한국심리학회지: 일반
Korean Journal of Psychology: General
2024, Vol. 43, No. 3, 199-230
http://dx.doi.org/10.22257/kjp.2024.9.43.3.199

구조방정식 모형의 전반적인 평가 및 효과크기와 연속성에 대한 숙고

유 소 현　　김 수 영†

이화여자대학교 심리학과

잠재변수 간 관련성을 설명하는데 활발히 사용되고 있는 구조방정식 모형은 적합도를 통하여 그 유용성을 판단할 수 있다. 모형 적합도의 통계적 유의성을 평가하는 χ^2검정과 실질적 유의성을 평가하는 적합도 효과크기 지수는 각각 이분법적 해석과 연속적 해석 방식을 이용하여 모형의 유용성을 평가한다. 실질적 유의성의 경우 적합도의 수준을 연속적으로 평가하는 것이 목적이지만, 이를 위해 사용되는 효과크기 지수는 현재 대다수의 연구에서 이분법적으로 해석되고 있다. 본 연구는 모형 적합도의 실질적 유의성을 평가하기 위하여 효과크기 지수의 관점에서 적합도의 수준을 연속적으로 해석하는 방법과 올바른 가이드라인의 사용에 대해 다룬다. 먼저 χ^2검정을 이용한 통계적 유의성 평가의 의의에 대해 간단히 설명한 뒤, 구조방정식 모형의 맥락에서 효과크기의 개념을 정의한다. 이후 다양한 종류의 적합도 효과크기 지수를 소개하며, 해당 지수들의 해석에 사용되는 가이드라인의 특징에 대해 설명한다. 마지막으로, 추정된 효과크기 지수 값을 연속적으로 해석하고자 할 때 참고하기에 적절한 가이드라인의 올바른 예시를 제공하며, 연속성이 반영되지 않았을 때의 잘못된 모형 평가 사례와 가이드라인을 근소하게 만족하지 못하는 모형에 대한 올바른 해석 방식에 대해 논의한다.

주요어 : 구조방정식 모형, 모형 적합도, 효과크기, 연속성, 가이드라인

† 교신저자: 김수영, 이화여자대학교 심리학과, 서울시 서대문구 이화여대길 52
Tel: 02-3277-3792, E-mail: suyoung.kim@ewha.ac.kr

심리학을 포함한 사회과학 영역에서 구조방정식 모형은 관찰변수 또는 잠재변수 간 관련성을 설명하기 위해 보편적으로 사용되는 모형 중 하나이다. 구조방정식 모형을 이용할 때 연구자는 모형이 자료를 설명하는 정도를 나타내는 적합도의 개념을 통해 설정한 모형이 유용한가에 대한 판단을 내리게 된다. 모형의 유용성은 일반적으로 통계적 유의성(statistical significance)과 실질적 유의성(practical significance)을 통해 판단할 수 있으며, 구조방정식 모형에 대한 적절한 판단을 내리기 위해서는 적합도의 통계적 유의성과 실질적 유의성을 평가하는 도구들의 서로 다른 목적과 역량을 이해하고 이를 올바르게 사용하는 것이 중요하다. 그럼에도 불구하고 지금까지 출판된 국내외 논문들에서는 모형적합도의 평가 과정을 정확하게 이해하지 못하고 추정의 결과를 단편적으로 서술하는 경우가 매우 빈번하였다. 이에 본 연구는 모형 적합도의 유의성을 판단하는 기존의 다양한 평가 도구들의 목적을 바탕으로 전반적인 적합도 평가 과정을 통합 정리하고, 특히 실질적 유의성을 평가하는 단계에 효과크기의 개념을 반영하여 보다 실용적이며 적절한 모형 평가 방식에 대하여 논의한다. 이는 다양한 종류의 구조방정식 모형이 광범위하게 이용되고 있는 현재, 내용 영역의 연구자(substantive researchers)들이 모형의 유용성을 올바르고 적절하게 판단하기 위해 꼭 짚고 넘어가야 하는 주제이다.

적합도의 통계적 유의성을 평가하는 대표적인 도구인 χ^2검정(Jöreskog, 1969)의 경우, 모형이 자료를 완벽하게 설명한다는 완전 적합(exact fit) 가설을 설정하고 이에 대한 통계적 검정을 진행한다. 유의수준 α와 p값을 비교하여 적합도를 완전 적합 혹은 불완전 적합으로 해석하는 이분법적 방식을 통해 연구자는 모형이 자료를 완벽하게 설명하는지 아닌지에 대한 통계적 유의성의 정보를 얻게 된다. 반면 적합도의 실질적 유의성의 경우 모형과 자료 간 차이가 완전 적합으로부터 얼마나 떨어져 있는가를 연속선 상에서 평가한다. 일반적으로 통계 모형이나 모수의 실질적 유의성을 평가하기 위해 Cohen의 d, η^2, R^2과 같은 효과크기 지수들이 사용되는데, 구조방정식의 경우 적합도 지수를 이용해 적합도의 효과크기를 확인할 수 있다. 적합도 지수는 χ^2검정이 기각되었을 때 적합도가 완전 적합으로부터 얼마나 떨어져 있는지에 대한 효과크기를 연속적인 관점에서 평가한다. Yuan과 Marshall (2004)은 구조방정식 모형의 맥락에서 일반적인 효과크기의 형태와 유사한 새로운 적합도 평가 지수를 제안하였으며, Maydeu-Olivares (2017)는 통계적으로 유의한 차이 검정 결과에 대하여 효과크기를 이용해 그 차이를 질적으로 확인하듯, 유의한 적합도 검정 결과에 대하여 적합도 지수를 이용해 모형과 자료 간 차이의 효과크기를 확인할 것을 제안하였다. 나아가, Gomer 등(2019) 역시 다양한 종류의 구조방정식 모형의 효과크기를 제시하고 이를 이용한 시뮬레이션 결과들을 제공하였다.

χ^2검정과 적합도 지수는 이분법적 평가와 연속선 상에서의 평가라는 각각의 방식을 통해 모형의 유용성을 평가한다. 이는 곧 χ^2검정 결과를 연속적으로 해석하거나, 반대로 추정된 적합도 지수 값을 이분법적으로 해석하는 것은 잘못된 평가 방식임을 의미한다. 특히 적합도 지수의 경우, 그 해석에 사용되는 가이드라인을 적합도 평가의 절대적 기준으로 사용해서는 안 됨을 여러 연구에서 지속적으로 경고하였으나(Barrett, 2007; Gomer et al.,

2019; Markland, 2007; Marsh et al., 2004), 그럼에도 이를 모형의 유용성에 대한 이분법적 통과 기준으로 사용하는 관행은 사라지지 않고 있다. 본 연구에서는 총 여섯 개의 프리미엄 저널[1]에서 2020년부터 2022년까지 출판된 연구 가운데 적합도 지수를 이용해 구조방정식 모형을 평가한 250개의 연구를 분석한 결과, 대부분의 연구에서 적합도 지수와 그 가이드라인은 마치 χ^2검정의 p값과 α의 관계처럼 사용되고 있었다. p값이 α보다 작으면 영가설을 기각하고 α보다 크면 영가설 기각에 실패하듯, 적합도 지수가 가이드라인에서 제시하는 기준을 만족하면 모형이 좋거나(good) 적절하고(acceptable) 기준을 만족하지 못하면 나쁘거나(bad) 적절하지 않다(unacceptable)고 해석하는 현상이 대부분의 연구에서 확인되었다.

