한국심리학회지: 일반
Korean Journal of Psychology: General
2025, Vol. 44, No. 4, 321-343
http://dx.doi.org/10.22257/kjp.2025.12.44.4.321

차별기능문항의 원인 오명세화가 탐지 정확도에 미치는 영향: 요인혼합모형 기반 시뮬레이션 연구

최 은 진 이지앤웰니스(주)
이 찬 희 경북대학교
박 중 규† 경북대학교

본 연구는 공변인을 포함한 요인혼합모형을 활용하여, 하나의 검사 내에 관찰 가능한 집단에 의한 차별기능문항(Observed DIF: ODIF)과 잠재계층에 의한 차별기능문항(Latent DIF: LDIF)이 동시에 존재하는 경우, DIF 원인의 오명세화가 탐지 정확도에 미치는 영향을 시뮬레이션을 통해 분석하였다. 이를 위해, DIF 유형, 크기, 표본 크기를 조작한 조건에서 계층 수 추정률, 검정력, 제1종 오류율, 편향 등을 비교하였다. 분석 결과, LDIF와 ODIF를 모두 포함한 모형(LDIF&ODIF 모형)이 계층 수를 가장 정확히 추정하였으며, ODIF만을 포함한 모형(ODIF 모형)은 추정률이 매우 낮았다. LDIF의 검정력은 LDIF&ODIF 모형에서, ODIF의 검정력은 ODIF 모형에서 가장 높았다. LDIF&ODIF 모형은 대부분의 조건에서 낮은 제1종 오류율을 보였고, 편향 역시 허용 범위 내 또는 다소 높은 수준에 그쳤다. 이는 두 종류의 DIF가 함께 존재하는 경우, 이를 동시에 탐지할 수 있는 요인혼합모형의 활용이 탐지 정확도를 높일 수 있음을 시사한다.

주요어 : 요인혼합모형, 차별기능문항, 잠재계층분석

† 교신저자: 박중규, 경북대학교 심리학과 교수, 대구광역시 북구 대학로 80,
Tel: 053-950-7176, Email: jkp@knu.ac.kr

일반적으로 교육학을 비롯한 다양한 분야에서 시행되는 검사는 자격 검정, 적성 평가, 특정 주제에 대한 이해도 측정 등 여러 의사결정 과정에서 중요한 자료로 활용된다(진수정, 성태제, 2004; 이대용, 김석우, 길임주, 2014; 박상은, 노현종, 2019). 이러한 의사결정은 개인뿐만 아니라 사회 및 제도 전반에도 큰 영향을 미칠 수 있다. 따라서 검사는 타당하고 신뢰할 수 있어야 하며(진수정, 성태제, 2004), 무엇보다 평가 과정이 공정하게 이루어져야 한다.

만약 특정한 속성을 가진 피험자나 피험자 집단에게 검사가 유리하거나 불리하게 작용한다면, 해당 검사는 공정성을 갖추지 못했다고 볼 수 있으며, 결과적으로 올바른 의사결정을 내리는 데 기여하지 못할 것이다(안선영, 2022). 따라서 검사의 신뢰성과 평가의 공정성을 확보하기 위해, 특정 피험자 집단에게 편향적으로 작용하는 문항이 있는지 면밀히 검토할 필요가 있다. 이러한 편향을 포함하여 검사의 공정성을 저해하는 문항을 차별기능문항(Differential Item Functioning, DIF)이라고 한다(Mellenbergh, 1994). 검사 문항에 DIF가 존재할 경우, 동일한 능력을 가진 집단이라 하더라도 측정 결과에 차이가 나타날 수 있다. 또한, DIF 문항이 검사에 포함될 경우 특정 집단이 불공정한 이익 또는 불이익를 받게 되어 평가의 공정성이 훼손되고 검사의 타당도를 떨어뜨리는 결과를 초래할 수 있다. 이러한 이유로 심리학자들과 교육학자들은 검사의 공정성과 타당도를 유지하기 위해 DIF 문항을 탐지하고 관리하는 과정을 중요하게 여겨왔다.

전통적인 DIF 탐지 방법의 한계

지난 수십 년 동안 선행 연구들은 DIF 탐지를 위한 방법론적 틀을 구축해왔다. 전통적으로 DIF 분석은 성별, 인종 또는 기타 관찰 가능한 집단 간에 DIF가 존재하는지를 중심으로 수행되었다. 기존 DIF 분석에서는 동일한 관찰된 집단, 예를 들어 같은 성별이나 같은 인종에 속한 개인들은 모두 동질적인 특성을 가진다고 가정하는데, 이는 현실적으로 충족되기 어려운 가정이다. 이는 동일한 집단에 속하더라도 개인 간에 관찰되지 않는 이질성이 존재할 수 있기 때문이다(De Ayala et al., 2002; Cohen & Bolt, 2005; Webb et al., 2008). 또한, 여러 연구에서는 DIF가 발생하는 원인이 동일한 집단에 속해 있더라도 모든 개인에게 일관되게 적용되지 않으며, 개인 간의 이질성으로 인해 관찰 가능한 특성만으로는 DIF의 발생 원인을 완전히 설명하기 어려울 수 있음을 보여주었다(Cohen & Bolt, 2005; Samuelsen, 2008).

관련 연구들에 따르면, 실제 검사 환경에서 DIF의 원인은 단순히 관찰 가능한 특성뿐만 아니라 잠재적인 특성에서도 비롯될 수 있다(Meij et al., 2010; Tay et al., 2011). 이러한 결과는 DIF가 성별이나 인종과 같은 관찰 가능한 특성뿐만 아니라 성격, 반응 양상과 같이 응답에 질적인 차이를 일으킬 수 있는 개인의 보이지 않는 하위 집단과도 연관될 수 있음을 시사한다(De Ayala et al., 2002). 더 나아가, DIF의 원인이 잘못 식별될 경우 DIF 탐지의 정확성에 영향을 미칠 수 있다는 점도 파악되었다(Meij et al., 2010).

Mantel-Haenszel(Holland & Thayer, 1986)과 SIBTEST(Shealy & Stout, 1993)와 같은 전통적인 DIF 탐지 방법의 주요한 제한점 중 하나는 개인의 보이지 않는 잠재집단에 관한 정보를 제공하지 못한다는 점이다. 이러한 한계를 극

복하기 위해 연구자들은 대안으로 혼합 모형(mixture modeling)을 도입하였다.

혼합 모형은 기본적으로 관찰되지 않는 잠재집단, 즉 잠재계층(latent classes)의 존재를 가정한다. 이러한 잠재계층을 모형에 포함함으로써 연구자들은 잠재적인 집단 간 이질성을 효과적으로 포착하여 개인이 속하는 잠재계층을 보다 정밀하게 규명할 수 있을 뿐만 아니라, 이러한 잠재계층 간에 발생하는 DIF를 탐지할 수 있다(e.g., Cohen & Bolt, 2005; DeMars & Lau, 2011; Maij-de Meij et al., 2010; Yalcin, 2018). 만약 잠재계층에 따른 DIF(latent class DIF)가 존재한다면, 서로 다른 잠재계층에 속한 개인들이 문항에 대해 다르게 반응할 가능성이 높으며, 혼합 모형은 이러한 latent DIF를 탐지하는 데 유용한 도구로 활용될 수 있다.

DIF의 발생 원인에 따른 분류

DIF 탐지에 대한 최근의 방법론적 연구들은 DIF의 발생 원인에 따라 '관찰 가능한 DIF (observed DIF: ODIF)'와 '잠재 DIF (latent DIF: LDIF)'로 구분한다. 특정 문항에 대한 응답 반응이 같은 잠재계층 내에서 성별과 같은 관찰 가능한 집단에 따라 달라진다면, 해당 문항은 '관찰 가능한 집단에 의한 DIF(ODIF)'로 정의된다. 반면, 응답 반응이 잠재적인 특성을 통제한 상태에서 잠재계층의 소속 여부에 따라 차이를 보인다면, 이는 '잠재계층에 의한 DIF(LDIF)'로 간주된다.

일부 연구들은 DIF의 발생 원인으로 잠재계층과 관찰 가능한 집단을 모두 고려할 필요가 있음을 강조한다. 예를 들어, Bilir(2009)는 관찰 가능한 집단과 잠재계층 간에 큰 차이가 존재하는 경우, 단 하나의 요인(관찰 가능한 집단 또는 잠재계층)만을 고려하여 DIF를 분석하면 추정된 모수의 편향(parameter bias)이 발생할 수 있음을 지적하였다. 또한, Tay et al.(2011)은 DIF의 원인을 보다 정확하게 이해하기 위해서는 잠재계층과 관찰 가능한 특성을 DIF 탐지 과정에서 함께 고려할 필요가 있다고 제안하였다.

Oliveri 등(2013)의 연구는 국제 평가 데이터에서 관찰되는 DIF의 원인을 심층적으로 탐구하였다. 이 연구는 전통적인 DIF 탐지 기법이 주로 성별, 언어, 국가 등 관찰 가능한 집단 특성에 의존하는 반면, 잠재계층 접근은 응답 패턴 내 이질성을 반영하는 잠재적 하위집단의 존재 가능성을 전제로 한다는 점을 강조하였다. 연구는 PIRLS 2006 국제 평가 자료를 분석하여, DIF가 관찰 가능한 집단 보다는 학생의 학업 성취 수준이나 교사의 교수법과 같은 잠재적 요인에 의해 형성된 집단과 더 밀접하게 관련될 수 있음을 실증적으로 입증하였다. 특히, DIF로 판정된 문항들 중 상당수는 낮은 성취 집단과 높은 성취 집단이 선호하는 상이한 인지 과정 유형에 기인한 것으로 나타났으며, 동일한 집단 내에서도 응답 패턴의 상당한 이질성이 존재함을 확인하였다. 해당 연구 결과는 기존의 관찰 가능한 특성에 의존한 DIF 탐지 방법이 잠재 요인들을 충분히 반영하지 못함으로써 평가의 공정성에 부정적인 영향을 미칠 수 있음을 시사한다. 이에 따라 해당 연구에서는 DIF 분석 시 표면적 특성에 한정하지 않고, 잠재적 특성과 응답 패턴의 이질성을 동시에 고려하는 분석적 접근법의 필요성을 강조하였다.

