사 타당도의 근거로 활용할 수 있다. 나아가 학습자의 특성을 이해하고 이를 기반으로 한 교육적 처치나 지원을 제공하는데 필요한 하나의 정보로 활용할 수도 있을 것이다.

2. Saltus 모형

Saltus 모형은 문항반응이론 모형의 하나인 Rasch 모형의 확장으로서, 피아제의 발달 단계 이론과 같은 인지 발달의 불연속적 특성을 정량적으로 모형화하기 위해 개발되었다(Draney & Wilson, 2007; Wilson, 1989). 예컨대, 학습자가 전조작기에서 구체적 조작기로 단계적 발달이 발생하는 경우, 학습자의 잠재 구인(능력)은 연속성을 가진 척도가 아닌 분절(segmentation)된 불연속적 특성을 가지게 된다. Saltus 모형은 인지 발달을 측정하는 과정에서 관찰되는 이러한 비연속적인 변화를 잠재집단과 집단 간 문항 난이도 차이로 설명한다. 즉, 전조작기와 구체적 조작기라는 발달 단계는 서로 다른 잠재집단으로 구분할 수 있고, 학습자가 한 단계 도약함에 따라 발생하는 잠재 능력의 불연속성은 발달 단계(잠재집단) 간 문항 난이도의 차이로 나타낼 수 있다(Wilson, 1989).

Saltus 모형은 Rasch 모형의 확률적 구조에 기반하여, 개인의 능력(θ)과 문항 난이도(β)와 더불어 특정 잠재집단 간의 차이를 나타내는 모수(τ, 이하 Saltus 모수)를 추가함으로써 불연속적 발달 단계를 모형화한다. 개념적으로 Saltus 모형의 잠재집단은 문항군에 의해 정의되기 때문에 잠재집단의 수와 문항군의 수는 같다고 가정한다. 만약 한 검사가 전조작기에서 구체적 조작기 사이의 학습자 능력을 측정하는 목적으로 10개의 문항을 개발하였고, 처음 5개 문항은 모든 발달 단계의 학습자가 모두 정확하게 응답할 수 있는 문항이지만 나머지 5개 문항은 구체적 조작기에 도달한 학습자들이 정확하게 응답할 수 있는 문항으로 개발하였다면, 이 경우는 2개의 잠재집단(전조작기 단계와 구체적 조작기 단계)과 2개의 문항군(1~5번 문항과 6~10번 문항)이 존재하게 된다. 그리고 첫 번째 문항군에 속한 문항 난이도는 두 잠재집단 간 다르지 않지만, 두 번째 문항군에 속한 문항 난이도는 전조작기에 해당하는 잠재집단 보다 구체적 조작기에 해당하는 잠재집단에서 상대적으로 더 낮을 것(쉬울 것)으로 예측할 수 있다. 이러한 두 잠재집단 간

두 번째 문항군의 문항 난이도 차이는 Saltus 모수로 표현된다. Saltus 모형을 일반화하면 아래와 같이 정의된다(Jeon, 2018; Wilson, 1989).

$$P(y_{ijk}=1|\theta_{jh}) = \frac{\exp\left(\theta_{jh} - \beta_i + \sum_{h=1}^{H}\phi_{jh}\tau_{hk}\right)}{1+\exp\left(\theta_{jh} - \beta_i + \sum_{h=1}^{H}\phi_{jh}\tau_{hk}\right)} \quad \text{식(1)}$$

식(1)의 y_{ijk}는 문항군 k에 해당하는 문항 i에 대한 피험자 j의 이분 반응, θ_{jh}는 잠재집단 h에 속하는 피험자 j의 능력 모수, β_i는 문항 i의 난이도, ϕ_{jh}는 피험자 j가 잠재집단 h에 속함을 나타내는 지시자(소속 = 1, 미소속 = 0)를 의미한다. 그리고 τ_{hk}는 잠재집단 h와 문항군 k 간 난이도 차이를 정량화한 Saltus 모수를 의미한다.

Jeon(2018)은 식(1)을 로짓(logit) 모형의 형식으로 다음과 같이 재정의하였다.

$$logit\big(P(y_{ij}=1|\theta_{jh}, C_j=h)\big) = \theta_{jh} - \beta_{ih} \quad \text{식(2)}$$

$$P(y_{ij}=1|\theta_{jh}) = \sum_{h=1}^{H}\pi_h \times P(y_{ij}=1|\theta_{jh}, C_j=h) \quad \text{식(3)}$$

식(2)에서 C_j는 피험자 j가 속하는 잠재집단을 의미하는 범주형 잠재 변수(C_j = 1, ..., H)고, β_{ih}는 잠재집단 h에서 문항 i의 난이도 모수를 의미한다. π_h는 잠재집단 h의 비율을 의미하며, 피험자는 오직 하나의 잠재집단에 속할 수 있는 상호 배타적인 잠재집단을 가정하므로 $\sum_{h=1}^{H}\pi_h = 1$로 정의된다.

식(2)를 식(1)과 비교해 보면, β_{ih}는 $\beta_i - \sum_{h=1}^{H}\phi_{jh}\tau_{hk}$와 같으나, $\phi_{jh}\tau_{hk}$는 오직 피험자 j가 속한 잠재집단 h의 τ_{hk}의 선형조합(즉, h = 1일 때, $\tau^*_{hk} = \tau_{11} + \tau_{12} + \cdots + \tau_{1K}$)과 같다. 그리고 문항 i는 문항군 k에 의해 정의되므로 $\beta^*_{i(k)h} = \beta^*_{ih} = \beta_i - \tau^*_{hk}$로 표현할 수 있다. 이상의 내용을 적용하여 앞서 가정한 2개의 잠재집단과 2개의 문항군이 존재하는 상황(H = K = 2)에서 각 잠재집단에 속한 피험자의 응답 확률을 Saltus 모형으로 표현하면 다음과 같다.

잠재집단 1: $logit\big(P(y_{ij}=1|\theta_{jh}, C_j=1)\big) = \theta_{j1} - \beta^*_{i1} = \theta_{j1} - \beta_i + \tau_{11} + \tau_{12}$ 식(4)

잠재집단 2: $logit\big(P(y_{ij}=1|\theta_{jh}, C_j=2)\big) = \theta_{j2} - \beta^*_{i2} = \theta_{j2} - \beta_i + \tau_{21} + \tau_{22}$ 식(5)

Saltus 모형은 일반적으로 모형 식별(model identification)을 위해 $\tau_{1k} = 0$과 $\tau_{h1} = 0$

제약이 적용되며(Draney & Wilson, 2007; Jeon, 2018), 이 제약에 따라 $\beta_{11}^{*} = \beta_i - \tau_{11} - \tau_{12} = \beta_i$, $\beta_{i2}^{*} = \beta_i - \tau_{21} - \tau_{22} = \beta_i - \tau_{22}$ 로 정의된다. 결과적으로 $\tau_{22} = \beta_{i1}^{*} - \beta_{i2}^{*}$가 되며, 이는 잠재집단 2의 문항군 2에 해당하는 문항 난이도가 잠재집단 1과 얼마나 차이가 있는지를 정량적으로 나타내는 모수를 의미한다. 잠재집단 h에 속하는 피험자 j의 능력 모수 θ_{jh}는 정규분포를 따르며($\theta_{jh} \sim N(\mu_h, 1)$), 모형 식별을 위해 첫 번째 잠재집단의 평균(μ_1)은 0으로 고정한다(Jeon, 2018).

Wilson(1989)의 연구에서 Saltus 모형은 원래 불연속적인 단계적 변화의 특성을 모형화하는 목적으로 개발되기는 하였으나, 기본적으로 Rasch 모형에 잠재집단 개념을 도입하였다는 점에서 혼합 문항반응이론 모형과 맥을 같이 한다. 다만, 사전에 정의되지 않는 잠재집단의 문항 모수를 탐색적으로 확인하는 혼합 문항반응이론모형과 다르게 Saltus 모형은 기존에 알고 있는 정보를 바탕으로 문항군과 잠재집단을 정의하고 잠재집단 간 문항 난이도의 차이를 확인하는 접근이라는 점에서 차이가 있다. 더불어 잠재집단별 모든 문항 난이도를 다르게 추정하는 대신 같은 문항군에 포함된 문항의 잠재집단 간 난이도 차이를 같다고 가정하므로 Saltus 모형은 제약된 확인적 혼합 문항반응이론(constrained confirmatory mixture IRT) 모형이라 할 수 있다(Jeon, 2018).

Ⅲ. 연구 방법

1. 분석 자료

Saltus 모형을 활용한 문항 특성에 따른 잠재적 차별기능문항군 탐색이라는 연구의 목적을 위하여 국제학업성취도평가인 PISA 2022 과학 검사 자료를 활용하였다. 가장 최근에 수행된 PISA 2022의 공개자료로 수학과 읽기 검사의 자료도 제공되었으나, 과학 검사와 달리 두 교과의 검사는 단계적 적응형 검사(multistage adaptive testing, MSAT)로 설계되었다는 차이가 있다(김성경 외, 2023; OECD, 2024). 단계적 적응형 검사 설계 기반의 수학과 읽기 검사의 경우, 학생의 반응에 따라 제공되는 문항이 달라지며 그로 인해 시스템 결측률이 매우 높은 문제가 발생할 수 있다. 따라서 이 연구에서는 단계적 적응형 검사가 적용되지 않은 PISA 2022 과학 검사의 응답 자료를 사용하였다. PISA 2022 과학 검사를 구성하는 측정 문항은 다양한 맥락 및 내용을 포함하고, 여러 핵심역량을 측정한며, 선다형 및 구성형 등 문항의 유형이 다양하므로 문항 특성에 따른 문항군 단위 차별기능 탐색이라는 연구의 목적에 부합하는 자료라 할 수 있다.