나아가, 250개의 연구 가운데 설정된 모형에 대하여 적합도 지수 값이 가이드라인의 기준에 근접하나 만족은 하지 못한 적합도(marginal fit)를 나타낸 32건의 사례 가운데 절반 이상의 연구에서 해당 모형은 자료를 충분히 설명하고 있지 못한 것으로 보고되었다. 특히 Tyler 등(2020)의 경우 적합도 지수가 기준에 매우 근접함에도 불구하고(TLI=.89, CFI=.91, RMSEA=.06) 제시된 모형을 배제하는 등, 현재 적합도 지수는 많은 연구에서 단순히 특정한 값을 기준으로 적합도에 대한 이분법적 판단을 내리는 통계적 검정과 유사한 방식으로 사용되고 있었다.

적합도 지수가 본래의 목적에 맞게 연속적으로 사용되기 위해서는 적합도 지수를 통계적 검정이 아닌 효과크기 지수로서 인식하고 사용할 필요가 있다. 통계적 검정이 α를 영가설 기각여부의 절단점으로 사용하는 반면, 효과크기 지수의 가이드라인은 어디까지나 적합도의 수준이 얼마나 높아졌거나 낮아졌는가를 나타내는 일종의 알림판 역할에 불과하다. 가이드라인의 기준값은 적합도를 이분법적으로 평가하는 근거가 될 수 없으며, 단순히 연속선 상에서 모형 적합도의 위치를 해석하는 과정에서 참고하는 보조적인 지표에 불과하다.

적합도 지수를 효과크기 지수로 사용하는 과정에 이와 같은 가이드라인의 연속성(continuity)이 제대로 고려되지 않을 경우, 특히 효과크기 지수의 특성에 대해 완벽히 이해하지 못하고 있는 연구자는 적합도의 효과크기를 이분법적으로 해석하는 잘못을 저지를 수 있다. 실제로 적합도 지수를 해석할 때 인용되는 저명한 가이드라인들(Bentler & Bonett, 1980; Browne & Cudeck, 1993; Hu & Bentler, 1999)은 적합도 지수의 연속적인 성격을 제대로 반영하지 못하는 절단적인 기준과 해석을 제시하며, 이와 같은 문제는 비교적 최근 제안된 새로운 가이드라인(Asparouhov&Muthén, 2018; Shi et al., 2018)까지 이어지고 있다. 그러나 적합도를 평가하고 모형을 수정하는 과정에서 연속성이 반영되지 않은 가이드라인이 이용될 경우, 가이드라인의 기준값을 아주 조금이라도 만족하지 못한다는 이유 하나만으로 충분히 유용한 모형을 배제하는 비효율적이며 실용적이지 못한 모형의 평가를 하는 문제가 발생할 수 있다.

본 연구의 목적은 적합도 효과크기 지수의 종류 및 연속선 상의 해석 방식에 대한 논의

1) Journal of applied psychology, Journal of applied developmental psychology, Journal of educational psychology, Journal of counseling psychology, Journal of abnormal psychology, Journal of personality and social psychology

를 제시하는 것이다. 이를 통하여 모형 자체의 유용성에 대해 지금 당장 평가를 내려야 하는 내용 영역 연구자들이 직접적으로 참고할 수 있는 실용적인 모형 평가 및 해석 방법을 제안하고자 한다. 적합도 효과크기 지수의 경우 이미 알려진 여러 적합도 지수 이외에도 Maydeu-Olivares(2017) 및 Gomer 등(2019)의 최근 연구들을 통해 제안된 새로운 지수들이 존재하여 이를 소개하고자 한다. 구조방정식 모형의 맥락에서 사용할 수 있는 다양한 효과크기 지수들을 소개하는 것은 적합도의 실질적 유의성의 평가 방법에 대한 이해를 넓힐 수 있을 것이다. 또한, 본 연구는 적합도의 효과크기 지수가 본래의 쓰임에 맞게 사용될 수 있도록 가이드라인의 연속성과 임의성이라는 특징을 바탕으로 각 지수의 기준값과 해석 방식을 재정리한 가이드라인의 예시를 제공한다. 해당 가이드라인을 통해 효과크기 지수를 연속적으로 해석함에 따라 기존에는 배제되었던 모형이 유용한 모형으로 평가될 수 있는 상황에 대해 논의하고 적합도 평가도구의 올바른 해석 방식의 중요성을 재고한다.

이와 같은 목적을 달성하기 위하여 본 연구에서는 우선 적합도의 첫 번째 평가단계에 해당하는 χ^2검정 과정에 대해 간략히 리뷰한다. 이후, χ^2검정의 대안으로 제시된 적합도 지수를 적합도의 효과크기 지수로 적용하는 것에 대한 가능성과 함께 다양한 종류의 효과크기 지수들을 정리하여 제시한다. 마지막으로, 적합도의 효과크기를 해석하기 위한 기준값과 가이드라인의 올바른 예시 및 사용 방식을 논하고, 실제 연구에서 효과크기의 연속성을 반영하지 못함으로 인해 발생하는 잘못된 적합도 평가 사례를 확인한다. 특히 아주 근소한 차이로 기준값을 만족하지 못하는 모형(marginal fit model)에 대하여 연속성이 반영된 가이드라인을 이용해 해석할 경우 모형의 유용성과 무용성에 대한 결과가 다르게 나타날 수 있음을 실제 사례를 통해 제시한다.

적합도의 통계적 유의성

적합도의 효과크기 개념에 대해 본격적으로 논의하기 전, 구조방정식 발전 초기 대표적인 모형 평가 방법으로 사용되었던 χ^2검정을 이용해 적합도의 통계적 유의성을 확인하는 과정을 검토한다. 또한, 실제 연구에서 표본크기와 관련된 χ^2검정 결과의 유용성에 대해 논의하고 검정 결과가 적합도 평가 과정에서 갖는 의미에 대하여 재고한다.

완전 적합가설의 검정

공분산 구조 분석(covariance structure analysis)의 맥락에서 Jöreskog(1969)가 소개한 χ^2검정은 수집된 자료의 공분산 행렬과 추정된 공분산 행렬(모형 함의 공분산 행렬) 간의 차이, 즉 적합도의 통계적 유의성을 평가한다. χ^2검정의 영가설은 $\Sigma = \Sigma(\theta)$로, 모집단의 공분산 행렬(Σ)과 모수 기반의 모형 함의 공분산 행렬($\Sigma(\theta)$) 간 차이가 없음을 가정한다. 이때 모집단의 수준에서 두 행렬 간 차이는 직접적으로 계산할 수 없기 때문에, 자료와 모형의 차이는 표본의 수준에서 두 공분산 행렬 간 차이를 최소화하는 합치함수 $F(S,\Sigma(\hat{\theta}))$을 통해 추정된다. 다양한 종류의 합치함수 가운데 가장 대표적으로 사용되는 최대우도 (maximum likelihood, ML) 합치함수 F_{ML}은 아래와 같다.

