한국심리학회지: 일반
Korean Journal of Psychology: General
2026, Vol. 45, No. 4, 687-716
http://dx.doi.org/10.22257/kjp.2026.9.45.4.687

OPEN ACCESS

잠재계층분석과 잠재프로파일분석에서 외부변수와 지표변수 간 직접효과 누락이 외부변수 효과 추정에 미치는 영향: 2단계 접근법과 3단계 접근법의 비교

이 찬 희 김 한 규 박 중 규†
경북대학교 경북대학교 경북대학교

본 연구는 잠재계층분석과 잠재프로파일분석에서 외부변수와 지표변수 간의 직접효과가 모형에서 누락된 경우 관찰될 수 있는 추정 편향을 시뮬레이션 연구를 통해 검증하였다. 외부변수와 혼합모형의 종류에 따라 먼저 연구 1에서는 공변인이 잠재계층분석 모형의 지표변수에 직접적인 영향을 미치는 상황을 가정하였다. 연구 2에서는 잠재프로파일분석의 지표변수가 결과변수에 직접적인 효과를 미치는 상황을 가정하였다. 두 연구 모두에서 직접효과를 모형에 포함한 경우와 포함하지 않은 경우에 대해 2단계 접근법과 3단계 접근법을 사용하여 모수를 추정하고 그 결과를 비교하였다. 그 결과, 2단계 접근법을 바탕으로 공변인의 직접효과를 모형에 포함한 경우, 그렇지 않은 경우에 비해 전반적인 편향이 낮았으며, 공변인의 직접효과 크기가 클수록 더 편향이 발생하였다. 또한, 누락된 직접효과의 개수가 증가할수록 편향이 증가하였다. 마찬가지로 지표변수가 결과변수에 미치는 직접효과를 포함한 모형은 포함하지 않은 모형에 비해 현저히 낮은 상대적 편향 평균과 1종 오류율을 보였다. 해당 결과들은 외부변수가 지표변수에 미치는 영향을 모형 내 직접효과로 포함하는 것이 모수 추정의 정확성을 향상시킬 수 있음을 시사한다.

주요어 : 잠재계층분석, 잠재프로파일분석, 외부변수, 직접효과, 2단계 접근법

* 이 논문은 2026년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2026S1A5C4021468).

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

Copyright © 2026, The Korean Psychological Association. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution and reproduction in any medium, provided the original work is properly cited.

잠재계층분석(Latent class analysis; LCA)과 잠재프로파일분석(Latent profile analysis; LPA)은 혼합 모형(mixture models)의 대표적인 방법이다(Collins & Lanza, 2009). 혼합 모형(mixture models)은 모집단이 관찰되지 않은 이질적인 하위집단들의 혼합으로 구성되어 있다고 가정하고, 각 하위집단을 식별하기 위한 목적으로 활용한다. LCA와 LPA는 이러한 혼합 모형에 속하는 분석법으로, 전자는 범주형 변수를, 후자는 연속형 변수를 지표변수로 활용한다. 연구자들은 이 기법들을 사용하여 유사한 응답 패턴을 보이는 개인들을 동일한 하위집단으로 묶고, 상이한 응답 패턴을 보이는 개인들은 별개의 잠재계층으로 구분해 왔다. 최근에는 이러한 모형들을 이용하여 단순히 개인을 잠재계층으로 분류하는 것을 넘어서, 외부변수와 도출된 잠재계층 간의 관계를 살펴보는 분석에도 관심이 증가하고 있다(Vermunt, 2010). 예를 들어, 잠재계층에 영향을 미치는 공변인과 잠재계층이 예측하는 결과변수를 분석에 포함시킴으로써 잠재계층의 선행요인과 후속결과를 함께 살펴볼 수 있다(Lanza et al., 2013).

이와 같은 방법론의 발전에 따라서 잠재계층과 외부변수와의 관계를 보다 효율적이고 정확하게 추정하기 위한 다양한 방법들이 개발되어 왔으며, 대표적인 방법으로는 1단계 접근법(one-step approach), 3단계 접근법(three-step approach), 2단계 접근법(two-step approach) 등이 있다. 최초로 제안된 1단계 접근법은 잠재계층을 식별하고 외부 변수와 잠재계층 간의 관계를 동시에 추정하는 방법으로, 문항 응답 확률, 잠재계층 확률, 그리고 공변인 및 결과변수와 잠재계층 간의 관계를 나타내는 모수가 하나의 모형 내에서 동시에 추정된다(Bolck et al., 2004). 이 방법은 모수 추정의 부정확성, 수행 방식의 논리적 문제 등의 단점이 있어, 이를 보완하기 위해 3단계 접근법이 제안되었다(Bolck et al., 2004; Vermunt, 2010). 3단계 접근법은 먼저 지표변수를 통해 문항 응답 확률과 잠재계층 확률을 추정한 뒤, 사후확률을 기반으로 개인을 각 잠재계층에 할당하고, 이때 계산된 사후확률을 고정하여 잠재계층과 외부변수 간의 관계를 추정하는 세 단계로 이루어진다(Asparouhov & Muthén, 2014). 이후 3단계 접근법의 두 번째 단계인 개인 할당 단계를 생략하여 모형 해석의 직관성과 단순성을 높인 2단계 접근법이 새롭게 제안되었다(Bakk & Kuha, 2018).

지표변수와 외부변수 간 조건부 독립

외부변수와 잠재계층 간의 관계를 분석하기 위해 앞서 살펴본 방법들은 공통적으로 지표변수와 외부변수 간에 관련이 없다는 조건부 독립을 기본 가정으로 한다(Bakk et al., 2013; Janssen et al., 2019; Lanza et al., 2013). 이러한 조건부 독립 가정은 측정불변성(measurement invariance) 또는 차별기능문항(differential item functioning)이 존재하지 않는다는 가정과 동일하며(Masyn, 2017), LCA나 LPA의 기본적인 지역독립성(local independence) 가정과 유사한 개념이다. 만약 외부변수와 지표변수 간의 조건부 독립 가정이 충족되지 않았음에도 불구하고 이를 간과한 채 후속 분석을 수행할 경우, 모수 추정에 편향이 발생할 뿐만 아니라 잠재계층의 개수 역시 부정확하게 결정될 수 있다(Bakk et al., 2022; Di Mari et al., 2020; Janssen et al., 2019; Masyn, 2017). 예를 들어, Janssen 등(2019)은 공변인이 지표변수의 문항 응답 확률에 직접적인 영향을 미치는 경우, 이러한

직접효과를 모형에 반영하지 않으면 공변인과 잠재계층 간 관계의 추정에 상당한 편향이 발생할 수 있음을 보고하였다. 또한 Di Mari 등(2020)은 결과변수가 지표변수의 응답 확률에 미치는 직접 효과를 모형에 반영하지 않을 경우, 잠재계층별 결과변수 평균의 추정에 편향이 발생할 수 있음을 보여주었다.

이러한 결과는 외부변수와 직접적으로 연관된 지표변수를 식별하고, 해당 직접효과를 모형에 적절히 반영하는 과정이 정확한 모수 추정을 위해 중요함을 시사한다. 그럼에도 불구하고, 연구자들이 널리 사용하는 3단계 접근법은 모형 내에서 외부변수와 지표변수 간의 직접효과를 설정할 수 없다는 한계가 있다(Janssen et al., 2019). 이는 3단계 접근법의 최종 단계에서 이전 단계에서 추정된 개인의 잠재계층 소속에 대한 사후확률을 바탕으로 외부변수와 잠재계층 간의 관계를 포함하는 구조모형을 추정하며, 이 과정에서 잠재계층을 구성하는 지표변수 자체는 최종 분석 모형에 직접 포함되지 않기 때문이다. 따라서 3단계 접근법에서는 외부변수가 특정 지표변수의 응답에 직접적인 영향을 미치는 경우에도 이러한 직접효과를 최종 모형에 명시적으로 포함하여 추정하기 어렵다. 이는 외부변수와 지표변수 간의 조건부 독립 가정이 위배되는 상황에서 특히 중요한 제한점이 될 수 있다. 반면, 2단계 접근법은 3단계 접근법과 달리 구조모형을 추정하는 단계에서도 지표변수를 사후확률로 대체하지 않고 원래의 측정모형에 포함된 형태로 유지한다. 이에 따라 외부변수와 개별 지표변수 간에 직접적인 관계가 존재하는지를 탐색할 수 있으며, 조건부 독립성 위반이 확인되는 경우 해당 직접효과를 모형에 추가하여 이를 명시적으로 반영할 수 있다는 장점이 있다(Di Mari et al., 2020; Janssen et al., 2019).

또한 기존의 관련 연구들은 주로 외부변수가 지표변수의 응답 확률에 미치는 직접효과에 초점을 맞추어 왔다. 특히 공변인은 개인의 잠재계층 소속 확률을 예측하는 변수로서, 개념적으로 잠재계층보다 선행하는 위치에 있다(Lanza et al., 2013). 따라서 공변인이 잠재계층을 구성하는 개별 지표변수의 응답 확률에 직접적인 영향을 미치는지를 탐색하는 것은 변수 간 이론적 방향성 측면에서도 자연스럽다. 반면, 결과변수는 일반적으로 잠재계층에 의해 예측되는 변수로 설정되므로, 잠재계층이 결과변수에 선행하는 구조를 형성한다(Lanza et al., 2013). 이러한 관계를 고려할 때, 결과변수가 지표변수의 응답 확률에 영향을 미치는 방향보다는 잠재계층을 구성하는 특정 지표변수가 결과변수에 직접적인 영향을 미치는 구조를 고려하는 것이 시간적·논리적 방향성 측면에서 보다 타당할 수 있다. 그럼에도 불구하고, 지표변수가 결과변수에 미치는 직접효과를 탐색하거나 이를 분석 모형에 반영한 연구가 상대적으로 제한적이었다. 더 나아가 외부변수와 개별 지표변수 간의 관계를 체계적으로 탐색하고, 확인된 직접효과를 모형에 반영함으로써 조건부 독립성 위반에 대응하려는 연구 역시 아직 충분히 이루어지지 않은 실정이다. 추가로, 외부변수와 지표변수 간의 관계를 검토한 선행 연구들은 모두 LCA를 기반으로 수행되어 왔다. 그러나 심리학 및 사회과학 분야에서는 대부분 5점 또는 7점 척도의 리커트형 문항이 사용되고 있으며(이상우, 2016; Joshi et al., 2015), 이러한 자료의 특성으로 인해 LPA 역시 실제 연구에서 매우 빈번하게 활용되고 있다. 따라서 외부변수와 지

표변수 간의 관계에 대한 검토를 범주형 지표변수를 사용하는 LCA에 한정하기보다는, 연속형 지표변수를 사용하는 LPA 상황으로도 확장하여 살펴볼 필요가 있다. 또한 성별이나 인종과 같은 범주형 변수가 공변인으로 활용되는 경우가 많은 것과 달리, 결과변수의 경우에는 특정 검사 점수와 같은 연속형 변수가 사용되는 경우도 빈번하다. 이에 따라 범주형 지표변수와 공변인 간의 관계뿐만 아니라, 연속형 지표변수와 연속형 결과변수 간의 관계를 함께 탐색할 필요가 있다.

따라서 본 연구의 연구 1에서는 LCA를 기반으로 잠재계층의 소속을 예측하는 공변인이 지표변수의 응답 확률에 미치는 직접효과를 가정하였다. 반면, 연구 2에서는 LPA를 기반으로 특정 지표변수가 결과변수에 미치는 직접효과를 가정하였다. 이를 바탕으로 각 연구에서는 해당 직접효과를 모형에 포함한 경우와 포함하지 않은 경우를 설정하고, 2단계 접근법과 3단계 접근법을 적용하여 그 결과를 비교하고자 한다. 구체적으로 모수 추정치의 편향 정도와 함께 1종 오류율 및 검정력을 평가함으로써, 직접효과의 존재 여부와 분석 접근법에 따라 추정 결과가 어떻게 달라지는지를 살펴보고자 한다. 더 나아가 2단계 접근법을 활용하여 외부변수와 지표변수 간의 관계를 확인하고, 확인된 관계를 모형에 적절히 반영하기 위한 단계적 분석 절차를 제안하고자 한다.