PISA 2022 과학 검사에 대한 한국 학생들의 문항 응답 자료는 OECD 홈페이지(https://www.oecd.org/en/data/datasets/pisa-2022-database.html)를 통해 제공되며, 함께 제시되는 코드북 및 기술보고서(OECD, 2024)를 통해 PISA 2022 과학 검사 설계 및 측정 영역 등에 대한 기본적인 정보를 확인할 수 있다. 더불어 김성경 외(2023)에서는 한국에서 실시된 PISA 2022 과학 검사의 문항 수준에 대한 정보를 상세하게 제공하고 있다. 이러한 자료와 문헌을 바탕으로 PISA 2022 과학 검사에 대한 기본 정보를 정리하면 다음과 같다. PISA 2022 과학검사는 6개 문항군으로 구성되며, 각 문항군에는 18~20개 문항이 포함되어 총 115개 문항이 사용되었다(김성경 외, 2023). 기본적으로 PISA 2022에 참여한 모든 학생이 115개 문항에 응답하는 것은 아니며, 무선적으로 2개 문항군을 할당하는 방식으로 평가가 설계되었다. 이러한 설계 특성으로 인해 모든 문항에 대한 응답자료를 분석 자료로 사용할 경우, 결측이 높게 발생하는 문제가 있다. 이에 이 연구에서는 PISA 자료를 활용한 선행연구(신효정, 2021; 장윤선, 이주연, 2023; 함은혜, 2022)와 같이 하나의 문항군을 독립적인 분석단위로 가정하여 한 개의 문항군을 선정하여 분석 자료로 사용하였다.

기본적으로 문항군에 포함된 문항 수, 해당 문항군이 할당된 학생 수를 고려하여 문항군을 선정하였다. 아울러 본 연구는 문항 특성에 따른 차별기능을 탐색하는 것이 목적이므로 문항 특성 정보로 활용할 수 있는 PISA 2022 과학 검사 평가틀의 네 가지 차원(맥락, 역량, 지식, 인지적 요구)과 문항 유형도 문항군을 선정하는 기준으로 함께 고려하였다. PISA 2022 과학 검사의 평가틀을 구성하는 네 가지 차원은 각각 3개의 하위 범주로 구분된다. 맥락 차원은 '개인적, 지역적/국가적, 전 세계적'으로 구분되고, 역량 차원은 '현상에 대한 과학적 설명, 과학 탐구의 평가 및 설계, 자료 및 증거의 과학적 해석'의 세 범주로 구분된다. 지식 차원은 '내용 지식, 절차적 지식, 인식론적 지식'으로 구분되고, 인지적 요구 차원은 상, 중, 하로 구분된다(김성경 외, 2023). 문항 유형은 크게는 선다형과 구성형으로 구분되며, 선다형은 단순선다형과 복합선다형으로 세분된다. 구성형 문항은 학생의 문항 응답이 자동으로 채점되는 구성형(C)와 채점기준표를 바탕으로 채점자가 문항에 대한 응답을 채점하는 구성형(H)로 구분된다.

분석에 사용할 문항군을 선정하기 위하여 PISA 2022 과학 검사의 문항군별 문항 특성 분포를 살펴보았다. 문항군에 따라 문항 유형 분포의 차이가 있고, 특히 구성형(C) 문항은 문항군 S5와 S6에 각 한 문항씩만 포함되었다. 맥락 차원도 문항군별 분포가 고르지 않았는데 일부 문항군은 개인적, 전세계적 맥락 문항은 포함되지 않았다. 역량 차원은 상대적으로 분포가 고른 편이었고, 지식 차원은 인식론적 지식 문항이 다른 범주에 비해 적었다. 인지적 요구가 상 수준인 문항은 5개로 S1에 가장 많이 포함되었고 나머지 문항군에는 포함되지 않거나 최대 2개였다.

<표 1> 한국 PISA 2022 과학 검사의 문항군 구성

문항군	문항 수(개)	문항 유형(개)				맥락(개)			역량(개)			지식(개)			인지적 요구(개)		
		단순	복합	구성C	구성H	개인	지역	세계	설명	설계	해석	내용	절차	인식	상	중	하
S1	20	8	7	0	5	0	18	2	10	6	4	12	7	1	5	11	4
S2	18	6	5	0	7	7	8	3	6	8	4	6	8	4	1	12	5
S3	20	8	7	0	5	1	19	0	9	4	7	8	10	2	0	12	8
S4	18	3	9	0	6	0	6	12	5	3	10	5	9	4	1	15	2
S5	20	5	6	1	8	0	11	9	8	5	7	8	10	2	2	15	3
S6	19	4	13	1	1	3	8	8	11	4	4	10	3	6	1	9	9

출처: 김성경 외(2023: 149~152)와 PISA2022 코드북(https://www.oecd.org/en/data/datasets/pisa-2022-database.html)을 바탕으로 작성함.

문항군별 문항 특성의 분포를 고려하여 대체로 문항이 고르게 분포한 S2의 응답 자료를 최종 분석 자료로 선정하였다. PISA 2022 과학 검사의 S2 문항군은 18개 문항으로 구성되었으며, 각 문항의 특성 정보는 〈표 2〉와 같다. 구성형 문항이 7개, 단순선다형 문항은 6개, 복합선다형 문항이 5개 포함되었다. 문항의 맥락 차원에서는 지역적/국가적 관련 문항이 8개로 가장 많았고, 개인 관련 문항은 7개, 전세계적 관련 문항은 3개 포함되었다. 역량 차원에서 가장 많은 문항이 과학 탐구의 평가 및 설계를 요구하였고, 현상에 대한 과학적 설명을 요구하는 문항은 6개, 자료 및 증거의 과학적 해석을 요구하는 문항은 4개 문항이다. 지식 차원에서 절차적 지식 관련 문항이 8개로 가장 많고, 내용 지식 관련 문항 6개, 인식론적 지식 관련 문항이 4개였다. 다만, 인지적 요구 차원은 평가도구가 측정하고자 하는 과학 역량의 수준과 직접적으로 연관되므로 본 연구의 분석에 활용하는 문항 특성 정보에서 제외하였다.

<표 2> 한국 PISA 2022 과학 검사 S2 문항 특성

단위문항	문항 ID	문항 유형	맥락	역량	지식
펭귄 섬	CS425Q02S	단순선다형	지역적/국가적	자료 및 증거의 과학적 해석	절차적
	DS425Q03C	구성형	지역적/국가적	현상에 대한 과학적 설명	내용
	DS425Q04C	구성형	전 세계적	과학 탐구의 평가 및 설계	인식론
	CS425Q05S	단순선다형	지역적/국가적	과학 탐구의 평가 및 설계	절차적
녹색 공원	CS438Q01S	복합선다형	지역적/국가적	과학 탐구의 평가 및 설계	절차적
	CS438Q02S	단순선다형	지역적/국가적	과학 탐구의 평가 및 설계	절차적
	DS438Q03C	구성형	지역적/국가적	과학 탐구의 평가 및 설계	인식론
암모나이트	CS608Q01S	복합선다형	지역적/국가적	현상에 대한 과학적 설명	내용
	CS608Q02S	복합선다형	전 세계적	자료 및 증거의 과학적 해석	인식론

단위문항	문항 ID	문항 유형	맥락	역량	지식
	CS608Q03S	단순선다형	전 세계적	현상에 대한 과학적 설명	내용
	DS608Q04C	구성형	지역적/국가적	자료 및 증거의 과학적 해석	절차적
뇌로 조공하는 로봇공학	DS610Q01C	구성형	개인적	현상에 대한 과학적 설명	내용
	CS610Q02S	단순선다형	개인적	현상에 대한 과학적 설명	내용
전구 비교	CS643Q01S	단순선다형	개인적	자료 및 증거의 과학적 해석	절차적
	CS643Q02S	복합선다형	개인적	과학 탐구의 평가 및 설계	절차적
	DS643Q03C	구성형	개인적	현상에 대한 과학적 설명	내용
	CS643Q04S	복합선다형	개인적	과학 탐구의 평가 및 설계	절차적
	DS643Q05C	구성형	개인적	과학 탐구의 평가 및 설계	인식론

출처: 김성경 외(2023: 150).

PISA 2022 과학 검사에 대한 문항별 응답은 기본적으로 '0(No credit)' 또는 '1(Full credit)'의 이분 변수로 제공된다. 이와 더불어 유효하지 않은 응답에 대해서는 'Not reached', 'Not applicable', 'Invalid', 'No Response'로 구분하여 제공한다. 본 연구에서는 유형에 상관없이 유효하지 않은 응답은 모두 0으로 변환하여 분석 자료로 사용하였다. 단, S2 문항군의 18개 문항 가운데 하나 이상의 문항에 유효 응답이 있는 경우로 한정하였다. 즉, 모든 문항에서 'No credit'인 경우는 포함하였으나, 모든 문항의 응답 값이 'No response'로 입력된 경우는 분석 자료에서 제외하였다.

PISA 2022 과학 검사 중 S2 문항군이 할당된 한국 학생은 총 842명이었으며, 응답 분포는 〈표 3〉과 같다. 18개 문항에서 Invalid에 해당하는 경우는 없었고, 그 외 유효하지 않은 응답(Not Reached, Not Applicable, No Responses) 비율은 대부분 1% 미만이었으나 일부 문항은 10%를 초과하기도 하였다. 대체로 무응답 비율은 구성형 문항 문항이 높았다. 문항별 응답 분포를 살펴본 결과, 모든 문항에서 유효 응답이 없는 경우가 2건 있었고 최종적으로 이를 제외한 총 840명의 문항 응답을 분석 자료로 사용하였다.