여러 연구자들은 이러한 이질적인 응답 양식을 가진 하위 집단에서 다양한 유형의

DIF를 탐지하는 데 있어 유용한 도구로서 혼합 모형을 활용해왔다. 예를 들어, Tay et al. (2011)은 혼합문항반응 이론(Mixture Item Response Model, MixIRT)을 적용하여 네 가지 유형의 DIF－균일적 ODIF, 비균일적 ODIF, 균일적 LDIF, 비균일적 LDIF－를 확인하였으며, 실제 데이터를 활용한 DIF 탐지 절차를 제시하였다. 또한, Lee et al.(2021)은 영과잉(zero-inflated) 데이터 분석에서 MixIRT 모형이 높은 검정력과 낮은 1종 오류율을 보임을 실증적으로 확인하며, MixIRT 모형의 DIF 탐지 정확성을 뒷받침하였다.

DIF 탐지에서 FMM-MIMIC 모형의 활용

DIF 탐지의 또 다른 대안적 접근법으로는 요인혼합모형(Factor Mixture Model, FMM)과 다중지표다중원인(Multiple Indicator Multiple Causes Model, MIMIC) 모형을 통합한 요인혼합-다중지표다중원인 모형(Factor Mixture Multiple Indicator Multiple Causes Model, FMM-MIMIC)이 있다. 이 모형의 가장 큰 특징은 하나의 모형 내에서 두 가지 상이한 원인(잠재계층과 관찰 가능한 특성)에 의해 발생하는 DIF를 동시에 탐지할 수 있다는 점이다(Lee & Beretvas, 2014; Lee et al., 2021; Tay et al., 2011).

Lee와 Beretvas(2014)는 다집단 구조방정식 모형(Multiple-group Structural Equation Modeling) 접근을 기반으로 DIF 탐지의 정확도를 연구한 바 있다. 그들의 연구에 따르면, 두 종류의 DIF가 모두 검사에 포함된 상황에서 LDIF 혹은 ODIF 중 하나만을 포함한 모형보다 두 DIF를 모두 포함한 FMM 모형이 DIF 탐지 성능에 있어 더욱 정확한 탐지 결과를 보였다. 이에 따라, 연구자들은 DIF 탐지의 정확성을 높이기 위해 DIF의 원인을 적절하게 설정한 FMM 모형의 활용을 권장하였다. 또한, Wang et al.(2021)은 공변인(covariate)을 포함한 FMM 모형의 DIF 탐지 성능을 분석하였으며, 잘못 설정된 공변인이 부정확한 DIF 탐지와 편향된 결과를 초래할 수 있음을 지적하였다. 이를 통해, 연구자들은 DIF 탐지 과정에서 적절한 공변인 설정의 중요성을 강조하였다. 전반적으로, 이러한 혼합 모형 기반 접근법은 DIF의 본질을 보다 심층적으로 이해하는 데 기여하며, 다양한 원인에 의해 발생하는 DIF를 탐지할 수 있는 유연한 통계적 도구로 활용될 수 있다.

몇몇 연구에서는 단일 검사 내에서 여러 DIF 원인이 존재할 경우, 실제 검사 환경에서 DIF의 원인을 적절히 설정하고, 이를 정확하게 식별하는 것이 어려울 수 있음을 지적한다(Tay et al., 2011). 예를 들어, ODIF와 LDIF가 모두 존재하는 검사에서 연구자들이 단일 DIF 원인만을 반영한 모형을 사용할 가능성이 있으며, 반대로 오직 하나의 DIF 원인만 포함된 검사에서 연구자들이 복수의 원인을 반영한 DIF 탐지 모형을 적용할 수 있다. 이러한 상황들은 실제 검사 환경에서 다양한 DIF 유형을 설정하고 이를 정확하게 탐지하는 과정이 복잡하고 도전적인 과제임을 보여준다.

추가적으로, Bilir(2009), Lee와 Beretvas(2014)가 약 2,000명의 대규모 표본을 기반으로 연구를 수행한 것과 달리, 최근 연구들은 문항 수와 표본 크기가 모두 적은 상황에서의 DIF 탐지 연구가 증가하는 추세에 있다. 이는 특히 임상 연구와 같은 특정 상황에서 대규모 표본을 수집하는 것이 현실적으로 어렵다는 점을 고려할 때, 더 작은 표본 크기를 대상으로 한 연구의 필요성이 강조되고 있음을 보여

준다. 또한, 심리학과 같은 사회과학의 일부 분야에서는 종종 10개 이하의 문항을 포함하는 검사를 사용하는데, 이런 경우 낮은 신뢰도와 높은 측정 오차를 초래할 가능성이 크다. 따라서, 이런 문제를 해결하기 위해 단순히 관찰된 점수에 의존하기보다 측정 오차를 고려한 정교한 분석을 가능하게 하는 구조방정식 모형(structural equation modeling)의 틀을 DIF 탐지에 활용하는 것이 바람직하다.

더 나아가, 기존의 많은 시뮬레이션 연구들은 균일적 ODIF(uniform observed DIF)와 LDIF (uniform latent DIF)를 탐지하는 것에 초점을 두고 있지만(Lee & Beretvas, 2014; Lee et al., 2021), 몇몇 연구들은 단일 검사 내에서 균일적 DIF와 비균일적 DIF(nonuniform DIF)가 동시에 발생할 수 있음을 제시하였다(Tay et al., 2011). 이는 다양한 조건에서 두 가지 유형의 DIF가 공존하는 경우의 탐지 정확도와 수행정도를 체계적으로 조사할 필요가 있음을 시사한다.

앞선 연구들은 주로 ODIF 또는 LDIF를 개별적으로 탐지하는 데 중점을 두었으나, 두 가지 상이한 DIF 원인이 공존하는 상황에서의 탐지 오류에 관한 체계적인 연구는 미흡한 실정이다. 이에 본 연구는 FMM의 틀 안에서 ODIF와 LDIF가 공존하는 조건 하에서 DIF 원인 오명세화가 DIF 탐지의 수행력과 정확성에 미치는 부정적 영향을 분석하며 이로 인해 발생할 수 있는 잠재적 결과를 규명하는 데 중점을 둔다. 이를 위해, 시뮬레이션을 통해 균일적 ODIF와 LDIF, 비균일적 ODIF와 LDIF가 동시에 존재하는 상황에서, 단일 DIF 원인(ODIF 또는 LDIF)만을 고려한 탐지 방식의 정확도를 평가하였다. 이러한 접근은 다중 DIF 원인을 고려하지 않는 기존 방법론의 한계를 보완함과 동시에, 균일적 DIF 중심의 선행 연구를 확장하여 비균일적 DIF 탐지에 관심 있는 연구자들에게 실증적 근거를 제공할 것으로 기대된다.

본 연구는 다음과 같은 순서로 진행된다. 먼저, FMM-MIMIC 모형에서 DIF 탐지 절차를 소개한 후, 관찰 가능한 공변인이 설정된 FMM-MIMIC 모형의 특징을 기술한다. 이후, 본 연구에서 적용한 시뮬레이션 조건과 결과를 제시하며, 마지막으로 다양한 유형과 원인을 가진 DIF 탐지 연구의 논의점을 정리하고, 실증 연구에서 DIF 탐지를 위한 시사점과 제안을 제시하였다.

차별기능문항의 탐지

DIF의 원인에 따라 구분되는 ODIF와 LDIF는 각각 MIMIC 모형 또는 FMM을 통해 모수화할 수 있다. 그림 1에 제시된 바와 같이, MIMIC 모형은 잠재 평균의 차이를 검정하고

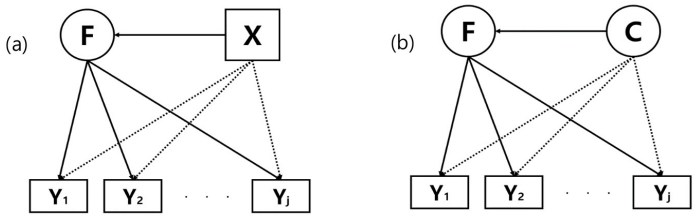

그림 1. MIMIC 모형과 FMM 모형

측정 불변성을 평가하기 위해 고안된 구조방정식 모형이다(e.g., McCarthy, Pedersen, & D'Amico, 2009; Woods, Oltmanns, & Turkheimer, 2009). 이 모형은 더미 코딩된 관찰 가능한 집단변수(X)를 포함하며, 이는 잠재 변수와 지표변수 모두에 영향을 미치는 공변인으로 기능한다. 혼합 모형에서는 이러한 변수를 보조변수(auxiliary variable)라고 부르며, 보조변수는 잠재 집단의 분류뿐만 아니라, 측정 지표의 반응에도 영향을 줄 수 있는 예측 변수로 모형에 포함된다.

ODIF는 그림 1(a)에 제시된 것처럼, MIMIC 모형 내에서 잠재변수(F)를 거치지 않고 X로부터 관찰 가능한 지표 변수 Y_j로의 직접 효과를 추가함으로써 확인할 수 있다(Woods & Grimm, 2011; Kim, Yoon, & Lee, 2012). 그림 1(a)에서 점선으로 표시된 이러한 직접 효과는 잠재 요인을 통제한 상태에서도 관찰 가능한 집단 간 차이가 존재함을 의미하며, 이는 ODIF의 존재를 시사한다.

한편, FMM은 잠재계층(C)가 Y_j에 미치는 직접 효과를 포함할 수 있으며(Wang et al., 2021), 이러한 효과는 그림 1(b)에서 점선으로 표시된 경로로 나타난다. 이와 같은 직접 효과는 잠재계층에 따라 문항의 기능이 다르게 나타나는 LDIF를 의미한다.