$$F_{ML} = \log|\Sigma(\hat{\theta})| + tr(S\Sigma^{-1}(\hat{\theta})) - \log|S| - p \quad (1)$$

위에서 S와 $\Sigma(\hat{\theta})$은 각각 표본의 공분산 행렬과 추정된 공분산 행렬을, $\hat{\theta}$은 추정치 벡터를, p는 변수의 개수를 의미한다. 식 1을 통해 계산된 F_{ML}을 이용해 χ^2검정통계량 T_{ML}을 아래와 같이 계산할 수 있다.

$$T_{ML} = (n-1)F_{ML} \text{ 또는 } nF_{ML} \quad (2)$$

F_{ML}의 계산에 사용되는 표본의 공분산 행렬 S가 어떤 분포를 따르는가에 따라 F_{ML}에 n을 곱할지 $(n-1)$을 곱할지가 달라진다. 예를 들어, Mplus의 경우 S가 다변량 정규분포를 따른다고 가정함에 따라 F_{ML}에 n을, LISREL의 경우 S가 Wishart 분포를 따른다고 가정(Hayduk, 1987)함에 따라 F_{ML}에 $(n-1)$을 곱한다. 두 계산 방식은 표본크기가 증가함에 따라 점근적으로 유사한 결과를 제공한다. 식 1에서 표본크기가 충분히 크고, 내생변수들이 다변량 정규성을 만족하며, 모형이 올바르게 설정되었다는 가정 하에 T_{ML}은 점근적으로 χ^2분포를 따른다.

앞서 언급한 바와 같이, χ^2검정의 영가설 $(\Sigma = \Sigma(\theta))$은 모형이 자료를 완벽하게 반영하는 완전 적합을 가정한다. 이는 중요한 함의점을 가지고 있는데, χ^2검정의 기각이 연구자가 설정한 모형이 자료를 설명하는 데 무조건 실패했음을 가리키는 것이 아니라는 것이다. 다만 χ^2검정의 기각은 모형이 자료를 완벽하게 설명하고 있지는 않음을 의미한다. 즉, χ^2검정은 연속선 상에 놓인 적합도의 여러 수준 가운데 완전 적합이라는 하나의 기준에 대한 이분법적 판단을 통해 적합도를 평가하는 통계적 기법인 것이다. 모형 적합도 평가 도구로 소개된 이래 여러 연구에서 적합도 평가 시 반드시 χ^2검정 결과를 보고할 것을 지속적으로 제안하였으며(Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne, 2007; Kline, 2016; Markland, 2007), 실제로 본 연구에서 검토한 250개의 연구 가운데 218개의 연구가 χ^2검정의 결과를 보고하였다.

χ^2검정과 표본크기

연구자가 적합도 평가를 위해 통계적 검정을 이용할 경우, 결과에 대한 해석은 검정의 영가설에 대하여 이루어져야 한다. 이는 곧 χ^2검정 결과의 해석이 완전 적합을 기준으로 이루어져야 함을 의미한다. 그러나 χ^2검정 결과를 보고한 218개의 연구 가운데 기각된 검정 결과에 대하여 모형이 자료에 완벽하게 적합하지는 않음을 설명하는 연구는 찾을 수 없었다. 나아가, χ^2검정의 p값조차 보고하지 않은 사례도 많아, 전반적으로 χ^2검정 결과에 대한 보고와 해석이 제대로 이루어지고 있지 않음을 확인할 수 있었다.

χ^2검정의 완전 적합 영가설 기각 결과를 무시하는 근거들을 제시한 연구들 가운데 대다수는 표본크기를 그 근거로 제시하였다. χ^2검정에 사용되는 검정 통계량 T_{ML}은 점근적으로 χ^2분포를 따르며, 이에 따라 표본크기가 충분히 크다면 검정 통계량이 χ^2분포를 따르지만 표본크기가 작을 경우 χ^2분포를 따르지 않을 수 있다. 또한, 표본크기가 클 경우 검정 통계량이 커짐에 따라 모형과 자료의 실제 차

이가 작더라도(즉, 합치함수 F_{ML}의 값이 작더라도) 모형은 기각될 수 있다. 실제로 표본크기가 200 혹은 400만 넘어가도 대부분의 모형에서 χ^2검정은 기각되는 것으로 알려져 있다(Barrett, 2007; Kenny, 2020).

표본크기가 검정의 결과에 영향을 미치는 문제는 여러 연구에서 꼽은 χ^2검정의 대표적인 한계에 해당하며(Fan, Thompson, & Wang, 1999; Hu & Bentler, 1995; Marsh & Balla, 1994), 그에 따라 χ^2검정 결과를 그대로 받아들여서는 안 된다거나(Goffin, 2007), 혹은 χ^2검정을 대체할 수 있는 다른 종류의 통계적 검정을 제안하는 연구(Browne & Cudeck, 1993; MacCallum, Browne, & Sugawara, 1996)들이 발표되었다. 그러나 χ^2검정이 t검정, 혹은 F검정과 같은 일반적인 통계적 검정의 종류 중 하나임을 생각했을 때, 표본크기의 문제는 χ^2검정뿐 아니라 일반적인 통계적 검정 전반에서 나타나는 대표적인 한계점이다(Thompson, 1996). 통계적 검정이 갖는 한계들을 언급한 다양한 연구들은(Kirk, 1996; Meehl, 1967; Tukey, 1991; Wilkerson & Olson, 1997) 영가설의 비현실성 또는 유의수준을 이용한 이분법적 판단의 문제와 함께 표본크기가 검정 결과에 영향을 미치는 문제를 언급하였으며, 특히 Fan(2001)은 통계적 검정이 갖는 여러 문제 가운데 가장 대표적인 한계점으로서 표본크기를 꼽았다. 하지만 표본크기를 근거로 t검정이나 F검정, 혹은 통계적 검정 자체가 의미 없다고 주장하는 연구는 없으며, 일반적으로 표본크기의 영향을 받지 않고 검정 결과를 해석할 수 있는 대안적인 도구(예, 효과크기)를 개발하는 방향으로 발전되었다.

나아가 χ^2검정의 경우 현재 모형 적합도를 평가하는 여러 방법 중 거의 유일하게 적합도의 통계적 유의성을 평가한다(Barrett, 2007). 가장 일반적으로 사용되는 적합도 지수 중 하나인 RMSEA(root mean square error of approximation)를 이용한 근사 적합(close fit) 검정이 있기는 하지만, 이를 제외한 TLI (Tucker-Lewis index), CFI(comparative fit index), SRMR(standardized root mean square residual) 등의 경우 RMSEA와 동일하게 비중심 χ^2분포를 이용하거나 새로운 분포 기반의 불편향 추정치(Maydeu-Olivares, 2017)가 제안되었음에도 아직 이를 이용해 적합도에 대한 통계적 검정을 진행하는 과정은 대중화되지 못하였다. 즉, 현재 χ^2검정은 적합도의 통계적 유의성을 확인할 수 있는 대표적인 평가 도구에 해당하며, 표본크기를 비롯한 몇 가지 문제들이 χ^2검정의 결과를 무시하고 제대로 보고하지 않는 근거가 될 수는 없다. χ^2검정을 통해 얻게 되는 통계적 유의성 결과에 더하여 모형과 자료 간 차이의 수준이라는 실질적 유의성에 대한 해석이 보충된다면 χ^2검정은 그 자체로 충분히 의미 있는 적합도 평가에 해당한다(Maydeu-Olivares, 2017; Steiger, 1989).

