

## 비내재관계 모형 선택을 위한 적합도 지수, $\Delta BIC$ , Vuong 검증의 비교·평가\*

김 인 혜      김 미 립<sup>†</sup>      이 순 목  
성균관대학교      성신여자대학교      성균관대학교

본 연구의 목적은 구조방정식모형을 이용한 연구에서 비내재관계 모형 선택 시 활용 가능한 다양한 지수들의 특성과 한계를 종합적으로 검토하는 데 있다. 이를 위해 적합도 지수(CFI, TLI, RMSEA, SRMR), 정보기반지수( $\Delta BIC$ ), 그리고 Vuong 검증을 대상으로 몬테카를로 시뮬레이션을 실시하였다. 연구 설계는 요인계수 차이, 오차상관 크기, 요인상관 크기, 표본크기를 포함한  $3 \times 3 \times 3 \times 4$  요인설계로 총 108개 조건을 구성하고, 각 조건에서 1,000개의 자료를 생성하였다. 주요 결과를 보면, CFI와 TLI는 구조적 차이에 예민하게 반응하여 표본크기가 커질수록 안정적인 모형 변별력을 보였다. 반면 RMSEA는 표본크기와 자유도에 영향을 크게 받아 엄격한 기준으로 작동하는 경향을 보였으며, SRMR은 요인상관이 높은 조건에서 두 모형을 명확히 구분하지 못하는 한계가 나타났다.  $\Delta BIC$ 는 모형 간 차이의 크기를 수치적으로 제시할 수 있었지만, 자유도가 동일한 상황에서는 해석적 제약이 확인되었다. 마지막으로 Vuong 검증은  $\Delta BIC$ 가 제공하지 못하는 통계적 유의성 정보를 보완함으로써, 모형 선택의 판단 근거를 강화하였다. 본 연구는 다양한 조건에서 지수별 성능을 비교·평가함으로써, 연구자가 비내재관계 모형 선택 시 지수의 특성과 한계를 균형 있게 고려할 수 있도록 비교 전략에 대한 제언을 제공한다. 이를 통해 구조방정식모형을 이용하는 연구자가 지수 간 상호보완적 해석에 기반하여 보다 타당한 모형 선택을 수행할 수 있도록 기여하고자 한다.

주요어 : 적합도 지수,  $\Delta BIC$ , Vuong 검증, 비내재관계 모형 비교, 시뮬레이션 연구

\* 본 연구는 김인혜의 박사학위논문 연구의 일부임.

† 교신저자: 김미림, 성신여자대학교 심리학과, 성신여자대학교 인간중심 AI 연구소, 서울특별시 성북구 보문로 34다길 2, Tel: 02-920-7069, E-mail: mirimkim@sungshin.ac.kr



Copyright © 2026, The Korean Psychological Association. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial Licenses(<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.

## 서론

구조방정식모형은 변수 간의 복잡한 관계를 비교적 자유롭게 설정하고 검증할 수 있기 때문에, 연구자는 이론에 대한 상이한 가설을 구현한 경쟁모형들을 동시에 검토하기도 한다(Levy & Hancock, 2007; Levy & Hancock, 2011; Merkle et al., 2016). 이는 경쟁모형방식(competition model approach)을 활용한 최종 모형의 탐색과 해석으로 볼 수 있으며(MacCallum, 1986; Kline, 2016), 하나의 자료에 대해 여러가지 경쟁모형을 설정하고 이들 간의 상대적 적합도를 비교하여 최적의 모형을 선택하는 분석적 접근이다.

구조방정식모형의 비교는 크게 내재관계 모형의 비교와 비내재관계 모형의 비교로 나눌 수 있다. 내재관계는 하나의 모형이 다른 모형에 완전히 포함된 관계이며, 자유도 차이에 기반한 카이제곱 차이( $\Delta\chi^2$ ) 검증으로 모형 간 우열을 평가할 수 있다(Jöreskog, & Sörbom, 1981). 그러나 실제 연구에서 비교하는 경쟁모형들은 한 모형이 다른 모형을 완전히 포함할 수 없는 비내재관계에 있는 경우가 많다. 이 경우  $\Delta\chi^2$ 는 활용할 수 없기 때문에 대안적으로 적합도 지수(fit indices; 예: CFI, TLI, RMSEA, SRMR)나 정보기반지수(information criteria; 예: AIC, BIC) 등을 활용하여 모형을 비교한다(Hu & Bentler, 1999; Burnham & Anderson, 2004).

비내재관계에 있는 경쟁모형의 요인구조 혹은 경로의 차이는 가설의 차이를 나타내므로, 경쟁모형의 적합도 차이는 가설모형이 자료를 얼마나 잘 설명하는지를 비교하는 근거로 활용할 수 있어야 한다. 그러나 보다 높은 모형 적합도는 모형의 반복 수정을 통한 우연의 자

산화(capitalization on chance) 때문일 수도 있으며, 이 경우 최종 모형의 통계적 적합도는 우수할지라도 이론적 타당성은 부족할 수 있다(MacCallum et al., 1992).

또한 모형 선택이 적합도 지수에 의존적일 경우, 모형이 특정 자료에만 지나치게 적합하여 일반화 가능성을 저해하는 과적합(overfitting) 현상이 나타날 수 있다. 과적합은 모형이 복잡할수록 빈번히 발생하는 경향이 있으며(Browne & Cudeck, 1993), 요인분석 모형을 예시로 든다면 불필요한 오차상관의 추가나 교차요인계수(cross-loading)의 허용 등 모형 수정이 추가될수록 모형의 통계적 적합도는 향상될 수 있다. 그 결과 최종모형은 적합도 지수는 높지만 이론적으로는 타당하지 않은 요인구조를 나타낼 수 있다(Marsh et al., 2004). 적합도 지수의 성능을 살펴보기 위해서는 모형 간 복잡도 차이를 통제하는 연구가 필요함을 알 수 있다. 모형 간 복잡도 차이로 인해 발생할 수 있는 적합도 지수의 차이를 최소화한다면, 관찰된 지수의 차이는 모형의 구조적 차이에서 나타나는 영향을 직접적으로 반영할 가능성이 높아진다(Hancock & Mueller, 2011; Marsh et al., 2004). 예를 들어, 문항  $i$ 가 각각  $F_1$  이나  $F_2$ , 단일 요인에만 회귀하는 두 경쟁모형은 한 개의 요인에만 부하한다는 면에서 복잡도가 같지만, 상이한 의미를 갖는 요인구조가 될 것이다. 각 모형에 대한 통계적 지수의 차이는 구조방정식모형의 구조에 대한 민감성을 드러낸다고 예상할 수 있다.

위와 같은 학문적 의의를 바탕으로 본 연구는 적합도 지수에 해당하는 CFI, TLI, RMSEA, SRMR, 그리고 정보기반지수에 해당하는 BIC에 기반하여 경쟁모형의 구조적 차이를 탐색한다. 구체적으로 모형 간 자유도를 동일하게

유지한 상태에서 교차요인계수의 귀속 방식이나 오차상관의 설정 여부와 같은 구조방정식 모형의 구조적 차이를 다양화하고, 비내재관계 경쟁모형의 비교를 실시함으로써 구조적 차이를 탐색하는 통계적 지수의 성능을 살펴보고자 한다. 추가로, 경쟁모형 간의  $\Delta BIC$ 를 산출하여 이에 대한 Vuong 검증도 실시하여 구조방정식모형에 대한 Vuong 검증의 실용성도 함께 살펴본다. 현재 연구는 모형의 복잡도를 통제하여 진행하므로, 다양한 통계적 지수가 구조방정식모형의 구조적 차이에 대해 갖는 민감도를 보다 엄밀하게 평가할 수 있을 것으로 기대한다(유소현 & 김수영, 2024; Hu & Bentler, 1999; McNeish et al., 2018; Xia & Yang, 2019).

## 이론적 배경

MacCallum(1986)은 구조방정식모형 간 비교를 타당한 이론모형 간의 비교로, Burnham과 Anderson(2004)은 모형 선택을 ‘최적의 모형(best model)’이 아닌 ‘상대적으로 괜찮은 모형(relatively better model)’을 찾는 것으로 보았다. 이처럼 모형 간의 비교는 연구자가 실제로는 참모형(true model)을 식별할 수 없다는 점에서, 모형 간의 상대적 평가라고 볼 수 있다.

본 연구에서 주목하고 있는 비내재관계의 모형을 비교하고 선택하는 경우, 연구자는 적합도 지수의 상대적 우세나(예: CFI · TLI 값이 더 크거나 RMSEA · SRMR 값이 더 작은 모형을 선택) 정보기반지수의 최소화 원리에(AIC · BIC가 더 작은 모형을 선택) 따라 모형을 판단한다(Kline, 2016; Burnham & Anderson, 2004). 다만 이러한 판단 방식은 지수 간 해석 기준

이 다르고, 차이의 크기에 대한 통계적 유의성을 직접 제공하지 않는다는 한계를 지니므로 Vuong 검증을 통해 통계적 유의성을 확인하기도 한다. 가장 최근에는 Merkle 등(2016)이 구조방정식모형에 적용가능한 Vuong 검증 절차를 제안하였다. Vuong 검증은 적합도 지수 및 정보기반지수와 병행하여 경쟁모형의 비교 결과를 구조방정식모형 맥락에서도 보조적 도구로 활용할 수 있게 되었다.

## 적합도 지수

적합도 지수는 모형과 자료의 전반적 일치 정도를 수치화하여 연구자가 모형의 수용 가능성을 판단할 수 있도록 근거를 제공한다(Hu & Bentler, 1999; Kline, 2016). 다양한 적합도 지수 중, 본 연구에서는 구조방정식모형을 이용한 연구에서 가장 빈번하게 보고되면서, 서로 다른 평가 논리를 대표하는 지수들에 주목하여 CFI와 TLI, 그리고 절대적 오차 수준을 반영하는 RMSEA와 SRMR을 분석 대상으로 선정하였다(김미림 & 김예지, 2024).