외부변수의 효과 추정 방법

1단계 접근법

잠재계층을 예측하는 공변인의 효과를 추정하기 위해 처음 제안된 방법은 1단계 접근법이다. 이 방법은 관찰 가능한 집단 간에 문항반응 확률과 잠재계층 확률이 동일한지를 검증하기 위해 Clogg와 Goodman(1984)이 제안한 다집단 잠재계층분석(multi-group LCA)에 기초한다. 이후 이 접근법은 공변인을 모형에 직접 포함하여 공변인에 따른 잠재계층 소속 확률의 차이를 추정하는 방식으로 활용되어 왔다(Chung et al., 2006; Lanza et al., 2007). 이 방법은 LCA에 공변인을 통합한 최초의 시도로 평가되며, 이후 Dayton과 Macready(1988)가 로지스틱 연결 함수를 활용하여 잠재계층 확률에 대한 공변인의 영향을 모형화하는 일반적인 방식으로 확장하였다. 공변인의 영향을 반영한 모형의 수식적 표현은 (식 1)과 (식 2)에 나타나 있으며(Agresti, 2013; Vermunt, 2010), <그림 1>은 1단계 모형의 구조를 도식화한 것이다.

$$P(Y_i = y \mid Z_{pi}) = \sum_{k=1}^{K} P(C_i = k \mid Z_{pi}) \prod_{j=1}^{J} P(Y_{ij} = y_j \mid C_i = k) \quad \text{(식 1)}$$

$$P(C_i = k \mid Z_{pi}) = \frac{\exp(\alpha_k + \sum_{q=1}^{Q} \beta_{qk} Z_{piq})}{1 + \sum_{h=1}^{K-1} \exp(\alpha_h + \sum_{q=1}^{Q} \beta_{qh} Z_{piq})} \quad \text{(식 2)}$$

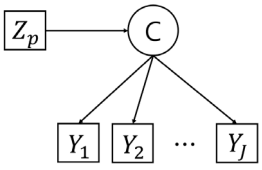


그림 1. 1단계 모형의 도식화

먼저, 범주형 문항 $j(j=1,\ldots,J)$에 대한 개인 i의 반응은 Y_{ij}, 문항 j에 대한 특정 응답 패턴은 y_j로 표기한다. 또한, 범주형 잠재변수 C_i는 개인 i가 할당된 잠재계층 k $(k=1,\ldots,K)$로 정의할 수 있다. 또한 Z_{piq}는 Q개의 공변인 중 하나이고, Z_{pi}는 개인 i의 전체 공변인 벡터이다.

(식 1)의 문항 응답 확률을 나타내는 $P(Y_{ij}=y_j|C_i=k)$는 잠재계층 k에 속한 개인 i가 특정 문항 j에 대해 특정 응답 y_j을 보일 확률을 의미하며, 지역독립성 가정을 바탕으로 $P(Y_i=y|C_i=k)$는 특정 잠재계층 k에 속한 개인의 개별 문항 응답 확률의 곱으로 나타낼 수 있다.

LCA의 구조모형은 잠재계층의 분포를 결정하는 잠재계층 확률(latent class proportion)을 모수로 가진다. LCA에서 잠재계층 확률인 $P(C_i=k)$는 전체 인원 중 특정 잠재계층에 속한 사람들의 비율을 나타내며, 일반적으로 잠재계층의 상대적인 크기를 의미한다. 반면, (식 2)에는 잠재계층 C_i에 영향을 주는 공변인 Z_p가 포함되어 있으며, 공변인이 잠재계층 확률보다 선행하기에 $P(C_i=k|Z_{pi})$로 정의한다. 1단계 접근법은 기본적으로 Z_{pi}가 모형에 포함된 상태에서 잠재계층의 수를 결정하고 이를 기반으로, (식 2)에 제시된 $P(Y_{ij}=y_j|C_i=k)$와 $P(C_i=k|Z_{pi})$을 최대우도법(Maximum Likelihood; ML)을 활용하여 동시에 추정한다.

초기에 제안된 1단계 접근법에는 몇 가지 한계가 존재한다. 먼저, 탐색적 연구에서 고려해야 할 공변인의 수가 많은 경우 비효율적일 수 있다. 연구자는 공변인을 포함한 모형과 포함하지 않은 모형 중 어느 모형을 기준으로 잠재계층 수를 결정할 것인지에 대한 어려움을 겪을 수 있다. 또한 구조모형의 명세(specification)가 측정모형의 모수 추정에 영향을 미치기 때문에(Vermunt, 2010) 공변인이 추가되거나 제거될 때마다 측정모형을 다시 추정해야 하며, 그 과정에서 잠재계층의 구성이나 질적 의미가 달라질 가능성이 있다. 아울러 공변인의 수가 증가할수록 분할표의 칸 수가 기하급수적으로 증가하여 희소성 문제가 발생할 수 있으며, 모형이 복잡해짐에 따라 추가적인 모형 설정 및 추정상의 어려움도 나타날 수 있다. 마지막으로, 1단계 접근법과 같이 구조모형과 측정모형을 동시에 추정하는 방식은 일반적인 응용연구의 분석 절차와도 다소 부합하지 않을 수 있다. 실제로 많은 연구자는 먼저 측정모형을 통해 잠재계층의 수와 특성을 확정한 후, 공변인을 추가하여 잠재계층과의 관계를 분석하는 방식을 보다 자연스럽게 받아들이는 경향이 있다. 이러한 1단계 접근법의 한계를 보완하기 위해 다양한 단계별 접근법(stepwise approach)이 제안되었으며, 현재는 그중 3단계 접근법이 가장 널리 활용되고 있다.

3단계 접근법

3단계 접근법은 외부변수의 효과를 분석하는 연구에서 널리 사용되고 있으며(Vermunt & Magidson, 2021), 이 방법은 1단계 접근법과 달리 측정모형과 구조모형을 분리하여 추정한다. 3단계 접근법의 첫 번째 단계에서는 (식 3)을 이용해 외부변수를 포함하지 않은 LCA 모수를 추정한다(Vermunt, 2010).

$$\log L_{STEP1} = \sum_{i=1}^{N} \log P(Y_i = y) = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} P(C_i = k) \prod_{j=1}^{J} P(Y_{ij} = y_j \mid C_i = k) \right]$$

(식 3)

두 번째 단계에서는 개인들의 전체 응답 패턴 벡터(Y_i)와 첫 번째 단계에서 추정된 모수를 바탕으로 이들을 잠재계층에 할당하고, 이에 대한 분류 오류(classification error)를 산출한다. 이 과정에서 개인 i의 실제 할당된 잠재계층은 W_i로 표기한다. 계층 할당에서 가장 중요한 요소는 응답 패턴 벡터(Y_i)가 주어졌을 때 계층 k에 속할 사후확률인 $P(C_i = k | Y_i = y)$이다. 사후확률은 베이즈 정리를 이용해 계산할 수 있으며, 자세한 식은 (식 4)와 같다.

$$P(C_i = k \mid Y_i = y) = \frac{P(Y_i = y \mid C_i = k) P(C_i = k)}{P(Y_i = y)}$$

(식 4)

분류 오류를 계산하기 위해서는 개인 i가 사후에 분류된 계층인 W_i가 필요하며, 사후분류 규칙 중 가장 널리 쓰이는 두 가지는 최대확률 할당(modal assignment)과 확률 가중 할당(proportional assignment)이다. 최대확률 할당은 W_i를 사후확률이 가장 높은 계층 k값으로 추정하며, 사후확률이 가장 큰 값에 대해 가중치($P(W_i = k | Y_i = y)$ 또는 w_{it})를 1로, 나머지 w_{it}에 대해 0으로 부여하는 단일 분류 방식(hard partitioning)이다. 반면, 확률 가중 할당은 불확실성을 반영하는 분류 방식(soft partitioning)으로 가중치 w_{ik}의 값을 $P(C_i = k | Y_i = y)$값 자체로 설정한다. 분류 오류는 $P(W_i = s | C_i = k)$로 나타내며, 실제 개인의 잠재계층은 k이지만 s로 잘못 할당될 확률을 의미한다. 분류 오류를 산출하는 식은 (식 5)와 같다(Vermunt, 2010).

$$P(W_i = s \mid C_i = k) = \frac{\sum_{Y} P(Y_i = y) P(C_i = k \mid Y_i = y) P(W_i = s \mid Y_i = y)}{P(C_i = k)}$$

(식 5)

최대확률 할당을 적용한다면 $P(W_i = s | Y_i = y)$가 0 또는 1이 되며, 확률 가중 할당을 적용할 경우 $P(W_i = s | Y_i = y)$는 $P(C_i = s | Y_i = y)$가 된다. 따라서, 분류 오류의 전체 비율은 (식 6)을 통해 계산할 수 있다. 두 번째 단계에서는 분류 오류를 고려하여 개인을 잠재 계층에 할당한 후, 다음 단계에서는 공변인과의 효과를 추정하는 방법과 결과변수와의 관계를 추정하는 방법으로 나뉘게 된다.

$$\sum_{k=1}^{K} P(C_i = k) \sum_{s \neq k} P(W_i = s | C_i = k)$$

(식 6)

공변인의 효과 추정

마지막 세 번째 단계에서 공변인의 효과를 추정하는 과정은 앞서 설명한 1단계 접근법과 개념적으로 거의 동일하나, 초기 3단계 접근법에서는 분류 오류를 고려하지 않고 단순히

(식 2)에서 C_i 대신 W_i를 대입하여 사용하였다(Vermunt, 2010). 그러나, Bolck 등(2004)은 $P(W_i = s|Z_i)$와 $P(C_i = k|Z_i)$의 관계에서 단순히 W_i가 C_i를 대체할 수 없음을 밝혀냈다. 이후 이러한 문제를 보완하기 위해 잠재계층 분류의 불확실성, 즉 분류 오류를 세 번째 단계의 분석에 반영하는 수정된 3단계 접근법(bias-adjusted three-step approach)이 개발되었다. Bolck 등(2004)이 제안한 BCH(Bolck, Croon & Hagenaars) 방법은 분류 오류를 고려한 최초의 3단계 접근법이라는 측면에서 의의가 있다. 그러나 이 방법은 공변량이 범주형인 경우에만 적용할 수 있고, 표준오차를 정확하게 추정하지 못한다는 한계가 있다. 이후 수정된 BCH 방법이 제안되었으나 표준오차가 과소추정될 가능성이 여전히 존재하였다. 이를 보완하기 위해 Vermunt(2010)는 최대우도(maximum likelihood, ML) 방법을 제안하였다. 이에 따라 기존의 단순 3단계 접근법은 분류 오류를 보정하는 형태로 수정되었으며, 그 모형과 세 번째 단계의 최대우도식은 각각 (식 7)과 (식 8)에 제시하였다.

$$P(W_i = s \mid Z_{pi}) = \sum_{k=1}^{K} P(C_i = k \mid Z_{pi}) P(W_i = s \mid C_i = k) \quad \text{(식 7)}$$

$$\log L_{ML} = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} P(C_i = k \mid Z_{pi}) P(W_i = s \mid C_i = k) \right] \quad \text{(식 8)}$$

결과변수와의 관계 추정

3단계 접근법에서 잠재계층과 결과변수 간의 관계를 추정하는 방법으로는 앞서 언급한 ML 접근법과 BCH 접근법 두 가지가 있으며, 이 두 방법은 동일한 (식 9)을 사용하지만, 적용하는 방식에 차이가 있다(Bakk & Vermunt, 2016).

$$P(W_i = s, Z_{oi}) = \sum_{k=1}^{K} P(C_i = k) f(Z_{oi} \mid C_i = k) P(W_i = s \mid C_i = k) \quad \text{(식 9)}$$

(식 9)는 결과변수와 잠재계층 간의 관계를 추정하는 식이다. Z_o는 결과변수를 의미하며, 결과변수와의 관계를 나타내는 모수는 개인 i가 계층 k에 속할 때, 개인 i의 결과변수 값인 Z_{oi}의 확률분포 $f(Z_{oi}|C_i = k)$이다.

먼저 ML 접근법에서는 최대우도 추정을 통해 (식 9)를 직접 추정한다. 분류 오류인 $P(W_i = s|C_i = k)$는 추정된 값으로 고정되며, $f(Z_{oi}|C_i = k)$는 집단 간 비교를 위해 추정된다. 집단 간 비교를 위해서는 $f(Z_{oi}|C_i = k)$의 분포를 특정해야 하는데 이 과정에서 연속형 결과변수는 정규분포로, 명목형 혹은 순서형 결과변수는 다분형 분포로, 가산형 결과변수는 포아송 분포로 가정하는 것이 일반적이다. Z_{oi}를 연속형으로 가정하면 해당 확률분포의 형태를 결정하는 분산은 계층 간 동일 조건 또는 비동일 조건 중에서 선택 가능하다. 자유 모수인 평균과 분산에 대한 표준오차는 강건 추정기(robust estimator)를 사용하여 추정하며, 강건 추정기는 확률 가중 할당 방법을 적용하는 경우 필수적이다(Bakk et al., 2014). 그러나 연속형 결과변수의 분포가 정규성을 위배하는 경우, 초기 구조와 다

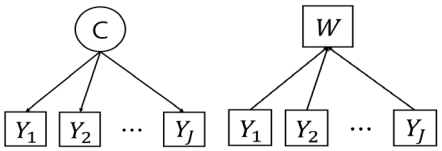


그림 2. 단계 1(왼쪽)과 단계 2(오른쪽) 모형

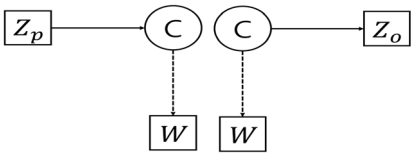


그림 3. 단계 3의 모형

른 잠재계층이 도출될 수 있으며, 이러한 문제를 방지하기 위해 Mplus와 Latent GOLD 프로그램에서는 첫 번째 단계에서 추정된 잠재계층 구조를 세 번째 단계에서 초기값(starting values)으로 고정하는 옵션을 제공하여 계층 구조가 변하지 않도록 설정할 수 있다(Asparouhov & Muthén, 2014). 이후 Wald 검정을 사용하여 각 계층 간 결과변수의 평균 차이가 유의한지 검정한다.