적합도의 실질적 유의성

연구자는 χ^2검정을 통해 완전 적합에 대한 통계적 유의성을 검정하고, 검정이 기각되면 적합도의 효과크기 지수를 통해 모형이 완전 적합으로부터 얼마나 떨어져 있는지에 대한 실질적 유의성을 확인함으로써 적합도를 종합적으로 평가할 수 있다. 그림 1은 이와 같은 적합도의 전반적인 평가 과정을 도식화하여 제시한다.

구조방정식 모형의 효과크기

실질적 유의성의 확인

일반적인 통계적 검정이 끝나고 통계적 유의성을 확인한 뒤, 연구자는 효과크기 지수와 같은 도구를 이용해 검정의 실질적 유의성에 대하여 확인한다. 통계적 유의성이 검증되었음에도 불구하고 추가적으로 실질적 유의성을 확인해야 하는 이유는 가설검증 결과가 표본크기의 영향을 받기 때문이다. 두 집단의 평균 차이가 동일하더라도, 표본크기에 따라 통계적 유의성의 결과는 달라진다. 만일 두 집단의 평균 차이가 매우 작음에도 불구하고 표본크기가 크다면, 표집분포의 표준오차 값은 줄어들며 결과적으로 검정통계량은 매우 큰 값으로 계산되어 통계적으로 유의한 결과를 제시할 확률이 올라간다. 그러나 이는 실질적인 유의미성을 의미하는 것이 아닌, 표본크기에 의해 왜곡된 통계적 유의미성에 지나지 않는다. 표본크기의 영향을 배제하고 두 집단 간의 실질적인 차이를 확인하기 위해서는 검정의 실질적 유의성을 확인할 필요가 있다.

완전 적합 가설 검정에 대한 효과크기

모형의 실질적 유의성을 해석하는 대표적 평가도구인 효과크기는 '해당 현상이 모집단에 존재하는 정도(the degree to which the phenomenon is present in the population)' 또는 '영가설이 잘못된 정도(the degree to which the null hypothesis is false)'(Cohen, 1988)를 나타내는 지수이다. 이와 같은 맥락에서 구조방정식 모형 적합도의 효과크기는 χ^2검정의 영가설인 완전 적합 가설이 잘못된 정도, 즉 모형의 적합도가 연속선상에서 완전 적합으로부터 떨어진 정도를 나타내는 개념으로 정의할 수 있다. 이는 곧 적합도의 수준과 효과크기의 부적 관계를 의미하는데, 구체적으로 모형과 자료 간 차이가 증가할수록 효과크기의 값은 커지게 되며, 그에 따라 적합도의 수준은 낮아지고, 연구자의 모형은 점점 지지할 수 없게 됨을 의미한다.

구조방정식 모형 적합도의 효과크기는 일반

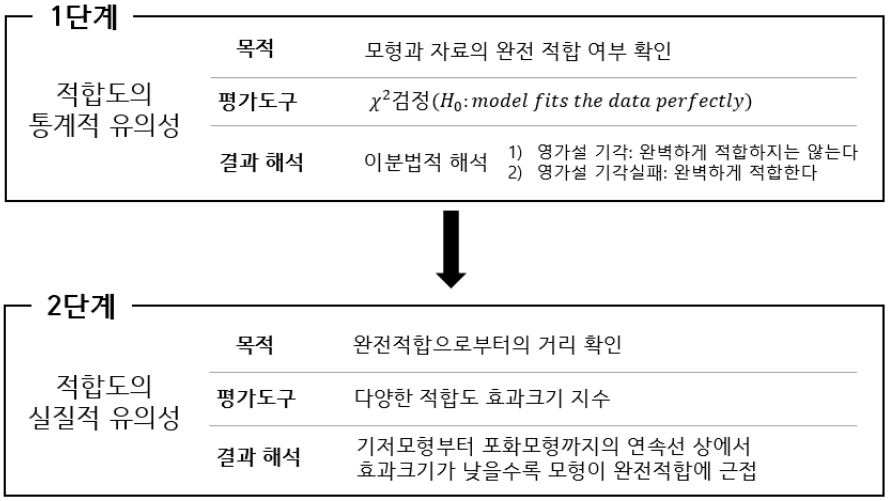

그림 1. 모형적합도의 전반적인 평가 과정

적으로 값이 커질수록 연구자가 주장하고자 하는 가설을 지지하는 전통적인 효과크기(예, Cohen의 d)와 달리 그 값이 작아질수록 모형을 지지한다는 점에서 구분되는 해석상의 차이를 갖는다. 이는 일반적인 차이 검정에서 사용하는 영가설과 달리 구조방정식 모형의 완전 적합 검정은 영가설을 기각하지 않아야 연구가설이 지지 되는 수용-지지 검정(accept-support test)[2]에 해당하기 때문이다(Kline, 2016). 완전 적합 가설을 기각하는 데 실패하는 것이 연구자의 모형을 지지하는 일이 되며, 완전 적합 가설에서 멀어짐에 따라 효과크기는 증가하고 모형의 설명력은 낮아진다.

적합도 지수를 이용한 효과크기의 평가

χ^2검정이 갖는 표본크기 등의 한계를 보완함과 동시에 이분법적 프레임을 벗어난 적합도의 평가를 위해 발전된 적합도 지수는 연속성이라는 특징을 바탕으로 적합도의 효과크기 지수로서 사용될 수 있다. 1980년대부터 최근까지 다양한 적합도 지수들이 제시되었으며, 어떠한 관점에서 적합도를 정의하는가에 따라 몇 가지 범주(예, 상대적 적합도 지수, 절대적 적합도 지수 등)로 분류될 수는 있으나, 모든 지수는 공통적으로 적합도의 수준을 연속적으로 확인한다는 점에서 효과크기와 동일한 목적을 갖는다.

적합도 지수와 일반적인 효과크기 지수 간의 유사성, 또는 적합도 지수에 효과크기의 개념을 적용할 수 있다는 주장은 이전부터 지속적으로 제기되었다(Hu & Bentler, 1999; Maydeu-Olivares, 2017; Yuan & Marshall, 2004). 적합도 지수와 효과크기는 자료와 모형 간의 관계, 또는 독립변수와 종속변수 간의 관계의 정도를 기술적으로 나타내며, 그와 동시에 모수를 기반으로 하는 분포(예, 비중심 χ^2분포) 하에서 점추정치 및 구간 추정치의 형태로 정의될 수 있다. 또한, Jöreskog와 Sörbom(1981)은 초기 적합도 지수인 GFI(goodness of fit index)를 일반 선형 모형의 효과크기 지수인 R^2과 유사한 역할을 하는 지수로 소개하였으며, 이후 GFI는 구조방정식 모형의 결정계수로 사용되기도 하였다(Tanaka & Huba, 1989). NFI(normed fit index), CFI(comparative fit index), TLI(Tucker-Lewis index) 역시 선형회귀분석에서의 R^2과 같은 역할을 하는 것으로 해석할 수 있다(Laar & Braeken, 2022). Hu와 Bentler(1999)도 적합도 지수는 R^2과 같이 자료의 분산 가운데 모형에 의해 설명된 분산의 양을 측정하는데 사용해야 하며, 적합도 지수를 통계적 검정의 도구처럼 사용하자는 의견(Maiti & Mukherjee, 1991)에 대하여 적합도 지수의 목적과 부합하지 않음을 주장했다.