<표 3> 한국 PISA 2022 과학 검사 S2의 응답 분포(N=842)

문항 ID	No credit		Full Credit		Not Reached		Not Applicable		No Responses	
	명	(%)	명	(%)	명	(%)	명	(%)	명	(%)
CS425Q02S	275	(32.66)	559	(66.39)	3	(0.36)	0	(0.00)	5	(0.59)
DS425Q03C	443	(52.61)	363	(43.11)	2	(0.24)	5	(0.59)	29	(3.44)
DS425Q04C	519	(61.64)	259	(30.76)	3	(0.36)	5	(0.59)	56	(6.65)
CS425Q05S	205	(24.35)	632	(75.06)	2	(0.24)	0	(0.00)	3	(0.36)
CS438Q01S	210	(24.94)	627	(74.47)	3	(0.36)	0	(0.00)	1	(0.12)
CS438Q02S	267	(31.71)	564	(66.98)	4	(0.48)	0	(0.00)	6	(0.71)
DS438Q03C	447	(53.09)	295	(35.04)	4	(0.48)	5	(0.59)	91	(10.81)

문항 ID	No credit		Full Credit		Not Reached		Not Applicable		No Responses	
	명	(%)	명	(%)	명	(%)	명	(%)	명	(%)
CS608Q01S	415	(49.29)	412	(48.93)	4	(0.48)	0	(0.00)	11	(1.31)
CS608Q02S	291	(34.56)	536	(63.66)	5	(0.59)	0	(0.00)	10	(1.19)
CS608Q03S	393	(46.67)	436	(51.78)	6	(0.71)	0	(0.00)	7	(0.83)
DS608Q04C	309	(36.70)	511	(60.69)	6	(0.71)	5	(0.59)	11	(1.31)
DS610Q01C	453	(53.80)	316	(37.53)	14	(1.66)	5	(0.59)	54	(6.41)
CS610Q02S	151	(17.93)	673	(79.93)	18	(2.14)	0	(0.00)	0	(0.00)
CS643Q01S	152	(18.05)	671	(79.69)	6	(0.71)	0	(0.00)	13	(1.54)
CS643Q02S	253	(30.05)	546	(64.85)	7	(0.83)	0	(0.00)	6	(0.71)
DS643Q03C	449	(53.33)	262	(31.12)	6	(0.71)	5	(0.59)	120	(14.25)
CS643Q04S	459	(54.51)	369	(43.82)	7	(0.83)	0	(0.00)	7	(0.83)
DS643Q05C	616	(73.16)	193	(22.92)	11	(1.31)	5	(0.59)	17	(2.02)

PISA 2022 과학 검사 평가틀의 문항 특성 정보에 따라 정의된 문항군별 신뢰도는 〈표 4〉와 같다. 문항군별 신뢰도 확인을 위해 급내상관계수(intraclass correlation coefficient: ICC)를 R 패키지 psych 2.5.3.(Revelle, 2025)를 통해 계산하였다. 문항군을 구성하는 문항이 3개로 가장 작은 맥락 특성의 전 세계적 문항군을 제외한 나머지 문항군의 ICC는 .464에서 .736으로 양호한 수준이었다(Cicchetti, 1994).

<표 4> 문항군별 신뢰도

문항 특성	ICC[lower, upper]		
문항 유형	단순선다	복합선다	구성형
	.612 [.507, .651]	.556 [.506, .602]	.661 [.624, .695]
맥락	전 세계적	개인적	지역적/국가적
	.395 [.320, .462]	.736 [.708, .762]	.616 [.576, .654]
역량	현상에 대한 과학적 설명	자료 및 증거의 과학적 해석	과학 탐구의 평가 및 설계
	.590 [.546, .632]	.642 [.601, .680]	.639 [.601, .675]
지식	내용적	인식론적	절차적
	.590 [.546, .632]	.464 [.402, .521]	.733 [.705, .760]

2. 분석 방법

가. 잠재적 차별기능문항군 분석을 위한 Saltus 모형 설정

이론적 배경에서 기술한 Saltus 모형의 특징을 고려해 볼 때, 인지 발달 단계 검증의 목적이 아닌 집단 간 능력 또는 수행에 차이가 있을 것으로 가정하는 특정 잠재집단을 분류하고, 각 집단의 특성을 문항 모수를 통해 파악하는 목적으로도 Saltus 모형을 활용할 수 있다. 예를 들어, 영어 능력 평가에서 추가적인 지원이 필요한 학생과 그렇지 않은 학생을 구분하는 목적에서 Saltus 모형이 적용 가능하다(Jeon, 2018). 이 연구에서는 문항 특성에 따라 문항군이 특정 잠재집단에서 차별적으로 기능하는지를 탐색하는 잠재적 차별기능문항군 분석을 목적으로 Saltus 모형을 사용하고자 하였으며, 이를 위해 다음의 절차를 통해 Saltus 모형을 설정하였다.

첫째, 하나의 Saltus 모형에는 하나의 문항 특성을 반영하였다. 이 연구에서는 문항 유형, 맥락, 역량, 지식차원이라는 문항 특성에 따른 잠재적 차별기능문항군을 탐색하고자 하였으므로, 각 문항 특성을 반영한 4개의 Saltus 모형을 설정하여 분석하였다.

둘째, 전술한 바와 같이 기본적으로 Saltus 모형은 문항 특성에 따라 정의된 문항군 수와 같은 수의 잠재집단을 가정한다. 이 연구에서 사용한 문항 특성은 모두 3개의 하위 범주로 문항을 구분하므로 각 Saltus 모형은 3개의 잠재집단을 갖는 것으로 설정하였다. 이를테면 문항 유형을 반영한 Saltus 모형은 단순선다형, 복합선다형, 구성형 문항군으로 구분되고, 이에 따라 잠재집단의 수를 3개로 설정하여 각 피험자가 세 잠재집단에 속할 확률을 추정하였다. 일반적인 잠재집단 모형에서는 모형 적합도와 같은 통계적 근거를 바탕으로 최적의 잠재집단 수를 결정하지만, Saltus 모형은 문항군 수가 잠재집단 수를 정의하는 기준이 되므로 정보함수지수 및 우도비 검증 등의 모형적합도 비교는 별도로 수행하지 않았다.

셋째, 각 Saltus 모형의 첫 번째 문항군에 해당하는 문항 모수는 모든 잠재집단에서 같다고 가정하였다. 따라서 첫 번째 문항군은 일종의 잠재집단 간 가교 문항(anchor item)의 역할을 하게 된다. 그리고 나머지 문항군의 문항 모수는 첫 번째 잠재집단을 기준으로 나머지 잠재집단에서 어느 정도 차이가 있는지를 Saltus 모수(τ_{hk})를 통해 파악하였다. 만약, Saltus 모수가 통계적으로 유의미하다면 이는 해당 문항특성이 첫 번째 잠재집단과 다르게 기능하였다고 해석하였다. 즉, 첫 번째 잠재집단이 전통적 차별기능문항 분석 상황에서 참조집단의 역할을 하고, 나머지 잠재집단은 관심집단 역할을 하게 된다. 그리고 Saltus 모수의 추정값을 바탕으로 각 문항군의 잠재적 차별기능 여부와 정도를 파악할 수 있다.

문항 유형을 반영한 Saltus 모형에서 단순선다형을 첫 번째 문항군으로 설정한 경우로 예를 들어 설명하면 다음과 같다. 모든 문항에 대한 첫 번째 잠재집단의 문항 모수가 추정되고, 단순선다형 문항군에 해당하는 8개 문항의 모수는 세 잠재집단에서 모두 같다고 가정하였으므로 첫 번째 잠재집단의 단순선다형 문항의 모수는 나머지 두 집단에서도 같은 값으로 고정된다. 그리고 복합선다형 문항군이 첫 번째 잠재집단에 비해 두 번째 잠재집단에게 유리(또는 불리)하게 기능한 정도(τ_{22})와 구성형문항군이 두 번째 잠재집단에게 유리(또는 불리)한 정도(τ_{23})가 추정된다. 마찬가지로 복합선다형 문항군이 첫 번째 집단보다 세 번째 잠재집단에 더 유리(불리)하게 기능한 정도(τ_{32})와 구성형 문항이 세 번째 잠재집단에게 유리(불리)한 정도(τ_{33})가 Saltus 모수로 추정된다. 이렇게 추정된 Saltus 모수와 첫 번째 잠재집단의 문항 모수 추정값에 따라 두 번째와 세 번째 잠재집단의 복합선다형 및 구성형 문항의 모수가 결정된다. 정리하면, 이 연구에서는 모든 집단의 문항 변별도는 같게 고정한 1모수 모형을 설정하였으므로, 각 Saltus 모형을 통해 문항 변별도 1개, 문항 난이도 18개, Saltus 모수 4개, 잠재집단의 능력모수 평균(μ_h) 2개, 잠재집단의 비율(π_h) 2개의 모수가 추정된다. Saltus 모형의 모수 및 개별 피험자가 잠재집단에 속할 확률 추정을 위해 ***Mplus*** 8.0(Muthén & Muthén, 1998-2017)를 활용하였고, 이를 위한 ***Mplus*** 코드는 Jeon(2018)을 참고하여 생성하였다.

나. Saltus 모형의 가교 문항군 설정

잠재적 차별기능문항 분석에서 가교 문항의 역할을 하는 문항군을 무엇으로 설정할 것인지를 결정해야 한다. 만약 차별기능문항이 존재하는 문항군을 가교 문항으로 설정하게 된다면, 모형의 적합도와 문항 모수 추정의 타당성이 위협받게 된다. 그러나 이 연구에서는 비차별기능문항에 대한 사전 정보가 없기 때문에 Wilson(1989)에서와 같이 발달 단계상 하위 수준에 해당하는 문항군을 가교 문항군으로 설정하였다.