### CFI

CFI(Comparative Fit Index: Bentler, 1990)는 모든 측정변수 간 공분산을 0으로 가정하는 기저모형을 비교 기준으로 하여 기저모형보다 연구모형이 공분산 구조를 얼마나 더 설명하는지를 나타내는 지수이다. 값이 클수록 모형의 적합도가 우수함을 의미하며,  $CFI \geq .90$ 은 수용 가능,  $CFI \geq .95$ 는 우수한 적합도로 해석한다(Hu & Bentler, 1999). CFI는 RMSEA와 같이 자유도를 분모에 두어 모형의 복잡도에 직접 벌칙(penalty)을 부과하는 지수가 아니며, 상대적으로 표본크기에 비교적 덜 민감한 지수로

알려져 있다(Hu & Bentler, 1999).

그러나 CFI는 모형의 구조적 제약에 민감하게 반응하여 자유모수가 추가된 덜 제약적인 모형에 유리하게 작동하는 경향이 있으며, 이론적 근거 없는 자유모수의 허용은 모형의 적합도를 과대평가하고, 반대로 필요한 구조적 요소를 과도하게 제약할 경우 적합도를 과소평가할 가능성이 있다(Hu & Bentler, 1999; MacCallum et al., 1992).

중요한 점은 이러한 반응이 자유도의 크기 자체에 의해 결정되는 것이 아니라, 교차요인 계수와 같이 특정 모수에 대한 제약이 부과되는 위치와 그 제약이 자료의 공분산 구조와 충돌하는 정도에 따라, 동일한 자유도 조건에서도 달라질 수 있다는 것이다. 즉, 두 모형이 동일한 자유도를 갖더라도 교차요인 계수가 어느 요인에 귀속되는지에 따라 모형과 자료 간 불일치의 크기( $\chi^2$ )는 달라질 수 있으며, CFI는 이러한  $\chi^2$  차이를 기저모형과의 상대적 비교를 통해 반영한다.

#### TLI

TLI(Tucker - Lewis Index: Tucker & Lewis, 1973) 또한 CFI와 유사하게 값이 클수록 모형의 적합도가 높은 것을 나타내며,  $TLI \geq .90$ 은 수용 가능,  $TLI \geq .95$ 는 우수한 적합도로 해석한다. 그러나 TLI는 CFI와 달리 단순한 적합도의 크기보다는 자유도를 고려했을 때 적합도가 얼마나 개선되었는지를 함께 반영하는 지수이다.

즉, 구조적 제약을 완화하여 적합도가 증가하더라도 그 증가폭이 자유도 감소를 충분히 상쇄하지 못할 경우, TLI는 해당 모형을 낮게 평가하는 경향을 보인다. 이러한 특성으로 인해 TLI는 자유도를 감소시키는 불필요한 구조

적 완화(예: 교차요인 계수나 오차상관의 허용 등)가 이루어질 경우 오히려 값이 낮아질 수 있다(Hu & Bentler, 1999). 또한 표본크기가 작은 조건에서는 값이 불안정하거나 음수가 나올 가능성도 보고되고 있다(Marsh et al., 2004).

본 연구와 같이 경쟁모형의 복잡도 차이를 통제된 조건에서도, TLI는 모형의 구조적 제약 방식에 따라 서로 다른 반응을 보일 수 있으며, 이에 따라 경쟁모형 간 구조적 차이를 CFI보다 엄격하게 구분하는 경향이 나타날 수 있다.

#### RMSEA

RMSEA(Root Mean Square Error of Approximation: Steiger, 1990)는 모형이 모집단 공분산 구조를 얼마나 근사하는지를 평가하는 절대적 기준의 적합도 지수로, 값이 낮을수록 모형이 자료에 더 잘 부합함을 의미하며,  $RMSEA \leq .08$ 은 수용 가능,  $RMSEA \leq .05$ 는 우수한 적합도로 해석한다(Browne & Cudeck, 1993).

RMSEA는 연구모형과 자료 간의 불일치를 자유도와 표본크기를 고려하여 산출하는 지수이다. 이러한 계산 특성으로 인해 누락된 모수나 잘못된 구조적 차이에 민감하게 반응하는 경향이 있는 것으로 알려져 있다. 예를 들어, 실제 존재하는 오차상관을 모형에 반영하지 않을 경우 잔차(residual) 구조의 차이는 모형 적합도를 낮게 추정할 수 있으며(Brown, 2015; Kline, 2016), 이러한 차이는  $\chi^2$  및 RMSEA와 같은 적합도 지수에 영향을 미친다(Browne & Cudeck, 1993; MacCallum et al., 1996). 이러한 특성에 따라 RMSEA는 잔차 구조 전반에서 발생하는 불일치에 민감하게 반응한다.

따라서 본 연구의 다양한 조건에서 RMSEA

가 모형이 자료의 공분산 구조를 전체적으로 얼마나 잘 설명하는지 비교적 엄격하게 반영하는 지수로 작동할 수 있다.

Kenny 등(2015)은 RMSEA가 표본크기 뿐 아니라 모형의 자유도와 상호작용하여 영향을 받는다고 보고하였다. 특히 자유도가 낮고 표본이 작은 조건에서는 RMSEA가 양호한 적합 정도에서도 과대평가되는 경향이 있어, 모든 조건에서 RMSEA를 절대적 적합도 판단 기준으로 해석하는 데 주의가 필요하다고 하였다.

#### SRMR

SRMR(Standardized Root Mean Square Residual: Bentler, 1995)은 실제 자료의 상관행렬과 연구 모형 기반의 상관행렬 간 차이를 평균한 지수로, 값이 작을수록 적합도가 우수하며 일반적으로  $SRMR \leq .08$ 이면 수용 가능한 적합도로 해석한다(Hu & Bentler, 1999).

SRMR은 측정변수와 요인 간의 관계를 나타내는 교차요인계수보다는 측정변수 간 관계를 나타내는 오차상관에 상대적으로 민감하게 반응한다(Kline, 2016).

한편 요인상관이 큰 조건(요인간 상관이 .7 이상으로 높게 설정된 상황)에서는 측정변수 간 상관이 전반적으로 커지기 때문에, 구조적 차이가 크지 않은 경우 SRMR 값이 상대적으로 작게 산출될 가능성도 제기되어 왔다(Kline, 2016; Marsh et al., 2004). 이는 SRMR이 상관행렬의 잔차를 기반으로 산출되는 지수이기 때문이며, 그 결과 요인상관보다는 오차상관 수준에서 발생하는 불일치에 보다 직접적으로 반응하는 특성과 관련이 있다.

본 연구에서는 이러한 적합도 지수의 특성을 고려하여, 오차상관 및 요인상관의 크기 조건에 따라 SRMR이 경쟁모형 간 차이를 어

떻게 반영하는지를 다른 적합도 지수들과 함께 비교하고 해석하였다.

요약하면, 적합도 지수들은 각각 다른 정보에 기반하여 모형의 적합 정도를 평가하므로, 모형간 복잡도 차이가 통제된 조건에서도 동일한 판단을 제공하지 않을 가능성이 있다. 그러나 이러한 반응 차이가 실제로 어떤 조건과 방향에서 나타나는지 체계적으로 검토된 바가 거의 없다.

구체적으로 CFI와 TLI는 기저모형 대비 연구모형의 적합도 개선 정도에 민감하게 반응하는 반면, RMSEA는 모형과 자료의 전반적 불일치의 크기를 표본크기와 자유도를 고려하여 평가하고, SRMR은 측정변수간 상관 구조가 잔차 수준에서 얼마나 재현되었는지에 민감하게 반응한다.

주의할 점은, Hu와 Bentler(1999)가  $CFI \geq .95$ ,  $RMSEA \leq .08$ 과 같은 기준값(cutoff criteria)을 제시한 이후, 대다수의 연구자들이 해당 기준값을 모형 적합성을 판단하기 위한 표준처럼 사용하고 있다는 것이다. 이에 대하여 최근 연구들은 이러한 고정적 기준값에 대해 비판적 시각을 제기하고 있다.

예를 들어, Jackson 등(2009)은 동일한 기준값을 이용하여 적합도 지수를 판단하는 것이 모든 연구 맥락에서 타당한 것은 아니며, 기준값은 모형 복잡성이나 표본 특성에 따라 달라질 수 있다고 지적하였다. McNeish(2022, 2023) 또한 전통적 적합도 지수 기준값이 표본크기, 모형 구조나 모형 유형에 따라 과도하게 엄격하거나 관대하게 작동할 수 있음을 보고하며 고정된 기준값을 일률적으로 적용하는 한계를 지적하였다.

이에 따라 본 연구는 단일 적합도 지수의 기준 충족 여부에 근거한 판단을 반복하는데

그치지 않고, 모형 간 복잡도 차이를 통제된 조건에서 CFI, TLI, RMSEA, SRMR과 같은 적합도 지수들이 구조적 차이의 형태와 표본크기에 따라 어떻게 다른 패턴을 보이는지 검토하고자 한다. 아울러 다양한 적합도 지수들의 판단이 서로 엇갈릴 수 있는 상황에서, 정보기반지수인 BIC와 Vuong 검증이 경쟁모형 간 상대적 우열을 해석하는데 어떤 추가적 정보를 제공할 수 있는지 함께 살펴보려고 한다. 이를 통해 비내재관계 모형 비교에서 보다 정교한 판단의 틀을 제시하고자 한다. 정보기반지수인 BIC와 Vuong 검증에 대해서는 다음에 제시하였다.