두 모형은 관찰 가능한 집단을 나타내는 공변인을 포함한 FMM-MIMIC 모형으로 통합될 수 있다. 해당 모형은 공변인과 잠재계층을 동시에 포함하며, 이를 통해 관찰 가능한 집단 또는 잠재 집단에 의해 발생하는 DIF를 다음의 식에서 동시에 탐지할 수 있다.

$$Y_{jk} = \tau_{jk} + \Lambda_{jk}\eta_k + \beta_{jk}X + \omega_j\eta_k X + \epsilon_{jk} \quad (1)$$

$$\eta_k = A_k + \Gamma_{\eta_k}X_i + \xi_k \quad (2)$$

j = 1, …, J이고, k = 1, …, K일 때, 식 1의 Y_{jk}는 계층 k의 지표변수 j에 대한 응답이다. X는 잠재변수와 지표변수 모두에 직접적으로 영향을 미치는 공변인이며 이러한 효과들은 회귀 계수로 모형화 할 수 있다(Lubke & Muthén, 2005). 또한, Λ_{jk}는 지표변수 j의 계층에 따른 요인 부하량을 나타내며 β_{jk}는 계층 k에서 공변인이 지표변수 j에게 영향을 미치는 회귀 계수를, ω_{jk}는 공변인과 요인 간 상호작용 항을 의미한다.

이 모형에서 잠재계층 간에는 동일하다는 가정 하에 유의한 회귀 계수 β_j는 절편의 불일치성을 나타내며, 이는 집단 간 평균에 유의한 차이가 존재함을 의미한다. 따라서, 해당 변수 j는 균일적 ODIF로 간주될 수 있다. 또한, 집단에 따라 잠재변수의 평균에 차이가 있는 경우를 고려했을 때, 각 잠재계층 내에서 관찰 가능한 집단 간 차이를 측정할 수 있다는 점에서 이 접근은 전통적인 MIMIC 모형과 구별된다(Lee, Han, & Choi, 2022).

MIMIC 모형은 본래 균일적 ODIF를 탐지하기 위한 목적으로 사용되어 왔지만, Woods와 Grimm(2011)은 비균일적 ODIF를 확인하기 위해 요인과 공변인 간 상호작용 항을 포함한 MIMIC-interaction 모형을 제안하였다. 식 1에서 상호작용 항 ω_j의 유의한 회귀 계수는 기울기의 불일치성(slope non-invariance)을 나타내므로, 이 계수가 유의할 경우 비균일적 ODIF의 존재를 시사하는 지표가 된다.

η_k의 영향을 통제한 상태에서 잠재계층 변수가 특정 지표 변수의 절편에 유의한 영향을 미친다면, 해당 지표변수는 균일적 LDIF로 간

주한다. 즉, τ_{jk}가 잠재 집단 k에 따라 다르다면, 변수 j는 균일적 LDIF이다. 만약, Λ_{jk}가 잠재계층 별로 달라진다면, 이는 계층 간 기울기의 불일치성을 나타내며, 해당 지표 변수 j가 비균일적 LDIF임을 시사한다. 잠재계층 간의 차이에 의해 발생하는 LDIF를 확인하기 위해 Lord(1980)가 제안한 카이제곱 통계치가 사용할 수 있다.

$$\chi_j^2 = \frac{(\hat{\tau}_{j1} - \hat{\tau}_{j2})^2}{\sigma^2_{\hat{\tau}_{j1}} + \sigma^2_{\hat{\tau}_{j2}}} \quad (3)$$

$\hat{\tau}_{j1}$와 $\hat{\tau}_{j2}$는 j번째 문항에 대해 관찰가능한 집단 또는 잠재계층의 절편 추정치이고 $\sigma^2_{\hat{\tau}_{j1}}$와 $\sigma^2_{\hat{\tau}_{j2}}$는 문항 j에 대한 절편 추정치의 분산을 의미하며 χ^2통계치는 자유도가 1인 분포를 따른다. 만약, χ^2통계치가 5% 유의수준에서 유의하다면 DIF가 존재한다고 판단할 수 있다. 식 3의 통계치는 본래 균일적 DIF를 탐지하기 위해 고안되었지만, 특정 문항의 요인부하량과 요인부하량의 분산으로 대체하여 비균일적 DIF를 탐지에도 활용할 수 있다 (Lord, 1980; Maij-de Meij et al., 2010).

요약하면, 균일적 ODIF와 비균일적 ODIF는 각 계층별로 식 1에 포함된 τ_{jk}와 Λ_{jk}가 동일할 때 β_j와 ω_j의 통계적 유의성을 검증함으로써 탐지할 수 있다. 한편, 균일적 LDIF와 비균일적 LDIF는 집단별로 β_j와 ω_j가 동일할 때 잠재계층 간 τ_{jk}와 Λ_{jk} 간의 차이를 비교하여 확인할 수 있다.

방 법

자료 생성

본 연구는 다양한 조건에서 DIF의 원인을 잘못 설정하였을 때 나타날 수 있는 결과를 확인하기 위해 시뮬레이션 연구를 수행하였다. 데이터 생성을 위한 모집단은 그림 2(a)에 제시된 모형, 즉 LDIF와 ODIF가 동시에 포함된 FMM-MIMIC 모형으로 설정하였으며, 생성된 데이터는 그림 2에 제시된 세 가지 분석 모형을 통해 평가하였다.

첫 번째 분석 모형은 모집단 모형과 동일한 LDIF와 ODIF를 모두 포함하는 FMM 모형이며, 두 번째와 세 번째 모형은 그림 2(b)와 2(c)에 제시된 바와 같이 각각 LDIF만 또는 ODIF만을 포함하는 단일 원인 기반 모형이다. 이 두 모형은 DIF의 원인을 잘못 지정했을 때 DIF 탐지의 정확도, 수행력, 그리고 모수 추정의 신뢰성을 정확한 모형(참모형)과 비교하기 위한 목적으로 사용되었다.

또한, 본 연구는 두 가지 DIF 유형(균일적

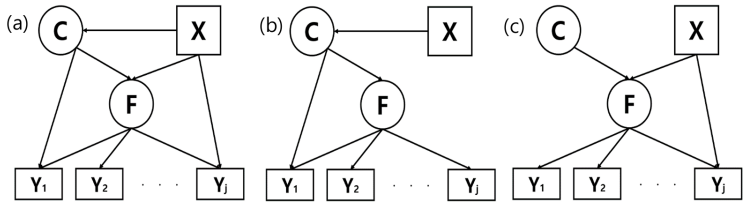

그림 2. 분석에 사용된 모형

DIF, 비균일적 DIF)을 모두 포함하고 있기 때문에, 총 2개의 모집단 모형과 이에 대응하는 6개의 분석 모형으로 구성되었다.

연구 조건

본 연구에서는 DIF 원인을 잘못 설정하였을 때 나타날 수 있는 부정적 영향을 확인하기 위해, DIF의 종류, DIF의 크기, 표본 크기의 세 가지 요인을 조작하여 시뮬레이션을 실시하였다.

균일적 DIF는 관찰 가능한 집단 또는 잠재계층 간에 문항의 절편이 다를 때 발생하며, 선행 연구들에서는 집단 간 차이가 0.2에서 0.5 사이일 경우를 작은 DIF, 0.5에서 1.5 사이일 경우를 큰 DIF로 간주하였다(윤수철, 이순묵, 2013; Lee와 Beretvas, 2014; Stark 외, 2006; Wang 외, 2021). 이를 바탕으로, 본 연구에서는 공변인에서 Y6 문항으로 향하는 직접 경로 계수를 0.6으로 설정한 경우를 작은 균일적 ODIF, 1.2로 설정한 경우를 큰 균일적 ODIF로 구성하였다. 마찬가지로, 균일적 LDIF의 경우, 참조 계층인 계층 2와 비교하여 계층 1의 Y5 문항 절편을 각각 0.6(작은 균일적 LDIF)과 1.2(큰 균일적 LDIF)만큼 크게 설정하였다.

비균일적 DIF 조건에서는 Mplus의 상호작용항 생성 기능을 활용하여 공변인과 요인 간의 상호작용 항을 포함시켰으며, Y6 문항으로 향하는 경로 계수를 각각 0.2와 0.4로 설정하여 작은 ODIF와 큰 ODIF 조건을 구성하였다. 또한, 비균일적 LDIF는 잠재계층에 따라 Y5 문항의 요인 부하량이 다르도록 설정하였다. ODIF와 마찬가지로, 잠재계층 간 요인 부하량 차이를 0.2(작은 LDIF) 또는 0.4(큰 LDIF)로 설정하여 계층 1의 요인 부하량이 각각 0.7 또는 0.9가 되도록 하였다. 이때 Y5 문항의 잔차 분산은 LDIF 크기에 따라 달라지며, 작은 LDIF 조건에서는 0.51, 큰 DIF 조건에서는 0.19로 설정되었다.

표본 크기는 500, 1,000, 2,000의 세 가지 수준으로 조작하였다. 기존의 시뮬레이션 연구들은 보통 표본 크기를 2,000으로 고정하여 LDIF와 ODIF를 동시에 측정하였으나(Bilir, 2009; Lee와 Beretvas, 2014), 임상 연구와 같이 표본 확보가 어려운 상황을 고려할 때, 소규모 표본 조건에서의 탐지 성능 또한 확인할 필요가 있다. 이에 따라 본 연구에서는 작은 표본 조건도 포함하였으며, 두 집단의 표본 분포는 동일하게 설정하여 각각 250, 500, 1,000으로 설정하였다.