ML 방법이 (식 9)를 직접 추정하는 것과는 달리, BCH 접근법은 분산분석(ANOVA) 모형을 활용한다. 이 방법은 사후계층인 W_i에 대한 가중치로 분류 오류의 역수를 사용하여 실제 잠재계층의 분류를 위한 가중치로 활용한다(Bolck et al., 2004; Vermunt, 2010). 이렇게 구성된 모형은 ML 방법과는 달리 의사 최대우도(Pseudo Maximum Likelihood; PML) 추정 방식을 통해 계산된다.

ML 접근법과 비교할 때, BCH 접근법의 중요한 장점 중 하나는 결과변수의 분포가 잘못 설정되더라도 잠재계층의 정의가 변하지 않는다는 점이다. 또한, 강건 추정기는 대부분의 명세화 오류를 보정하기 때문에 오차 분포에 대한 명세화 오류 역시 수정할 수 있다. 예를 들어, 잠재 계층별 오차에 등분산성(homoscedasticity) 또는 이분산성(heteroscedasticity)을 가정하더라도 Wald 검정을 통해 추정되는 계층별 평균은 동일하다. 그러나 잠재계층 간 분리도가 낮고 표본 크기가 작은 상황에서 계층 간 오차 분산이 서로 다를 경우, 일부 계층의 오차 분산이 음수로 추정될 수 있다. 이러한 경우에는 이분산성을 가정한 BCH 접근법보다 등분산성을 가정한 BCH 접근법을 사용하는 것이 권고된다(Bakk & Vermunt, 2016). 3단계 접근법의 단계별 모형은 <그림 2>와 <그림 3>에 제시하였으며, <그림 2>의 왼쪽과 오른쪽 도식은 각각 1단계와 2단계를 나타낸다.

2단계 접근법

2단계 접근법은 3단계 접근법의 두 번째 단계에서 발생할 수 있는 분류 오류 문제를 보완하기 위해 제안되었다. Bakk과 Kuha(2018)는 단계별 접근법의 핵심이 측정모형과 구조모형을 구분하여 추정하는 데 있으며, 분류 단계를 별도로 설정할 필요는 없다고 보았다. 이에 따라 제안된 2단계 접근법은 분류 오류를 별도로 추정하지 않아도 되며, 수정된 3단계 접근법에 비해 분석 절차가 보다 직관적이라는 장점이 있다(Bakk & Kuha, 2018). 이 접근법은 Dayton과 Macready(1988)의 공변인과 잠재계층 간 관계에 대한 연구와 Bandeen-Roche 등(1997)이 제안한 방법을 기초로 하며, 이후 Xue와 Bandeen-Roche(2002)에 의해 발전되었다. Bakk과 Kuha(2018)는 이러한 선행 연구의 개념을 바탕으로 이를 하나의 단계별 접근법으로 체계화하였다.

2단계 접근법의 첫 번째 단계는 3단계 접근법과 마찬가지로 측정모형을 추정하는 것이며, 이에 대한 공식은 (식 3)을 따른다. 2단계 접근법에서 공변인의 효과를 추정할 때는 $P(C_i=k)$이 공변인에 의해 직접 영향을 받기 때문에, 첫 번째 단계에서 추정된 $P(C_i=k)$을 고정한 채로는 다음 단계로 진행할 수 없다. 따라서 해당 모수는 두 번째 단계로 넘어가기 전, 모형에서 탈락하게 된다. 반면, 결과변수는 공변인과 달리 잠재계층 확률에 영향을 미치는 변수가 아니므로, 첫 번째 단계에서 추정된 $P(C_i=k)$을 고정하여 다음 단계에서 사용한다. 즉, 첫 번째 단계에서 추정할 전체 모수를 θ_1, 두 번째 단계에서 추정할 전체 모수를 θ_2라고 하면, 공변인과 잠재계층 간의 관계를 추정할 경우 θ_1은 $P(Y_i=y|C_i=k)$만 포함한다. 반면 결과변수와 잠재계층 간의 관계를 추정할 경우, θ_1은 $P(Y_i=y|C_i=k)$와 $P(C_i=k)$가 된다. 이후, 두 번째 단계에서는 이전 단계에서 추정된 모수를 고정하여 θ_2를 추정하게 된다.

$$P(Y_i=y_i \mid Z_{pi}) = \sum_{k=1}^{K} P(C_i=k \mid Z_{pi}) \prod_{j=1}^{J} P(Y_{ij}=y_i \mid C_i=k)$$

(식 10)

(식 10)은 두 번째 단계에서 공변인과 잠재계층 간 관계를 추정하기 위한 식이며, $P(Y_i=y|C_i=k)$가 고정된다. 공변인이 존재할 때 잠재계층 k에 속할 확률인 $P(C_i=k|Z_{pi})$는 다중회귀 모형으로 표현할 수 있으며, 그 식은 앞서 제시한 (식 2)와 같다.

(식 11)은 결과변수와 잠재계층 간 관계를 추정하기 위한 식이다. (식 11)에서는 (식 10)과 달리 잠재계층 확률이 결과변수에 영향을 받지 않으며, 이에 따라 이전 단계에서 추정된 $P(Y_i=y|C_i=k)$와 $P(C_i=k)$가 고정된다.

$$P(Z_{oi}, Y_i=y) = \sum_{k=1}^{K} \left[P(C_i=k) P(Z_{oi} \mid C_i=k) \prod_{j=1}^{K} P(Y_{ij}=y_i \mid C_i=k) \right]$$

(식 11)

2단계에는 모수를 ML 방식으로 추정하며 BCH 접근의 의사-최대우도와 동일한 틀을 따른다. 2단계 추정치의 표준오차는 1단계와 2단계의 불확실성을 모두 포함한 전체 분산(표준오차)을 활용하는 것이 권고된다(Bakk & Kuha, 2018). <그림 4>, <그림 5>에는 2단계 접근법의 단계별 모형을 제시하였다.

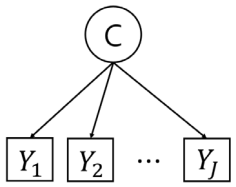


그림 4. 2단계 접근법의 단계 1 모형

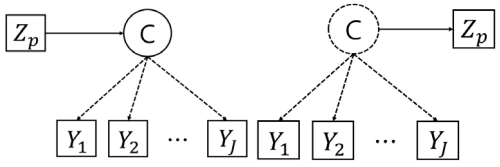


그림 5. 2단계 접근법의 단계 2 모형

주. 점선은 고정된 값을 나타냄.

지표변수와 외부변수 간 조건부 독립의 위배 탐지

혼합모형 분석에서는 잠재계층을 조건으로 지표변수와 외부변수가 서로 독립적이라고 가정한다. 그러나 이러한 조건부 독립성 가정이 위배될 경우, 모형에 포함된 모수 추정치에 편향이 발생할 수 있다(Masyn, 2017). 이러한 편향이 발생하는 원인은 다양하지만, 그중 하나는 지표변수와 외부변수 간의 관계가 모형에 반영되지 않은 채 잔존하기 때문이다(Masyn, 2017). 이와 같이 지표변수와 외부변수 간에 잠재계층을 통해 설명되지 않는 관계가 존재하는 경우, 두 변수 간에 직접효과(direct effect)가 존재한다고 볼 수 있다. 이러한 직접효과를 모형에 적절히 반영하지 않으면 문항응답 확률과 잠재계층 확률뿐만 아니라, 외부변수와 잠재계층 간의 관계를 나타내는 모수 추정치에도 편향이 발생할 수 있다. 따라서 지표변수와 외부변수 간에 잔존하는 관계의 존재 여부를 확인하고, 이를 모형에 적절히 반영하는 절차가 필요하다(Janssen et al., 2019; Oberski et al., 2013; Vermunt & Magidson, 2021). 다만 이러한 잔존 관계는 특정 방향의 직접효과뿐 아니라 다른 형태의 모형 명세 오류에서도 발생할 수 있으므로, 조건부 독립성 가정의 위배를 곧바로 특정 직접효과의 존재로 해석해서는 안 된다.

(식 12)에 제시된 BVR(Bivariate Residual; Vermunt & Magidson, 2005)은 이러한 잔존 관계를 확인하는 지표 중 하나로, 관찰변수 쌍에서 설명하지 못한 이변량 잔차의 크기를 나타낸다. 높은 BVR은 두 변수 사이에 설명되지 않은 관계가 남아 있거나 또는 다양한 형태의 모형 명세 오류가 존재할 가능성을 시사한다(Oberski et al., 2013). 따라서 BVR이 큰 변수쌍은 이론적 타당성을 바탕으로 직접효과의 설정을 포함한 모형 명세 오류의 가능성을 검토할 필요가 있다.

$$BVR_j = \frac{\sum_{z\in\{0,1\}} \frac{(O_{lz}-E_{lz})^2}{E_{lz}}}{\sum_{l\in\{0,1RIGHT}} \quad \text{(식 12)}$$

(식 12)의 O_{lz}는 외부변수 Z와 지표변수 y_j의 교차표(cross-table)에서 관찰된 값이며, E_{lz}는 교차표에서 기대되는 값을 의미한다. 선행연구에서는 범주형 변수 쌍에 대해 BVR 값 3 이상을 경험적인 기준(rule of thumb)으로 제시해 왔다.

두 번째 지표는 EPC(Expected Parameter Change; Oberski & Vermunt, 2018)이다. EPC는 문항반응이론(Item Response Theory)과 구조방정식(Structural Equation Modeling: SEM)에서 널리 활용하는 잔차 기반 통계량이다(Janssen et al., 2019). EPC 지수는 고정된 특정 모수를 자유롭게 추정하였을 경우 예상되는 모수 변화의 크기를 나타내며, 이를 통해 잠재적인 모형 명세 오류의 위치와 크기를 탐색하는 데 활용한다(Whittaker, 2012). Oberski 등(2013)은 이를 혼합모형에 적용하는 방법을 제안하였다. 특히 외부변수와 지표변수 간 조건부 독립성이 위배될 가능성이 있는 경우, EPC는 어떠한 변수쌍에서 직접효과를 추가적으로 고려할 필요가 있는지를 탐색하는 데 유용한 지표로 활용될 수 있다. 즉, EPC 값이 상대적으로 큰 경로를 확인함으로써 모형에 포함되지 않은 직접효과의 가능성을 탐색할 수 있다. 다만 BVR과 마찬가지로 EPC가 특정 직접효과의 존재를 확정

하는 것은 아니므로, 해당 관계의 이론적 타당성과 모형의 구조를 함께 고려한 후 직접효과의 포함 여부를 결정하는 것이 적절하다. 외부변수와 지표변수 간 관계에 대한 EPC의 계산은 (식 13)과 같다(Oberski et al., 2013).

$$EPC_j = \frac{s_j^2}{Var(s_j)} \quad \text{(식 13)}$$

s_j는 조건부 독립의 위배에 대한 점수를 의미하며, $\frac{\partial L(\theta)}{\partial \psi_j}$로 나타낼 수 있다. $L(\theta)$는 외부변수와 지표변수 간 직접효과를 설정한 모형에 대한 로그우도값을 나타내며, ψ_j는 직접효과에 대한 모수이다.

직접효과 모형화

외부변수와 지표변수 간 조건부 독립 가정이 위배되어 두 변수 사이에 모형으로 설명되지 않은 관련성이 존재하는 경우, 이를 모형에 반영하고 관련 모수의 추정 편향을 줄이기 위한 하나의 방법으로 직접효과를 설정할 수 있다.

직접효과를 모형화하기 위해서는 1단계 접근법 또는 2단계 접근법을 활용할 수 있다. 앞서 논의한 바와 같이 일반적인 3단계 접근법의 최종 단계에서는 원 지표변수가 모수 추정 과정에서 제외되므로, 이 단계에서 지표변수와 외부변수 간 직접효과를 설정하기 어렵다(Di Mari et al., 2020; Janssen et al., 2019). 따라서 3단계 접근법에서 직접효과를 설정하려면 특정 지표변수와 외부변수 간 관계를 사전에 알고 있어야 한다는 점에서 실제 연구에서 활용하기에는 어려움이 있다(Janssen et al., 2019).

1단계 접근법에서 직접효과를 설정하는 경우, 측정모형과 구조모형 모두를 재추정해야 하며, 직접효과를 추가하여 모형을 다시 추정하면 모형 추정치와 도출된 잠재계층의 정의가 바뀔 수 있다. 반면 2단계 접근법은 1단계 접근법과 달리 측정모형과 구조모형을 분리하여 추정하므로 직접효과가 설정된 모수만 재추정한다. 대부분의 측정모형 모수를 고정할 수 있다는 점은 더 견고한 모형을 유지하고 추정 시간을 단축할 수 있다는 장점으로 작용한다(Janssen et al., 2019).