#### 정보기반지수: BIC와 $\Delta BIC$

정보기반지수는 모형과 자료의 합치 정도를 반영하지만, 모형의 절대적 적합 여부보다는 비내재관계의 대안모형 간 상대적 비교에 초점을 둔다는 점에서 적합도 지수와 구분된다.

정보기반지수로는 AIC, BIC(Bayesian Information Criterion: Schwarz, 1978), CAIC 등이 있으며, 이들 지수는 연구자가 가정한 자료생성구조가 관측된 자료에 얼마나 부합하는지를 평가하는 것과 동시에 모형의 복잡도에 대한 벌칙을 고려한다. 따라서 모형이 자료를 잘 설명할수록, 그리고 모형이 간명할수록 작은 값을 산출하는 경향이 있다.

본 연구에서는 BIC를 중심으로 정보기반지수의 수행을 검토하였다. BIC는 다른 정보기반지수에 비해 표본크기에 비례하는 복잡도 벌칙을 포함하므로 표본크기가 증가할수록 미세한 구조적 차이에 보다 민감하게 반응하는 특성을 지니기 때문이다. 이러한 특성은 모형 간 BIC 차이값( $\Delta BIC$ )을 이용하여 모형의 상

대적 우세를 해석하는 데 적합하다(Nylund et al., 2007; Raftery, 1995). Raftery(1995)는  $\Delta BIC$  크기에 따른 해석 기준<sup>1)</sup>을 제시하여,  $\Delta BIC$ 를 효과크기와 유사하게 해석할 수 있도록 하였다. 그러나 이러한 해석 기준은 표본크기와 모형 규모에 민감하게 달라질 수 있다는 점에서, 표본크기와 모형의 규모가 다양한 구조방정식모형에서 활용하기에는 한계를 지녔다고 볼 수 있다.

또한  $\Delta\chi^2$ 와는 다르게  $\Delta BIC$ 에 대한 통계적 유의성은 직접적으로 판단하는 것이 어렵다는 제한점을 지닌다(Hancock & Mueller, 2011). 이에 대하여 본 연구는 Vuong 검증(Merkle et al., 2016; Vuong, 1989)을 활용하여 구조방정식모형 기반  $\Delta BIC$ 의 통계적 유의성을 검토하고자 한다.

#### Vuong 검증

Vuong(1989)은 두 경쟁모형의 관측치별 로그우도 차이를 Kullback-Leibler (K-L) 정보량에 기반하여 비교함으로써, 각 모형이 모집단 분포에 상대적으로 얼마나 근접한지를 통계적으로 평가하는 검증 방법을 제안하였다. K-L 정보량은 동일한 자료에 대해 각 모형이 유도하는 확률분포(추정된 모수에 기반한 결합확률 밀도함수) 간의 차이를 측정하는 정보 측정치로, Vuong 검증은 이를 표본 자료에 기반하여 추정함으로써 모형 간 적합도 차이(예:  $\Delta BIC$ )

1) Raftery(1995)가 제시한  $\Delta BIC$  해석 기준은 표본크기와 모형의 규모에 따라 민감하게 달라질 수 있으며, 본 연구에서는 해당 기준을 모형 선택의 규범적 판단 규칙으로 적용하지 않았다. 표에 제시된 기준은  $\Delta BIC$  값의 상대적 크기를 이해하기 위한 참고 정보로만 활용하였다.

가 우연에 의한 것인지 여부를 검증한다.

Vuong 검증을 적용하기 위해서는 두 경쟁모형이 통계적으로 구분 가능한지의 여부를 확인하는 절차가 필요하다. 이를 위해 분산 검증(variance test)으로 두 모형의 관측치 별 로그우도 차이의 분산이 0과 유의하게 다른지를 검증하여 두 모형이 사실상 동일한 적합도를 갖는지 또는 통계적으로 구분이 가능한 대안 모형인지를 판단한다. 분산 검증 결과가 유의하지 않은 경우 두 모형을 통계적으로 구분할 수 없는 것으로 간주하며, 이 경우 후속  $z$  검증에 따른 상대적 적합도 차이 판단은 의미있는 정보를 제공하지 않는다(Vuong, 1989). 따라서 구조방정식모형에 대하여 Vuong 검증을 적용할 때는 로그우도 차이 기반  $z$  검증 결과만을 단독으로 해석하기보다는 분산 검증 결과를 함께 보고할 필요가 있으며, 이에 기반하여 경쟁모형이 구분 가능한 상황에서만 상대적 적합도 차이에 대한 통계적 유의성을 해석할 필요가 있다.

이와 같은 선행연구에 기반하여 본 연구는  $\Delta BIC$ 의 상대적 크기를 해석하는 데 있어 Vuong 검증이 중요한 보완적 근거로 활용할 수 있다고 판단하였다.

## 연구 문제

모형 비교와 평가에는 서로 다른 계산 논리에 기반한 다양한 통계적 지수를 활용한다. 대표적으로 적합도 지수는 연구모형이 자료의 공분산 구조를 얼마나 잘 재현하는지를 평가하는 데 초점을 두는 반면, 정보기반지수는 모형의 적합도와 복잡도를 동시에 고려하여 상대적 우수성을 판단한다(Schwarz, 1978; Burnham & Anderson, 2002). 또한 관측치별 로

그우도 차이에 기반한 Vuong(1989) 검증은 비내재관계 모형 간 상대적 적합성을 통계적으로 판별하는 절차를 제공한다. 이처럼 통계적 지수들은 서로 다른 정보와 기준에 기반하여 모형을 평가하므로, 각 지수가 구조적 차이에 대해 어떠한 방식으로 반응하는지를 비교하고 검토할 필요가 있다.

통계적 지수는 모형이 포함하는 모수의 수와 자유도, 즉 구조적 복잡도에 의해 영향을 받는 것으로 알려져 있으며, 동일한 수준의 구조적 차이가 존재하더라도 복잡한 모형일수록 상대적으로 더 양호한 평가를 받을 가능성이 있다(Browne & Cudeck, 1993; Kline, 2016; Marsh et al., 2004). 특히 적합도 지수는 자유도와 모형 규모에 민감하게 반응하여, 복잡도가 서로 다른 모형을 비교할 경우 구조적 우수성보다는 복잡도 차이를 반영할 위험이 있다(Hancock & Mueller, 2011; MacCallum et al., 1992; Marsh et al., 2004).

한편 정보기반지수 역시 모형의 복잡도를 벌칙 형태로 반영하여 모형을 평가한다는 점에서 복잡도와 밀접한 관련을 갖는다(Schwarz, 1978; Burnham & Anderson, 2002). 따라서 복잡도가 통제되지 않은 상태에서 모형을 비교할 경우, 구조적 차이에 대한 각 통계적 지수의 반응을 명확히 해석하기 어려울 수 있다(Chen, 2007; MacCallum et al., 1996; McNeish et al., 2018). 이는 모형의 구조적 차이에 대한 각 적합도 지수의 기능을 탐색하기 위해서는 모형 간 복잡도 차이를 통제한 연구가 필요함을 시사한다.

이에 본 연구는 모형 간 복잡도 차이를 통제된 조건에서 비내재관계 경쟁모형을 비교함으로써, 적합도 지수와 정보기반지수의 상대적 민감도와 판별 특성을 평가하고자 한다.

비교 경쟁모형이 갖는 요인구조와 세부 연구 문제는 다음과 같다.

첫째, 모형의 구조적 차이를 나타내는 요인 계수 차이, 오차상관, 요인상관과 표본크기를 다르게 조작하여 시뮬레이션 연구를 실시함으로써 각 조건에 따른 적합도 지수(CFI, TLI, RMSEA, SRMR)의 변화를 살펴보고자 한다. 연구 조건의 다양화에 따른 각 지수값의 변화 양상과 실증연구에 많이 활용하는 적정 기준 값(CFI · TLI > .90, RMSEA · SRMR < .08)에 대한 충족률을 검토함으로써 지수별로 민감하게 반응하는 모형의 조건에 대해 탐색하고자 한다.

둘째, 조건별  $\Delta BIC$  값의 분포를 분석하여, 정보기반지수가 뚜렷하게 탐지할 수 있는 경쟁모형 간 차이가 무엇인지 탐색하고자 한다. 또한 해당 결과를 다른 적합도 지수의 반응 양상과 비교하여  $\Delta BIC$ 가 제공하는 정보의 성격을 평가하고자 한다.

셋째,  $\Delta BIC$ 의 통계적 유의성을 평가하기 위한 보완적 도구로 Vuong 검증을 활용하여 구조방정식모형 연구에서의 Vuong 검증 활용과 해석에 대한 정보를 제공하고자 한다.

## 연구 방법

### 연구 모형 설정

본 연구는 Hu와 Bentler(1998, 1999)의 접근을 참고하여, 확인적 요인분석 모형을 기반으로 자료생성모형과 두 개의 경쟁모형을 설정하였다. 자료생성모형은 모집단 구조를 가정하는 기준 모형이며, 경쟁모형은 자료생성모형을 바탕으로 설계된 두 개의 대안적 이론모형이다. 두 경쟁모형은 연구자가 실제 분석에

서 고려해 봄 직한 모형이 되도록 구성하였다.