고정 조건

본 연구에서 가정한 모집단 모형은 2개의 잠재계층과 1개의 요인으로 구성되며, 6개의 연속형 변수(Y1부터 Y6까지)를 지표 변수로 포함한다. 잠재계층의 비율은 균등하게 설정하였으며, 이는 각 관찰치가 두 계층에 속할 확률이 동일하도록 설정되었음을 의미한다. 요인 평균은 계층 1에서는 1, 계층 2에서는 0으로 설정하였으며, 요인의 분산은 두 계층 모두 1로 고정하였다. 모형의 식별을 위해 첫 번째 지표 변수(Y1)의 요인 부하량을 1로 고정하였고, 나머지 문항들의 요인 부하량은 DIF를 연구한 선행연구를 바탕으로 각각 0.8, 0.7, 0.6, 0.5, 0.5로 설정하였다(윤수철, 이순묵(2013), 0.6~1.0; Stark et al.(2006), 0.58~0.9; Wang et al.(2021), 0.5~0.8). 요인 부하량의 제곱과 잔차 분산의 합이 1이 되도록 하기 위

해, 각 지표 변수에 대한 잔차 분산은 순서대로 0.36, 0.51, 0.64, 0.75, 0.75로 설정하였으며, 모든 지표 변수의 절편은 0으로 고정하였다. 관찰 가능한 집단 변인은 공변인으로 포함되었으며, 두 개의 집단으로 구성하고 0과 1로 코딩하였다. 모형에는 두 개의 DIF 문항이 포함되었으며, Y5 문항은 LDIF 문항으로, Y6 문항은 ODIF 문항으로 설정하였다. 균일적 DIF 검증 조건에서는 Y5문항은 집단별로 β_j는 동일하지만, τ_{jk}가 계층에 따라 상이한 값을 가지며, Y6는 계층별로 τ_{jk}가 동일하지만 β_j가 집단에 따라 다른 값을 가지도록 설정하였다. 반면, 비균일적 DIF 검증 조건에서 Y5는 집단별로 ω_j는 동일하지만 Λ_{jk}가 잠재계층에 따라 상이한 값을 가지며, Y6는 계층별로 Λ_{jk}가 동일하나 ω_j는 집단별로 상이한 값을 가지도록 설정하였다.

분석

본 연구의 시뮬레이션은 총 12개의 조건으로 설계되었으며, 선행 연구를 참고하여 각 조건별로 100개의 데이터를 생성하였다(Lee & Beretvas, 2014; Bilir, 2009; Lee et al., 2021; Woods & Grimm, 2011).

생성된 데이터는 DIF 원인이 서로 다른 세 가지 모형, 즉 LDIF와 ODIF를 모두 포함한 모형(LDIF&ODIF 모형), LDIF만 포함한 모형(LDIF 모형), ODIF만 포함한 모형(ODIF 모형)을 통해 분석하였다. 본 연구에서는 잠재계층 수의 추정 정확성, DIF 탐지의 효과성, 모형 내 모수 추정의 정확성을 DIF 탐지의 수행력을 판단하는 기준으로 설정하였다.

잠재계층 수 추정의 정확성은 세 가지 정보지수(AIC, BIC, 조정된 BIC)를 활용하여, 전체 데이터 중 실제 잠재 계층 수를 정확히 추정한 비율로 평가하였다. 각 반복에서는 잠재계층 수가 1개에서 3개까지 설정된 FMM-MIMIC 모형을 추정하였으며, 가장 낮은 정보 지수 값을 산출한 모형을 가장 적합한 모형으로 간주하였다. 이후, 정보 지수가 참모형을 얼마나 잘 예측했는지를 나타내는 비율을 산출하였다.

1종 오류율은 DIF가 존재하지 않는 문항을 DIF로 잘못 탐지한 비율로 정의하였다. 구체적으로 LDIF의 경우, Y1 - Y4 및 Y6 문항에 대해 식 (3)을 적용하여 Λ_{jk} 와 τ_{jk} 를 각각 검증하고, 유의한 경우 1종 오류로 판정하였다. 반면, ODIF의 경우 Y1 - Y5 문항에 대해 식 (1)에 포함된 β_j와 ω_j가 유의한 경우, 1종 오류로 판정하였다. 또한, 검정력은 실제 DIF가 존재하는 문항을 정확히 탐지한 비율로 정의하였다. 이에 따라 Y5 문항에 대해서는 LDIF 탐지 결과를 바탕으로 검정력을 계산하였으며, Y6 문항은 β_j와 ω_j의 유의성을 평가하여 ODIF의 검정력을 산출하였다. 일반적으로 1종 오류율이 5% 미만이고 검정력이 80% 이상일 경우, 해당 모형의 수행력이 허용 가능한 수준으로 간주된다.

모수 추정의 정확성은 상대적 편향을 통해 평가하였다. 상대적 편향은 추정된 모수값과 참값의 차이를 참값으로 나누어 계산되며, 그 수식은 다음과 같다.

$$Bias(\hat{\theta}) = \frac{\hat{\theta} - \theta}{\theta} \quad (4)$$

LDIF의 상대적 편향은 계층 1의 절편 또는 요인부하량과 계층 2의 절편 또는 요인부하량

추정치 간 차이로 계산되며 ODIF의 상대적 편향은 공변인 또는 상호작용 항의 경로 계수추정치를 이용하여 계산된다. Hoogland & Boomsma (1998)에 따르면, 상대적 편향이 ±5% 미만이면 모수 추정치가 정확하다고 판단할 수 있다. 데이터 생성과 분석은 Mplus 8.2 (Muthén, & Muthén,, 2019)를 사용하였으며 모수 추정을 위해 EM 알고리즘을 통한 최대우도 추정을 사용하였다.

결 과

정확한 계층의 수 추정

그림 3~5는 2개 계층 모형이 참모형일 때, LDIF, ODIF, LDIF&ODIF 모형이 계층의 수를 정확히 2개로 추정한 비율을 나타낸다. 대부분의 모형에서 AIC가 가장 높은 비율로 2개 계층을 선택했으며 표본의 크기가 커질수록 비율이 더 높아지는 결과를 보였다.

모형별로 살펴본 결과에 따르면 LDIF 모형으로 계층의 수를 추정한 경우, DIF의 종류, DIF의 크기, 표본의 크기 IC의 종류에 관계없이 거의 모든 조건에서 90%이상의 정확도를 보였다. 그러나 ODIF 모형의 경우, DIF의 크기가 큰 비균일적 DIF 조건을 제외하면 AIC가 BIC와 a-BIC보다 높은 정확한 층의 수 추정 비율을 보였으나 모든 조건에서 계층의 수를 정확히 추정해내지 못하였다.

참모형인 LDIF&ODIF 모형으로 계층의 수를 추정하였을 경우, 균일적 DIF의 경우 AIC가 BIC와 a-BIC보다 더 나은 수행을 보였다. BIC는 작은 DIF 조건에서 표본의 수에 상관없이 0%의 정확성을 보였으며 그 비율은 큰 DIF와 표본의 크기가 큰 조건일 때 증가하였다.

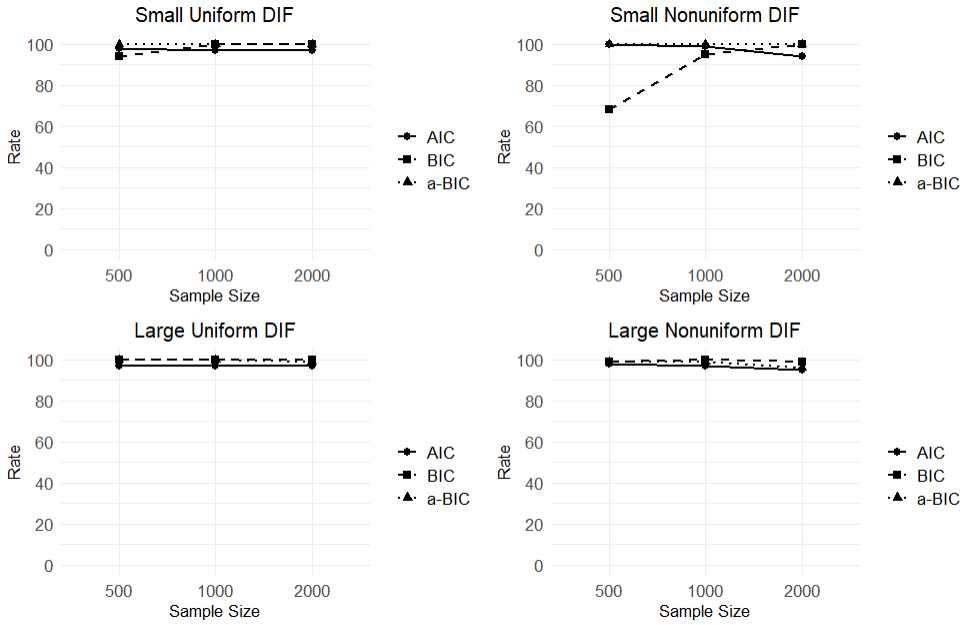

그림 3. 정확한 계층 수 추정의 결과(LDIF 모형)

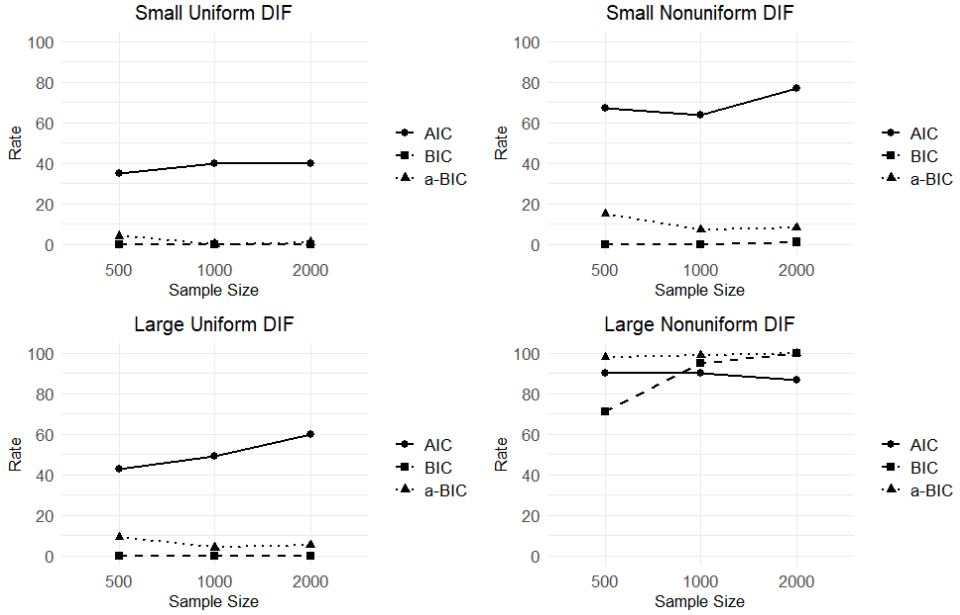

그림 4. 정확한 계층 수 추정의 결과(ODIF 모형)