직접효과를 모형화하는 것은 문항 응답 확률 $P(Y_i = y | C_i = k)$에 대한 조건부 독립 가정을 완화하는 의미를 가진다. 이는 특정 응답을 보일 확률이 잠재계층뿐만 아니라 공변인에게도 영향을 받는다고 가정하기 때문이다. 따라서, 공변인의 영향을 설정하여 (식 14)와 같이 표현할 수 있다(Janssen et al., 2019).

$$P(Y_i = y \mid Z_{pi}) = \sum_{k=1}^{K} P(C_i = k \mid Z_{pi}) P(Y_i = y \mid C_i = k, Z_{pi}) \quad \text{(식 14)}$$

2단계 접근법을 활용할 경우, 먼저 BVR, EPC 등의 지표를 이용하여 어떤 지표변수가 공변인과 연관이 있는지 탐색한 후, 해당 변수 간 직접효과를 설정하여 두 변수 간의 관계를 모형에 포함할 수 있다. 이 과정에서 직접효과가 없다고 판단된 지표변수들의 문항 응답 확률은 1단계에서 추정된 모수치로 고정하고, 직접효과가 있는 지표변수의 문항 응답 확률은 두 번째 단계에서 추정한다. 공변인이

지표변수에 미치는 직접효과는 (식 15)의 다항 로지스틱 회귀식에 포함된 γ_{qk}로 모수화할 수 있다(Janssen et al., 2019).

$$\log it P(Y_i = y \mid C_i = k, Z_{pi}) = \eta_{jk} + \sum_{q=1}^{Q} \gamma_{qk} Z_{piq} \quad \text{(식 15)}$$

지표변수가 결과변수에 영향을 미치는 경우에는 잠재계층별 결과변수의 조건부 분포를 나타내는 $P(Z_{oi}|C_i = k)$가 잠재계층뿐 아니라 특정 지표변수의 값에도 의존하게 된다. 이 경우 관심모수는 지표변수의 값을 통제했을 때의 계층별 조건부 결과수준과 지표변수가 결과변수에 미치는 직접효과로 구분하여 해석할 필요가 있으며, 이러한 구조는 (식 16)과 같이 나타낼 수 있다.

$$P(Y_i = y, Z_{oi}) = \sum_{k=1}^{K} P(C_i = k) P(Z_{oi} \mid C_i = k, Y_{ij} = y_j) \; P(Y_i = y \mid C_i = k) \quad \text{(식 16)}$$

이와 같이 결과변수와 지표변수 사이에 남아 있는 관계를 직접효과로 모형화하면, 지표변수의 영향을 통제한 계층별 조건부 절편 즉 평균을 추정할 수 있다. 2단계 접근법의 첫 번째 단계에서는 지표변수의 측정모형과 잠재계층 확률을 추정하고, 두 번째 단계에서는 이들 모수를 고정한 상태에서 계층별 조건부 평균과 지표변수가 결과변수에 미치는 직접효과를 함께 추정한다.

3단계 접근법은 분류 오류를 고려하면서 측정모형과 외부변수 간 관계를 분리하여 추정할 수 있어 LCA와 LPA의 응용연구에서 널리 활용되어 왔다(Asparouhov & Muthen, 2014; Vermunt & Magidson, 2021). 그러나 앞서 살펴본 바와 같이 이 방법의 최종 단계에서는 원래의 지표변수가 제외되므로 외부변수와 지표변수 간 직접효과를 추가적으로 모형화하는 데 제약이 있다. 이에 본 연구에서는 3단계 접근법을 비교 대상으로 포함하여, 직접효과가 존재하지만 이를 반영하지 않은 경우의 추정 결과를 2단계 접근법과 비교하고 직접효과의 탐색 및 모형화의 중요성을 살펴보고자 하였다.

방법: 연구 1

연구 모형

본 연구에서는 공변인과 지표변수 간 직접효과가 존재하는 경우, 직접효과의 누락 여부가 관심모수 추정에 미치는 영향을 평가하기 위해 시뮬레이션 연구를 수행하였다. <그림 6>에 나타난 연구 1의 연구모형은 하나의 공변인이 총 6개의 이분형 지표변수 중 일부에 직접효과를 미치는 LCA 모형이다. 해당 모형에 대해 직접효과를 설정한 2단계 접근법, 직접효과를 설정하지 않은 2단계 접근법, 그리고 3단계 접근법을 적용하여 추정한 결과를 비교하였다. 데이터 생성과 모형 추정은 Latent

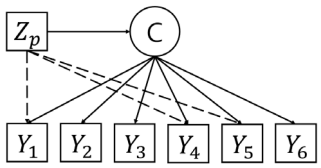


그림 6. 연구 1의 연구 모형

GOLD 6.0(Vermunt & Magidson, 2020)을 활용하였다.

본 연구에서 공변인이 존재할 때 계층 k에 속할 확률은 (식 17)을 이용하여 정의되었다.

$$\log\left(\frac{P(C_i = k \mid Z_{pi})}{P(C_i = 1 \mid Z_{pi})}\right) = \alpha_k + \beta_k Z_\pi \quad \text{(식 17)}$$

첫 번째 계층을 참조 계층으로 지정하였으며, α_k은 공변인의 값이 0일 때 계층 k에 속할 로그오즈의 절편을, β_k는 공변인의 값이 1일 때 잠재계층에 미치는 공변인의 효과를 의미한다.

(식 18)과 (식 19)는 각 계층에 속한 개인이 문항에 응답할 확률을 나타낸다.

$$logit(P(Y_{ij} = 1 \mid C_i = k, Z_{pi})) = \eta_{jk} + \gamma_j Z_{pi} \quad \text{(식 18)}$$

$$logit(P(Y_{ij} = 0 \mid C_i = k, Z_\pi)) = -\eta_{jk} - \gamma_j Z_{pi} \quad \text{(식 19)}$$

(식 18)을 사용하여 문항에 대해 1로 응답할 확률을, (식 19)를 통해 문항에 대해 0으로 응답할 확률을 각각 조작하였다. η_{jk}는 각 계층에 속한 개인이 가지는 문항 응답 확률의 절편이며, γ_j는 각 문항에 대하여 가지는 공변인의 직접효과를 의미한다.

고정 조건

연구 1의 참모형에서는 잠재계층의 수를 3개로 설정하였고, 지표변수는 6개의 이분형 변수로 구성하였다. 또한 3개의 잠재계층별로 문항에 대해 0 또는 1로 응답할 확률을 다르게 지정하였다. 계층 1은 모든 지표변수에 대해 0으로 응답하는 특성을 가지며, 계층 2는 모든 지표변수에 1로 응답하도록 설정하였다. 계층 3은 6개 문항 중 앞의 3개 문항에는 1로, 나머지 3개 문항에는 0으로 응답하는 특성을 부여하였다. 각 계층의 모집단 비율은 동일하게 설정하였다.

공변인은 단일 변수로 이루어져 있으며, 0과 1로 구성된 이분형 변수로서 0과 1의 비율을 50:50으로 고정하였다. 공변인의 효과를 설정하기 위해, 계층 1을 참조 계층으로 설정하였고 α_k에 대하여 α_2을 0.4, α_3를 -0.6으로 설정하였다(Janssen et al., 2019). 또한, 공변인의 효과인 β_k에 대해 β_2은 -1, β_3는 1로 설정하여 공변인이 1일 때, 계층 1에 속할 로짓은 -0.6만큼, 계층 2에 속할 로짓은 0.4만큼 변화하도록 각각 설정하였다(Bakk & Kuha, 2018; Janssen et al., 2019).

조작 조건

연구 1의 시뮬레이션에서 조작한 요인은 (1) 표본 크기(500, 1000, 2000), (2) 1에 응답할 확률(0.8, 0.9), (3) 지표변수에 미치는 직접 효과의 개수(1개, 2개, 3개), (4) 직접 효과의 크기(0.4, 0.7)이다.

표본 크기는 참고문헌을 바탕으로 판단한 500을 최소 기준으로 설정하였으며(Asparouhov & Muthén, 2014; Di Mari et al., 2020; Janssen et al., 2019), 선행 연구에서 널리 사용된 2000을 최대 기준으로 선택하였다(Bakk & Kuha, 2018). 잠재계층 분류의 질은 항목에 대해 '1'로 응답할 확률을 다르게 설정하여 분리도를

조작하였다. 구체적으로 중간 수준(0.8)은 로짓 절편 약 1.386[=log(0.8/0.2)], 높은 수준(0.9)은 약 2.197[=log(0.9/0.1)]로 설정하여 계층 간 응답반응의 차이를 다르게 설정하였으며, 이를 통해 3단계 접근법의 분류 오류가 달라지도록 설계하였다. 직접효과 조작은 공변인이 지표에 미치는 직접 경로의 개수(1개, 2개, 3개)와 크기(0.4=중간, 0.7=강함)를 변화시켜, 직접효과의 개수가 많고 크기가 클수록 공변인이 잠재계층에 미치는 효과 추정치가 편향된다는 점을 체계적으로 검증할 수 있도록 하였다.

연구 1의 시뮬레이션은 4가지 요인으로부터 도출된 총 36개 조건을 바탕으로 데이터 생성부터 분석까지 100회 반복 수행하였다.

평가 항목

연구 1에서는 공변인이 잠재계층에 미치는 효과인 β_2와 β_3을 정확하게 추정하였는지 평가하기 위해 상대적 편향 평균(Average Relative Bias)을 사용하였다. 상대적 편향 평균은 연구 조건에서 설정한 모수와 추정치 간의 편향 정도를 확인할 수 있으며, (식 20)과 같이 연구자가 설정한 모수와 실제 추정치 값의 차이를 모수로 나누어 계산된다(Harwell, 2019).

$$Bias = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{2}\left(\frac{(\widehat{\beta_{2r}}-\beta_2)}{\beta_2}+\frac{(\widehat{\beta_{3r}}-\beta_3)}{\beta_3}\right) \quad \text{(식 20)}$$

R은 모의실험 반복의 횟수이고, β_2와 β_3는 자료를 생성할 때 연구 조건에 따른 모수이며, -1과 1로 고정하였다. 반면 $\hat{\beta}_2$와 $\hat{\beta}_3$는 모형을 통해 추정한 추정치를 나타낸다. 상대적 편향 평균이 ±10%를 초과하지 않는다면 수용할 만한 수치라고 판단하였으며, ±5% 미만인 경우 편향이 작거나 거의 없다고 판단하였다(Muthén & Muthén, 2002; Wolf et al., 2013).

연구 결과: 연구 1

상대적 편향 평균

공변인이 지표변수의 문항 응답 확률에 영향을 미치는 경우에 직접효과를 포함한 모수 추정의 정확도를 파악하기 위해, 공변인이 잠재계층에 미치는 영향에 대한 상대적 편향 평균을 산출하였다. 또한, 이를 표본 크기 조건별로 <그림 7>, <그림 8>, <그림 9>에 제시하였다.

직접효과를 포함한 모형에서는 대부분의 조건에서 상대적 편향 평균이 5% 미만으로 나타났다. 구체적으로, 표본 크기가 500이고 분류의 질이 중간 수준인 조건에서는 상대적 편향 평균이 -3.1%에서 -11.6% 사이였으며, 분류의 질이 높은 조건에서는 -0.3%에서 -1.8% 사이로 편향이 더 적었다. 반면, 직접효과를 포함하지 않은 2단계 및 3단계 접근법은 각각 최대 25%, 24%에 달하는 높은 상대적 편향 평균을 보였다. 특히, 직접효과의 수가 3개이고 크기가 0.4인 조건에서는 2단계 접근법이 -16.6%, 3단계 접근법이 -15.8%의 편향을 나타냈으며, 직접효과 크기가 0.7인 조건에서는 각각 -25.1%, -24.1%로 편향이 더욱 커졌다. 다만, 분류의 질이 높은 조건에서는 직접효과를 포함하지 않은 2단계 및 3단계 접근법 모두 ±5% 수준의 비교적 양호한 편향을 보였다.