### 자료생성모형

그림 1은 본 연구에서 모집단 모형으로 가정한 자료생성모형이다. 해당 모형은 2개의 요인과 8개의 측정변수로 이루어진 확인적 요인분석 모형으로, 각 요인은 각각 4개와 5개의 측정변수와 연결되어 있다. 본 모형에서는 측정변수 X4에 두 요인이 모두 부하하도록 교차요인계수를 설정하여, 하나의 문항이 두 요인과 동시에 관련성을 가질 수 있는 현실적인 측정 상황을 반영하였다(Brown, 2015). 또한 X4와 근접한 측정변수 X5와의 측정오차상관( $e4 \leftrightarrow e5$ )을 설정하여 실제 자료 분석에서 자주 발생하는 측정오차 간의 상관 가능성을 모형에 반영하였다. 본 연구에서 모형의 구조적 차이는 자료생성모형에 존재하는 교차요인계수를 경쟁모형에서 제거함으로써 발생하는 모형 간 부분적 구조적 불일치를 의미한다. 이는 이론적 간명성을 이유로 교차요인계수를 배제하는 과정에서 빈번히 발생할 수 있는 현실적인 구조적 불일치를 반영한 것이다(Fabrigar et al., 1999). 이와 달리, 오차상관은 모든 모형에서 동일하게 포함하여 교차요인계수 제거에 따른 구조적 차이만을 비교할 수 있도록 설계하였다. 이에 따라 본 연구의 자

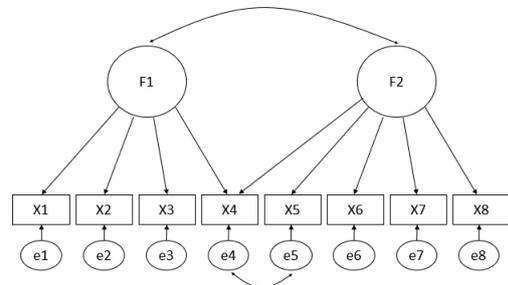


그림 1. 연구의 자료생성모형

료생성모형은 교차요인계수( $X_4$ )와 오차상관( $e_4 \leftrightarrow e_5$ )을 모두 포함함으로써, 요인분석 연구에서 빈번히 마주하는 자료 구조를 구현하고 있다(김미림 & 임예지, 2024; Brown, 2015).

### 경쟁모형

분석에 사용된 2개의 경쟁모형은 그림 2와 같다. 경쟁모형은 자료생성모형과 달리, 측정변수  $X_4$ 의 교차요인계수를 추정하지 않도록 설정하였다. 경쟁모형 1은  $X_4$ 가 요인  $F_1$ 에만 부하하는 모형이며, 경쟁모형 2는  $X_4$ 가 요인  $F_2$ 에만 부하하는 모형이다. 측정변수의 수, 요인상관, 오차상관은 자료생성모형과 동일하게 유지하였다. 이러한 설정을 통해 경쟁모형들은 복잡도 차이를 동일하게 유지하면서도, 교차요인계수 제거에 따른 해석적 차이가 적합

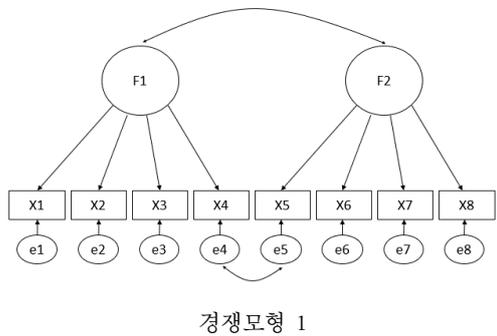
도 지수에 어떻게 반영되는지를 비교할 수 있도록 고안하였다. 이는 연구자들이 교차요인계수를 허용하지 않는 간명한 모형을 선호하는 실제 분석 관행을 반영한 것이다(Fabrigar et al., 1999; Hair et al., 2019; Henson & Roberts, 2006; Kline, 1994).

### 독립변수

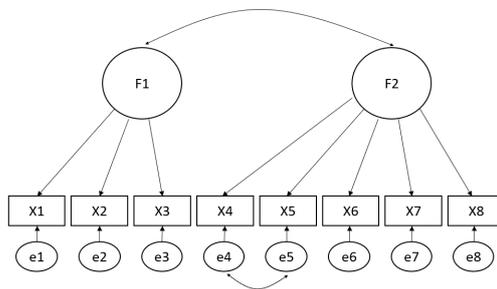
그림 2에서 살펴볼 수 있듯이 두 경쟁모형은  $x_4$ 에 부하하고 있는 요인이 다른 것이 차이점이다. 이에 기반하여 본 연구는  $X_4$ 에 부하하는 두 경쟁모형의 요인계수 차이, 오차상관 크기, 요인상관 크기, 그리고 표본크기를 독립변수로 설정하였다(표 1 참조). 해당 독립변수는 요인분석 모형을 평가하고 선택하는데 영향을 미치는 주요 변수들이라는 점에서 선택하였다(Brown, 2015).

### 요인계수 차이

Bollen(1989)은 요인계수 크기가 모형의 해석과 적합도 평가에서 중요한 의미를 가지며, Brown(2015)은 교차요인계수를 포함한 확인적 요인분석 모형에서 특정 문항이 어느 요인에 더 크게 기여하는지가 모형의 설명력과 적합도 해석에 영향을 줄 수 있음을 지적하였다.



경쟁모형 1



경쟁모형 2

그림 2. 연구의 경쟁모형들

표 1. 모수의 정도에 따른 실험조건의 기준

내용	상	중	하
요인계수 차이	.4	.2	0 <sup>a</sup>
오차상관 크기	.3	.1	.001
요인상관 크기	.7	.5	.3
표본크기	1000	500, 300	100

<sup>a</sup> 차이없음.

따라서 본 연구에서는 요인계수 차이를 핵심 연구 조건으로 설정하여 자료생성모형에서 X4에 대한 F1과 F2의 설명력에 차이가 발생하도록 하였으며, 구체적으로는 F1과 관계된 요인계수를 F2로부터의 요인계수보다 크게 설정하여 X4에 대한 F1의 설명력이 상대적으로 더 크게 하였다. 요인계수 차이가 클수록 F1의 설명력이 크다고 볼 수 있으므로, 경쟁모형 1이 경쟁모형 2보다 자료에 잘 부합하여 모형 적합도 차이가 발생할 가능성이 크며, 반대로 차이가 없을 경우 두 모형의 적합도 차이는 뚜렷하지 않을 수 있다.

#### 오차상관 크기

오차상관 크기는 X4와 X5의 오차항 간 상관 정도를 의미한다. 본 연구에서는 복잡도 차이가 동일하도록 유지하기 위해 오차상관이 전혀 없는 조건은 0에 가까운 .001로 설정하였다. 이는 사실상 상관이 거의 없는 상태를 의미하며, 이 경우 X4의 교차요인계수가 제거된 경쟁모형에서는 남은 요인이 설명하지 못한 부분에 대한 오차가 커져 전체 적합도가 낮아질 가능성이 있다. 반대로 오차상관이 커질수록 SRMR 값은 작아져 모형이 더 적합한 것처럼 보일 가능성이 있다. 이는 SRMR이 오차상관 변화에 민감하게 반응한다는 선행 연구(Marsh et al., 2004; Kline, 2016)와도 맥락을 같이 한다.

#### 요인상관 크기

요인상관 크기는 선행 연구(Hair et al., 2019; Kline, 2016; Tabachnick & Fidell, 2019)를 참고하여 설정하였다. 요인상관이 클수록 두 요인의 구분이 어려워질 수 있으며 경쟁모형 간의 적합도 차이 또한 적을 수 있다. 반대로 요인

상관이 작은 경우에는, 두 요인이 명확히 구분되고 경쟁모형 간 차이가 더욱 뚜렷하게 나타나 모형 적합도에 근거한 판별이 용이해질 수 있다. 따라서 요인상관은 모형 선택 과정에서 적합도 지수와 정보기반지수의 민감도를 검증하는데 중요한 조건이 된다.

#### 표본크기

표본크기는 통계적 분석의 신뢰도와 타당도를 결정짓는 핵심 요소이며(West et al., 2012), 특히 구조방정식모형을 이용한 연구에서는 모형의 수렴 여부, 추정의 정확성, 적합도 지수의 평가에 직접적인 영향을 미친다. Kline (2016)은 표본크기를 4단계로 분류하였으며, 본 연구에서는 이를 참고하여 100, 300, 500, 1000의 4수준을 설정하였다. 작은 표본(100)의 경우 추정치의 불안정성과 적합도 지수 왜곡의 가능성이 크며, 중간 표본(300, 500)은 연구 현장에서 자주 활용되는 크기로 안정성과 현실성을 동시에 확보할 수 있다. 큰 표본(1000)은 모형 간 차이를 가장 명확하게 드러낼 수 있지만, 동시에 작은 구조적 차이에도 지나치게 민감하게 반응할 가능성이 있다는 점에서 해석에 주의가 필요하다.

#### 종속변수

본 연구에서는 비내재관계 모형 비교에서 모형 선택의 기준으로 활용될 수 있는 결과를 종속변수로 설정하였다. 구체적으로, 적합도 지수(CFI, TLI, RMSEA, SRMR), 정보기반지수에 기반한  $\Delta BIC$ , 그리고 Vuong 검증에 따른 경쟁모형 판별 결과를 포함하였다. 이들 기준은 비내재관계 모형 비교에서 서로 다른 정보의 성격을 제공하며, 단일 지수에 의존하기보다

는 상호보완적으로 해석할 필요가 있다.

본 연구에서 평가기준의 성능 또는 판별력 이란, 조건 변화(요인계수 차이, 요인상관, 오차상관, 표본크기)에 따라 경쟁모형 간 상대적 우세를 얼마나 일관적으로 구분하는지를 의미한다. 이에 따라 본 연구는 지수의 평균값뿐 아니라 사전에 설정한 기준의 충족 여부와 경쟁모형 선택들을 함께 고려하여, 조건별로 기준의 반응 양상을 종합적으로 평가하였다(Hu & Bentler, 1999; Marsh et al., 2004).