교과 교육 연구에서 주로 사용하는 Bloom의 신교육목표분류학에 기반을 두어 지식 차원의 가교 문항군을 설정하였다. Bloom의 신교육목표분류학에 따르면 지식 차원은 사실적 지식, 개념적 지식, 절차적 지식, 메타인지 지식의 위계가 있다(김우중 외, 2024). 이에 따라 내용적 지식, 절차적 지식, 인지적 지식 가운데 내용적 지식을 지식 차원 문항 특성 Saltus 모형의 가교 문항군으로 설정하였다.

PISA 2022 과학 검사의 역량 차원은 현상에 대한 과학적 설명, 자료 및 증거의 과학적 해석, 과학의 탐구 평가 및 설계로 구분되는데, 신교육목표분류학에서 설명과 해석은 이해 수준의 해당하고, 평가와 설계는 이해보다 높은 수준으로 구분된다(김우중 외, 2024). 이 기준에 따르면, 현상에 대한 과학적 설명과 현상에 대한 과학적 설명이 인지

적 위계 상 같은 수준에 해당하므로 평가틀에 제시된 각 문항의 인지적 요구 수준을 추가로 비교하였다. 현상에 대한 과학적 설명 문항은 하 또는 중 수준에 해당하였고, 현상에 대한 과학적 설명 문항은 중 또는 상 수준의 문항이었다. 이를 토대로 현상에 대한 과학적 설명을 역량 차원의 가교 문항군으로 결정하였다.

문항 유형의 경우, 가장 단순한 형태인 단순선다형을 가교 문항군으로 사용하는 것으로 결정하였다. 마지막 문항 특성인 맥락 차원의 경우, 맥락의 수준을 결정하는 객관적 기준이 없는 관계로 역량 차원과 마찬가지로 평가틀에 제시된 인지적 요구 수준을 토대로 가교 문항을 설정하였다. 지역적/국가적 문항은 상, 중, 하 수준의 문항이 모두 포함되었으며, 개인적 문항은 중과 하 수준에 해당하는 문항이 포함되었다. 반면, 전세계적 문항은 하 수준의 문항은 없으나 모두 중 수준으로 다른 문항군과 달리 인지적 요구 수준이 일관되었다. 이에 하나의 문항군에서 일관된 인지적 요구 수준을 갖는 문항군을 가교 문항순으로 설정하는 것이 가장 타당하다고 판단하여 전세계적 문항군을 가교 문항군으로 설정하였다.

다. 제약된 Saltus 모형 설정

Saltus 모형을 통해 잠재적 차별기능문항군으로 판정된 문항군이 실제로 해당 문항군의 차별 기능을 반영한 것인지, 아니면 잠재집단의 능력 차이가 잠재적 차별기능문항군으로 나타난 것인지를 구분하고자 제약된 Saltus 모형을 추가로 설정하여 분석하였다. 일반적으로 잠재적 차별기능문항 연구에서는 차별기능문항이 아닌 문항(즉, 가교문항)에 대한 정보를 사전에 알고 있으며, 차별기능문항이 아닌 문항에 대한 문항 반응을 검사 도구가 측정하고자 하는 잠재 구인의 대리 변수(proxy)로 사용하여 잠재집단을 추출한다(Wallin, Chen, & Moustaki, 2024). 전술한 바와 같이, Saltus 모형은 하나의 문항군은 모든 잠재집단에게 동일하게 기능하는 비차별기능문항이 존재하는 것을 가정하므로 이 가교 문항군에 대한 응답을 대리 변수로 활용하여 잠재적 차별기능문항을 탐색하는 것이 가능하다.

하지만 혼합 문항반응이론 모형의 문항 모수 추정은 잠재집단의 능력 분포와 독립적이지 않으므로 혼합 문항반응이론 모형에서 활용한 차별기능문항 분석 결과에 잠재집단 간 능력 차이(group impact)와 문항의 차별적 기능이 혼재되어 나타날 가능성이 존재한다(DeMars & Lau, 2011; Frick, Strobl, & Zeileis, 2015). 그러므로 본 연구에서는 Saltus 모형을 통한 잠재적 차별기능문항군 분석 결과에 잠재집단의 능력 차이가 혼재되었는지를 확인하고자 한다. 이를 위해 Frick, Strobl, & Zeileis(2015)가 제안한 바와 같이 모든 잠재집단의 능력모수 분포 평균은 0, 표준편차는 1로 같게 설정한 제약된

Saltus 모형을 추가적으로 설정하여 분석하였다. 제약된 Saltus 모형의 결과를 잠재집단의 능력모수 분포에 동일성 제약을 가하지 않은 비제약 모형과 비교할 때 잠재적 차별기능문항군의 존재 및 정도를 의미하는 Saltus 모수의 통계적 유의미성과 추정값 크기가 달라진다면, 비제약 모형의 결과에 문항군에 따른 차별기능과 집단 간 능력 차이가 혼재된 것으로 유추할 수 있다.

라. 잠재집단 특성 분석

잠재적 차별기능문항군에 대한 해석을 위하여 문항 특성에 따른 문항군이 차별적으로 기능하는 잠재집단 특성을 분석할 필요가 있다. PISA 2022는 검사에 대한 응답자료와 함께 학생의 배경 및 학습 관련 특성에 대한 설문조사 자료도 제공하기 때문에 잠재집단 특성을 분석하는데 이러한 정보를 활용할 수 있다. 이 연구에서는 학생의 배경 변인으로 성별과 경제 · 사회 · 문화적 지위 지표를 사용하였고, 학생을 대상으로 수집한 교육맥락 변인 중 과학 과목의 학습과 관련된 9개 변인을 사용하여 차별기능문항군이 존재하는 잠재집단의 특성을 분석하였다. Saltus 모형을 통해 추출된 문항 특성별 잠재집단에서 각 학생 특성 변인의 평균을 독립표본 t-검정을 통해 비교하였다. 잠재집단 특성 분석에 사용된 학생 특성 변인에 대한 정보는 〈표 5〉에 제시하였다.

<표 5> 잠재집단 특성 분석을 위한 학생 특성 변인

변수명	변수설명	비고
성별	성별 (1:여학생, 2: 남학생)	이분
ESCS	경제 · 사회 · 문화적 지위 지표	연속척도
과학도서	가정에서 보유한 과학 분야 책 수	4점 척도
숙제 시간	과학 숙제를 위해 사용하는 시간	6점 척도
과목 선호도	나는 과학 과목을 좋아한다.	4점 척도
과목 자기 효능감	나는 과학 과목이 쉽다	4점 척도
과목 성취동기	나는 과학 과목을 잘하고 싶다	4점 척도
창의적 사고 자기 효능감	과학 실험을 위해 좋은 아이디어를 많이 생각할 수 있다	4점 척도
교내 과학 동아리	학교 과학 동아리 참여 정도	5점 척도
교외 과학 동아리	학교 밖 과학 동아리 참여 정도	5점 척도
ICT 사용	학교 과학수업에서 ICT 사용 정도	5점 척도

Ⅳ. 연구 결과

1. Saltus 모형의 잠재적 차별기능문항군 분석 결과

Saltus 모형을 활용한 PISA 2022 과학 검사의 문항 특성에 따른 잠재적 차별기능문항군 분석 결과는 〈표 6〉과 같다. 문항 유형을 반영한 Saltus 모형의 τ_{22}는 0.479(p 〈 .05)로 통계적으로 유의미하였으나, 나머지 Saltus 모수는 유의하지 않았다. 이는 복합선다형 문항군이 잠재집단 2에 유리하게 기능하였고, 구성형 문항군은 차별적으로 기능하지 않은 것으로 해석할 수 있다. 문항의 맥락을 반영한 Saltus 모형의 Saltus 모수 중 τ_{22}와 τ_{32}가 통계적으로 유의하여 개인적 맥락 문항군의 차별기능은 존재하나 지역적/국가적 맥락 문항군의 차별기능은 없음이 확인되었다. 각 모수의 추정값을 비교하면, τ_{22} = 0.694(p 〈 .05)로 개인적 맥락 문항군은 잠재집단 2에 유리하게 기능했다. 반면에 τ_{32} = −1.452(p 〈 .001)로 동일 문항군이 잠재집단 3에서는 불리하게 기능하였다.

<표 6> 문항특성별 Saltus 모수 추정값

Saltus 모수	유형[1]		맥락[2]		역량[3]		지식[4]	
	추정값	표준오차	추정값	표준오차	추정값	표준오차	추정값	표준오차
τ_{22}	0.479*	0.201	0.694*	0.309	0.129	0.437	0.227	0.454
τ_{23}	0.155	0.279	-0.337	0.328	0.827**	0.279	1.171***	0.241
τ_{32}	0.154	0.396	-1.452***	0.278	-1.037***	0.278	-0.163	0.302
τ_{33}	-0.397	0.460	-0.175	0.196	0.097	0.204	0.064	0.627

* p 〈 .05; ** p 〈 .01, *** p 〈 .001

1. 문항군 1=단순선다형, 문항군 2=복합선다형, 문항군 3=구성형
2. 문항군 1=전세계적, 문항군 2=개인적, 문항군 3=지역적/국가적
3. 문항군 1=현상에 대한 과학적 설명, 문항군 2=자료 및 증거의 과학적 해석, 문항군 3=과학 탐구의 평가 및 설계
4. 문항군 1=내용적 지식, 문항군 2=인식론적 지식, 문항군 3=절차적 지식

역량 차원의 잠재적 차별기능문항군 탐색 결과, τ_{23}와 τ_{32}가 통계적으로 유의미하였다. 따라서 자료 및 증거의 과학적 해석과 과학 탐구의 설계 및 평가 역량 문항군은 서로 다른 집단에게 차별적으로 기능했음을 알 수 있다. τ_{23} = 0.827(p 〈 .01)로 과학 탐구의 평가 및 설계 문항군은 잠재집단 2에 상대적으로 유리하게 기능하였다. 한편, τ_{32} = −1.037(p 〈 .001)로 자료 및 증거의 과학적 해석 문항군은 잠재집단 3에 상대적으로 불리하게 기능하였다. 지식 차원의 경우, τ_{23}만 통계적으로 유의미하였다. τ_{23} 추정값은

1.171(p < .001)로 절차적 지식과 관련된 문항군이 잠재집단 2에 상대적으로 유리한 것으로 확인되었다. 인식론적 지식 관련 문항군에서의 차별기능은 확인되지 않았다.