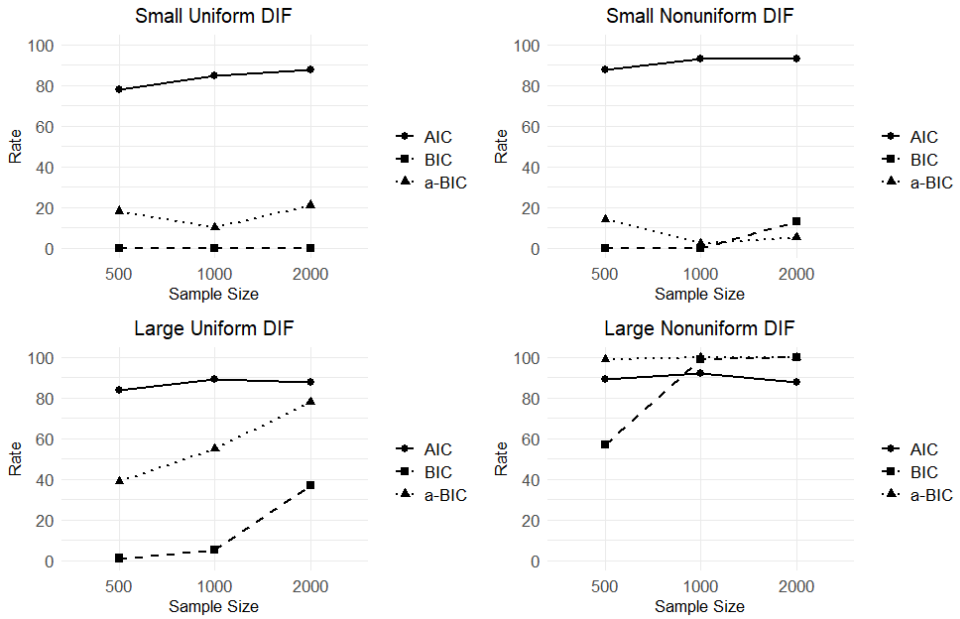

그림 5. 정확한 계층 수 추정의 결과(L&ODIF 모형)

a-BIC의 결과는 BIC의 결과와 거의 동일한 양상을 보였지만 큰 DIF와 표본의 크기가 큰 조건에서 우수한 결과값을 나타냈다. DIF의 크기가 작을 때, 비균일적 DIF의 결과 역시 균일적 DIF의 결과와 전반적으로 비슷한 양상을 보였지만 BIC의 정확도가 표본의 크기가 1,000 이상일 때 증가했으며 a-BIC의 정확도는 점차 감소하는 모습을 보였다. 반면, 비균일적이고 큰 DIF의 조건에서 BIC는 표본의 크기가 500일 때를 제외하고는 완벽하게 2개 계층을 추정하였다. a-BIC는 모든 조건에서 완벽하게 2개 계층을 추정하였으며, AIC 역시 모든 조건에서 우수한 수행력을 보였다.

결과를 종합하면 AIC가 모든 조건에서 약 80% 이상의 확률로 2개의 계층을 정확히 추정하였으며, 이는 분석에 활용된 정보지수 중 AIC가 잠재계층 수 결정에 있어 가장 일관되며 높은 정확도를 지닌 지표임을 시사한다. 이에 AIC를 기준으로 LDIF&ODIF 모형에서 잠재계층 수를 정확히 2개로 추정하지 못한 경우를 검토한 결과, 균일적 DIF 조건에서는 DIF의 크기가 클 때와 표본 크기가 1000개 이상인 경우에 계층 수를 3개로 과대추정하는 비율이 약 9~11% 수준으로 나타났다. 반면, 그 외의 조건에서는 계층 수를 1개로 과소추정하는 비율이 6~18% 수준으로 관찰되었다. 비균일적 DIF 조건에서는 DIF 크기가 작고 표본 크기가 500명인 경우에 계층 수를 과소추정하는 경향이 있었으나, 그 외 조건에서는 계층 수를 과대추정하는 경향이 나타났으며, 이 비율은 4~12% 수준이었다.

균일적 DIF의 검정력

DIF 탐지의 정확도를 평가하기 위해, LDIF 모형과 ODIF 모형을 LDIF&ODIF 모형과 비교

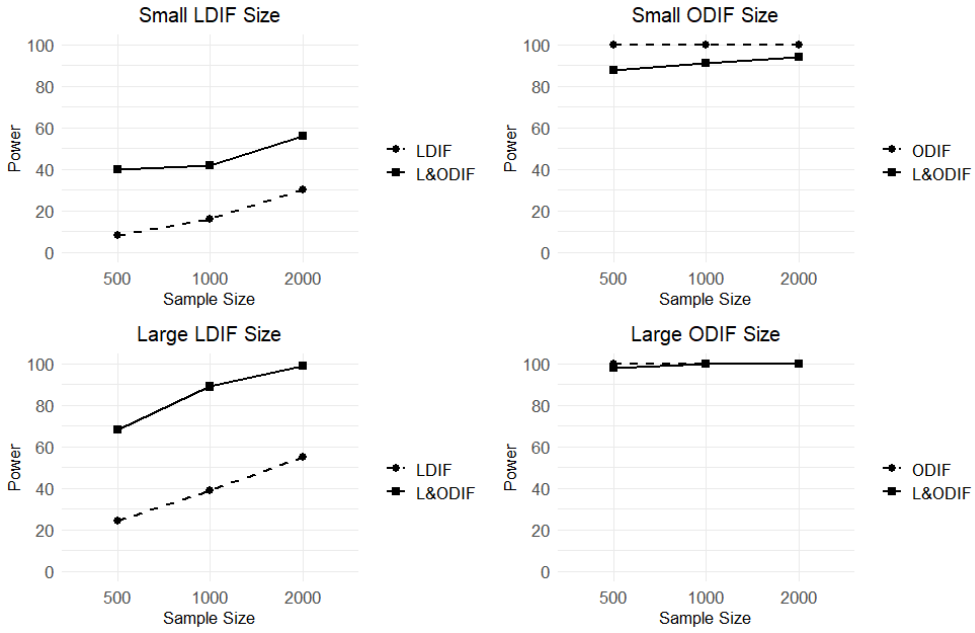

그림 6. 균일적 DIF의 검정력 결과

하였으며, 검정력이 80% 이상일 경우 DIF를 허용 가능한 수준으로 탐지한 것으로 간주하였다.

균일적 DIF 조건의 결과는 그림 6과 같다. 먼저, LDIF 모형은 LDIF를 거의 탐지해내지 못하였다. DIF의 크기와 표본의 크기가 증가할수록 검정력이 증가하는 경향을 보였지만, 전반적으로 최소 기준을 충족하지 못하는 수준의 검정력을 보였다. 반면, LDIF&ODIF 모형은 LDIF 모형보다 더 나은 검정력을 보였다. DIF의 크기가 큰 경우, 표본의 크기가 1,000 이상일 때 양호한 검정력을 나타냈다. ODIF 조건에서는, LDIF&ODIF 모형이 모든 조건에서 우수한 검정력을 보였으며 DIF의 크기가 커질수록 검정력이 크게 향상되었다. ODIF 모형은 모든 조건에서 완벽한 검정력을 보였다.

비균일적 DIF의 검정력

비균일적 DIF에 대한 결과는 그림 7과 같다. LDIF&ODIF 모형은 DIF의 크기가 크고 표본 크기가 클 때 높은 검정력을 보였다. 특히 DIF 크기의 영향이 강하게 나타나 표본의 크기가 2,000이더라도 DIF 크기가 작으면, DIF 크기가 크고 표본 크기가 500인 조건보다 검정력이 더 낮게 나타났다. LDIF 모형은 LDIF&ODIF 모형과 비슷한 양상을 보였지만 검정력은 더욱 낮게 나타났다. ODIF 조건에서는 LDIF&ODIF 모형이 DIF 크기가 작고 표본크기가 1,000 이하인 조건을 제외하고 모든 조건에서 80% 이상의 검정력을 보였다. 반면, ODIF 모형은 모든 조건에서 90%를 넘는 검정력을 보여주었으며 DIF가 큰 조건에서는 완벽한 검정력을 보였다.

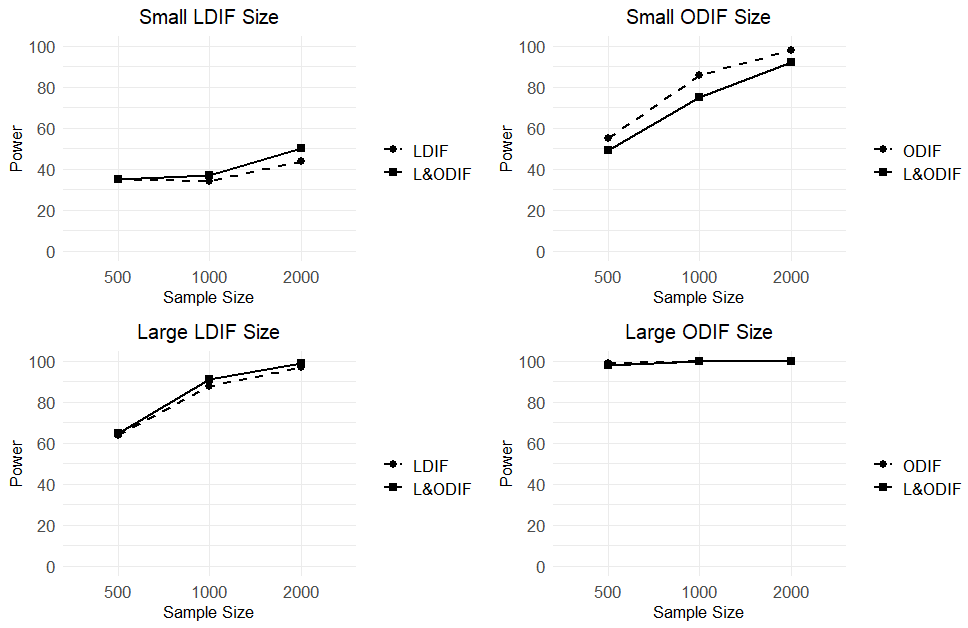

그림 7. 비균일적 DIF의 검정력 결과

균일적 DIF의 1종 오류율

1종 오류율은 DIF가 없는 문항에 대해 DIF가 있다고 판단하는 비율을 의미하며 그 비율이 5%보다 크면 허용 기준을 넘는 1종 오류율을 가진다고 간주하였다. 1종 오류율의 결과는 그림 8과 같다.