반면, 적합도 지수를 이용해 모형의 효과크기를 확인하는 것이 절대적으로 불가능한 것은 아니지만 선호되지도 않는다는 주장 또한 제기되었다. 대표적으로 Gomer 등(2019)은 현재 적합도 지수 가이드라인의 기준값이 가설 검정의 맥락에서 2종 오류의 통제(Hu & Bentler, 1999)를 위해 설정된 값에 불과하며, 또한 자료가 정규성을 만족하지 못하거나 모형 조건이 달라짐에 따라 적합도 지수가 편향될 수 있다는 문제점 등을 이유로 적합도 지수를 효과크기와 같은 개념으로 보는 것이 적절하지 않음을 주장하였다. 그러나 자료의 비정규성 문제의 경우, 합치함수 F_{ML}을 추정하

2) accept-support test는 통계 철학적으로 옳지 않은 표현이지만, 실질적으로 구조방정식 모형 적합도 검정의 특성을 잘 표현하고 있다.

는데 사용되는 최대우도 추정법이 본래 정규성 가정의 위반에 상당히 강건한 것으로 알려져 있으며(Schermelleh-Engel et al., 2003), 또한 현재 Mplus와 EQS 등의 통계 프로그램이 다양한 적합도 지수의 추정에 이용되는 검정 통계량 T_{ML}을 $T_{SB}(=\frac{T_{ML}}{\hat{c}}$, $\hat{c}$는 자료의 비정규성 수준을 고려한 척도화 계수)로 대체함으로써 비정규성에 대한 교정을 적용할 수도 있어(Brosseau-Liard & Savalei, 2014), 자료의 비정규성 문제로 인해 적합도 지수를 효과크기 지수로 사용하지 못할 이유는 없다고 볼 수 있다. 비록 여러 시뮬레이션 연구에서 적합도 지수가 표본크기나 지표변수의 개수와 같은 모형 조건에 의해 영향을 받는다는 사실이 밝혀졌으나(Ding, Velicer, & Harlow, 1995; Fan, Thompson, & Wang, 1999; Kenny & McCoach, 2003; Marsh, Hau, Balla, & Grayson, 1998), 해당 연구들은 공통적으로 적합도 지수의 편향과 가이드라인을 맹신하는 관행에 대하여 경고했을 뿐 적합도 지수를 이용해 모형이 잘못 설정된 정도를 확인하는 행위 자체에 대해 의문을 제기하지는 않았다. 본래의 목적에 맞게 연속선 상에서 모형과 자료의 차이를 확인하는 도구로 사용한다면 적합도 지수는 충분히 구조방정식 모형 적합도의 효과크기 지수로서 활용될 수 있다.

적합도 효과크기 지수의 종류

전통적 효과크기 지수

RMSEA, SRMR, CFI와 같이 오래전부터 사용되었던 적합도 지수들의 경우, 적합도의 수준을 연속선 상에서 파악한다는 점에서 구조방정식 모형의 효과크기 역할을 담당할 수 있다. 자유도에 의해 조정된 모형과 자료 간의 거리를 나타내는 RMSEA는 완전 적합 가설이 옳지 않다는 가정 아래의 분포인 비중심 χ^2 분포의 비중심 모수(noncentrality parameter) λ를 이용해 추정된다. 표본크기가 지수에 미치는 영향에 대한 교정과 모형의 복잡성에 대한 페널티, 그리고 단위의 통일 등을 거쳐 최종적으로 RMSEA는 다음과 같이 정의 된다(Steiger, 1989).

$$RMSEA = \sqrt{\frac{\lambda}{df(n-1)}} \tag{3}$$

위에서 n은 표본크기, df는 모형의 자유도를 나타낸다.

식 3의 RMSEA는 모집단의 공분산 행렬 Σ를 이용하여 정의되는 모수에 해당하기 때문에 실제로 값을 구할 수는 없으며, 일반적으로 RMSEA를 이용한 적합도의 평가는 점 추정치인 $\widehat{RMSEA}$과 구간 추정치인 90% 신뢰구간을 통해 이루어진다.

$$\widehat{RMSEA} = \sqrt{\frac{\hat{\lambda}}{df(n-1)}} = \sqrt{\frac{\chi^2 - df}{df(n-1)}} \tag{4}$$

RMSEA는 모형과 자료 간 차이의 크기를 나타내는 데 있어 추정치의 단위가 표준화 되어있지 않으며 원변수의 단위를 그대로 이용함에 따라(Maydeu-Olivares, Shi, Rosseel, 2018) 전통적 효과크기 지수 가운데에서도 대표적인 비표준화 지수에 해당한다. 비록 비표준화 지수의 경우 추정된 값의 해석이 모형의 구조와 크기에 따라 달라지는 한계를 갖지만(Chen, Curran, Bollen, Kirby, & Paxton, 2008; Kline,

2016), 그럼에도 다양한 통계 프로그램에서는 RMSEA의 추정치를 이용한 근사 적합(close fit) 검정 결과를 제공하고 있으며, 현재 가장 대표적인 적합도 효과크기 지수 중 하나로 사용되고 있다.

SRMR은 Σ와 $\Sigma(\theta)$ 간 차이인 잔차 행렬을 이용해 적합도의 수준을 확인하는 지수로, 아래와 같이 정의된다.

$$SRMR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left\{(\sigma_{ij}-\sigma_{ij}^{0})/(\sigma_{ii}\sigma_{jj})^{\frac{1}{2}}\right\}^{2}}{p(p+1)/2}} \quad (5)$$

위에서 σ_{ij}와 $\sigma^{0}{}_{ij}$는 각각 Σ와 $\Sigma(\theta)$의 요소를 의미하며, p는 변수의 개수를 의미한다. σ_{ij}와 $\sigma^{0}{}_{ij}$를 각각의 표준편차로 나눠주는 표준화 과정을 통해 SRMR은 자료와 모형 간의 차이를 표준화된 값으로 나타낸다.

적합도의 실제 평가에 사용되는 추정치 $\widehat{SRMR}$의 경우 S와 $\Sigma(\hat{\theta})$의 요소인 s_{ij}와 $\hat{\sigma}_{ij}$을 이용해 아래와 같이 정의된다.

$$\widehat{SRMR} = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left\{(s_{ij}-\widehat{\sigma_{ij}})/(s_{ii}s_{jj})^{\frac{1}{2}}\right\}^{2}}{p(p+1)/2}} \quad (6)$$

SRMR은 대표적인 표준화 효과크기 지수로, RMSEA와 비교하여 '잔차에 대한 표준화된 공분산 행렬의 평균', 혹은 '표준화된 효과크기의 평균' 등의 정의와 함께 추정치의 크기를 직접적으로 해석할 수 있다는 장점이 있다. 다만, 표준화된 효과크기 지수라 해서 SRMR이 반드시 0에서 1 사이의 값을 갖는 것은 아니다. 실제 연구에서는 대부분의 SRMR 값이 1 이하의 값을 나타내지만, 이론적인 SRMR의 범위는 1 이상의 값을 가질 수 있다(West et al., 2012). 또한 표준화 된 자료의 공분산 행렬과 모형 함의 공분산 행렬 간의 차이를 나타내는 SRMR의 정의를 고려했을 때, 일반적으로 추정된 값이 2를 넘는 것은 이론적으로 불가능할 것으로 예상할 수 있다.