〈표 7〉은 Saltus 모형을 통해 추출된 잠재집단의 비율과 각 잠재집단의 능력모수 추정값 평균을 제시한 것이다. 문항 유형을 반영한 Saltus 모형의 경우, 복합선다형 문항이 유리하게 기능한 잠재집단 2의 비율이 45.4%로 가장 높았다. 복합선다형 문항이 잠재집단 2에 유리하게 기능하였음에도 불구하고, 이 집단의 능력모수 추정값 평균은 −2.251(p < .001)로 잠재집단 1보다 낮았다. 참조집단에 해당하는 잠재집단 1로 분류된 학생의 비율은 34.5%이고, 잠재집단 3의 비율은 20.1%였다. 잠재집단 3과 관련된 Saltus 모수(τ_{32}와 τ_{33})는 통계적으로 유의하지 않았으므로 이 집단에 차별적으로 기능하는 문항군은 없었으나, 능력모수 추정값 평균은 −4.761(p < .01)로 가장 낮았다.

<표 7> 문항특성별 Saltus 모형의 잠재집단 추출 결과

모형	잠재집단	빈도(명)	비율(%)	능력모수 추정값 평균	표준오차
유형	잠재집단 1	290	34.5	0.000	-
	잠재집단 2(복합선다형 문항군 유리)	381	45.4	-2.251***	0.618
	잠재집단 3	169	20.1	-4.761**	1.409
맥락	잠재집단 1	443	52.7	0.000	-
	잠재집단 2(개인적 문항군 유리)	207	24.6	1.468**	0.485
	잠재집단 3(개인적 문항군 불리)	190	22.6	-1.982***	0.373
역량	잠재집단 1	349	41.5	0.000	-
	잠재집단 2(평가 및 설계 문항군 유리)	240	28.6	-1.553***	0.389
	잠재집단 3(과학적 해석 문항군 불리)	251	29.9	-3.010***	0.539
지식	잠재집단 1	265	31.5	0.000	-
	잠재집단 2(절차적 지식 문항군 유리)	304	36.2	-0.738	1.203
	잠재집단 3	271	32.3	-2.521***	0.641

* p < .05; ** p < .01, *** p < .001

맥락 차원의 잠재집단 분류 결과, 잠재집단 1의 비율이 52.7%로 가장 높았다. 그리고 개인적 맥락과 관련된 문항군이 유리하게 기능한 잠재집단 2의 비율은 24.6%, 같은 문항군이 불리하게 기능한 잠재집단 3의 비율은 22.6%로 비슷하였다. 잠재집단 2와 3의 능력모수 추정값 평균은 각각 1.468(p < .01), −1.982(p < .001)로 개인적 맥락의 문항군이 유리하게 기능한 잠재집단 2의 평균이 가장 높고, 해당 문항군이 불리하게 기능한 잠재집단 3의 평균이 가장 낮았다.

역량 차원의 잠재집단 추출 결과도 참조집단에 해당하는 잠재집단 1의 비율이 41.5%

로 가장 높고, 나머지 두 집단의 비율은 비슷하였다. 과학 탐구의 평가 및 설계 역량을 요구하는 문항군이 유리하게 기능한 잠재집단 2의 비율은 28.6%이고, 능력모수 추정값의 평균은 −1.553(p 〈 .001)이었다. 자료 및 증거의 과학적 해석을 요구하는 문항군이 불리하게 기능한 잠재집단 3의 비율은 29.9%였고, 능력모수 추정값의 평균은 −3.010(p 〈 .001)으로 잠재집단 2와 3 모두 잠재집단 1에 비해 능력모수 추정값 평균이 낮았다.

지식 차원을 반영한 Saltus 모형의 세 잠재집단의 크기는 대체로 비슷하였다. 잠재집단 1의 비율은 31.5%로 다른 집단의 비율에 비해 낮았고, 잠재집단 2와 3의 비율은 각각 36.2%, 32.3%였다. 절차적 지식에 대한 문항군이 유리하게 기능한 잠재집단 2의 능력모수 추정값 평균은 −0.738이나 통계적으로 유의하지는 않았다. 반면, 유리하거나 불리하게 기능한 문항군이 없는 잠재집단 3의 능력모수 추정값 평균은 −2.521(p 〈 .001)로 다른 집단보다 낮은 것으로 확인되었다.

2. 제약된 Saltus 모형의 잠재적 차별기능문항군 분석 결과

잠재집단 간 능력 차이를 통제하지 않은 Saltus 모형을 활용하여 잠재적 차별기능문항군을 분석한 결과, 일부 잠재집단에서는 특정 문항군이 유리하게 기능하였음에도 다른 잠재집단에 비해 능력 수준이 낮았고, 차별적으로 기능한 문항군이 없는 잠재집단의 능력 수준도 참조집단 차이가 있는 것으로 확인되었다. 이러한 결과를 통해 Saltus 모형의 분석 결과에 잠재적 차별기능문항군의 영향과 집단 간 능력 차이가 혼재되어 있음을 유추할 수 있다. 이에 따라 각 모형의 잠재집단 능력 수준을 같게 고정한 제약된 Saltus 모형을 활용하여 잠재적 차별기능문항군 분석을 수행하였고, 그 결과는 다음과 같다.

세 잠재집단의 능력모수 평균을 0으로 고정한 제약된 Saltus 모형에서의 Saltus 모수 및 잠재집단 추출 결과는 〈표 8〉과 같다. 먼저, 문항 유형을 반영한 모형의 Saltus 모수는 모두 통계적으로 유의하지 않았다. 다시 말해서 앞서 제시한 Saltus 모형의 τ_{22}는 실제로 집단 간 능력수준 차이를 의미하며, 복합선다형 문항군의 차별적 기능을 반영한 것은 아니라고 해석할 수 있다. 잠재집단 추출 결과는 일부 학생(8명, 0.1%)은 잠재집단 3으로 분류되었으나 대부분은 참조집단에 해당하는 잠재집단 1로 분류되었다. 이러한 결과 역시 문항 유형에 따른 잠재적 차별기능문항군은 존재하지 않음을 보여주는 또 다른 근거라 할 수 있다.

맥락을 반영한 제약된 Saltus 모형의 결과에 따르면 개인적 맥락의 문항군과 지역적/국가적 맥락의 문항군의 차별기능이 존재하는 것으로 나타났다. 개인적 맥락의 문항군의 경우, τ_{22} = 2.350(p 〈 .001), τ_{32} = 3.933(p 〈 .001)로 잠재집단 2와 3에 유리하게

기능하였다. 지역적/국가적 맥락의 문항군의 경우는 τ_{23} = 0.798(p 〈 .001)로 잠재집단 2에 유리하였다. 잠재집단 추출 결과를 살펴보면, 전체의 76.0%(638명)가 잠재집단 2로 분류되었고, 나머지 24.0%(202명)는 잠재집단 1로 분류되었다. 하지만 잠재집단 3으로 분류된 학생은 없었다. 즉, 개인적 문항군과 지역적/국가적 맥락 문항군이 모두 유리하게 기능한 집단으로 분류된 학생의 비율이 가장 높았고, 개인적 맥락 문항군만 유리하게 기능한 집단으로 분류된 학생은 없었다.

<표 8> 문항특성별 Saltus 모형의 잠재집단 추출 결과

Saltus 모수	유형[1]		맥락[2]		역량[3]		지식[4]	
	추정값	표준오차	추정값	표준오차	추정값	표준오차	추정값	표준오차
τ_{22}	0.011	0.056	2.350***	0.200	-0.255	0.479	-0.364	0.187
τ_{23}	-0.012	0.386	0.798***	0.134	-0.131	0.327	-1.295***	0.197
τ_{32}	0.186	0.525	3.933***	0.633	-1.389***	0.305	-0.003	0.007
τ_{33}	-2.360	1.494	0.021	0.415	-0.896	0.196	-0.002	0.086
잠재집단	**빈도(명)**	**비율(%)**	**빈도(명)**	**비율(%)**	**빈도(명)**	**비율(%)**	**빈도(명)**	**비율(%)**
잠재집단 1	832	99.0	202	24.0	0	0.0	0	0.0
잠재집단 2	0	0.0	638	76.0	551	65.6	704	83.8
잠재집단 3	8	0.1	0	0.0	289	34.4	136	16.2