균일적 LDIF의 경우, LDIF&ODIF 모형의 1종 오류율은 DIF 크기가 크고 표본 크기가 1,000 이상일 때 허용 범위 내의 값이 관찰되었다. 특히 DIF 크기가 작고 표본 크기가 2,000인 조건에서보다 DIF 크기가 크고 표본 크기가 500인 조건에서 더 낮은 1종 오류율이 나타났다. 반면, LDIF 모형은 모든 조건에서 25% 내외의 심각한 1종 오류율을 보였는데, 이는 ODIF 문항(Y6)을 LDIF로 탐지하였기 때문이다. 이러한 결과는 공변인이 직접효과를 가지는 경우 LDIF로 오탐지될 가능성이 있음을 시사한다.

반면, 균일적 ODIF 조건은 LDIF와 다른 양상을 보였다. LDIF&ODIF 모형은 LDIF 조건에서의 결과보다 상대적으로 더 나은 수준의 1종 오류율을 보였으며, 특히 DIF 크기와 표본 크기가 모두 큰 조건에서 준수한 수준을 보였다. ODIF 모형의 1종 오류율은 대부분 5% 이상의 값을 보고하였으며 DIF의 크기와 표본 크기가 커질수록 1종 오류율이 증가하는 경향을 보였다. 이는, 모형화되지 않은 LDIF의 영향이 DIF 크기와 표본 크기가 커질수록 더 강하게 나타나는 결과라 할 수 있다.

비균일적 DIF의 1종 오류율

비균일적 DIF의 1종 오류율 결과는 그림 9과 같다. 비균일적 LDIF를 확인하기 위해 LDIF&ODIF 모형을 적용한 경우, DIF 크기가

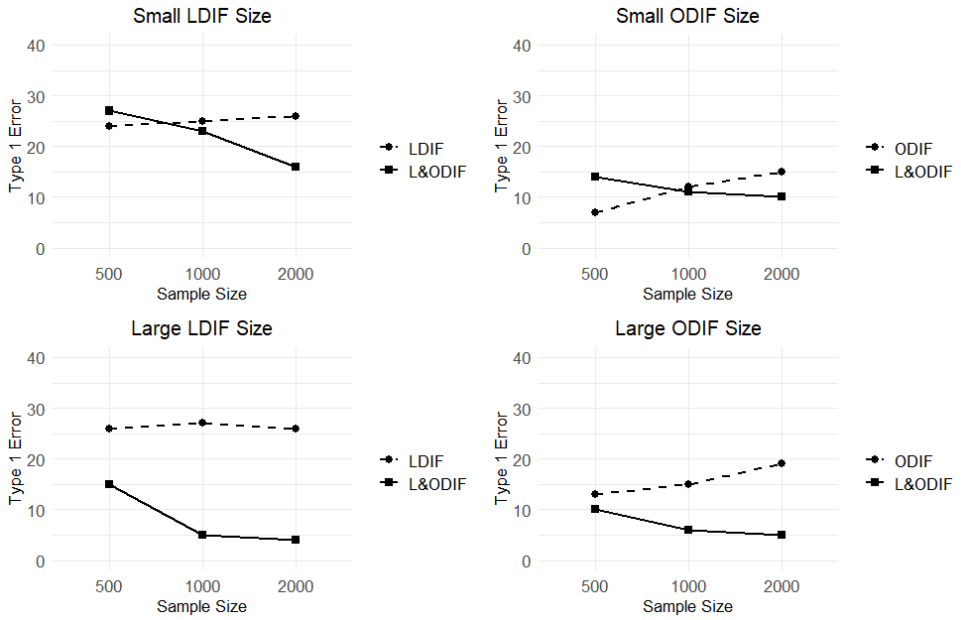

그림 8. 균일적 DIF의 1종 오류율 결과

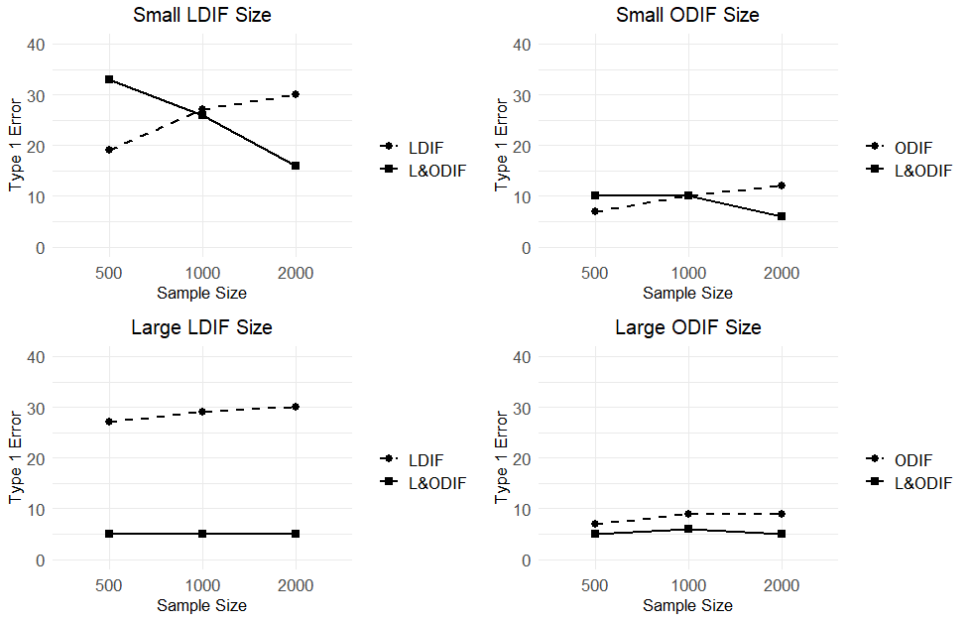

그림 9. 비균일적 DIF의 1종 오류율 결과

작은 조건에서 표본 크기가 커질수록 1종 오류율은 줄어드는 모습을 보였지만, 표본 크기가 2,000인 조건에서도 상당히 높은 값을 보고하였다. 반면, DIF 크기가 큰 조건에서는 일부의 결과가 5%가 넘는 1종 오류율을 보고하였지만 10%를 넘는 심각한 1종 오류율은 없었다. LDIF 모형을 적용한 결과, 최대 31%의 1종 오류율이 관찰되었는데, 이는 앞서 언급했던 바와 같이 모든 ODIF 문항이 LDIF로 탐지된 데 기인한다.

LDIF&ODIF 모형을 적용하여 ODIF를 탐지한 경우, 1종 오류율이 최대 10%로 허용 기준을 초과하는 경우도 있었지만 그 수치가 크지 않아 비교적 양호한 수준으로 나타났다. 특히, DIF의 크기가 큰 조건에서 낮은 1종 오류율을 보였으며 ODIF 모형의 경우 4~20%의 1종 오류율을 보여 LDIF&ODIF 모형에 비해 높은 1종 오류율이 관찰되었다.

균일적 DIF의 상대적 편향

균일적 DIF 탐지 시 모수 추정의 상대적 편향 결과는 그림 10과 같다. LDIF&ODIF 모형을 활용하여 LDIF를 탐지할 경우, 전반적인 모수 추정치는 부적으로 편향되었으나 LDIF 모형에 비해 편향이 줄어드는 경향을 보였다. 일부 조건에서는 상대적으로 낮은 편향이 관찰되었으나, 대부분의 조건에서 정확한 모수 추정이 이루어지지 못하였다. 이에 반해, ODIF를 탐지하기 위해 LDIF&ODIF 모형을 사용하는 경우에는 ODIF와 마찬가지로 모든 조건에서 허용 범위 내의 편향이 관찰되었다.