### 적합도 지수

적합도 지수는 구조방정식모형을 이용한 연구에서 가장 널리 사용되는 평가 기준으로, 본 연구에서는 CFI, TLI, RMSEA, SRMR을 분석 대상으로 하였다. 각 지수는 서로 다른 특성을 가지는데, CFI와 TLI는 요인계수 차이와 같은 구조적 모수의 차이를 탐지하는데 민감하며, 모형이 이러한 차이를 더 잘 설명할수록 값이 높아지는 경향을 보인다(Kline, 2016). RMSEA는 자유도에 대한 벌칙을 포함하여 모형의 간명성 및 근사적합 수준을 평가하고, 복잡한 모형이나 표본크기가 작은 경우에 불리하게 작용할 수 있다(MacCallum et al., 1996). SRMR은 관측치와 모형이 예측한 공분산 간 잔차 차이에 따라 값이 달라지며, 잔차가 클수록 값이 커지고 오차상관 변화에 민감하게 반응한다(Kline, 2016). 본 연구에서는 이들 지수에 대해 다음 네 가지 산출값을 종속변수로 설정하였다.

### 평균

각 조건별 반복(replication) 산출된 지수의 평균을 계산하여 적합도 지수의 분포를 확인하였다. 이는 지수가 전반적으로 높은지/낮은지

를 파악하는 1차적 기준으로 활용하였다. 또한 평균값의 상대적 차이를 통해 조건에 따라 어떠한 모형이 전반적으로 더 양호한 적합을 보이는지에 대한 기초적 경향을 파악하고자 하였다.

### 기준충족률

각 경쟁모형에 대한 적합도 지수가 전통적인 기준값을 충족하는지를 확인하였다. 기준충족률은 해당 모형이 ‘수용 가능한 수준’에 도달했는지를 나타내므로, 연구자가 분석 상황에서 모형을 선택할 수 있는 최소 조건을 반영한다.

CFI와 TLI는  $\geq .90$  이상이면 수용 가능한 수준이므로(Hu & Bentler, 1999), 본 연구에서는 각 경쟁모형이 CFI와 TLI가  $>.90$ 일 때 1, 아닐 경우에는 0으로 값을 부여하여 기준충족률을 계산하였다. RMSEA와 SRMR은  $\leq .08$ 일 때 수용 가능한 수준이므로(Hu & Bentler, 1999), 각 경쟁모형이 RMSEA와 SRMR가  $<.08$ 일 때 1, 아닐 경우에는 0으로 값을 부여하여 기준충족률을 계산하였다.

### 경쟁모형 선택률

각 반복에서 두 경쟁모형의 적합도 지수를 비교하여 기준을 충족하면서 동시에 상대적으로 더 적합한 모형을 선택한 비율을 산출하였다(CFI와 TLI의 경우 더 큰 값을 가진 모형, RMSEA와 SRMR의 경우 더 작은 값을 가진 모형). 이 값은 지수가 ‘두 모형을 얼마나 일관적으로 구분하는지’를 보여주는 기준이다.

예를 들어, 각 반복에서 경쟁모형 1의 CFI가 .9를 초과하고 경쟁모형 2보다 크면 1, 반대로 작으면 0으로 값을 부여하였으며, 두 모형의 값이 동일할 경우에는 어느 쪽도 선택하

지 않은 것으로 처리하였다. 이렇게 산출된 0/1 값의 평균을 해당 조건에서의 경쟁모형 1 선택률로 정의하였다. 경쟁모형 2의 선택률은 전체에서 경쟁모형 1의 선택률과 동일한 방식으로 계산되며, 두 모형의 선택률 합이 100%가 되지 않는 이유는 값이 동일하여 선택되지 않은 사례가 포함되기 때문이다. 여기서 경쟁모형 선택률은 특정 모형의 우수성을 확정하기 위한 판단 기준이라기보다는, 두 경쟁모형의 판별에 대한 해당 적합도 지수의 일관성을 요약한 값으로 해석할 수 있다.

#### △BIC

△BIC는 두 경쟁모형 간의 BIC 차이(△BIC = BIC(경쟁모형 2) - BIC(경쟁모형 1))를 의미한다. BIC는 값이 작을수록 우수한 모형을 의미하므로, 본 연구에서는 △BIC가 양수일 경우 경쟁모형 1이, 음수일 경우 경쟁모형 2가 우세한 것으로 판별하며, 다음 세 가지 산출값을 종속변수로 설정하였다.

#### 평균

각 조건별 반복에서 산출된 △BIC의 평균을 계산하였다. 이는 전반적으로 경쟁모형 1과 2의 △BIC 차이가 큰지/작은지, 경쟁모형 1과 2 중 어느 모형이 상대적으로 더 작은 값을 보이는지를 파악하는 1차적 기준으로 활용하였다.

#### 경쟁모형 선택률

△BIC가 양수인지 음수인지에 따라 경쟁모형 1 또는 2가 선택된 빈도를 기록하였다. 양수일 경우 1, 음수일 경우 0으로 값을 부여하여 △BIC가 특정 모형을 지지하는 정도를 살펴보았다.

#### 절대값 기준(차이에 대한 효과크기 해석)

△BIC에 대한 Raftery(1995)의 해석 기준을 적용하여(△BIC 0 - 2=작은 차이=1, 2 - 6=보통 차이=2, 6 - 10=큰 차이=3, >10=매우 큰 차이=4), 단순한 방향성뿐 아니라 모형 간 차이가 얼마나 강력한 증거를 제공하는지를 효과크기 차이의 상대적 크기를 참고적으로 구분하여 제시하였다.

#### Vuong 검증

Vuong(1989)이 제안한 비내재모형 로그우도 차이 기반 검증은 두 모형의 상대적 우수성을 통계적으로 평가할 수 있는 방법이다. 본 연구에서의 Vuong 검증은 두 경쟁모형간 로그우도 차이가 통계적으로 유의한지를 평가하는 검증으로, 기각되지 않을 경우 두 모형은 동등하게 적합한 것으로 해석할 수 있다. 본 연구에서는 다음 두 가지 산출값을 종속변수로 설정하였다.

#### 통계적 기각률

Vuong 검증에서 산출된 로그우도 차이 기반  $z$  통계량의 유의확률( $p < .05$ )을 기준으로, 두 모형의 차이가 통계적으로 유의한지를 평가하였다. 유의할 경우 1, 유의하지 않을 경우 0으로 값을 부여하여 확률을 계산하였고, △BIC가 제공하지 못하는 통계적 유의성을 직접적으로 제공하는 장점이 있다.

#### 경쟁모형 선택률

Vuong 검증에서 통계적으로 유의한 경우, 어느 모형이 더 우수한 것으로 나타났는지를 비율로 계산하였다. 유의한 개수를 분모로 하고 경쟁모형 1이 우세할 경우 1, 반대일 경우 0으로 값을 부여하였다. 이는 단순히 ‘차이가

있다/없다'를 넘어, '어떤 모형이 자료와 더 부합하는지'를 판단하는 근거가 된다.

이처럼 본 연구의 종속변수는 단순히 지수의 평균값을 비교하는 수준을 넘어서, 기준 충족률, 경쟁모형 선택률, 통계적 기각률 등 다양한 산출 기준을 포함하였다. 이를 통해 각 조건에서 지수가 어떤 양상으로 반응하는지를 보다 정교하게 분석하고, 판별에 영향을 미친 모형의 특성을 파악할 수 있도록 설계하였다.

### 자료 생성 및 분석 방법

전체 조건은 108개(요인계수 차이:3 × 오차상관 크기:3 × 요인상관 크기:3 × 표본크기:4)로, 연구 설계에 따라 Mplus 8.4 프로그램 (Muthén & Muthén, 2023)을 사용하여 조건별로 1,000개의 시뮬레이션 자료를 생성하였다. 생성된 자료는 최대우도법으로 추정하였으며, 각 경쟁모형(모형 1과 2)에 대한 적합도 지수와 BIC 값을 산출하였다. 이후 종속변수의 산출은 R 패키지MplusAutomation (Hallquist & Wiley, 2018)과 lavaan (Rosseel, 2012)를 이용하여 추출 및 계산하고 Vuong 검증은 nonnest2 (Merkle et al., 2020) 패키지를 사용하였다.

### 연구 결과

연구 결과에서는 요인계수 차이, 오차상관 크기, 요인상관 크기, 그리고 표본크기 조건에 따른 지수들의 기능을 기술하고 비교하였다. 연구 조건 별로 각 통계 지수의 평균값, 기준 충족률, 그리고 경쟁모형 선택률에 대하여 기술하였으며, 자세한 결과는 부록 표 A-1부터 표 A-5를 통해 확인이 가능하다.

### CFI

대부분의 연구 조건에서 경쟁모형 1의 평균값이 경쟁모형 2보다 높은 값을 기록하였다. 특히 요인계수 차이와 오차·요인상관 크기가 모두 큰 조건에서 경쟁모형 2의 CFI 평균값은 .790으로 현저히 낮게 나타났다.

CFI 기준충족률 또한 경쟁모형 1이 보다 높은 값을 보였다. 요인계수 차이가 크고 표본크기 $\geq 300$ 인 조건에서는 경쟁모형 1이 100% 충족한 반면, 경쟁모형 2는 0% 수준에 머물렀다.

경쟁모형 선택률 또한 경쟁모형 1이 전반적으로 더 높은 값을 보였으며, 요인계수 차이와 오차상관 크기가 모두 큰 조건에서는 거의 100%에 도달하였다. 반면 요인계수 차이가 작고 요인상관이 작은 조건에서는 두 모형의 선택률이 근접하였다.

### TLI

CFI와 마찬가지로 경쟁모형 1의 TLI 평균값이 경쟁모형 2보다 대체로 높았으나, 일부 조건(예: 요인계수 차이가 작고 오차·요인상관이 작은 경우)에서는 두 모형이 동일하게 .918을 기록하였다.