* p 〈 .05; ** p 〈 .01, *** p 〈 .001

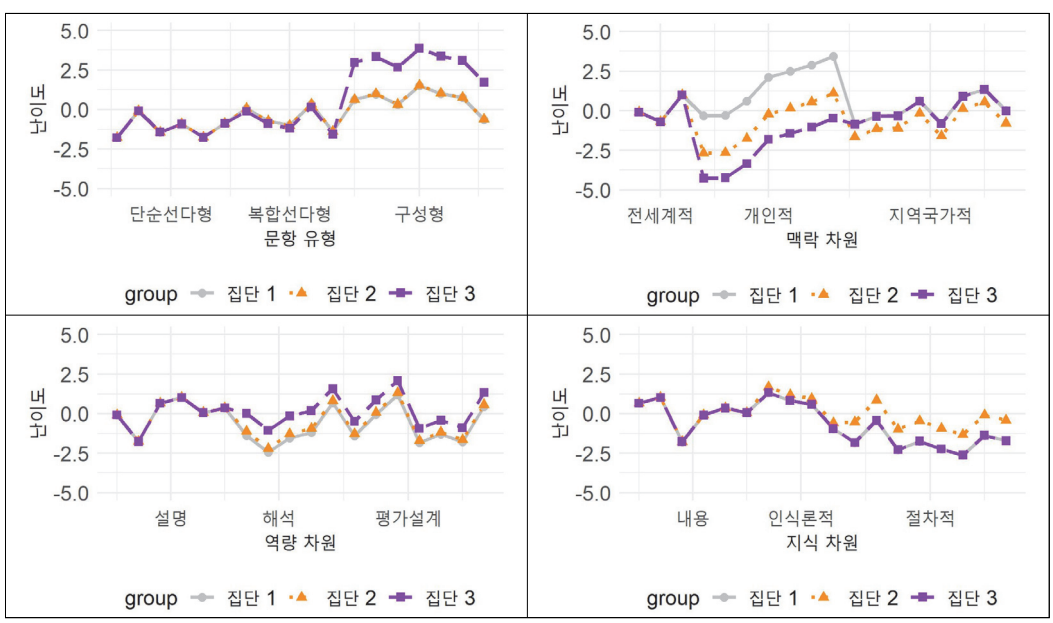
1. 문항군 1=단순선다형, 문항군 2=복합선다형, 문항군 3=구성형
2. 문항군 1=전세계적, 문항군 2=개인적, 문항군 3=지역적/국가적
3. 문항군 1=현상에 대한 과학적 설명, 문항군 2=자료 및 증거의 과학적 해석, 문항군 3=과학 탐구의 평가 및 설계
4. 문항군 1=내용적 지식, 문항군 2=인식론적 지식, 문항군 3=절차적 지식

역량 차원의 경우, 자료 및 증거의 과학적 해석이 잠재집단 3에게 불리하게 기능하였고(τ_{32} = −1.389, p 〈 .001), 앞서 제시한 기본 Saltus 모형에서 나타났던 과학 탐구의 평가 및 설계 문항군의 차별기능(τ_{23})은 없었다. 잠재집단 2로 분류된 학생의 비율은 65.6%(551명)로 가장 높았는데 잠재집단 2와 관련된 Saltus 모수(τ_{22}와 τ_{23})는 모두 통계적으로 유의하지 않으므로 잠재집단 2는 참조집단에 해당하는 잠재집단 1과 질적으로 차이가 없는 집단이라 할 수 있다. 나머지는 학생(289명, 34.4%)은 모두 자료 및 증거의 과학적 해석 문항군이 불리하게 기능한 잠재집단 3으로 분류되었다.

제약된 Saltus 모형을 적용한 지식 차원에 따른 잠재적 차별기능문항군 탐색 결과, 기본 Saltus 모형의 분석 결과에서와 같이 절차적 지식 문항군의 차별기능이 있는 것으로 나타났으나 그 방향은 반대였다. 제약된 Saltus 모형에는 τ_{23} = −1.295 (p 〈 .001)로 절차적 지식 문항군이 잠재집단 2에 불리하게 기능한 것으로 확인되었다. 잠재집단 추출

결과, 절차적 지식 문항군이 물리하게 기능한 잠재집단 2로 분류된 학생의 비율이 83.8%(704명)이고, 나머지는 모두 잠재집단 3(136명, 16.2%)으로 분류되었다. 역량 차원의 모형에서와 같이, 지식 차원을 반영한 모형에서도 잠재집단 1로 분류된 학생은 없으나, 잠재집단 3과 관련된 Saltus 모수(τ_{32}와 τ_{33})는 통계적으로 유의미하지 않으므로 잠재집단 3을 참조집단으로 볼 수 있다.

제약된 Saltus 모형을 통해 추정된 각 모형의 잠재집단 별 문항 난이도 추정값을 [그림 1]과 같이 비교하였다. [그림 1]은 각 모형에서 가교문항 역할을 하는 문항군의 난이도 가장 먼저 제시되어 있기 때문에 첫 번째 문항군에서 세 잠재집단의 문항 난이도는 같다. 이어서 각 모형의 문항군 2, 3의 잠재집단별 난이도 추정값이 제시되었고, 잠재집단 간 난이도 크기는 해당 모형에서 추정된 Saltus 모수의 크기와 같다. 문항 유형에 따른 Saltus 모형의 경우, 구성형 문항군에 대한 잠재집단 3의 난이도가 다른 집단보다 크게 추정되었으나 앞서 기술한 바와 같이 τ_{33}는 통계적으로 유의하지 않고, 추정된 난이도의 값도 모두 극단적으로 높은 점을 고려할 때, 이러한 결과는 일부 학생의 극단적인 응답 양상에 의한 결과로 해석하는 것이 타당할 것이다. 나머지 세 모형의 난이도 추정값을 비교하면, 각 모형에서 추정된 Saltus 모수의 크기가 가장 큰 역량 차원의 모형에서 개인적 문항군에 대한 잠재집단별 난이도 차가 가장 컸다. 그리고 Saltus 모수가 양의 값으로 추정된 잠재집단의 문항군 난이도는 다른 집단보다 작게 추정되어 유리하게 기능한 집단에서 해당 문항이 더 쉽게 기능하였음이 그림에서도 확인되었다.



[그림 1] 제약된 Saltus 모형을 통해 추정된 잠재집단 별 문항 난이도

3. 잠재집단 특성

제약된 Saltus 모형을 통해 문항 특성별 문항군이 차별적으로 기능하는 잠재집단의 특성을 파악하기 위해 PISA 2022 검사에서 수집된 학생 특성 변인을 분석하였다. 제약된 Saltus 모형 분석 결과, 문항 유형에 따른 잠재적 차별기능문항군은 존재하지 않았으므로 맥락, 역량, 지식 특성에 따른 잠재집단 간 비교만 수행하였다.

각 잠재집단의 성별 분포는 대체로 비슷하였는데, 현상에 대한 과학적 설명 문항군이 불리하게 기능한 역량 특성의 잠재집단 3에는 남학생(54.3%)이 여학생에 비해 다소 많았다. 지식의 경우, 절차적 지식 문항군이 불리하게 기능한 잠재집단 2로 분류된 남학생의 비율(53.3%)이 여학생의 비율보다 높으나, 참조집단에 해당하는 잠재집단 3에서는 여학생의 비율(57.4%)이 남학생보다 높았다. 학생 특성 변인에 대한 잠재집단 간 평균 차이를 독립표본 t 검정을 통해 분석한 결과는 〈표 9〉에 제시되었다.

맥락 차원에서 개인적 맥락과 지역적/국가적 맥락 문항군이 유리하게 기능한 잠재집단 2는 참조집단에 해당하는 잠재집단 1과 비교할 때, 경제 · 사회 · 문화적 지위 지표가 높고, 가정에서 보유하는 과학도서가 더 많은 배경적 특징을 보였다. 과학 과목에 대한 학습 관련 교육맥락 변인의 경우, 과학에 대한 선호도, 과학 과목에 대한 자기효능감, 성취동기 수준이 잠재집단 1의 평균보다 잠재집단 2의 평균이 통계적으로 유의하게 높았고, 과학 수업에서의 ICT 사용 정도도 잠재집단 2가 더 많았다. 하지만 과학 과목의 숙제에 사용하는 시간, 과학 실험에서의 창의적 사고에 대한 자기효능감, 교내 및 교외 과학 동이리 참여 정도는 두 잠재집단 간 유의미한 차이가 없는 것으로 나타났다.

역량의 경우, 학생의 배경 특성인 경제 · 사회 · 문화적 지위 지표와 가정 내 과학도서 수는 참조집단에 해당하는 잠재집단 2의 평균이 현상에 대한 과학의 설명을 요구하는 문항군이 불리하게 기능한 잠재집단 3의 평균보다 통계적으로 유의하게 높았다. 마찬가지로 과학 과목 선호도, 과학 과목에 대한 자기효능감도 잠재집단 2의 평균이 잠재집단 3의 평균보다 높았다. 그러나 과학 과목에 대한 성취동기와 과학 수업 중 ICT 사용 정도는 두 잠재집단 간 유의한 차이는 없었다. 교외 과학 동아리 참여 정도는 오히려 잠재집단 2보다 잠재집단 3이 더 높게 나타났다. 한편, 지식 차원에 따른 차별기능문항군이 존재하는 두 잠재집단이 추출되기는 하였으나 두 집단 간 학생 특성 변인의 평균은 모두 유의한 차이가 없었다.