비균일적 DIF의 상대적 편향

비균일적 DIF 탐지 시 상대적 편향 결과는 그림 11과 같다. LDIF를 확인하기 위해

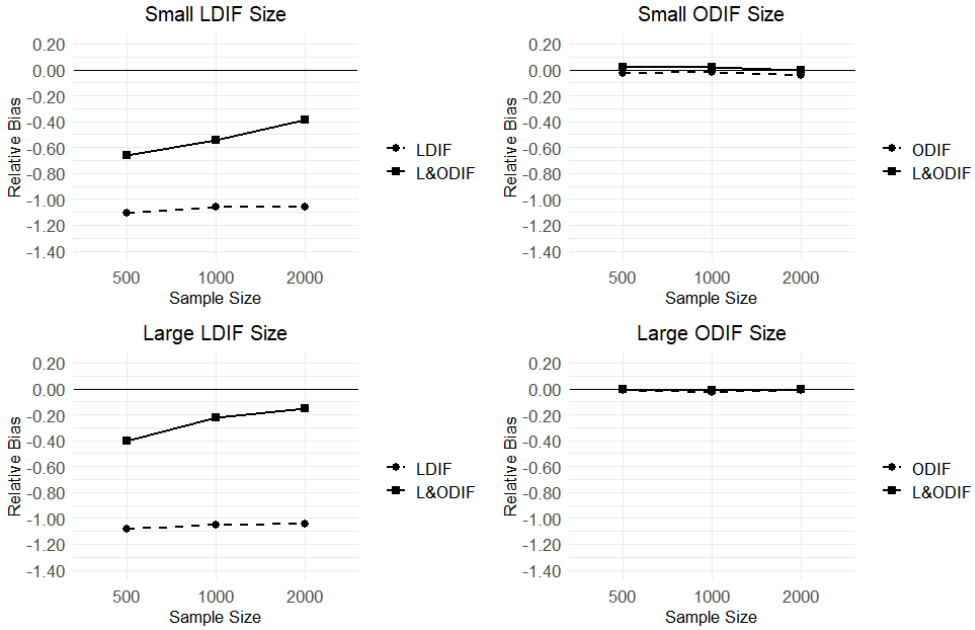

그림 10. 균일적 DIF의 편향 결과

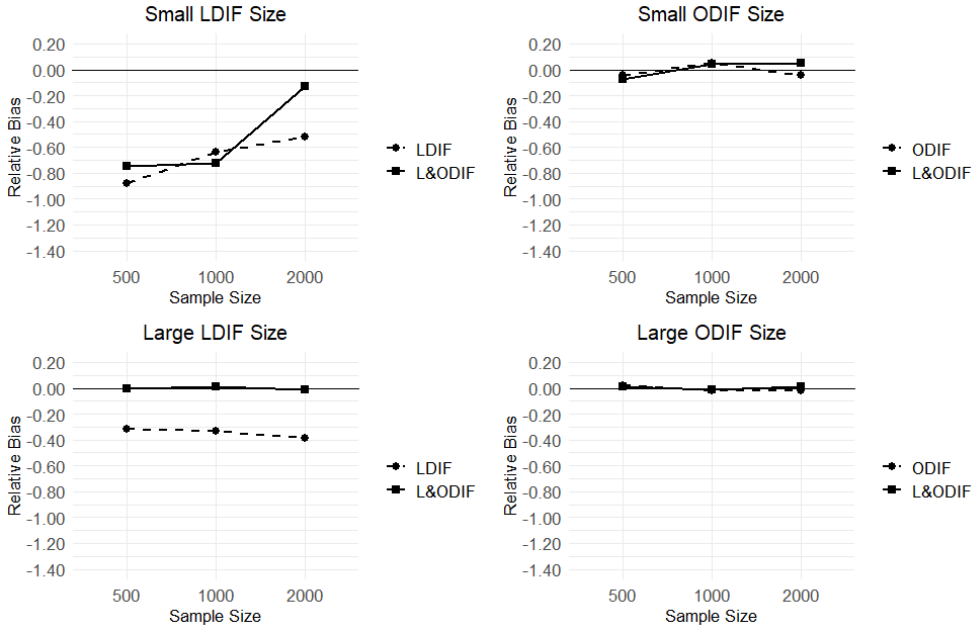

그림 11. 비균일적 DIF의 편향 결과

LDIF&ODIF를 적용한 경우, DIF 크기가 큰 조건에서는 편향이 거의 나타나지 않았다. 그러나 DIF의 크기가 작은 조건에서는 편향이 허용 기준을 부적으로 넘는 값이 관찰되었으며, 표본의 크기가 작아질수록 편향의 크기는 커졌다. 특히, 표본의 크기가 2,000에서 1,000으로 줄어들 때 편향의 증가폭이 상당히 크게 나타났다. 비균일적 ODIF를 탐지하였을 때, LDIF&ODIF 모형의 편향은 일부 특정 조건을 제외하고는 거의 발견되지 않았으며, 최대 편향은 11%로 허용 기준인 5%를 크게 초과하지 않는 수준이었다. DIF의 크기가 작을수록 편향이 더 많이 발생하는 경향을 보였으나, DIF 크기가 큰 조건에서는 편향이 거의 나타나지 않았다. 한편, ODIF 모형의 편향은 LDIF&ODIF 모형과 비교하여 유의미한 차이를 보이지 않았다.

논 의

본 연구는 FMM-MIMIC 모형을 활용하여 DIF의 원인을 잘못 설정하였을 경우, DIF 탐지의 수행력을 측정하고자 하였다. 연구에 사용된 데이터는 잠재계층의 비율, 잠재계층과 요인에 공변인이 미치는 영향, 균일적 DIF와 비균일적 DIF의 크기, 표본의 크기를 고려하여 생성되었다. 각 조건 당, 1개의 정확 추정 모형과 2개의 오추정 모형으로 분석을 실시하였으며 연구 결과를 정리하면 다음과 같다.

첫째, 정보 지수 중 AIC를 사용하는 것이 잠재계층의 개수를 선정하는 데 있어서 가장 정확하고 일관된 결과를 보였으며, 이러한 결과는 Lee와 Beretvas(2014)의 연구와 유사하였다. LDIF 모형의 경우, 정보 지수들은 모든 조건에서 거의 90% 이상의 비율로 2개 계층을 선택하였다. LDIF&ODIF 모형에서는 AIC가 잠재계층의 수를 올바르게 선택하였으며 ODIF 모형의 계층 수 선택은 LDIF&ODIF 모형의 결과보다 좋지 못하였다. LDIF&ODIF 모형과 ODIF 모형에서 BIC와 a-BIC의 수행능력은 상당히 저조하였는데, 특히 BIC는 DIF의 크기와 표본 크기에 영향을 받았으며 a-BIC는 DIF의 크기에 민감한 것으로 나타났다. 이러한 결과는 잠재계층에 대한 공변인의 효과를 설정하지 않은 경우, BIC와 a-BIC가 측정단위동일성과 절편동일성 조건에서 계층의 수를 정확하게 추정하지 못했던 Wang과 동료들(2021)의 결과와 유사하다. 또한, 잠재계층의 개수를 올바르게 결정하기 위해서는 잠재계층을 잘 설명하는 공변인을 모형에 포함하는 과정이 필요함을 시사한다. 잠재계층의 수를 정확하게 결정하기 위해 Tay와 동료들(2011)이 제시한 순서에 따라 LDIF 모형과 AIC를 사용하는 것이 권고된다.

둘째, 검정력 측면에서 LDIF와 ODIF는 상이한 결과를 보였다. 먼저, LDIF를 탐지할 경우, LDIF&ODIF 모형의 검정력과 LDIF 모형의 검정력은 모두 DIF의 크기와 표본 크기에 따라 차이를 보였으며, LDIF&ODIF 모형이 더 나은 수행력을 보였다. LDIF 탐지 결과와는 반대로, ODIF 탐지에서는 LDIF&ODIF 모형과 ODIF 모형 모두 우수한 검정력을 보였으며, ODIF 모형이 더 나은 검정력을 보였다. 이러한 경향은 균일적 DIF와 비균일적 DIF 모두에서 일관되게 나타났다. LDIF 모형과 비교하였을 때 LDIF&ODIF의 모형이 LDIF를 더 잘 탐지하는 것은 Bilir(2009)의 결과와 유사했지만,

LDIF&ODIF 모형의 ODIF 탐지가 ODIF 모형보다 더 저조하게 나타나 선행연구와 다른 결과를 보였다. 이러한 결과가 나타난 이유는 데이터 생성 방법의 차이 때문이라고 볼 수 있다. Bilir(2009)의 연구는 다집단 데이터를 생성한 다음 MIMIC 모형으로 분석을 실시하였지만, 본 연구는 데이터 생성과 분석 모두 MIMIC 모형을 사용하였다. 데이터 생성에 있어서 이러한 차이점은 LDIF와 ODIF의 탐지 정확도 차이에도 영향을 준 것으로 보인다. LDIF는 두 잠재계층의 모수 간 차이를 통해 DIF의 유의성을 평가하는 반면, ODIF는 회귀계수를 통해 DIF의 유의성을 확인하는 직접적인 방식을 채택하기 때문이다.

셋째, 균일적 DIF 조건에서 DIF의 크기와 표본 크기 모두 클 경우 LDIF&ODIF 모형은 양호한 1종 오류율을 보인 반면에 LDIF 모형과 ODIF 모형의 1종 오류율은 모든 조건에서 허용 기준을 초과하는 결과를 보였다. 특히, LDIF 모형은 25% 내외의 1종 오류율을 보였는데, 이는 ODIF 문항을 LDIF로 잘못 탐지했기 때문이다. 비균일적 DIF의 1종 오류율은 균일적 DIF의 1종 오류율보다 더 양호한 결과를 보였다. LDIF&ODIF 모형을 사용하였을 때, 일부 조건에서 1종 오류율이 허용 기준을 넘었지만 극단적인 값은 아니었다. 반면, LDIF 모형을 사용하여 비균일적 DIF를 탐지할 때, 일부 조건에서 균일적 DIF와 마찬가지로 ODIF 문항이 LDIF 문항으로 오탐지되었다. ODIF 모형의 1종 오류율은 많은 조건에서 허용 기준을 넘는 값을 보고하였지만, 균일적 DIF 조건일 때보다는 낮은 수준이었다.

마지막으로 LDIF&ODIF 모형과 ODIF 모형을 사용하여 ODIF를 탐지할 때, 대부분의 조건에서 편향이 거의 관찰되지 않았으며 일부 조건에서도 기준치를 크게 넘지 않는 부적 편향을 보였다. 이에 반해, LDIF&ODIF 모형 또는 LDIF 모형을 사용하여 균일적 DIF를 탐지할 때는 심각한 부적 편향이 발생하였다. 비균일적 LDIF 조건에서 DIF의 크기가 작을 경우에는 LDIF&ODIF 모형이 심각한 부적 편향을 보였지만 DIF의 크기가 큰 경우 편향이 크게 줄어 들었다. LDIF 모형은 조건과 상관없이 심각한 부적 편향을 보였다.