기준충족률과 경쟁모형 선택률 또한 대부분의 조건에서 경쟁모형 1이 우세함을 보였으며, 특히 표본크기가 작은 조건( $N=100$ )에서의 경쟁모형 2의 충족률은 30% 미만에 머물렀다.

### RMSEA

다른 적합도 지수들(CFI, TLI, SRMR)에 비해 RMSEA가 전반적으로 모형의 적합도를 낮게

평가하는 경향(값이 커지는)이 있음을 확인하였다. 경쟁모형 1은 요인계수 차이가 큰 조건에서만 .08 미만(.052~.076)의 RMSEA 평균을 기록하였으며, 경쟁모형 2는 대부분의 조건에서 .09 이상의 평균값을 보였다.

RMSEA가 전반적으로 모형을 낮게 평가하기 때문에 기준충족률(<.08) 또한 낮게 산출되었다. 거의 모든 조건에서 경쟁모형 2의 기준충족률은 0%에 머물렀으며, 경쟁모형 1의 기준충족률도 요인계수 차이가 작아지고 표본크기가 커질수록 급감하였다. RMSEA 기준을 충족한 경우, 경쟁모형 선택률은 경쟁모형 1이 거의 우세하였다.

#### SRMR

전반적으로 경쟁모형 1이 경쟁모형 2보다 낮은 SRMR 평균값을 보이고 높은 기준충족률을 보임으로써, SRMR은 경쟁모형 1 상대적으로 양호한 적합을 나타내는 경향을 보였다. 경쟁모형 선택률을 살펴보면, 요인계수 차이가 작은 조건을 제외하면 SRMR은 경쟁모형 2보다 경쟁모형 1을 선택하는 경향이 있었다. 특별히 요인계수 차이와 오차상관 크기가 크고, 요인상관이 작은 조건에서 SRMR이 경쟁모형 1을 선택한 비율은 100%였다. 그러나 요인계수 차이가 작은 조건에서는 경쟁모형 2의 선택률이 경쟁모형 1보다 대부분 높게 나타났다.

#### $\Delta$ BIC와 Vuong 검증

적합도 지수와는 다르게  $\Delta$ BIC는 기준값에 기반하여 모형을 선택하지 않는다. 따라서 본 연구는  $\Delta$ BIC에 대한 기준충족률 대신

Raftery(1995)의 해석 기준에 따라 경쟁모형 1과 경쟁모형 2의 차이에 대한 효과크기를 비교하고, Vuong 검증에 기반하여  $\Delta$ BIC에 대한 통계적 기각률과 경쟁모형 선택률을 살펴 보았다.

요인계수 차이, 오차상관, 그리고 요인상관의 크기가 큰 조건의 경우, 다른 연구 조건에 비하여  $\Delta$ BIC 평균값의 범위는 [162.90, 1637.51]로 산출되었다. Raftery(1995)의 기준에 따르면 모든 표본크기 조건에서 ‘매우 큰 차이(범주 4)’에 해당하였다. 이는 해당 조건에서  $\Delta$ BIC가 경쟁모형 간 차이에 대해 일관되게 강한 상대적 증거를 제공하였음을 의미한다. 반면 요인계수 차이와 오차상관 크기가 작은 조건에서는  $\Delta$ BIC 평균값이 낮아 두 모형 간 차이를 구분하지 못했으며, 이때의  $\Delta$ BIC 범위는 [0.28, 9.07] 수준으로 나타났다.

$\Delta$ BIC에 기반한 경쟁모형 선택률 또한 요인계수 차이와 오차상관 크기에 따라 달라졌다. 요인계수 차이와 오차상관 크기가 큰 조건에서는 거의 100%의 비율로 경쟁모형 1을 선택하였으나, 반대로 요인계수 차이와 오차상관 크기가 작은 조건에서는 경쟁모형 1을 선택하는 비율이 49.1~57.3%에 머물러 두 모형 간 변별력이 낮아졌다. 이와 같은  $\Delta$ BIC의 패턴을 Raftery(1995) 기준에 따라 해석해보면, 요인계수 차이와 오차상관 크기가 큰 조건에서 두 경쟁모형은 매우 큰 차이, 반대의 조건에서는 작은 차이로 볼 수 있다.

본 연구에서 Vuong 검증의 기각률은 두 경쟁모형 간 로그우도 차이가 통계적으로 유의한 경우의 비율을 의미하므로, 해당 비율이 높을수록 두 모형 간 차이가 통계적으로 유의미하게 나타나는 경향이 강함을 의미한다. 요인계수 차이와 오차상관 크기가 크거나 중간

인 조건에서는 대부분 100%의 기각률을 보여 경쟁모형 1은 경쟁모형 2와 다르다고 판별되었다. 그러나 요인계수 차이와 오차상관 크기가 작은 조건에서는 경쟁모형 1과 2의 차이가 유의미하지 않아 기각률이 3.6~12.1% 수준에 머물렀다. 이와 같은 패턴은 Vuong 검증에 기반한 경쟁모형 선택률과도 동일하여, 요인계수 차이와 오차상관 크기가 큰 조건에서는 경쟁모형 1이 항상 선택되었다.

### 실증 자료를 사용한 결과

본 연구는 실증 자료를 이용하여 시뮬레이션을 통한 연구 결과를 확인하고자, R psych 패키지(Revelle, 2024)의 Holzinger와 Swineford(1939)의 인지검사 자료(이하 HS1937 자료)를 활용하였다. HS1937 자료는 확인적 요인분석 모형이나 구조방정식모형을 이용한 연구에서 가장 빈번히 사용되는 공개 자료로, 요인 간 이론적 구분이 비교적 명확하면서도 특정 문항이 다수의 요인과 개념적으로 연관될 수 있는 특성을 지닌다. 그중 수리적 퍼즐(numerical puzzles) 문항은 선행연구에서 추론 요인과 언어 요인 중 어느 요인에 더 적절히 배치되는지를 두고 지속적으로 거론되고 있는 문항으로(Brown, 2015; Reise et al., 2010), 이 문항은 개념적으로 두 요인과 모두 연관성을 갖지만 실제 구조는 명확히 단정되기 어렵다고 논의되고 있다. 이러한 특성은 구조적 차이가 크지 않아 적합도 지수가 서로 유사하게 나타나는 상황을 실제 자료에서 재현할 수 있다는 점에서, 본 연구가 시뮬레이션에서 설정한 비내재관계 경쟁모형과 개념적으로 잘 대응된다.

Holzinger와 Swineford(1939)의 14개 검사 점수

에 대해 공간(Spatial), 언어(Verbal), 속도(Speed), 추론(Reasoning) 요인을 포함한 4요인 확인적 요인분석 모형을 설정하고 수리적 퍼즐 문항의 요인 배치만을 달리한 두 경쟁모형(그림 3 참조)을 비교한 결과, 두 모형은 전반적으로 양호하고 매우 유사한 적합도를 보였다. 그러나 적합도 지수의 미세한 차이(CFI=.904 vs .899; TLI=.877 vs .871; RMSEA=.087 vs .089; SRMR=.070 vs .075)와 함께, Vuong의  $\chi^2$  기반 분산 검증( $w^2=0.019$ ,  $p<.001$ ), 로그우도 차이 기반 z 검증( $z=1.738$ ,  $p=.041$ ) 결과는 두 경쟁모형이 통계적으로 구별 가능한 수준의 차이를 보이며, 자료가 경쟁모형 1, 즉 수리적 퍼즐 문항을 추론 요인에 배치한 모형에 상대적으로 더 일관되게 반응하고 있음을 시사하였다. 이는 수리적 퍼즐 문항이 언어적 이해 요소와도 개념적으로 연관될 수 있다는 선행 논의(Brown, 2015)를 고려하더라도, 실제 자료에서는 해당 문항을 추론 요인에 배치한 모형이 적합도 지수와 검정 통계량 전반에서 상대적으로 안정적인 패턴을 보였음을 보여주는 결과이다.

중요한 점은 두 모형의 적합도 지수가 매우 유사함에도 불구하고, 분산 검증과 Vuong 검증 결과를 토대로 실제 자료에서 미세한 구조적 차이를 탐지해냈다는 사실이다. 이는 본 연구의 시뮬레이션 결과, 특히 구조적 차이가 작고 적합도 지수가 서로 근접한 상황에서도

표 2. 실제 자료 사용 적합도 지수 결과

모형	CFI	TLI	RMSEA	SRMR	BIC
경쟁 모형 1	.904	.877	.087	.070	10678.880
경쟁 모형 2	.899	.871	.089	.075	10687.285

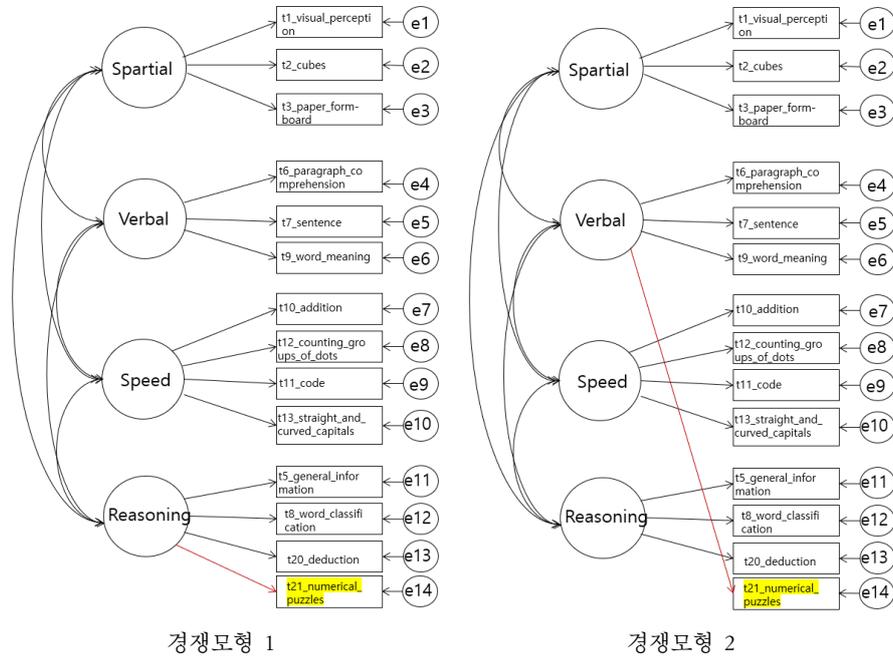


그림 3. Holzinger와 Swineford(1939) 자료를 사용한 연구의 경쟁모형들

$\Delta$ BIC와 Vuong 검증은 서로 다른 변별력을 통해 모형을 구별해낼 수 있다는 결과가 실제자료에서도 일관되게 재현된다는 점을 보여준다.