<표 9> 잠재집단 간 학생 특성 변인 평균 비교

변인	전체			맥락					역량					지식				
				잠재집단 1: 참조집단		잠재집단 2: 개인적 & 지역적/국가적 차원 문항군 유리		*t*	잠재집단 2: 참조집단		잠재집단 3: 자료 및 증거의 과학적 해석 문항군 불리		*t*	잠재집단 2: 절차적 지식 문항군 불리		잠재집단 3: 참조집단		*t*
	사례 수	평균	표준 편차	평균	표준 편차	평균	표준 편차		평균	표준 편차	평균	표준 편차		평균	표준 편차	평균	표준 편차	
ESCS	827	0.252	0.839	-0.040	0.976	0.340	0.772	-4.957***	0.348	0.763	0.063	0.944	4.361***	0.257	0.860	0.227	0.725	0.376
과학도서	697	2.660	1.137	2.308	1.113	2.764	1.124	-4.500***	2.792	1.111	2.403	1.146	4.329***	2.659	1.125	2.667	1.209	-0.068
숙제 시간	808	1.979	1.105	1.880	1.105	2.008	1.104	-1.378	2.011	1.119	1.914	1.074	1.172	1.991	1.114	1.917	1.056	0.708
과목 선호도	822	2.386	0.984	2.005	0.911	2.500	0.977	-6.451***	2.495	0.987	2.170	0.944	4.514***	2.389	0.989	2.370	0.960	0.197
과목 자기 효능감	821	2.127	0.862	1.889	0.832	2.198	0.858	-4.376***	2.185	0.856	2.011	0.863	2.751**	2.133	0.880	2.096	0.762	0.493
과목 성취동기	821	3.476	0.751	3.116	0.932	3.584	0.651	-6.439***	3.590	0.641	3.251	0.891	5.614	3.464	0.769	3.541	0.655	-1.091
창의적 사고 자기 효능감	403	2.400	0.882	2.301	0.884	2.433	0.880	-1.315	2.419	0.893	2.359	0.860	0.643	2.408	0.878	2.354	0.909	0.455
교내 과학 동아리	757	1.871	1.194	1.776	1.134	1.898	1.210	-1.167	1.888	1.198	1.835	1.187	0.577	1.874	1.199	1.853	1.173	0.186
교외 과학 동아리	743	1.373	0.875	1.482	0.950	1.340	0.849	1.753	1.323	0.829	1.474	0.953	-2.123*	1.396	0.897	1.254	0.745	1.858
ICT 사용	784	2.365	1.428	2.145	1.255	2.430	1.469	-2.559*	2.400	1.426	2.294	1.431	0.985	2.363	1.419	2.375	1.474	-0.088

* $p < .05$; ** $p < .01$; *** $p < .001$

V. 결론 및 제언

검사 문항(군)의 차별기능 탐색 결과는 검사의 타당성 근거로 활용할 수 있고, 나아가 차별기능의 원인을 파악하여 특정 문항(군)이 불리하게 기능하는 집단을 위한 교육적 정보로도 활용할 수 있다. 본 연구는 사전에 결정된 특정 관찰집단에서 차별적으로 기능하는 문항의 존재 여부를 판정하는데 중점을 두는 대신, 차별기능의 원인에 대한 유의미한 정보를 얻고자 검사의 평가틀을 통해 제공되는 문항 특성 정보를 활용하여 잠재적 차별기능문항군을 탐색하고자 하였다. 이를 위해 확인적 혼합문항반응이론 모형의 하나인 Saltus 모형을 활용하였고, 문항 특성에 따른 PISA 2022 과학 검사의 잠재적 차별기능문항군을 탐색하였다. 분석을 위해 사용한 문항 특성은 PISA 2022 과학 검사 평가틀에 제시된 문항 유형, 문항이 측정하는 맥락, 역량, 지식 차원이고, 각 문항 특성을 바탕으로 문항군을 정의하였다. 연구를 통해 도출된 주요 결과와 그에 대한 논의는 다음과 같다.

첫째, Saltus 모형을 활용한 잠재적 차별기능문항군 분석 결과 단순선다형, 복합선다형, 구성형 문항 가운데 복합선다형 문항군이 유리하게 기능하는 잠재집단이 존재하였다. 문항의 맥락 차원에서는 개인적 맥락의 문항군이 유리하게 기능하는 집단과 불리하게 기능하는 2개의 서로 다른 잠재집단이 존재하는 것으로 나타났다. 역량 차원의 경우, 과학 탐구의 평가 및 설계를 요구하는 문항군이 유리하게 기능한 잠재집단과 자료 및 증거의 과학적 해석이 불리하게 기능한 또 다른 잠재집단이 존재하였다. 지식 차원을 반영한 모형의 분석 결과, 절차적 지식에 대한 문항군이 유리하게 기능하는 잠재집단이 존재하였다. 이를 통해 PISA 2022 과학 영역의 문항 특성에 따른 문항군의 차별적 기능이 존재함을 확인하였다.

둘째, Saltus 모형을 통해 추출된 잠재집단이 오로지 차별기능문항군에 의해 추출된 것인지 아니면 잠재집단의 능력 수준 차이와 문항군의 차별적 기능이 혼재된 것인지를 확인하기 위해 잠재집단의 능력수준에 동일성 제약을 추가한 제약된 Saltus 모형을 활용하여 추가적인 분석을 수행하였고, 그 결과 문항 유형에 의한 잠재적 차별기능문항군은 없는 것으로 나타났다. 한편, 맥락 차원은 개인적 문항군 뿐 아니라 지역적/국가적 문항군도 차별적으로 기능하였다. 역량 차원에서는 과학 탐구의 평가 및 설계를 요구하는 문항군의 차별기능은 확인되지 않았고, 자료 및 증거의 과학적 해석 문항군의 차별기능만 존재하는 것으로 확인되었다. 지식 차원의 경우, 동일성 제약을 반영하지 않은 Saltus 모형의 분석 결과에서처럼 절차적 지식 문항군의 차별기능이 확인되었으나 해당 문항군이 불리하게 기능하는 잠재집단이 도출되는 차이가 있었다. 이는 만약 집단 간 능력 차이를

고려하지 않았을 때 추출된 차별기능문항군의 일부는 실제로는 차별적으로 기능하지 않았음에도 잘못 해석될 가능성이 존재함을 보여주는 경험적 근거라 할 수 있다. 따라서 Saltus 모형을 비롯한 잠재집단모형을 통한 잠재적 차별기능문항(군) 탐색에서 차별기능과 집단 간 능력 차이에 의한 효과를 분리할 수 있는 모형 설정이 결과의 타당성 확보를 위해 필수적임을 본 연구 결과를 통해 확인할 수 있었다.

셋째, 제약된 Saltus 모형을 적용한 분석 결과, 개인적 맥락과 지역적/국가적 맥락 문항군 모두 차별적으로 기능하고, 차별기능의 정도로 해석할 수 있는 Saltus 모수가 다른 문항 특성을 반영한 모형에 비해 더 큰 값으로 추정되었다. 따라서 PISA 2022 과학 검사에서 문항이 다루는 맥락적 특성이 주요한 차별기능의 원인이라 할 수 있다. 맥락적 특성은 문항이 다루는 소재 또는 상황으로 학생이 다양한 상황을 직 · 간접적으로 얼마나 경험하였는지에 따라 같은 지식과 역량을 요구하더라도 어느 상황에서 이를 활용하는가에 따라 문항에 대한 응답은 달라질 수 있을 것이다. 더욱이 다양한 상황에서 과학 지식과 역량을 활용하여 문항 해결을 요구하는 PISA 검사의 경우, 특정 집단에 친숙하거나 낯선 소재나 상황을 사용하지 않도록 주의가 필요할 것이다. 또한 차별기능의 원인을 교육적 지원의 정보로 활용하는 측면에서 다양한 맥락의 소재나 상황을 접할 수 있는 과학 교과 수업 및 평가 방식을 통해 학생의 직 · 간접 경험을 확대할 수 있는 기회를 제공하는 노력이 요구된다.

아울러 역량 차원에서 차별적으로 기능한 자료 및 증거의 과학적 해석 문항군은 대체로 절차적 지식을 측정하는 문항이고, 절차적 지식 문항군은 지식 차원에서 차별적으로 기능하였다. 이와 같은 결과를 고려해 볼 때, 여러 문항 특성의 상호작용에 의한 잠재적 차별기능문항군 탐색에 대한 추가적인 분석이 필요하다. 예컨대, 역량과 지식 차원 문항 특성의 조합을 반영한 Saltus 모형을 구축할 수도 있다. 다만, 이러한 상호작용을 모형에 반영하기 위해 가정하는 잠재집단의 수가 최대 9개(내용적 지식과 과학적 설명을 측정하는 문항, 내용적 지식과 과학적 해석을 측정하는 문항, 내용적 지식과 평가 및 설계를 측정하는 문항 등)로 증가하고, 그에 따라 추정해야 할 모수도 증가하게 된다. 따라서 여러 문항 특성을 하나의 모형에 반영할 경우, 안정적인 모수 추정을 위한 충분한 표본이 확보되어야 한다.

넷째, 맥락 및 역량에 따른 잠재적 차별기능문항군이 존재하는 것으로 확인된 잠재집단 간 학생의 배경 변인과 과학 과목 학습 관련 교육맥락 변인이 차이가 있었다. 이러한 결과를 통해 학생의 배경 특성과 학습에 대한 심리적 특성 등이 복합적으로 문항군의 차별적 기능이 발생할 수 있음을 확인하였다. 또한, 잠재집단의 특성과 차별기능의 양상을 비교해보면 학생의 배경특성과 긍정적인 학습 특성 변인에 해당하는 선호도, 자기효능감, 성취동기가 유리한 문항군이 존재하는 잠재집단에서 더 높았고, 불리하게 기능하

는 문항군이 존재하는 잠재집단에서는 낮은 경향을 보였다. 맥락에 따른 차별기능문항군이 존재하는 잠재집단의 배경 특성에서 유의한 차이가 나타났다는 것은 학교의 과학 교과 수업과 평가 방식이 학생의 직·간접적 경험을 확대하는 방향으로 개발되고 제공되어야 한다는 시사점을 지지하는 결과라 할 수 있다.

그렇지만 역량 및 지식에 따른 잠재적 차별기능문항군의 원인을 해석하고 이를 교육적 지원으로 활용하기에는 잠재집단의 특성이 충분하게 파악되지는 못하였다는 한계가 있다. 특히 역량에 따른 차별기능문항군의 원인을 파악하기 위해서는 학생이 주로 사용하는 학습 전략이나 학생의 학교에서 사용하는 수업 및 평가 방법 등에 대한 교육맥락적 정보가 더 필요할 것이다. 따라서 차별적으로 기능하는 문항(군) 분석 결과의 유용한 활용을 위해서 학습자의 특성에 대한 다양한 정보를 함께 수집할 수 있는 검사 설계가 중요하게 고려될 필요가 있다.