CFI는 변수 간 관계를 설정한 연구모형이 변수 간 어떠한 관계도 존재하지 않는 영모형(null model)에 비하여 자료를 얼마나 더 잘 설명할 수 있게 되었는지를 나타내는 상대적 효과크기 지수 가운데 가장 범용적으로 사용되는 지수이다. 상대적 효과크기 지수는 증분적합도 지수라고도 하는데, 일반적인 증분 적합도 지수의 모수 Δ는 아래와 같이 정의된다(Bentler, 1990).

$$\Delta = 1 - \frac{\lambda_M}{\lambda_N} \quad (7)$$

위에서 λ_M은 연구모형의 비중심 모수, λ_N은 영모형의 비중심 모수를 가리킨다. λ의 값이 커진다는 것은 모형이 자료를 제대로 설명하지 못함을 의미하며, 이에 따라 λ_N에 비하여 λ_M이 작아질수록 Δ는 커지게 된다. CFI는 Δ의 다양한 추정치 가운데 값의 범위를 0에서 1 사이로 고정하여 구하는 추정치로, 아래와 같이 정의된다.

$$\begin{aligned} CFI &= 1 - \frac{\widehat{\lambda_M}}{\widehat{\lambda_N}} \\ &= 1 - \frac{Max(\chi^{2_M} - df_M,\ 0)}{Max(\chi^{2_M} - df_M,\ \chi^{2_N} - df_N)} \end{aligned} \quad (8)$$

CFI는 부스트래핑을 이용해 신뢰구간을 추정하는 것이 가능하며(Cheng & Wu, 2017; Lai, 2019; Zhang & Savalei, 2016), 현재 TLI와 함께 대표적인 상대적 효과크기 지수로 사용되고 있다. 다만, TLI와 CFI는 상관이 높기 때문에 두 지수 중 하나만 보고할 것이 제안된다(Kline, 2016). 또한, CFI는 프로그램에 따라 조금 다른 값이 제시되기도 하는데, 이는 각 프로그램이 정의하는 영모형이 다르기 때문이다(예, Mplus와 EQS).

새로운 효과크기 지수

최근 모형의 적합도를 효과크기의 관점에서 평가해야 한다는 주장들(Gomer et al., 2019; Maydeu-Olivares, 2017; Maydeu-Olivares & Shi, 2017)과 함께 새로운 종류의 적합도 효과크기 지수들이 제안되었다. 기존의 전통적 효과크기 지수들과 비교하여 새로운 지수들은 한층 더 발전된 형태의 추정치를 가지며, 효과크기의 성격을 잘 나타내고 있다. 대표적으로 Maydeu-Olivares(2017) 및 Maydeu-Olivares와 Shi (2017)는 전통적 효과크기 지수인 $\widehat{SRMR}$이 모수 SRMR을 특히 작은 표본에서 과대추정하고 있음을 밝히며, 이에 따라 모수에 대한 불편향 추정치인 $\widehat{SRMR}_u$(unbiased SRMR)을 새롭게 제안하였다. 대부분의 통계 소프트웨어에서 식 6을 통해 추정하는 $\widehat{SRMR}$의 경우 일반적으로 실제 모수를 과대추정하고 있으며, 이에 따라 소프트웨어를 통해 추정된 모형의 $\widehat{SRMR}$은 실제보다 낮은 수준의 적합도를 나타낸다(Shi, Maydeu-Olivares, & DiStefano, 2018). 이와 같은 문제를 해결하기 위하여 Maydeu-Olivares(2017)는 정규분포 하에서 정의되는 $\widehat{SRMR}_u$을 이용해 적합도의 효과크기를 확인할 것을 새롭게 제안하였다. 모수 SRMR에 대한 불편향 추정치 $\widehat{SRMR}_u$은 아래와 같이 정의된다.

$$\widehat{SRMR}_u = \widehat{k^{-1}}\sqrt{\frac{\max(e'e - tr(\hat{\Xi}),0)}{t}} \quad (9)$$

위에서 $\sqrt{\frac{\max(e'e - tr(\hat{\Xi}),0)}{t}}$는 잔차의 제곱합을 의미하는 $e'e$의 기댓값을 이용하여 구한 SRMR의 추정치로, e는 자료와 모형 간의 표준화된 잔차를, $\hat{\Xi}$은 e의 공분산 행렬을 나타내며, t는 공분산 행렬의 독립적인 정보의 개수를 의미한다. $\hat{k}^{-1}$은 식 10과 같이 추정되며, 식 9를 통해 얻은 SRMR의 추정치가 편향되지 않도록 조정 해주는 역할을 한다.

$$\widehat{k^{-1}} = 1 - \frac{tr(\widehat{\Xi_s^2}) + 2e_s\hat{\Xi}e_s}{4(e_s'e_s)^2} \quad (10)$$

$\widehat{SRMR}_u$의 경우 정규분포를 바탕으로 신뢰구간을 추정할 수 있으며, 근사 적합에 대한 검정 역시 가능하다. 또한, $\widehat{SRMR}_u$이 새롭게 제안된 이후 시뮬레이션을 통해 $\widehat{SRMR}_u$을 이용해 구하는 구간 추정치와 근사 적합 검정 결과가 RMSEA를 통해 얻는 결과보다 더 정확하다는 연구(Maydeu-Olivares, Shi, & Rosseel, 2018)가 제시됨에 따라, $\widehat{SRMR}_u$이 적합도의 효과크기 수준을 나타내는 데 비교적 우월한 지수일 가능성이 확인되었다. 한편, Asparouhov와 Muthén(2018)은 $\widehat{SRMR}_u$이 Mplus에서 제공하는 $\widehat{SRMR}$과 다

른 방식으로 정의되어 있으며, $\widehat{SRMR}$은 큰 표본크기에서 사용될 수 있는 반면 $\widehat{SRMR}_u$의 경우 작은 표본크기에서의 SRMR의 추정에 핵심을 두고 있다도 주장하였다.