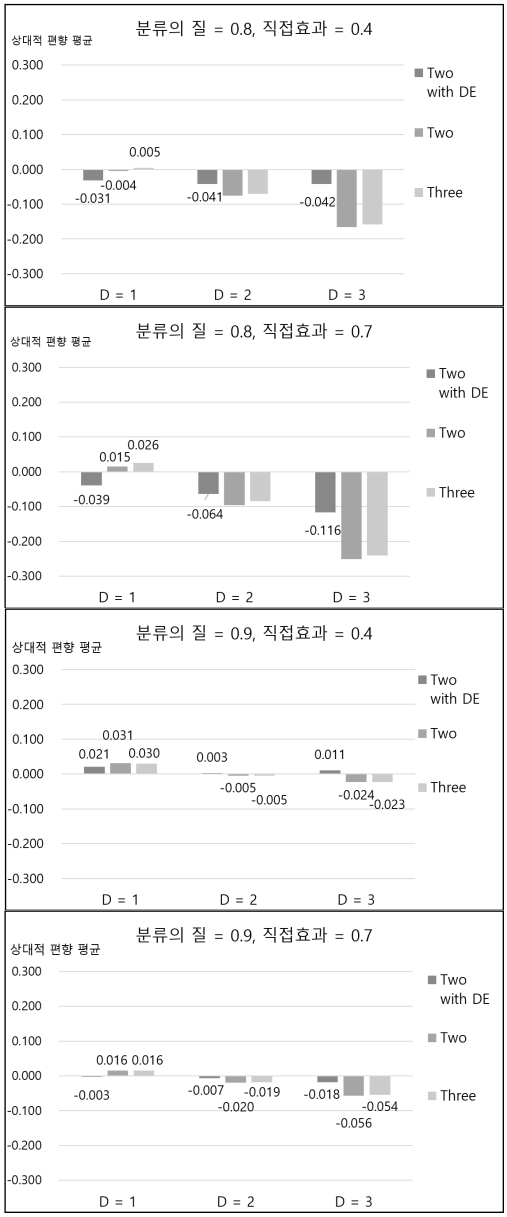


그림 7. 표본 크기가 500인 조건에서 공변인의 효과에 대한 상대적 편향 평균

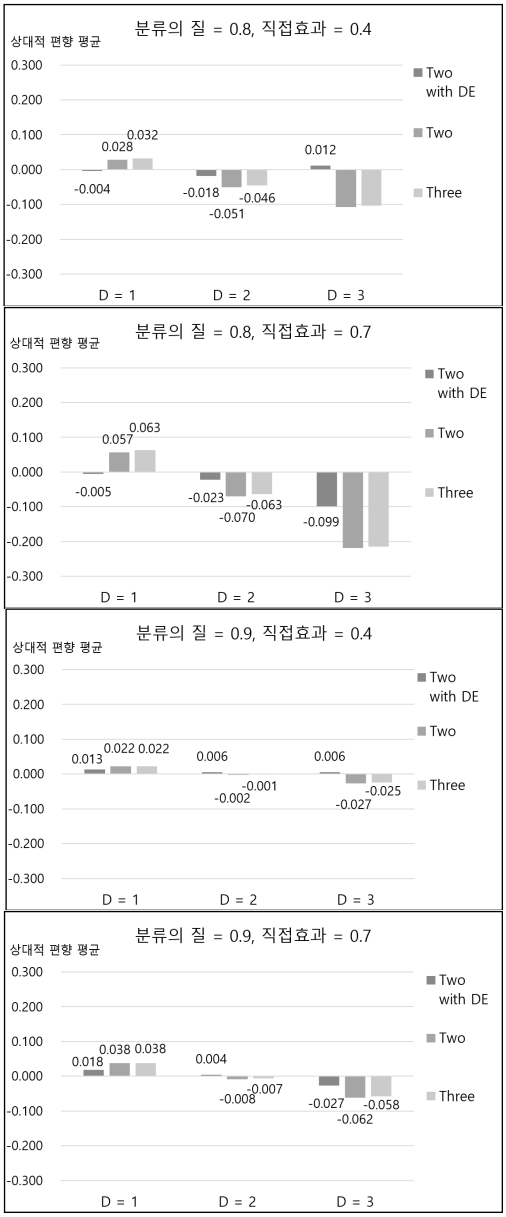


그림 8. 표본 크기가 1000인 조건에서 공변인의 효과에 대한 상대적 편향 평균

또한, 표본 크기가 커질수록 상대적 편향 평균은 전반적으로 감소하는 경향을 보였다. 표본 크기가 1000인 경우, 직접효과를 포함한 모형은 최대 -9.9%의 편향을 보였으며, 모든 조건에서 ±10% 이내의 상대적 편향을 유지하였다. 분류의 질이 중간 수준인 조건에

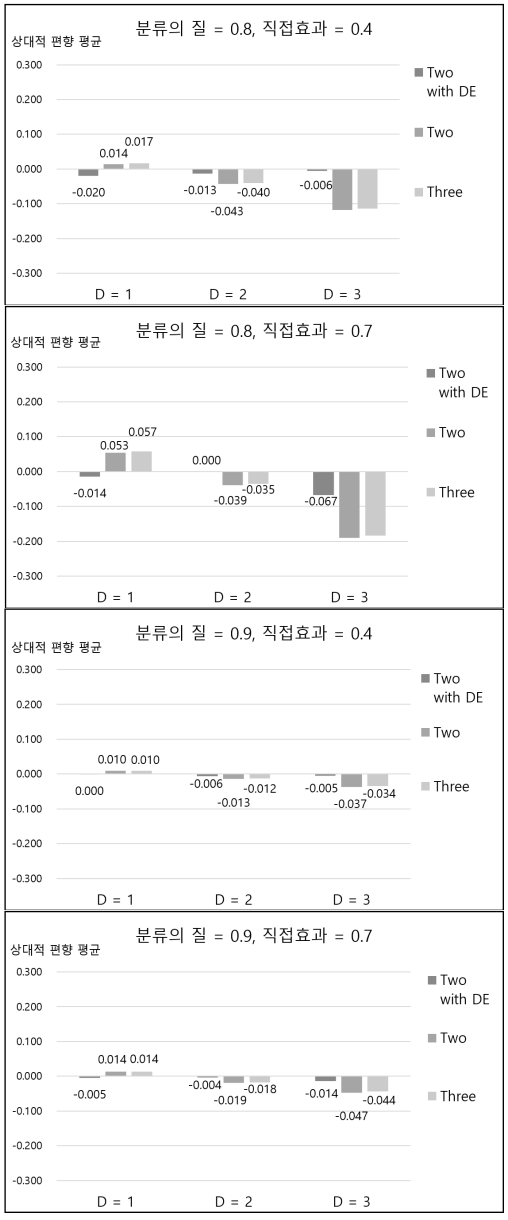


그림 9. 표본 크기가 2000인 조건에서 공변인의 효과에 대한 상대적 편향 평균

서는 -0.4%에서 -9.9% 사이의 편향 평균이 산출되었고, 분류의 질이 높은 조건에서는 0.4%에서 -2.7% 사이로 모두 5% 미만의 편향을 보였다. 반면, 직접효과를 포함하지 않은 2단계 및 3단계 접근법은 유사한 경향을 나타냈다. 분류의 질이 중간 수준이고 직접효과의 수가 3개인 조건에서는 직접효과 크기가 0.4일 때 2단계 접근법이 -10.8%, 3단계 접근법이 -10.3%로 10%를 약간 초과하였으며, 직접효과 크기가 0.7일 때는 2단계 접근법이 -21.9%, 3단계 접근법이 -21.5%로 심각한 편향이 확인되었다.

표본 크기가 2000인 조건에서는 직접효과를 포함한 모형이 최대 -7%를 넘지 않았고, 특히 분류의 질이 높은 경우 모든 조건에서 1% 미만의 편향을 보였다. 직접효과를 포함하지 않은 2단계 및 3단계 접근법은 분류의 질이 중간 정도인 경우 직접효과 수가 3개인 조건을 제외하고 모두 양호한 편향을 나타냈다. 분류의 질이 높은 경우에는 직접효과를 포함하지 않은 2단계 및 3단계 접근법 모두 ±5% 미만의 상대적 편향을 보였다.

방법: 연구 2

연구 모형

연구 2의 연구 모형(그림 10)은 4개의 연속형 지표변수 중 일부가 연속형 결과변수에 직접효과를 가지는 잠재프로파일모형으로 설정하였다.

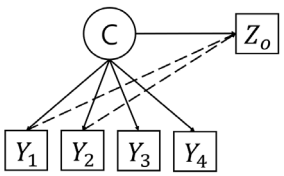


그림 10. 연구 2의 연구 모형

모형에 포함된 결과변수 Z_{oi}는 (식 21)로 정의하였다.

$$Z_{oi} = \mu_k + \beta Y_{1i} + \beta Y_{2i} \quad \text{(식 21)}$$

μ_k는 Y1과 Y2의 값을 모두 0으로 고정했을 때 계층 k에서 기대되는 조건부 평균을 의미하며, 1번 문항과 2번 문항의 직접효과를 β로 설정하였다.

해당 모형에 대해 직접효과를 포함한 2단계 접근법, 직접효과를 포함하지 않은 2단계 접근법, 그리고 3단계 접근법을 적용하여 추정한 결과를 비교하였다. 데이터 생성은 R-4.4.3(R Core Team, 2025)과 Mplus 8.5(Muthén & Muthén, 2017)를 사용하였고, 모형 추정은 Latent GOLD 6.0(Vermunt & Magidson, 2020)을 활용하였다.

고정 조건

연구 2의 모형은 계층 수를 2개로 설정하였고, 지표변수는 4개의 연속형 변인으로 구성하였다. 결과변수는 하나의 연속형 변수로 설정하였고, 분산은 각 계층에서 모두 1인 정규분포로 설정하였다. 직접효과가 설정된 지표변수는 Y_1과 Y_2로 설정하였으며, 모두 계층과 관계없이 일정한 크기의 직접효과를 가진다고 가정하였다.

조작 조건

연구 2의 시뮬레이션에서 조작한 요인은 총 네 가지로, (1) 표본 크기(500, 1000, 2000), (2) 계층 간 지표변수의 평균 차이(계층 간 평균 차이가 큰 조건과 작은 조건), (3) 계층별 결과변수의 평균 차이(계층별 평균이 동일한 조건과 다른 조건), (4) 지표변수의 직접효과 크기(0, 0.2, 0.5)이다. 표본 크기는 선행 연구를 참고하여 최소 500, 중간 1000, 최대 2000으로 설정하였고, 4개의 지표변수 평균을 조작하여 계층별 분리도를 '작은 차이' 조건(계층1=0.24, 계층2=0.74)과 '큰 차이' 조건(계층1=3, 계층2=7)으로 구분하였다. 결과변수 평균은 계층별로 동일 조건(μ_1=1, μ_2=1)과 상이 조건(μ_1=1, μ_2=5)으로 대비하여 결과변수의 평균 차이가 결과에 미치는 영향을 살펴보았다. 계층 간 평균 차이의 경우, 본 연구와 직접적으로 관련된 LPA 연구가 부족하여 경험연구를 바탕으로 본 연구의 설계에서 적정하다고 판단한 수준으로 계층 간 평균 차이를 정의하였다. 마지막으로 지표변수와 결과변수 간 누락된 직접 경로의 크기를 회귀계수 0(없음), 0.2(중간), 0.5(큼)으로 조작하여 직접효과의 존재와 크기에 따른 편향 및 성능 변화를 평가하도록 설계하였다. 연구 2는 이 네 가지 요인을 바탕으로 총 36개의 데이터 세트를 생성하여 100회 반복 실행하였다.

평가 항목

상대적 편향 평균

연구 2에서도 연구 1과 마찬가지로 잠재계층별 결과변수의 평균인 μ_1와 μ_2을 정확히 추정하였는지 평가하기 위해 연구 1과 동일하게 상대적 편향 평균을 사용하였다.

$$Bias = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{2}\left(\frac{(\hat{\mu}_1 - \mu_1)}{\mu_1} + \frac{(\hat{\mu}_2 - \mu_2)}{\mu_2}\right) \quad \text{(식 22)}$$

상대적 편향 평균은 (식 22)를 사용하여 계산하였으며 연구 조건에 따라 고정된 μ_1와 μ_2와 모형을 통해 추정한 $\hat{\mu}_1$와 $\hat{\mu}_2$을 바탕으로 계산하였다. 상대적 편향 평균이 ±10%를 초과하지 않으면 수용할 만한 수준이라고 판단하였으며, ±5% 미만인 경우 편향이 작거나 거의 없다고 판단하였다(Muthén & Muthén, 2002; Wolf et al., 2013).

1종 오류율

연구 2에서는 두 계층 간 결과변수 평균 차이 검정의 정확성을 평가하기 위해 1종 오류율(type I error rate)을 활용하였다. 지표변수가 결과변수에 직접적인 영향을 미치는 경우, 유의한 계층 간 평균 차이는 순수한 계층 간 차이일 수도 있지만, 지표변수와 결과변수 간의 선형적 효과가 잠재계층의 고유 특성으로 잘못 반영되어 결과변수 평균이 부정확하게 추정되는 현상으로 인해 나타날 수도 있다. 본 연구에서는 잠재계층별 결과변수 평균의 동일성을 검정하기 위해 Latent GOLD에서 제공하는 Wald 검정을 사용하였다. 이때 영가설은 잠재계층별 결과변수의 모집단 평균이 동일하다는 것이다. 따라서 계층별 평균이 동일한 조건에서 Wald 검정이 유의하게 나타난 비율을 경험적 1종 오류율, 즉 잘못된 유의판정 비율로 산출하고 명목 유의수준인 5%와 비교하였다(Muthén & Muthén, 2002).

검정력

연구 2의 조건에서 잠재계층별 결과변수의 평균 차이가 존재할 때, 올바르게 평균 차이가 존재한다고 판단한 비율을 검정력(power)으로 활용하였다. 평균 비교를 위해 Latent GOLD에서 제공하는 Wald 검정을 사용하였으며, 검정통계치가 유의한 경우 평균 차이를 정확히 탐지하였다고 판단하였다. 전체 자료 중 평균 차이를 올바르게 탐지한 자료의 비율을 계산하였으며, 전통적으로 1종 오류율을 5%로 설정하였을 때 통상적으로 2종 오류율(type II error rate)을 20%로 설정한다. 이에 따라, 본 연구에서도 2종 오류율을 20%로 설정하여 검정력이 0.8 이하일 경우 검정력이 낮다고 판단하였다(Cohen, 1992).

연구 결과: 연구 2

상대적 편향 평균

지표변수가 결과변수에 영향을 미치는 경우, 직접효과를 포함한 모형의 추정 정확성을 평가하기 위해 결과변수 평균의 상대적 편향 평균을 산출하였다. 결과는 <그림 11>, <그림 12>, <그림 13>에 제시하였다.