따라서 HS1937 자료를 활용한 실제 분석은 단순한 예시가 아니라, 본 연구가 제안하는 모형 비교 전략의 실증적 근거로 기능한다. 즉, 경쟁모형 간 구조적 차이가 미세하거나 적합도 지수만으로는 우열 판단이 어려운 상황에서, 다양한 적합도 지수(특히  $\Delta$ BIC와 Vuong) 기반의 종합적 판단이 필요하다는 본 연구의 핵심 주장을 실제 자료를 통해 다시 한 번 확인하였다. 이는 연구자가 실증 분석에서 마주하게 되는 ‘모형이 모두 적합해 보이는 상황’에서 어떤 적합도 지수를 우선적으로 고려해야 하는지에 대한 구체적 가이드를 제공한다는 점에서 방법론적 기여를 가진다.

## 논 의

본 연구는 동일한 자유도 조건에서 비내재 관계 모형 간 구조적 차이(요인계수 차이, 오차상관, 요인상관, 표본크기)에 따라 적합도 지수(CFI, TLI, RMSEA, SRMR),  $\Delta$ BIC 및 Vuong 검증이 어떠한 양상을 보여주는지 비교하였다. 연구 조건에 따라 각 통계적 지수가 선호하는 모형이 달라지는 현상이 다수 관찰되었다. 이는 특정 지수는 구조적 오류에 민감하게 반응하는 반면, 다른 지수는 동일한 오류를 거의 탐지하지 못하거나 왜곡된 판단을 내릴 수 있음을 실증적으로 보여주는 결과이다(Fan & Sivo, 2007; Chen et al., 2008). 좀 더 자세히 설명하면, CFI와 TLI는 요인계수 차이가 큰 조건과 중간 이상 표본크기에서 경쟁모형 1을 일관되게 선택하여 구조적 차이를 안정적으로

탐지한 반면, RMSEA는 동일 조건에서도 오차 상관이나 요인상관이 큰 조건에서는 두 경쟁 모형을 모두 부적합하게 평가하는 과도한 보수성을 보였다. SRMR은 반대로 오차상관이나 요인상관이 큰 조건에서 두 경쟁모형을 모두 적합하다고 판단하는 경향을 보여 구조적 차이를 상대적으로 덜 반영하였다. 이러한 지수 간 판단 불일치는 적합도 지수가 본질적으로 조건에 민감하게 반응한다는 Browne과 Cudeck (1993), Kline(2016)의 연구 결과와 일치하며, 단일 지수 사용을 경계해야 한다는 Hancock과 Mueller(2011)의 주장을 실증적으로 강화하였다.

지수별로 살펴보면, CFI와 TLI는 구조적 차이에 상대적으로 민감한 지수로, 요인계수 차이가 큰 조건에서 가장 높은 변별력을 보였으며, 이는 Marsh 등(2004)과 Brown(2015)이 보고한 상대적 적합도 지수의 장점과 부합한다. 반면 RMSEA는 자유도와 표본크기, 오차상관에 영향을 크게 받아 작은 표본 또는 오차상관이 큰 조건에서 두 모형 모두를 부적합으로 평가하여 구조적 차이의 변별력이 제한적으로 나타났다. SRMR은 요인상관 크기에 민감하여 요인상관이 큰 조건에서는 잘 변별하지 못하였으며, 이는 요인 간 분리가 어려울 때 SRMR이 구조적 차이를 탐지하지 못한다는 Kline (2016)의 해석과 일치한다. 흥미롭게도 RMSEA는 기준을 충족한 일부 조건에서 경쟁모형 1을 거의 100% 선택하여 강한 변별력을 보였는데, 이는 후속 연구에서 정교한 이론적 설명이 요구되는 부분이다.

$\Delta BIC$ 는 모형의 복잡도에 대한 벌칙을 포함한 정보기반지수로서, 구조적 차이가 크고 복잡도 차이가 실제로 의미 있는 상황에서 강하게 특정 모형을 선호하는 반면, Vuong 검증

은 사례별 우도 차이를 기반으로 두 모형 자료의 적합도 차이가 통계적으로 유의한지 여부를 직접 검증하는 방식으로 작동한다. 따라서  $\Delta BIC$ 는 간명성에 대한 벌칙을 통해 모형을 분리하고, Vuong 검증은 자료 설명력의 실제 차이를 통해 모형을 분리한다는 점에서, 두 기준은 서로 다른 변별력을 가진다. 이로 인해 동일한 자료에서도 두 기준이 상이한 결론을 제공할 수 있으며, 이는 연구자가 비내재관계 모형을 비교할 때 두 지수를 병행하여 해석해야 하는 중요한 이유를 제시한다.  $\Delta BIC$ 는 요인계수 차이가 큰 조건과 오차상관이 높은 조건에서 표본크기와 무관하게 경쟁모형 1을 안정적으로 지지하였다. 이는  $\Delta BIC$ 가 큰 구조적 차이에만 민감하고 작은 차이에는 둔감하다는 Raftery(1995), Burnham과 Anderson (2004)의 해석과 정확히 일치한다. 반면 Vuong 검증은 표본크기에 매우 민감한 검증 기반 접근으로써, 표본크기가 500 이상인 조건에서는 요인계수 차이가 중간 수준의 구조적 차이도 유의하게 탐지하였으나, 표본크기가 작은 조건에서는 요인계수 차이가 큰 조건에서도 차이를 탐지하지 못하는 등 표본 기반 검증력의 한계를 명확히 보여주었다. 이처럼  $\Delta BIC$ 는 명확한 차이의 안정적 탐지, Vuong 검증은 큰 표본에서 미세한 차이의 정교한 탐지라는 상호보완적 기능을 지니므로, 연구자가 실제 분석에서 두 기준 중 하나만을 단독 적용할 경우 구조적 오류를 과소·과대 평가할 위험이 있다. 따라서  $\Delta BIC$ 와 Vuong 검증을 병행하여 해석하는 전략은 비내재관계 모형 비교에서 단일 지수가 갖는 한계를 보완하고 보다 신뢰할 수 있는 모형 선택을 가능하게 한다.

실제 자료를 사용한 결과에서 주목할 점은

두 모형의 적합도가 전반적으로 매우 유사하다. 더라도,  $\chi^2$  기반 분산 검증,  $\Delta BIC$ , 그리고 Vuong 검증은 미세한 구조적 차이를 실제 자료가 지지하는 방향으로 민감하게 포착할 수 있다는 것이다. 즉, 실제 자료 사용 분석 결과는 경쟁모형이 거의 동일한 적합도를 보이더라도, 표면적 적합도 지수만으로는 모형 선택이 불가능한 상황에서 어떠한 보조적 판단 근거를 제공할 수 있는지를 실증적으로 보여준다. 이는 본 연구의 시뮬레이션 결과에서 확인된 바와 같이, CFI, TLI, RMSEA, SRMR과 같은 적합도 지수만으로는 판단이 모호한 경우에도  $\Delta BIC$ 와 Vuong 검증은 구조적 차이가 실제로 존재하는지를 판별하는 데 중요한 역할을 수행할 수 있음을 실제자료에서도 확인한 사례이다. 또한 개별 문항이 복수의 이론적 가능성을 갖는 수리퍼즐검사의 예시와 같이, 모형 간 미세한 구조적 차이를 평가하는데  $\chi^2$  기반 분산 검증,  $\Delta BIC$ , 그리고 Vuong 검증을 유용하게 활용할 수 있음을 시사한다.

이러한 결과를 바탕으로 연구자가 모형 비교를 수행할 때 고려해야 할 실질적 지침을 제시할 수 있다. 첫째, 구조적 차이가 명확한(예: 요인계수 차이가 큰 조건) 경우 CFI/TLI 및  $\Delta BIC$ 가 가장 일관되고 강력한 정보를 제공하므로 이들 지수를 우선적으로 고려하는 것이 바람직하다. 둘째, 구조적 차이가 미세한(예: 요인계수 차이가 작은 조건) 경우에는 CFI, TLI의 기준 충족률이 낮아질 수 있으며  $\Delta BIC$  또한 구조적 차이를 거의 탐지하지 못하므로, RMSEA의 엄격한 판정과 Vuong 검증의 표본 의존적 민감성을 함께 고려하는 다중기준 접근이 필요하다. 셋째, 요인상관이 높은 경우 SRMR은 구조적 차이를 전혀 반영하지 못하므로 단독 사용을 피해야 하며, CFI와 TLI

또는  $\Delta BIC$ 와 결합하여 해석하는 것이 유리하다. 넷째, 오차상관이 큰 조건에서는 RMSEA와 SRMR의 판단이 극단적으로 상충하므로, 이 두 지수는 보조적 지수로만 활용하고  $\Delta BIC$ , Vuong 검증, CFI/TLI의 결합을 통해 보다 안정적인 결론을 도출해야 한다. 각 지수가 어떤 구조적 차이에 더욱 민감한지에 대한 이해를 기반으로 경쟁모형을 비교하고 선택한다면, 모형 평가에서 발생할 수 있는 오판을 최소화하고 보다 정교한 해석이 가능할 것이라고 기대한다.