$\widehat{SRMR}_u$과 $\widehat{SRMR}$이 모두 SRMR에 대한 추정치로 사용되는 적합도 효과크기 지수라면, Gomer 등(2019)의 ϵ[3]는 두 집단 간의 평균 차이를 표준화한 값을 나타내는 Cohen의 d(Cohen, 1988)의 개념을 바탕으로 제안된 적합도 효과크기 지수이다. d는 검정의 종류에 따라 다양한 형태로 정의되는데, 이 가운데 독립표본 t검정의 맥락에서 d는 아래와 같다.

$$d = \frac{\overline{X_1} - \overline{X_2}}{s_p} \tag{11}$$

위에서 $\overline{X}_1$과 $\overline{X}_2$는 각 표본의 평균치를, s_p는 통합된 표준오차를 의미한다. 해당 수식을 모집단의 수준에서 정의할 경우 d의 모수 δ는 아래와 같이 정의된다.

$$\delta = \frac{\mu_1 - \mu_2 - 0}{\sigma_p} \tag{12}$$
$$= \frac{E[\overline{X_1} - \overline{X_2} \mid H_1] - E[\overline{X_1} - \overline{X_2} \mid H_0]}{SD}$$

위에서 $E[\overline{X}_1 - \overline{X}_2|H_1]$과 $E[\overline{X}_1 - \overline{X}_2|H_0]$는 각각 대립가설과 영가설 하에서의 표본 평균의 차이를 나타낸다. 구조방정식의 경우 기본적으로 다변량 구조를 따르고 있으며 공분산 행렬을 이용하기 때문에 두 표본 평균의 차이를 상수의 형태로 직접 변환할 수 있는 개념은 존재하지 않으나, 그 대신 자료와 모형의 차이를 나타내는 F_{ML} 혹은 T_{ML}을 이용하여 해당 개념을 대신할 수 있다. T_{ML}을 이용하여 식 12를 구조방정식의 맥락에 맞게 변형한 형태는 아래와 같다(Gomer et al., 2019).

$$\epsilon = \left(\frac{E[T_{ML}|H_1] - E[T_{ML}|H_0]}{N-1} \right)^{1/2} \tag{13}$$

ϵ는 Gomer 등(2019)이 d의 개념을 바탕으로 제안한 다양한 효과크기 지수 가운데 가장 표본크기의 영향을 적게 받음과 동시에 모형과 자료 간 차이를 제대로 탐지하는 것으로 밝혀졌다. ϵ의 추정치인 $\hat{\epsilon}$은 식 14와 같이 구할 수 있으며, 부스트래핑을 이용해 추정되는 신뢰구간과 함께 사용할 것이 추천된다.

$$\hat{\epsilon} = \left(\frac{T_{ML} - \overline{T^*_{ML}}}{N-1} \right)^{1/2} \tag{14}$$

위에서 $\overline{T}^*_{ML}$는 부스트래핑을 이용해 추정되는 T_{ML}의 값에 해당한다(Yuan & Marshall, 2004). 표 1은 전통적 적합도 효과크기 지수와 새로운 적합도 효과크기 지수의 모수와 추정치, 그리고 각 지수가 따르고 있는 분포를 종합적으로 제시하고 있다. 현재 대다수의 통계 프로그램에서는 적합도 효과크기 지수의 추정치로 $\widehat{RMSEA}$, $\widehat{SRMR}$, 그리고 CFI를 제공하며, 대부분의 연구자들은 해당 지수를 가이드라인과 비교하여 모형의 적합도를 평가한다.

3) Gomer 등(2019)은 해당 연구에서 ϵ를 '입실론'이 아닌 영어 알파벳 'E'로 명명하였다.

표 1. 전통적 적합도 효과크기 지수와 새로운 적합도 효과크기 지수

모수	추정치	분포
$RMSEA = \sqrt{\frac{\lambda}{df(n-1)}}$	$\widehat{RMSEA} = \sqrt{\frac{\hat{\lambda}}{df(n-1)}} = \sqrt{\frac{\chi^2 - df}{df(n-1)}}$	비중심 χ^2분포
$SRMR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left\{(\sigma_{ij} - \sigma_{ij}^0)/(\sigma_{ii}\sigma_{jj})^{\frac{1}{2}}\right\}^2}{p(p+1)/2}}$	$\widehat{SRMR} = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left\{(s_{ij} - \hat{\sigma}_{ij})/(s_{ii}s_{jj})^{\frac{1}{2}}\right\}^2}{p(p+1)/2}}$ $\widehat{SRMR}_u = \hat{k}^{-1}\sqrt{\frac{\max(e'e - tr(\hat{\Xi}), 0)}{t}}$	- 정규분포
$\Delta = 1 - \frac{\lambda_M}{\lambda_N}$	$CFI = 1 - \frac{Max(\chi^{2_M} - df_M,\ 0)}{Max(\chi^{2_M} - df_M,\ \chi^{2_N} - df_N)}$	비중심 χ^2분포
$\epsilon = \left(\frac{E[T_{ML}\|H_1] - E[T_{ML}\|H_0]}{N-1}\right)^{1/2}$	$\hat{\epsilon} = \left(\frac{T_{ML} - \overline{T}^*_{ML}}{N-1}\right)^{1/2}$	-

주. 표에 제시된 분포는 각 적합도 효과크기 지수를 정의하는데 이용된 분포를 의미한다.

가이드라인을 이용한 적합도의 효과크기 해석

효과크기 지수를 바탕으로 적합도의 실질적 유의성을 평가하는 과정에서 연구자는 가이드라인을 통해 적합도의 수준을 연속적으로 해석해야 한다. 이는 곧, 가이드라인이 제시하는 기준값이 적합도의 좋고 나쁨에 대한 절대적인 절단점이 아닌, 연속선 상에서의 해석을 위한 지표로 사용되어야 함을 의미한다.

적합도 효과크기 지수 가이드라인의 특징

기준값의 연속성과 임의성

적합도의 효과크기는 모형과 자료 간 차이를 나타내는 모수로 정의되며, 모수에 대한 추정치를 구하여 그 수준을 평가한다. 연구자는 적합도 추정치와 관습적으로 제시하는 가이드라인의 기준값들을 비교하여 연속선 상에서 모형과 자료 간의 거리를 파악할 수 있다.

적합도 지수의 가이드라인을 포함하여 통계학에서 사용되는 대다수의 관습적인 가이드라인들의 기준값을 정의하는 과정에는 연속성과 임의성이 반영된다. 대표적인 효과크기 지수 d(Cohen, 1988)의 경우, 추정된 d값의 해석을 위한 가이드라인에 따라 d= .50이 중간 정도의 효과크기를 나타낸다고 해석한다. 이때 추정된 d값이 .49 혹은 .48이라고 해서 처치의 효과가 작다고 해석하는 경우는 없다. 독립표본 t검정에서의 d는 두 모집단 간 차이의 수준을 연속적으로 나타내는 지표이며, .50이라는 기준값은 임의적으로 결정된 값이기 때문이다. RMSEA를 이용한 근사 적합 검정의 경우 일반적으로 사용되는 기준은 .05이다. 그러

나 근사 적합의 개념 역시 연속적인 적합도 수준의 한 지점을 의미할 뿐이며, .05는 임의로 정해진 상수로 모형의 설정이나 표본크기에 따라 기준값은 변할 수 있다(Chen et al., 2008).

나아가 RMSEA, SRMR, CFI와 같은 대표적인 적합도 효과크기 지수들의 가이드라인이 제시하는 기준값의 경우, 대부분 엄격한 수리적 과정을 통해 도출된 것이 아닌 연구자의 오랜 경험을 기반으로 제시된 값에 해당한다. Browne과 Cudeck(1993)은 다양한 모형의 RMSEA를 추정한 결과 .05 이하의 값이 나오는 모형은 자료와 상당히 근접한 것으로 평가할 수 있음을 제안했다. Bentler와 Bonett(1980)은 TLI를 비롯한 여러 증분적합도 지수들의 초기 형태를 제안하며, 본인들의 경험을 바탕으로 .90이라는 기준값을 제시하였다. 참고로 TLI의 경우 본래 신뢰도 지수로서 소개되었는데, 신뢰도 지수의 가이드라인 역시 다양한 경험과 직관을 바탕으로 합의된 임의적인 기준에 해당한다(Helmstadter, 1964).