우선, 직접효과가 없는 조건에서는 직접효과를 포함한 2단계 접근법과 직접효과를 포함하지 않은 2단계 및 3단계 접근법 모두 상대적 편향 평균이 1% 미만으로 매우 우수한 수준을 나타냈다. 그러나 직접효과가 존재하는 조건에서는 직접효과 설정 여부에 따라 편향 크기에 큰 차이가 나타났다.

직접효과를 포함한 2단계 접근법은 대부분

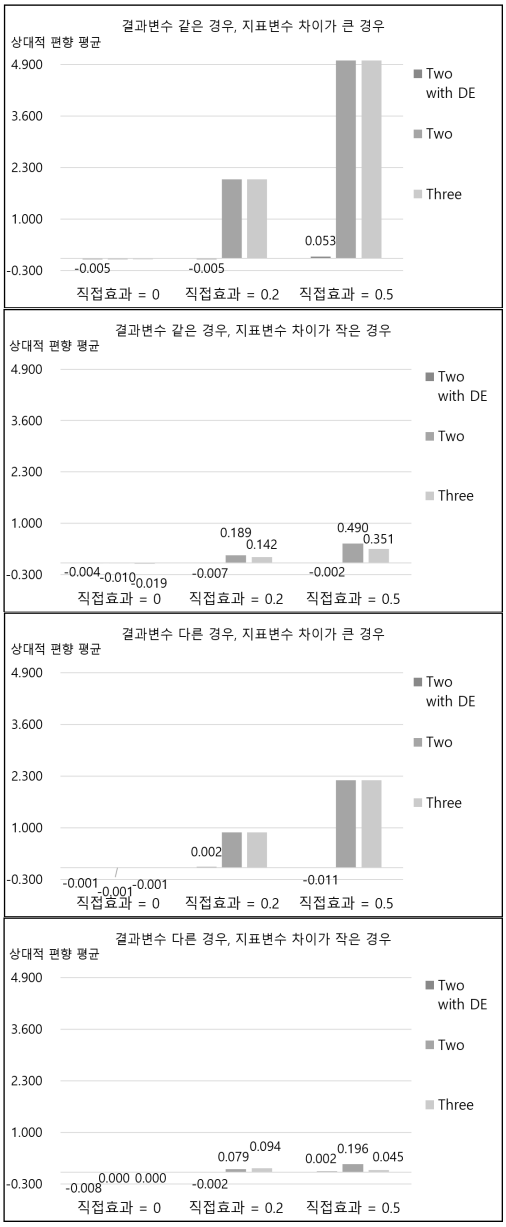


그림 11. 표본 크기가 500인 조건에서 결과변수 평균에 대한 상대적 편향 평균

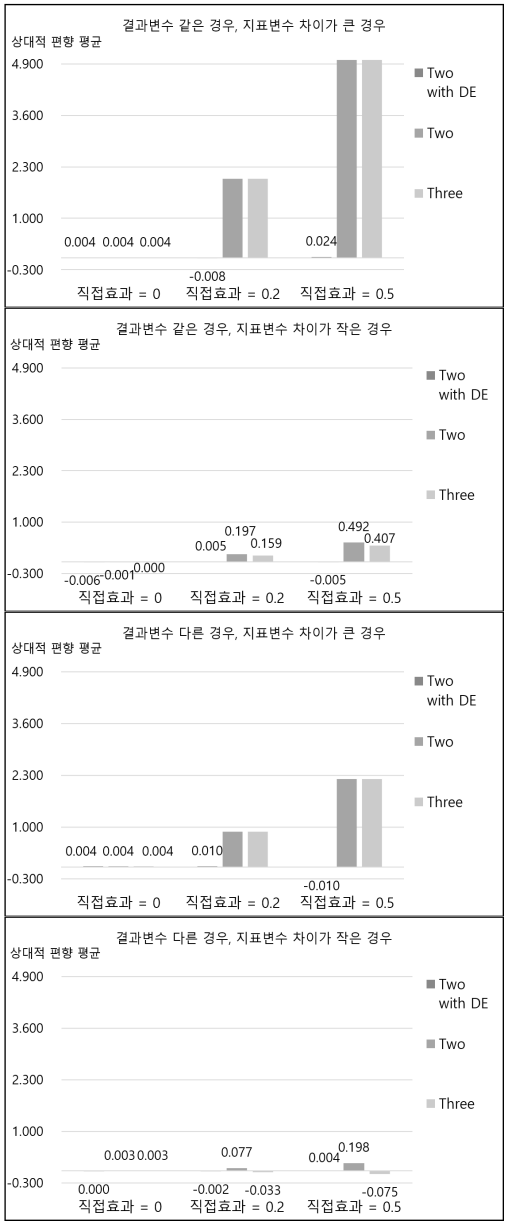


그림 12. 표본 크기가 1000인 조건에서 결과변수 평균에 대한 상대적 편향 평균

의 조건에서 상대적 편향 평균이 5% 미만으로 안정적인 경향을 보였다. 구체적으로, 잠재계층별 결과변수 평균이 동일한 경우, 지표변수 간 차이가 큰 조건에서는 -0.1%에서 -5.3%, 차이가 작은 조건에서는 0에서 -0.8% 수준의 편향 평균이 관찰되었다. 반면, 직접효

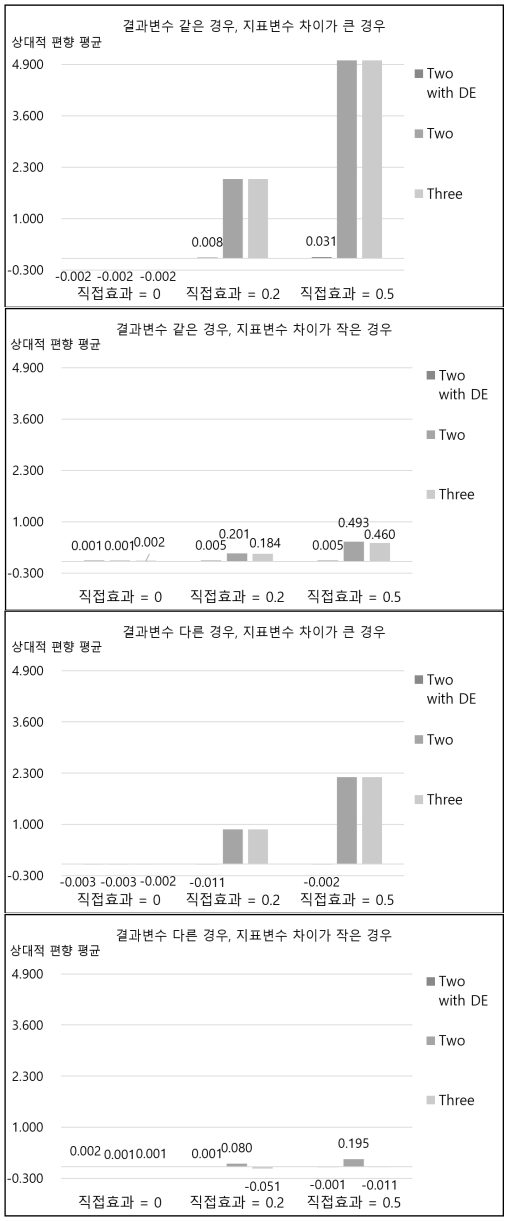


그림 13. 표본 크기가 2000인 조건에서 결과변수 평균에 대한 상대적 편향 평균

과를 포함하지 않은 2단계 및 3단계 접근법은 지표변수 간 차이가 큰 조건에서 상대적 편향 평균이 200%에서 500% 수준까지 증가하였다. 지표변수 간 차이가 작은 조건에서도, 직접효과를 포함하지 않은 2단계 접근법은 직접효과 크기가 0.2일 때 18.9%에서 20.1%, 직접효과 크기가 0.5일 때는 49.0%에서 49.2% 범위의 편향을 보였다. 3단계 접근법의 경우 각각 14.2%에서 18.4%, 35.1%에서 46.0% 범위의 편향이 관찰되었다.

잠재계층별 결과변수 평균이 다른 경우에도 유사한 경향이 나타났다. 지표변수 간 차이가 큰 조건에서 직접효과를 포함한 2단계 접근법은 최대 2% 수준의 상대적 편향을 보였으나, 직접효과를 포함하지 않은 2단계 및 3단계 접근법은 직접효과 크기에 따라 88.4%에서 220.4%까지 편향이 증가하였다. 다만, 지표변수 간 차이가 작은 조건에서는 직접효과 크기가 0.5인 경우를 제외하고, 직접효과를 설정하지 않은 2단계 접근법이 19%의 상대적 편향을 보인 경우를 제외하면, 나머지 조건에서는 2단계 및 3단계 접근법 모두 상대적 편향 평균이 10% 미만으로 비교적 양호한 결과를 나타냈다.

1종 오류율

먼저, <그림 14>는 계층 간 지표변수의 평균 차이가 큰 조건에서 결과변수의 평균 차이에 대한 1종 오류율을 산출한 결과이다.

표본 크기가 500일 때, 직접효과가 없을 경우 직접효과를 포함한 모형의 1종 오류율은 3%였으며, 직접효과를 포함하지 않은 2단계 및 3단계 방법은 모두 7%의 1종 오류율을 나타냈다. 직접효과 크기가 0.2인 경우 직접효과를 포함한 2단계 방법은 6%, 0.5인 경우에는 5%의 1종 오류율을 보였으나, 직접효과를 고려하지 않은 두 모형은 모든 경우에서 잠재계

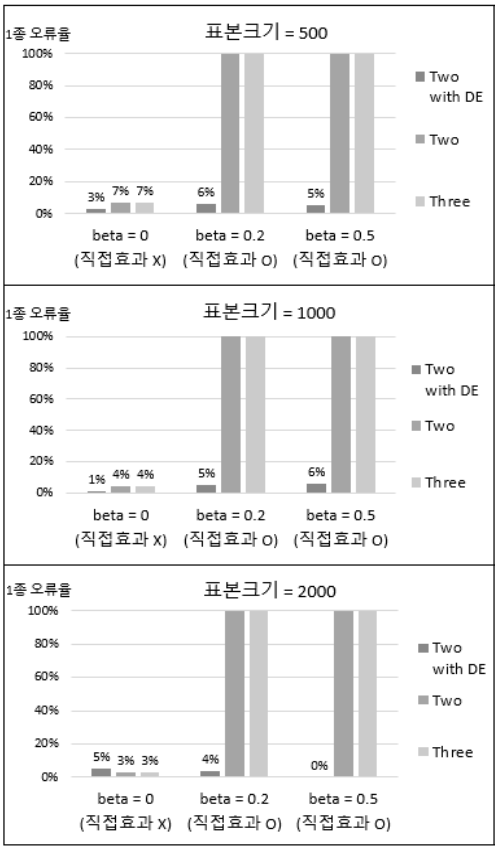


그림 14. 지표변수의 평균 차이가 큰 경우 세 모형의 1종 오류율

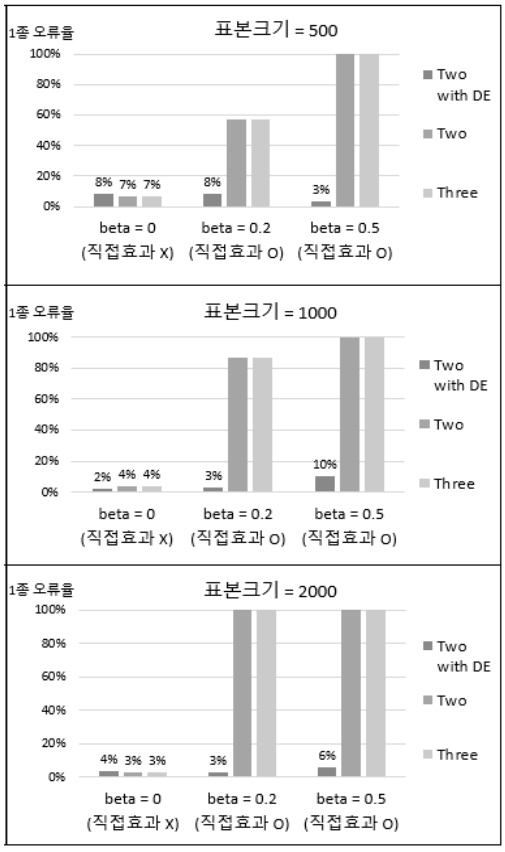


그림 15. 지표변수의 평균 차이가 작은 경우 세 모형의 1종 오류율

층별 결과변수의 평균 차이가 있다고 잘못 판단했다.

표본 크기가 1000일 때, 직접효과가 없으면 직접효과를 포함한 모형은 1%로 매우 낮은 1종 오류율을 보였고, 직접효과를 포함하지 않은 두 모형도 각각 4%로 양호한 수준이었다. 직접효과 크기가 0.2인 경우 직접효과를 포함한 모형은 여전히 5%의 적절한 1종 오류율을 유지했으며, 0.5인 경우에는 6%를 기록했다. 반면, 직접효과를 반영하지 않은 두 모형은 모두 100%의 1종 오류율을 보여 모든 상황에서 잘못된 결론을 내렸다.