이 연구는 기존 차별기능문항 연구에서 일반적으로 사용하지 않았던 확인적 혼합 문항반응이론 모형의 하나인 Saltus 모형을 사용하여 잠재적 차별기능문항군을 탐색하였다는데 그 의미가 크다. 특히 Saltus 모형은 기존에 잠재적 차별기능문항(군) 탐색에 주로 사용된 혼합 문항반응이론에 비해 모형이 간단하여 모수 추정의 부담이 적다는 장점이 있다. 또한, 검사 개발의 기본 정보인 문항 특성을 활용하여 차별기능문항군을 탐색하였으므로 문항 단위의 차별기능문항 탐색에 비해 결과 해석이 더 수월할 수 있고, 그 결과를 검사 도구의 타당성 검증뿐 아니라 문항 특성에 따른 학습자의 학습 경향 분석, 교수적 지원 방안 도출 등의 목적으로도 활용이 가능하다.

하지만 이 연구는 경험적 분석만 수행한 연구로써 잠재적 차별기능문항군 탐색을 위한 Saltus 모형의 통계적 강건성과 결과의 타당성 근거를 확보하기에는 한계가 존재한다. 그렇기에 다음과 같이 후속 연구를 위한 제언을 남기고자 한다. 이 연구에서 제안한 Saltus 모형이 다양한 검사 조건에서 잠재적 차별기능문항군을 안정적이고 타당하게 탐색할 수 있는지에 대한 모의실험 연구가 요구된다. 예컨대, 전체 검사 및 문항군 길이, 피험자 수, 차별기능문항군 비율 등 다양한 조건에서 Saltus 모형의 검정력과 오류 수준을 확인하고, 잠재적 차별기능문항군 탐색에 Saltus 모형을 적용할 수 있는 타당한 조건을 파악할 필요가 있다.

더불어 가교문항의 수가 차별기능문항군 탐색 결과에 미치는 영향에 대한 추가적인 연구도 수행되어야 할 것이다. 이 연구 결과에서 가교문항의 수가 가장 적었던 맥락 차원 모형은 다른 두 문항군 모두 차별기능문항군으로 판정되었다. 이러한 결과가 실제로 맥락 차원에 따른 잠재적 차별기능문항군의 존재로 인한 것인지 상대적으로 적은 수의 가교 문항이 모형의 모수 추정에 어떠한 영향을 주었는지를 확인할 수 있는 후속 연구도 수행되어야 할 것이다. 아울러, 본 연구에서는 Saltus 모형의 기본 개념과 PISA 2022

평가틀의 정보를 활용하여 이론적 근거에 의해서만 가교문항군을 설정하였다. 하지만 가교문항군 내 차별기능이 존재할 경우, 차별기능문항군 탐색 결과가 왜곡될 가능성이 있으므로 가교문항군을 설정하는 과정에서 각 문항군 내 차별 기능 존재에 대한 통계적 검증도 후속 연구에서는 고려되어야 할 것이다.

참고문헌

김성경, 김명화, 김인숙, 이신영, 백혜선(2023). **OECD 국제 학업성취도 평가 연구: PISA 2022 결과보고서**(연구보고 RRE 2023-10). 충북: 한국교육과정평가원.

김우중, 김동석, 신영준, 권난주, 오필석(2024). Bloom의 신교육목표분류에 따른 2022 개정 과학과 교육과정 초등학교 3~4학년군 성취기준 분석 및 2015 개정 교육과정과의 비교. **초등과학교육, 43**(3), 353-364.

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

신효정(2021). 프로세스 데이터를 활용한 응답속도와 정확도의 관계 분석: PISA 2015 미국과 한국의 사례를 중심으로. **교육평가연구, 34**(3), 587-614.

윤지영, 이윤선(2013). TIMSS(2007) 수학 검사의 차별기능에 관한 연구: 한국, 미국, 싱가폴 학생을 대상으로. **교육평가연구, 26**(2), 415-439.

장윤선, 이주연(2023). Rasch-tree 방법을 이용한 PISA 2015 과학검사의 차별기능문항 탐색. **교육평가연구, 36**(1), 83-110.

함은혜(2022). 프로세스데이터를 활용한 과제참여도 측정가능성 탐색: PISA2015 과학성취도평가의 문항응답시간을 중심으로. **교육평가연구, 35**(1), 23-48.

Bolsinova, M., Tijmstra, J., Rutkowski, L., & Rutkowski, D. (2024). Generalizing Beyond the test: Permutation-based profile analysis for explaining DIF using item features. *Journal of Educational and Behavioral Statistics, 49*(2), 204-240.

Bundsgaard, J. (2019). DIF as a pedagogical tool: analysis of item characteristics in IICILS to understand what students are struggling with. *Large-scale Assessments in Education, 7*(9), 1-14.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instrument in psychology. *Psychological Assessment, 6*(4), 284-290.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*(2), 133-148.

Dai, Y. (2013). A mixture Rasch model with a covariate: A simulation study via Bayesian Markov Chain Monte Carol estimation. *Applied Psychological Measurement,*

36(5), 375-396.

De Boeck, P., Cho, S-. J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement, 35*(8), 583-603.

DeMars, C. E., & Lau, A. (2011). Differential item functioning detection with latent classes: How accurately can we detect who is responding differentially? *Educational and Psychological Measurement, 71*(4), 597-616.

Draney, K., & Wilson, M. (2007). Application of the Saltus model to stagelike data: Some applications and current development. In M. von Davier & C. H. Carastensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications (pp. 119-130) New York: NY: Springer.*

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Finch, W. H. (2012). The MIMIC model as a tool for differential bundle functioning detection. *Applied Psychological Measurement, 36*(1), 40-59.

Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational Psychological Measurement, 75*(2), 208-234.

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issue and Practice, 20*(2), 26-36.

Jeon, M. (2018). A constrained confirmatory mixture IRT model: Extensions and estimation of the Saltus model using Mplus. *The Quantitative Methods for Psychology, 14*(2), 120-136.

Kang, T., & Cohen, A. S. (2006). A mixture IRT approach to differential item functioning. *Journal of Education Evaluation, 19*(3), 165-192.

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69*(2), 232-244.

Latifi, S., Bulut O., Gierl, M., Christie, T., & Jeeva, S. (2016). Differential performance on national exams: Evaluating item and bundle functioning methods using

english, mathematics, and science assessments. *Sage Open, 6*(2).

Li, Y. L., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4*(2), 115-136.

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessment: PISA trend 2022 and 2003. *Applied Measurement in Education, 22*(2), 164-184.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén

Nandakumar, T. (1993). Simultaneous DIF amplification and cancellations: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*(4), 293-311.

OECD (2024). *PISA 2022 Technical Report*, PISA, OECD Publishing, Paris

Oliveri, M. E., Ercikan, K., Lynos-Thomas, J., & Holtzman, S. (2016). Analyzing fairness among linguistic minority populations using a latent class differential item functioning approach. *Applied Measurement in Education, 29*(1), 17-29.

Oliveri, M. E., Ercikan, K., & Zumbo, B. (2013). Analysis of sources of latent class differential item functioning in international assessment. *International Journal of Testing, 13*(3), 272-293.

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. (1998). Differential bundle functioning using DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education, 11*(4), 353-369.

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research.* R package version 2.5.3, https://CRAN.R-project.org/package=psych.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika, 80*(2), 289-316.

Tay, L., Vermunt, J. K., & Wang, C. (2013). Assessing the item response theory with covariate (IRT-C) procedure for ascertaining differential item functioning. *International Journal of Testing, 13*(3), 201-222.

Tay, L., Newman, D. A., & Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods, 14*(1), 147-1763.

Wallin, G., Chen, Y., & Moustaki, I. (2024). DIF analysis with unknown groups and anchor items. *Psychometrika, 89*(2), 267-295.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development, *Quantitative Methods in Psychology, 105*(2), 276-289.

Xie, Y., & Wilson, M. (2008). Investigating DIF and extensions using an LLTM approach and also an individual differences approach: An international testing context. *Psychology Science, 50*(3), 403-416.

논문접수 : 2025.4.3. / 수정본접수 : 2025.4.24. / 게재승인 : 2025.5.8.

ABSTRACT

Detecting latent differential bundle functioning according item features using Saltus model

Yoonsun Jang
Assistant Professor, Daegu National University of Education

This study aimed to detect differential bundle functioning by identifying latent classes that respond heterogeneously to items according to item features. For this purpose, the response data of Korean students to the science assessment in PISA 2022 were analyzed using the Saltus model, a type of mixed item response theory model. Item bundles were defined by reflecting item features such as item type, context, competency, and knowledge dimension. According to the results of the study, there was no differential bundle functioning based on item type. In the context dimension, latent groups were found in which both the personal context and regional/national context item bundles functioned advantageously. In the competency dimension, the item bundle related to the scientific interpretation of data and evidence showed differential bundle functioning, and in one latent group, this bundle was found to function disadvantageously. Differential bundle functioning according to the knowledge dimension was identified only in the procedural knowledge item bundle. Moreover, differences in background and educational contextual variables were observed across latent classes in which differential bundle functioning was present. This study aimed to provide empirical evidence for various analysis methods of differential item (or bundle) functioning by exploring latent differential bundle functioning according to item features. Furthermore, the significance of this study lies in proposing an approach that utilizes the results of detecting differential item or bundle functioning not only as evidence for test validity but also as information for instructional support.

Key Words: Latent Differential Bundle Functioning, Item Features, Mixture Item Response Model, Saltus Model, PISA 2022 Science Assessment