마지막으로, 본 연구는 시뮬레이션 기반으로 동일한 자유도 조건에서 경쟁모형 비교를 수행하였다는 점에서 학술적 의의를 가지지만 몇 가지 한계도 존재한다. 첫째, 본 연구는 모형 복잡도를 통제하기 위해 비교적 간단한 구조(2요인 8측정변수)를 사용하였기에 실제 연구 상황에서 나타나는 다양한 상황(예: 교차요인계수, 정규성, 결측치, 모형 지정 오류 등)을 충분히 반영하지 못했다. 둘째, 본 연구는  $\Delta BIC$ 와 Vuong 검증을 중심으로 분석하였으나 AIC, CAIC, saBIC 등 다른 정보기반지수는 포함하지 않았으며, 이들 기준과의 비교는 후속 연구에서 중요하게 다루어야 할 주제이다. 셋째,  $\Delta BIC$ 와 Vuong 검증이 서로 다른 메커니즘을 지닌다는 사실은 본 연구에서 명확히 드러났지만, 두 기준이 충돌하는 상황에서 어떤 이론적 근거로 한 기준을 우선해야 하는가에 대한 세부적 안내는 추가 연구가 필요하다. 넷째, 본 연구는 시뮬레이션 기반으로 수행되었으므로 실제 심리측정 자료에서 동일한 지수 간 상충 패턴이 재현되는지를 검증하는 실증적 후속 연구가 요구된다. 이러한 확장 연구는 적합도 지수와 모형 선택 기준이 실제 연구 환경에서 어떻게 작동하는지를 보다 정

교하게 규명하는 데 기여할 수 있을 것이다.

### 참고문헌

- 김미림, & 임예지 (2024). 변안 심리검사 타당화 작업에 대한 체계적 검토: 검수와 보고 관행에 대한 검토와 제언. *한국심리학회지: 일반*, 43(3), 261-295.  
<https://doi.org/10.22257/kjp.2024.9.43.3.261>
- 유소현, & 김수영 (2024). 구조방정식 모형의 전반적인 평가 및 효과크기와 연속성에 대한 숙고. *한국심리학회지: 일반*, 43(3), 199-230.  
<https://doi.org/10.22257/kjp.2024.9.43.3.199>
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), 238-246.  
<https://doi.org/10.1037/0033-2909.107.2.238>
- Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.
- Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.  
<https://doi.org/10.1002/9781118619179>
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.
- Burnham, K. P., & Anderson, D. R. (Eds.). (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. New York, NY: Springer New York.
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2), 261-304.  
<https://doi.org/10.1177/0049124104268644>
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling: a multidisciplinary journal*, 14(3), 464 - 504.  
<https://doi.org/10.1080/10705510701301834>
- Chen, F. F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, 36(4), 462-494.  
<https://doi.org/10.1177/0049124108314720>
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272-299.  
<https://doi.org/10.1037/1082-989X.4.3.272>
- Fan, X., & Sivo, S. A. (2007). Sensitivity of Fit Indices to Model Misspecification and Model Types. *Multivariate Behavioral Research*, 42(3), 509-529.  
<https://doi.org/10.1080/00273170701382864>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.
- Hallquist, M. N., & Wiley, J. F. (2018). *MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable*

- Analyses in *Mplus. Structural Equation Modeling: A Multidisciplinary Journal*, 25(4), 621-638.  
<https://doi.org/10.1080/10705511.2017.1402334>
- Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, 71(2), 306-324.  
<https://doi.org/10.1177/0013164410384856>
- Henson, R. K., & Roberts, J. K. (2006). Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice: Common Errors and Some Comment on Improved Practice. *Educational and Psychological Measurement*, 66(3), 393-416.  
<https://doi.org/10.1177/0013164405282485>
- Holzinger, K. J., & Swineford, F. (1939). A study in factor analysis: The stability of a bi-factor solution. *Supplementary Educational Monographs*, 48, 1-91.
- Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3(4), 424-453.  
<https://doi.org/10.1037/1082-989X.3.4.424>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55.  
<https://doi.org/10.1080/10705519909540118>
- Jackson, D. L., Gillaspay, J. A., Jr., & Purc Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, 14(1), 6-23.  
<https://doi.org/10.1037/a0014694>
- Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods* (Version V) [Computer software manual]. Uppsala: Uppsala University, Department of Statistics; Chicago: International Educational Services.
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44(3), 486-507.  
<https://doi.org/10.1177/0049124114543236>
- Kline, P. (1994). *An easy guide to factor analysis*. Routledge, New York.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.
- Levy, R. & Hancock, G. R. (2007). A Framework of Statistical Tests For Comparing Mean and Covariance Structure Models. *Multivariate Behavioral Research*, 42(1), 33-66.  
<https://doi.org/10.1080/00273170701329112>
- Levy, R. & Hancock, G. R. (2011). An extended model comparison framework for covariance and mean structure models, accommodating multiple groups and latent mixtures. *Sociological Methods & Research*, 40(2), 256-278.  
<https://doi.org/10.1177/0049124111404819>
- MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, 100(1), 107-120.  
<https://doi.org/10.1037/0033-2909.100.1.107>

- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149.  
<https://doi.org/10.1037/1082-989X.1.2.130>
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504.  
<https://doi.org/10.1037/0033-2909.111.3.490>
- Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320-341.  
[https://doi.org/10.1207/s15328007sem1103\\_2](https://doi.org/10.1207/s15328007sem1103_2)
- McNeish, D. (2022). Limitations of the sum-and-alpha approach to measurement in behavioral research. *Policy Insights from the Behavioral and Brain Sciences, 9*(2), 196-203.  
<https://doi.org/10.1177/23727322221117144>
- McNeish, D. (2023). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behavior Research Methods, 55*(8), 4269-4290.  
<https://doi.org/10.3758/s13428-022-02016-x>
- McNeish, D., An, J., & Hancock, G. R. (2018). The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models. *Journal of Personality Assessment, 100*(1), 43-52.  
<https://doi.org/10.1080/00223891.2017.1281286>
- Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2020). *nonnest2: Tests of non-nested models* (R package version 0.5-6) [Computer software].  
<https://CRAN.R-project.org/package=nonnest2>
- Merkle, E. C., You, D., & Preacher, K. J. (2016). Testing nonnested structural equation models. *Psychological Methods, 21*(2), 151-163.  
<https://doi.org/10.1037/met0000038>
- Muthén, L. K., & Muthén, B. O. (1998-2023). *Mplus* (Version 8.10) [Computer software]. Muthén & Muthén.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535-569.  
<https://doi.org/10.1080/10705510701575396>
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-163. <https://doi.org/10.2307/271063>
- Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6), 544-559.  
<https://doi.org/10.1080/00223891.2010.496477>
- Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.4.x) [R package]. Northwestern University.  
<https://CRAN.R-project.org/package=psych>
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of*

- Statistical Software*, 48(2), 1-36.  
<https://doi.org/10.18637/jss.v048.i02>
- Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2), 461-464.  
<https://doi.org/10.1214/aos/1176344136>
- Steiger, J. H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research*, 25(2), 173-180.  
[https://doi.org/10.1207/s15327906mbr2502\\_4](https://doi.org/10.1207/s15327906mbr2502_4)
- Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1), 1-10.  
<https://doi.org/10.1007/BF02291170>
- Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2), 307-333.  
<https://doi.org/10.2307/1912557>
- West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209-231). The Guilford Press.
- Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, 51(1), 409-428.  
<https://doi.org/10.3758/s13428-018-1055-2>

1차원고접수 : 2025. 09. 11

2차원고접수 : 2026. 01. 30

최종게재결정 : 2026. 03. 13

## Comparison and Evaluation of Fit Indices, $\Delta$ BIC, and the Vuong Test for Non-Nested Model Selection in Structural Equation Modeling

INHYE KIM

MIRIM KIM

SOONMOOK LEE

Sungkyunkwan University

Sungshin Women's University

Sungkyunkwan University

This study aims to comprehensively examine the characteristics and limitations of various indices applicable to non-nested model selection in structural equation modeling (SEM). Monte Carlo simulations were conducted using fit indices (CFI, TLI, RMSEA, SRMR), the information-based index ( $\Delta$ BIC), and the Vuong test. The simulation design employed a  $3 \times 3 \times 3 \times 4$  factorial structure varying factor loading differences, error correlations, factor correlations, and sample sizes, yielding 108 conditions with 1,000 replications each. The findings indicate that CFI and TLI were particularly sensitive to structural misspecifications and demonstrated stable discriminative power with larger sample sizes. In contrast, RMSEA was substantially influenced by sample size and degrees of freedom, tending to impose stricter thresholds, while SRMR showed limitations in distinguishing models when factor correlations were high.  $\Delta$ BIC provided a quantitative measure of model differences but was constrained in interpretation under equal degrees of freedom. The Vuong test supplemented  $\Delta$ BIC by offering statistical significance, thereby reinforcing the basis for model selection. Overall, this study provides practical recommendations for comparison strategies, emphasizing that researchers should avoid reliance on a single index and instead adopt complementary interpretations across indices to achieve more robust and reliable model selection.

*Key words* : fit indices,  $\Delta$ BIC, Vuong test, nonnested model selection, simulation study