표본 크기 2000 조건에서는 직접효과가 존재하지 않을 경우 세 모형 모두 3~5% 범위 내에서 안정적인 1종 오류율을 나타냈다. 직접효과가 0.2인 경우 직접효과를 포함한 모형은 4%, 0.5인 경우에는 0%로 매우 낮은 1종 오류율을 유지했으나, 직접효과를 고려하지 않은 2단계 및 3단계 방법은 모든 상황에서 결과변수의 계층별 평균 차이가 있다고 판단했다.

<그림 15>는 계층 간 지표변수의 평균 차이가 작은 조건에서 결과변수의 평균 차이에 대한 1종 오류율 결과를 제시한다. 표본 크기

가 500이며 직접효과가 존재하지 않는 경우, 직접효과를 포함한 모형은 8%, 직접효과를 포함하지 않은 2단계 및 3단계 접근법은 모두 7%의 1종 오류율을 보였다. 그러나 직접효과가 0.2인 경우 직접효과를 포함한 2단계 접근법은 8%, 0.5인 경우에는 3%의 1종 오류율을 보였다. 직접효과가 0.2인 경우 직접효과를 포함하지 않은 두 접근은 57%에 달하는 높은 1종 오류율을 나타냈다. 직접효과의 크기가 0.5로 증가한 조건에서는 직접효과를 고려하지 않은 접근들이 모두 100%의 1종 오류율을 보였다.

표본 크기 1000 조건에서는 직접효과가 존재하지 않을 경우 직접효과를 포함한 모형이 2%, 직접효과를 포함하지 않은 두 모형이 각각 4%의 1종 오류율을 보였다. 그러나 직접효과가 0.2인 경우 직접효과를 포함한 모형은 3%의 적절한 1종 오류율을 유지하였으나, 직접효과를 포함하지 않은 2단계 및 3단계 접근법은 87%의 높은 1종 오류율을 기록하였다. 직접효과의 크기가 0.5로 증가한 조건에서는 직접효과를 포함한 모형의 1종 오류율이 10%로 다소 증가했으나, 직접효과를 고려하지 않은 두 모형은 모두 100%의 1종 오류율을 보였다.

표본 크기 2000 조건에서는 직접효과가 존재하지 않는 경우 모든 모형이 안정적인 1종 오류율을 나타냈다. 직접효과를 포함한 모형은 4%, 포함하지 않은 모형들은 각각 3%의 1종 오류율을 보였다. 하지만 직접효과가 0.2인 경우 직접효과를 포함한 모형은 3%, 0.5인 경우에는 6%의 1종 오류율로 비교적 잘 통제된 결과를 보였다. 반면, 직접효과를 포함하지 않은 2단계 및 3단계 방법은 두 조건 모두에서 100%의 1종 오류율을 나타냈다.

검정력

직접효과가 존재하지 않는 조건에서는, 직접효과를 포함한 모형과 포함하지 않은 접근 모두에서 검정력이 100%로 나타나 영가설을 정확히 기각할 수 있음을 확인하였다. 또한, 직접효과가 존재하는 경우에도 직접효과를 통제한 모형과 그렇지 않은 모형 모두 높은 검정력을 보여 평균 차이를 정확하게 탐지하였다.

결 론

본 연구는 외부변수와 지표변수 간에 다뤄지지 못한 관계로 인해 발생하는 편향을 확인하기 위하여 시뮬레이션 연구를 진행하였다. 본 연구에서는 기존에 연구된 공변인이 지표변수에 미치는 직접효과와 이를 효과적으로 통제하지 못하여 발생한 편향뿐만 아니라 잠재프로파일분석 내에서 지표변수가 결과변수에 미치는 직접효과와 이로 인해 야기되는 편향도 함께 살펴보았다. 이를 위해 연구과정을 연구 1과 연구 2로 나누어 진행하였으며, 각 연구에 36개의 조건을 생성하여 직접효과를 설정한 2단계 접근법, 직접효과를 설정하지 않은 2단계 접근법, 3단계 접근법 세 개의 모형을 통해 분석을 실시하였다. 이후, 연구 1에서는 상대적 편향 평균을, 연구 2에서는 상대적 편향 평균, 1종 오류율과 검정력을 사용하여 모형에 대한 평가를 진행하였으며, 잘못된 지표변수에 대해 직접효과를 설정한 오추정 오류는 고려하지 않았다. 연구 1과 연구 2는 외부변수의 종류만 달라지며 직접효과를 설정하는 것 자체는 모든 과정에서 동일한 절차를

거쳤다. 또한, 선행 연구에서도 동일한 절차로 직접효과를 검증하였기에 본 연구에서도 동일한 절차를 적용하였다.

연구 1에서는 공변인이 지표변수에 미치는 영향과 이로 인해 공변인이 잠재계층에 미치는 영향의 편향 정도에 대해 살펴보았다. 이를 위해, 표본의 크기, 잠재계층 분류의 질, 직접효과의 수, 공변인의 직접효과 크기를 연구 조건으로 설정하여 자료를 생성하여 그 결과를 비교하였다. 결과는 추정치의 편향을 계산하였으며, 공변인이 잠재계층의 분포에 미치는 영향의 추정치와 연구자가 설정한 모수치 간 차이를 모수치로 나눈 상대적 편향으로 확인하였다.

표본 크기가 작은 조건에서 대부분의 결과가 양호한 상대적 편향을 보고하였으며, 표본의 크기가 증가함에 따라 편향의 정도가 감소하는 양상을 보였다. 또한, 잠재계층 분류의 질이 높은 조건에서는 분류의 질이 중간 수준인 조건에 비해 상대적으로 편향치가 낮고 안정적인 모습을 보였으며, 이러한 경향은 표본 크기가 증가할수록 더욱 뚜렷하게 나타났다. 직접효과의 크기는 잠재계층 분류의 질이 높은 경우에 큰 영향을 미치지 않았다. 반면, 분류의 질이 중간 수준인 조건에서는 직접효과의 크기가 커질수록 더 큰 편향이 발생하는 양상이 보였다. 그럼에도 불구하고, 직접효과를 설정한 모형은 그렇지 않은 두 모형에 비해 상대적으로 낮고 안정적인 값을 보였다. 직접효과의 수가 1개인 경우 모형에 직접효과를 설정한 모형과 그렇지 않은 모형 간 편향의 정도에 큰 차이가 없었지만, 그 수가 증가할수록 직접효과를 다루지 않은 모형은 지속적으로 편향의 정도가 증가하였다. 하지만, 직접효과를 설정한 모형은 추정치를 더 안정적으로 산출하였으며, 편향의 정도가 거의 증가하지 않았다.

연구 2에서는 지표변수의 직접효과가 결과변수의 평균 차이 검정에 미치는 영향에 대해 살펴보았다. 이를 위해, 두 잠재프로파일 간 차이, 잠재프로파일별 결과변수의 평균 차이, 지표변수의 직접효과 크기, 표본 크기에 따라 자료를 생성하여 상대적 편향 평균, 1종 오류율과 검정력을 비교하였다.

연구 결과, 직접효과가 존재하지 않는 경우, 모든 접근법이 1% 미만의 상대적 편향으로 안정적인 추정치를 제공하였다. 그러나 직접효과가 존재하는 경우, 직접효과를 설정한 모형은 대부분의 조건에서 5% 미만의 편향을 유지한 반면, 이를 고려하지 않은 2단계 및 3단계 접근법은 특히 지표변수 간 차이가 크고 직접효과의 크기가 클수록 상대적 편향 등의 오차가 크게 증가하였다. 직접효과가 존재하지 않는 조건에서는 모든 모형이 1~8% 정도의 양호한 1종 오류율을 보였으며, 표본 크기가 증가함에 따라 1종 오류율이 낮아졌다. 하지만 직접효과가 존재하는 조건에서는, 직접효과의 설정 여부에 따라 1종 오류율에 큰 차이를 보였다. 직접효과를 모형에 포함한 접근법의 경우 약 6% 정도의 양호한 1종 오류율을 보였지만, 직접효과를 포함하지 않은 모형들은 100%의 심각한 1종 오류율을 보고하였다. 특히, 직접효과로 인해 결과변수의 평균 차이를 오인하는 상황은 직접효과가 커질수록 심화되는 양상을 보였으며, 이는 잠재프로파일 간 차이가 크지 않은 조건에 비해 차이가 큰 조건에서 더 뚜렷하게 나타났다. 이러한 결과는 LCA뿐만 아니라 LPA에서도 직접효과를 적절하게 다루어야 할 필요성을 시사하며, 두 연속형 변인 사이의 선형적인 관계를 모형

내에서 고려해야 함을 보여준다.

본 연구의 결과를 통해, 다양한 상황에서 직접효과를 설정한 모형이 그렇지 않은 모형에 비해 상대적으로 편향이 적었으며, 잠재계층별 평균 차이에 대한 해석을 더 정확하게 할 수 있음을 확인하였다. 또한, 연구 1의 결과를 통해 2단계 접근법이 3단계 접근법에 비해 작은 편향을 보고하고 있으며, 이는 여러 선행 연구의 결과와 일치한다(Di Mari et al., 2020; Janssen et al., 2019).

이러한 연구 결과는 외부변수와 잠재계층 간의 관계를 추정하는 과정에서 직접효과를 모형에 포함할 필요가 있음을 시사한다. 하지만, 이러한 절차를 구체적으로 제시한 연구는 거의 이루어지지 않았다. 따라서, 본 연구의 결과와 여러 선행 연구(Bakk et al., 2022; Di Mari et al., 2020; Janssen et al., 2019)를 바탕으로 본 연구에서는 LCA 또는 LPA를 실시하는 과정에서 직접효과를 탐색하고 이를 모형에 포함하는 실질적인 절차를 제안하고자 한다. 절차의 첫 번째 단계는 직접효과를 포함하지 않은 일반적인 모형을 분석하는 것이다. 앞서 제시한 2단계 접근법의 순서를 따라 우선 잠재계층모형을 분석하고, 여러 평가 지수(AIC, BIC, LMR-LRT 등)를 활용하여 최종 모형을 선택한 후, 외부변수와 잠재계층 간의 관계를 탐색한다. 두 번째 단계는 BVR, EPC와 같은 잔차 관련 지표들을 활용하여 정확히 어떠한 지표변수와 공변인 또는 결과변수 사이에 잔여 관계가 존재하는지 확인한다. 이러한 분석을 지원하기 위해 Latent GOLD 프로그램에서는 외부변수와 지표변수 사이에 남아있는 관계를 확인하기 위한 BVR과 EPC 등 여러 옵션을 제공하고 있으며, 이를 통해 잔여 관계가 존재하는지에 대하여 판단할 수 있다. 표 1과 표 2에서 예시로 제시된 바와 같이 BVR은 변수 쌍별로 프로그램상에서 제시된다.

지표변수(Y_1, Y_2)가 외부변수(Z_O)에 직접효과를 미칠 때, 표 1과 달리 표 2에서는 영향을 주고받는 두 변수 사이에는 3 이상의 큰 BVR 수치를 볼 수 있다. 이후, 세 번째 단계에서 공변인 또는 결과변수와 지표변수 간 관계를 추가하여 모형을 재추정한다. 공변인이 지표변수에 직접효과를 미치는 경우에는 직접효과가 존재하는 지표변수의 문항 응답 확률과 공변인이 잠재계층에 미치는 효과를 추정하고, 직접효과가 없는 지표변수의 문항 응답

표 1. 직접효과를 포함한 모형의 BVR

변인	Y_1	Y_2	Y_3	Y_4	Z_O
Y_1					
Y_2	1.04				
Y_3	0.89	1.00			
Y_4	1.06	0.99	1.05		
Z_O	0.55	0.50	0.53	0.50	

주. N=2000, β=0.5, 계층별 지표변수 차이가 큰 조건, 결과변수 차이가 큰 조건의 BVR 값임

표 2. 직접효과를 포함하지 않은 모형의 BVR

변인	Y_1	Y_2	Y_3	Y_4	Z_O
Y_1					
Y_2	1.04				
Y_3	0.89	1.00			
Y_4	1.06	0.99	1.05		
Z_O	165.92	168.05	1.05	1.06	

주. N=2000, β=0.5, 계층별 지표변수 차이가 큰 조건, 결과변수 차이가 큰 조건의 BVR 값임

확률은 첫 번째 단계의 추정치로 고정한다. 반면 지표변수가 결과변수에 직접효과를 미치는 경우에는 첫 번째 단계에서 추정한 문항 응답 확률과 잠재계층 확률을 고정하고, 두 번째 단계에서 잠재계층별 결과변수의 평균과 지표변수가 결과변수에 미치는 직접효과를 함께 추정한다. 이를 통해 외부변수와 지표변수 간의 잔여 관계를 반영한 상태에서 외부변수 관련 모수를 보다 정확하게 추정할 수 있다.