Tay와 동료들(2011)은 잠재계층 간 모수의 동일화 제약을 하지 않은 FMM으로 계층의 수를 결정하는 방법을 제시하였다. 더 나아가, Wang 외 5인(2021)의 연구는 공변인의 효과를 추정하는 것이 계층의 수를 정확히 추정하는데 도움이 된다고 언급했으며 Wang과 동료들(2023)의 연구에서 FMM을 활용하여 잠재계층의 수를 추정하는 경우, 공변인을 모형에 포함하지 않고 계층의 수를 추정한 후 공변인을 추가하는 3단계 접근법보다 공변인을 포함하여 계층의 수를 추정하는 1단계 접근이 더 유리하다는 결과를 제시하였다. 이에 더해, 본 연구의 결과에 따르면 FMM을 활용하여 ODIF와 LDIF를 탐지하고자 하는 경우, 정보 지수 중 AIC를 바탕으로 계층의 수를 결정하는 것이 가장 적절하며 이는 Lee와 Beretvas(2014)의 결과와 일치한다. FMM-MIMIC 모형을 사용하여 균일적 DIF를 탐지하고자 할 때, 최소 1,000개 이상의 표본을 확보할 필요가 있으며, 그 미만의 표본을 사용한다면 검정력, 1종 오류율, 편향에 심각한 영향을 미칠 수 있다. 비균일적 DIF의 경우에는 최소 2,000개의 표본이 권고되며 1,000개의 표본으로 비균일적 DIF를 탐지하고자 할 경우 모수 추정 능력에 대해서는 균일적 DIF와 큰

차이를 보이지 않으나 모형이 수렴되지 않을 수 있다.

Tay 등(2019)의 선행연구와 본 연구 결과를 토대로, FMM-MIMIC 모형을 활용한 DIF 탐지 절차를 다음과 같이 제안한다. 첫째, 공변인을 포함하고 문항의 요인 부하량과 절편을 잠재계층별로 자유롭게 추정하는 초기 모형을 적합한 후, BIC 등 정보기준을 활용하여 잠재계층과 요인의 개수를 결정한다. 둘째, 공변인이 요인 및 잠재계층에 미치는 영향을 평가하기 위해 공변인과의 경로를 순차적으로 모형에 추가하며, 이 과정에서 통계적으로 유의하지 않은 경로는 제거하여 모형을 단순화한다. 셋째, 각 문항의 잠재계층별 요인 부하량을 비교하여 비균일적 LDIF 여부를 검증한다. 요인 부하량의 차이가 없거나 통계적으로 유의하지 않은 문항에 대해서는 요인 부하량을 동일하게 제약한 상태에서 계층 간 절편의 차이를 검토하여 균일적 LDIF 여부를 검증한다. 넷째, LDIF로 탐지되지 않은 문항을 대상으로 공변인에 대한 상호작용 효과와 회귀계수의 유의성을 검증하여 균일적 ODIF와 비균일적 ODIF 여부를 평가한다. 이때 평가의 대상이 되는 문항은 LDIF 단계에서 동일화 제약이 적용된 상태임을 유의한다. 통계적으로 유의하지 않은 효과는 동일화 제약을 적용하여 모형을 간명하게 하며, 모든 DIF가 제거될 때까지 이 과정을 반복한다.

본 연구의 한계점은 다음과 같다. 첫째, 자료에 적합한 모형을 선정하는 방법 중 우도비 검정에 대해서는 확인하지 않았다. 둘째, 관찰 가능한 두 집단의 비율이 50:50으로 동일한 조건에 대해서만 확인하여 집단 간 비율이 동일하지 않은 조건에 대해서는 확인하지 않았다. 셋째, LDIF와 ODIF 모두를 가진 모집단에 대해서만 확인하였기 때문에 LDIF 혹은 ODIF 만 있거나 DIF가 없는 모집단에 대해서는 확인하지 않았다. 넷째, 잠재계층을 차별기능문항의 원인으로 보는 경우, 잠재계층은 응답자의 응답반응을 바탕으로 구분되기 때문에 응답자 표본에 따라 다른 결과를 가질 수 있다. 마지막으로, FMM-MIMIC 모형이 탐색적 방법이긴 하나 연구자가 사전에 설정한 DIF의 원인에 대해서만 그 영향을 확인할 수 있으며, 모든 DIF의 원인을 파악할 수 없다는 한계점이 존재한다.

참고문헌

박상은, & 노현종. (2019). 고등학생용 경제이해력 평가도구의 문항 양호도 및 성별에 따른 차별기능문항 분석. 경제교육연구, 26(2), 97-129.

안선영. (2022). Rasch 모형을 적용한 문항분석 및 차별기능문항 탐색 - 2021 학년도 물리인증제 1-2급을 바탕으로. 문화와융합, 44(10), 147-159.

이대용, 김석우, & 길임주. (2014). 성별에 따른 디자인 진로적성 검사의 차별기능문항군 탐색. 교육평가연구, 27(4), 945-964.

윤수철, 이순묵. (2013). 구조방정식모형을 이용한 차별문항기능의 탐지: MACS와 MIMIC의 비교. 한국심리학회지: 일반, 32(4), 1023-1052.

진수정, 성태제. (2004). MH 방법과 SIBTEST 방법을 이용한 문항 유형에 따른 차별기능문항의 탐색. 교육평가연구, 17(2), 215-236.

Bilir, M. K. (2009). *Mixture item response*

theory-MIMIC Model: Simultaneous estimation of differential item functioning for manifest groups and latent classes. The Florida State University.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42(2), 133-148. https://doi.org/10.1111/j.1745-3984.2005.00007

De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, 2(3-4), 243-276. https://doi.org/10.1080/15305058.2002.9669495

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. https://doi.org/10.1177/0013164411404221

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel Haenszel procedure. *ETS Research Report Series*, 1986(2), i-24. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, 26(3), 329-367. https://doi.org/10.1177/0049124198026003003

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing Measurement Invariance Using MIMIC: Likelihood Ratio Test With a Critical Value Adjustment. *Educational and Psychological Measurement*, 72(3), 469-492. https://doi.org/10.1177/0013164411427395

Lee, H., & Beretvas, S. N. (2014). Evaluation of two types of differential item functioning in factor mixture models with binary outcomes. *Educational and Psychological Measurement*, 74(5), 831-858. https://doi.org/10.1177/0013164414526881

Lee, S., Han, S., & Choi, S. W. (2021). DIF detection with zero-inflation under the factor mixture modeling framework. *Educational and Psychological Measurement*, 82(4), 678-704. https://doi.org/10.1177/00131644211028995

Lee Webb, M. Y., Cohen, A. S., & Schwanenflugel, P. J. (2008). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test－III. *Educational and Psychological Measurement*, 68(2), 335-351. https://doi.org/10.1177/0013164407308474

Lord, F. M. (1980). *Applications of Item Response Thoery to Practical Testing Problems*. Lawrence Erlbaum Associates. Inc, Hillsdale, NJ.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods*, 10(1), 21. https://doi.org/10.1037/1082-989X.10.1.21

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research*, 45(6), 975-999. https://doi.org/10.1080/00273171.2010.533047

McCarthy, D. M., Pedersen, S. L., & D'Amico, E. J. (2009). Analysis of item response and

differential item functioning of alcohol expectancies in middle school youths. *Psychological assessment*, 21(3), 444. https://doi.org/10.1037/a0016319

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, 115(2), 300. https://doi.org/10.1037/0033-2909.115.2.300

Muthén, B., & Muthén, L. (2019). Mplus: A general latent variable modeling program. *Muthén & Muthén*.

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

Samuelsen, K. M. (2008). Examining differential item functioning from a latent mixture perspective. *Advances in latent variable mixture models*, 177-197.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58(2), 159-194. https://doi.org/10.1007/BF02294572

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of applied psychology*, 91(6), 1292. https://doi.org/10.1037/0021-9010.91.6.1292

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), 147-176. https://doi.org/10.1177/1094428110366037

Yalcın, S. (2018). Determining differential item functioning with the mixture item response theory. *Eurasian Journal of Educational Research*, 18(74), 187-206.

Wang, Y., Cao, C., & Kim, E. (2023). Covariate inclusion in factor mixture modeling: Evaluating one-step and three-step approaches under model misspecification and overfitting. *Behavior Research Methods*, 55(6), 3281-3296. https://doi.org/10.3758/s13428-022-01964-8

Wang, Y., Hsu, H. Y., & Kim, E. (2021). Investigating the impact of covariate inclusion on sample size requirements of factor mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(5), 716-724. https://doi.org/10.1080/10705511.2021.1910036

Wang, Y., Kim, E., Ferron, J. M., Dedrick, R. F., Tan, T. X., & Stark, S. (2021). Testing measurement invariance across unobserved groups: The role of covariates in factor mixture modeling. *Educational and Psychological Measurement*, 81(1), 61-89. https://doi.org/10.1177/0013164420925122

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied psychological measurement*, 35(5), 339-361. https://doi.org/10.1177/0146621611405984

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the schedule for nonadaptive and

adaptive personality. *Journal of psychopathology and behavioral assessment*, 31, 320-330. https://doi.org/10.1007/s10862-008-9118-9

1차원고접수 : 2025. 04. 04
2차원고접수 : 2025. 07. 29
최종게재결정 : 2025. 09. 11

한국심리학회지: 일반
Korean Journal of Psychology: General
2025, Vol. 44, No. 4, 321-343
http://dx.doi.org/10.22257/kjp.2025.12.44.4.321

The Impact of Misspecifying DIF Sources on Detection Accuracy in Differential Item Functioning: A Simulation Study Based on a Factor Mixture Model

Eun Jin Choi EZN Wellness Co., Ltd.
Chan Hee Lee Kyungpook National University
Jung Kyu Park† Kyungpook National University

This study investigated the impact of misspecifying the source of differential item functioning on detection accuracy using a factor mixture model that incorporates covariates, when both observed DIF (ODIF), caused by known group membership, and latent DIF (LDIF), caused by latent class membership, coexist within a single test. DIF type, magnitude, and sample size were systematically varied to evaluate model performance in terms of class enumeration accuracy, detection power, Type I error rate, and parameter bias. The results showed that the model including both LDIF and ODIF (L&ODIF model) yielded the highest accuracy in estimating the number of latent classes, while the ODIF-only model showed very low estimation accuracy. The detection power for LDIF was highest in the L&ODIF model, whereas ODIF detection was most accurate in the ODIF-only model. The L&ODIF model demonstrated lower Type I error rates in most conditions, and parameter bias remained within or slightly above acceptable levels. These findings suggest that when both types of DIF are present, applying a factor mixture model capable of detecting LDIF and ODIF simultaneously can improve detection accuracy.

Key words : factor mixture model, differential item functioning, latent class analysis