앞서 제시한 결과를 바탕으로, 본 연구는 LCA 또는 LPA에서 외부변수와 지표변수 사이에 남아있는 관계가 존재하는 상황에서 발생하는 편향과 직접효과를 설정한 모형을 통해 이를 효과적으로 통제할 수 있음을 확인하였다. 공변인이 지표변수의 문항 응답 확률에 영향을 미치는 상황에서 직접효과를 모형에 포함하였을 때, 그렇지 않은 모형의 추정치에 비해 모수값과 근사한 값을 산출하였다. 이는, 공변인이 지표변수에 미치는 직접효과를 모형에 통제하지 않을 경우, 문항 응답 확률뿐만 아니라 잠재계층 확률의 편향을 발생시킬 수 있음을 의미한다. 지표변수가 결과변수에 직접효과를 미치는 상황에서도 이를 모형에 포함하여 분석한 결과와 포함하지 않은 모형의 결과 사이에 큰 차이를 보였다. 지표변수의 직접효과를 모형에 설정한 경우, 잠재계층별 결과변수의 평균 차이에 대한 결과가 올바르게 추정되었지만, 직접효과가 모형에서 누락된 경우에는 평균 차이에 대한 잘못된 결과를 추정하였다. 이는, 특정 지표변수와 결과변수 간의 남아있는 선형적 관계가 여러 지표변수로 구성된 잠재계층의 특성으로 오인될 가능성을 의미한다. 따라서, 연구자는 외부변수와 잠재계층 간의 관계를 탐색하는 과정에서 외부변수와 지표변수 간의 관계를 살펴볼 필요가 있다.

그러나 연구자가 외부변수와 지표변수 간의 조건부 독립성을 확인하고 보고할 수 있는 방법에 대한 선행 연구는 아직 충분히 제시되지 않은 실정이다. 이러한 한계로 인해, 현재 LCA와 LPA를 활용하여 외부변수와 잠재계층 간의 관계를 탐색하는 많은 연구에서 외부변수와 지표변수 간의 조건부 독립성 여부를 별도로 검토하지 않고 있다. 이러한 점에서, 본 연구는 구조적 관계뿐만 아니라 측정 수준에서 발생할 수 있는 직접효과의 가능성 또한 면밀히 검토할 필요가 있음을 시사한다.

본 연구의 한계점은 다음과 같다. 첫째, 공변인이 연속형 변수인 상황에 대한 논의가 이루어지지 않았다. 실제로 나이와 같은 연속형 변수는 여러 연구에서 공변인으로 자주 활용된다. Di Mari 등(2020)의 연구에서는, 연속형 결과변수가 이분형 지표변수의 문항 응답 확률에 영향을 미치는 구조를 제시하였으며, 직접효과를 다루지 않는 경우 문항 응답 확률의 추정에 편향이 발생할 수 있음을 밝혔다. 이러한 결과는 연속형 공변인의 경우에도 마찬가지로, 문항 응답 확률에 충분한 영향을 미칠 수 있을 것이며, 추정의 왜곡을 발생시킬 수 있기 때문에, 연속형 공변인의 효과에 대한 추가적인 연구가 필요하다. 둘째, 지표변수가 결과변수에 미치는 영향에 있어 정적인 관계만을 탐색하였다. 그러나 우울과 자아존중감과 같이 부적 관계가 있다고 알려진 변수의 경우, 각 변수에 부적인 영향을 줄 가능성도 존재한다. 이는 높은 우울감을 보이는 집단에서 더 낮은 자아존중감을 보고하도록 편향을 일으킬 수 있기에, 부적 관계에 대한 탐색이 추후 연구에서 다뤄질 필요가 있다. 셋째, 잘못된 지표변수가 직접효과를 주거나 받는 상

황을 고려하지 않았다. 실제와는 다른 문항에 대해 직접효과를 설정함으로써 발생할 수 있는 모형에 대한 오추정 오류 또한 추후 연구에서 고려할 필요가 있다. 넷째, 두 연속형 변수 간의 잔여 관계를 확인할 방법에 대한 기준이 명확하게 존재하지 않는다. 제안한 방법 중 두 번째 단계에서 사용되는 BVR은 Latent GOLD 프로그램에서는 명령문 'bivariateresiduals'을 통해 이분형-이분형 변수 간 관계뿐만 아니라 연속형-연속형 변수 간 관계도 확인할 수 있다. 이렇듯 BVR은 두 변수 간의 관계를 탐색할 수 있는 유용한 지표이지만, 대부분의 선행 연구가 이분형 변수 간 관계를 중심으로 이루어졌기 때문에, 연속형 변수에 적용하는 과정에서 해석상의 주의가 필요하며, 관계 유무를 판단하기 위한 기준점을 정하기 위한 추가 연구가 필요하다. 다섯째, 본 연구의 연구 2에서는 분류의 질에 따라 조건을 나누기보다 계층별 지표변수의 평균 차이로 그 조건을 대신하였다. 잠재프로파일모형을 실험 모형으로 다룬 선행 연구가 거의 존재하지 않았기에, 본 연구는 계층별 지표변수 값의 차이를 통해 이를 대체하였다. 추후 연구에서는, 분류의 질을 실험 조건으로 하여 이에 따른 지표변수의 값과 분류의 질이 결과 해석에 미치는 영향을 자세히 살펴볼 수 있을 것으로 예상된다. 마지막으로, 본 연구의 잠재프로파일모형은 잠재계층모형에 비하여 단순한 모형을 설정하였다. 이는 결과변수의 직접효과를 다룬 선행 연구와 모형적으로 비슷한 흐름을 가지면서 직접효과의 방향을 달리하였을 때 나타나는 결과를 보고하고, 잠재프로파일모형의 모의실험 연구에 대한 필요성을 제시하고자 하였기 때문이다. 따라서, 추후 연구에서는 잠재계층의 수를 증가시키는 등의 보다 복잡한 모형을 생성하여 결과를 비교할 수 있을 것으로 예상된다.

본 연구에서는 연구 목적에 대해 검증하기 위해 연구 조건을 최대한 단순화하였다. 구체적으로, 연구 1에서 잠재계층 간 분리도가 높은 이분형 지표변수를 사용하였으며, 연구 2에서는 분산 1의 정규분포를 따르는 연속형 지표변수를 활용하였다. 다수의 선행연구에서 지표변수의 분포 형태가 달라지거나 간접효과가 포함될 때, 모수 추정 안정성 및 잠재계층 분류 정확도 등이 변화할 수 있음이 보고된다 (Lubke & Neale, 2006; Padgett & Tipton, 2020; Asparouhov & Muthén, 2014). 따라서 향후 연구에서는 더욱 일반화된 결과를 얻기 위해서 다범주형 지표변수, 비정규 분포를 갖는 지표변수, 간접효과와 직접효과가 공존하는 구조 등을 포함하는 보다 확장된 연구 설계를 고려할 필요가 있다.

또한, 본 연구는 외부변수와 지표변수 간 조건부 독립성 위배에 따른 추정 편향을 검증하는 데 초점을 두어 직접효과의 존재와 크기를 중심으로 연구 조건을 설정하였다. 일부 조건에서는 직접효과와 잠재계층을 경유한 관계가 함께 존재하도록 설계하였으나, 직접효과만 존재하는 경우, 잠재계층을 경유한 관계만 존재하는 경우, 그리고 두 관계가 공존하는 경우를 모든 연구 조건에서 체계적으로 비교하지는 못하였다. 향후 연구에서는 두 관계의 존재 여부와 효과 크기 및 방향을 독립적으로 조작하여, 관계 구조에 따른 각 접근법의 추정 성능을 종합적으로 검증할 필요가 있다.

참고문헌

이상우. (2016). 리커트 척도의 변화와 강의평가결과. 상품학연구, 34(5), 33-40.

Agresti, A. (2013). Categorical data analysis. *John Wiley & Sons*.

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 329-341. https://doi.org/10.1080/10705511.2014.915181

Bakk, Z., Di Mari, R., Oser, J., & Kuha, J. (2022). Two-stage multilevel latent class analysis with covariates in the presence of direct effects. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(2), 267-277. https://doi.org/10.1080/10705511.2021.1980882

Bakk, Z., & Kuha, J. (2018). Two-step estimation of models between latent classes and external variables. *Psychometrika*, *83*(4), 871-892. https://doi.org/10.1007/s11336-017-9592-7

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating Latent Class Assignments to External Variables: Standard Errors for Correct Inference. *Political Analysis*, *22*(4), 520-540. https://doi.org/10.1093/pan/mpu003

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, *43*(1), 272-311. https://doi.org/10.1177/0081175012470644

Bakk, Z., & Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: a Multidisciplinary Journal*, *23*(1), 20-31. https://doi.org/10.1080/10705511.2014.955104

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*(440), 1375-1386. https://doi.org/10.1080/01621459.1997.10473658

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*(1), 3-27. https://doi.org/10.1093/pan/mph001

Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, *79*(388), 762-771. https://doi.org/10.1080/01621459.1984.10477093

Collins, L. M., & Lanza, S. T. (2009). Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences. *John Wiley & Sons*.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*(401), 173-178. https://doi.org/10.1080/01621459.1988.10478584

Di Mari, R., Bakk, Z., & Punzo, A. (2020). A random-covariate approach for distal outcome prediction with latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(3), 351-368. https://doi.org/10.1080/10705511.2019.1648186

Harwell, M. (2019). A strategy for using bias and

RMSE as outcomes in monte carlo studies in statistics. *Journal of Modern Applied Statistical Methods*, *17*(2) https://doi.org/10.22237/jmasm/1551907966

Janssen, J. H., Van Laar, S., De Rooij, M. J., Kuha, J., & Bakk, Z. (2019). The detection and modeling of direct effects in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(2), 280-290. https://doi.org/10.1080/10705511.2018.1541745

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, *7*(4), 396-403. https://doi.org/10.9734/BJAST/2015/14975

Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. S*tructural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 671-694. https://doi.org/10.1080/10705510701575602

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 1-26. https://doi.org/10.1080/10705511.2013.742377

Lubke, G., & Neale, M. C. (2006). Distinguishing Between Latent Classes and Continuous Factors: Resolution by Maximum Likelihood? Multivariate Behavioral Research, 41(4), 499-532. https://doi.org/10.1207/s15327906mbr4104_4

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 180-197. https://doi.org/10.1080/10705511.2016.1254049

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599-620. https://doi.org/10.1207/S15328007SEM0904_8

Muthén, L., & Muthén, B. (2017). Mplus user's guide (8th ed.). Muthén & Muthén.

Oberski, D. L., van Kollenburg, G. H., & Vermunt, J. K. (2013). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, 7, 267-279. https://doi.org/10.1007/s11634-013-0146-2

Oberski, D. L., & Vermunt, J. K. (2018). The expected parameter change (EPC) for local dependence assessment in binary data latent class models. *arXiv Preprint*. https://doi.org/10.48550/arxiv.1801.02400

Padgett, R. N., & Tipton, R. J. (2020). Identifying latent classes with ordered categorical indicators. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2009.07345

R Core Team. (2025). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

Vermunt, J. K., & Magidson, J. (2021). How to perform three-step latent class analysis in the presence of measurement non-invariance or

differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(3), 356-364. https://doi.org/10.1080/10705511.2020.1818084

Vermunt, J. K., & Magidson, J. (2005). Technical guide for Latent GOLD 4.0: Basic and advanced. Belmont Massachusetts: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2020). Upgrade manual for Latent GOLD 6.0. Statistical Innovations Inc.

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education*, *80*(1), 26-44. https://doi.org/10.1080/00220973.2010.531299

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*(6), 913-934. https://doi.org/10.1177/0013164413495237

Xue, Q. L., & Bandeen-Roche, K. (2002). Combining complete multivariate outcomes with incomplete covariate information: a latent class approach. *Biometrics*, *58*(1), 110-120. https://doi.org/10.1111/j.0006-341X.2002.00110.x

1차원고접수 : 2025. 10. 29
2차원고접수 : 2026. 07. 27
최종게재결정 : 2026. 08. 24

한국심리학회지: 일반
Korean Journal of Psychology: General
2026, Vol. 45, No. 4, 687-716
http://dx.doi.org/10.22257/kjp.2026.9.45.4.687

The Impact of Omitted Direct Effects Between External Variables and Indicators on Estimating External Variable Effects in Latent Class and Latent Profile Analysis: A Comparison of Two-Step and Three-Step Approaches

Chanhui Lee Kyungpook National University
Hangyu Kim Kyungpook National University
Jungkyu Park† Kyungpook National University

This study investigated estimation bias arising from the omission of direct effects between external variables and indicators in latent class analysis (LCA) and latent profile analysis (LPA). Two studies were conducted, each addressing a different type of external variable and mixture model. Specifically, Study 1 examined situations in which covariates directly affected indicators in LCA, whereas Study 2 examined situations in which indicators directly affected an outcome variable in LPA. In both studies, parameters were estimated using two-step and three-step approaches, and the results were compared across models with and without direct effects. In Study 1, the two-step approach yielded lower overall bias when covariate direct effects were included in the model than when they were omitted. Bias increased with the magnitude of the covariate direct effects and the number of omitted direct effects. Similarly, in Study 2, models that included direct effects of indicators on the outcome variable exhibited substantially lower mean relative bias and Type I error rates than models that omitted these effects. These findings suggest that explicitly modeling direct effects between external variables and indicators can improve the accuracy of parameter estimation.

Key words : latent class analysis, latent profile analysis, external variable, direct effect, two-step approach