임의적으로 설정된 기준의 경우 통계적 기반이 부족하다는 한계를 갖고 있기 때문에(McDonald and Marsh, 1990; Marsh & Balla, 1994; Marsh, Hau, & Grayson, 2005), 1980~1990년대의 다양한 연구들은 시뮬레이션을 이용하여 가이드라인을 설정하고자 시도하였다. 현재 적합도 지수 평가 과정에서 대표적인 참고문헌으로 사용되고 있는 Hu와 Bentler(1999)는 잘못 설정된 모형(misspecified model)을 이용한 시뮬레이션을 통해 가이드라인을 제시하였다. 구체적으로 TLI, CFI는 .95 이상일 때, RMSEA는 06, SRMR은 .08 이하일 때 2종 오류, 즉 잘못 설정된 모형을 기각하지 못하는 확률이 낮아지는 시뮬레이션 결과를 보고하였다. 그러나 해당 가이드라인 역시 모든 모형에 적용 가능한 통일된 기준이라 할 수는 없다. 일반적으로 시뮬레이션 연구의 경우 특정 조건을 지닌 모형을 이용하기 때문에 모든 조건에 대하여 일반화된 규칙으로 사용하기 어렵기 때문이다(Kline, 2016; Sivo, Fan, Wittal, & Willse, 2006). 연구자의 모형이 Hu와 Bentler(1999)의 시뮬레이션에서 사용된 모형과 차이가 클수록 해당 가이드라인의 정확성과 유용성은 낮아질 수밖에 없다.

이와 같은 한계에도 불구하고 대부분의 연구자들이 몇몇 참고문헌(예, Browne & Cudeck, 1993; Hu & Bentler, 1999; Kline, 2016)에서 제공한 적합도 지수 가이드라인을 절대적인 규칙으로 사용하는 이유는 어떠한 평가 대상에 대하여 하나의 정해진 규칙이 있으면 판단하기에 더 용이하기 때문이다(Marsh, Hau, & Wen, 2004). 모든 조건에 대하여 동일하게 적용되는 기준이 있을 경우 연구자는 적합도를 평가할 때 자신의 모형이 어떠한 특징과 조건을 갖고 있는지 고민할 필요가 없는 것이다. 그러나 모형의 종류와 상관없이 모든 상황에 통용될 수 있는 단 하나의 규칙은 존재할 수 없으며(Fan et al., 1999), Hu와 Bentler(1999)를 포함한 여러 연구자는 가이드라인을 엄격하게 지키는 방향보다는 단순히 해석에 도움이 될 수 있는 보조적인 역할로 사용할 것을 강조하였다(Lai & Green, 2016).

모형 조건에 따른 기준값의 조정

적합도 효과크기 지수의 가이드라인을 모든 상황에 적용 가능한 절대적인 규칙으로 사용할 수 없는 또 다른 이유는, 모형의 조건에 따라 효과크기 지수의 추정치에 편향이 생길 수 있기 때문이다. CFI, RMSEA, SRMR을 포함

한 거의 모든 적합도 효과크기 지수들은 표본크기나 변수의 개수 등과 같은 모형 조건에 따라 모수에 대한 편향된 추정치를 제공할 수 있다. 예를 들어, 자료의 표본크기는 적합도 지수가 탄생한 배경과 직접적으로 연결되는 요인으로, 적합도 지수는 본래 χ^2검정과는 다르게 표본크기의 영향을 받지 않을 것이라는 믿음 아래 발전되었다(Bentler & Bonett, 1980; Jöreskog & Sörbom, 1981, 1984). 그러나 적합도 지수의 사용이 대중화되고 적합도 지수 추정치의 편향에 관한 연구들이 증가하면서 적합도 지수 역시 표본크기의 영향을 받는다는 결과들이 제시되었다. 대표적인 초기 적합도 지수 GFI(goodness of fit index)와 AGFI(adjusted goodness of fit index)의 경우 발전 당시 표본크기로부터 독립적이라고 가정되었으나(Jöreskog & Sörbom, 1984), 이후 진행된 시뮬레이션 연구(Anderson & Gerbing, 1984)는 두 지수의 수리적 계산과정 안에 표본크기가 반영되지 않을 뿐 분포 자체는 표본크기의 영향을 받으므로, 표본크기가 증가할수록 GFI와 AGFI도 함께 증가함을 밝혔다. 이는 곧 표본크기가 클 경우 좋은 수준의 적합도를 나타내기 위하여 GFI와 AGFI의 가이드라인이 기준값보다 더 높은 값으로 설정되어야 함을 의미한다.

나아가, Marsh 등(1988)은 χ^2/df, GFI, AGFI, NFI, TLI 등을 포함한 30개 가량의 초기 적합도 지수들의 표본크기에 대한 편향을 확인한 결과, TLI만이 상대적으로 표본크기에 독립적임을 확인하였다. Marsh와 Balla(1994)는 CFI와 동일한 모수를 추정(Goffin, 1993)하는 RNI (relative noncentrality index; McDonald, 1989)가 비교적 표본크기에 독립적임을 발견하였으며, Fan 등(1999)은 TLI, CFI, RMSEA가 상대적으로 표본크기로 인해 발생하는 편향이 작음을 확인하고 해당 지수들을 중점적으로 이용할 것을 추천하였다. 다만 Curran 등(2003)의 경우 200 이하의 표본크기에서는 RMSEA 점추정치가 과대 추정되는 경향이 있음을 밝혀, 작은 표본크기의 자료에 대하여 모형을 추정하는 연구자들은 RMSEA 값을 해석할 때 기존의 가이드라인이 상대적으로 엄격한 기준이 될 수 있다.

적합도 지수의 편향에 대한 초기 연구가 표본크기를 중심으로 진행되었다면, 모형의 크기(model size)가 미치는 영향에 관한 연구는 90년대 중반까지 상대적으로 적은 비중을 차지하였다(Ding, Velicer, & Harlow, 1995). 확인적 요인분석 모형에서 전체 지표변수의 개수, 요인 당 지표변수의 개수, 혹은 자유도 등으로 정의(Shi, Lee, & Terry, 2017)되는 모형 크기의 효과란, 특히 작거나 중간 정도의 표본크기에서 모형의 크기가 증가할수록 T_{ML}이 정적으로 편향되며 1종 오류가 증가하는 현상을 의미한다(Herzog, Boomsma, & Reinecke, 2007). T_{ML}에 편향이 발생함에 따라 T_{ML}을 이용해 추정되는 TLI, CFI, RMSEA에도 모형 크기가 영향을 미치는데, 구체적으로 작은 표본크기(예, 200 이하)의 조건 아래에서 지표변수의 개수가 증가할수록 올바르게 설정된 모형임에도 불구하고 TLI와 CFI는 좋지 않은 적합도를, RMSEA는 반대로 좋은 적합도를 보여준다(Kenny & McCoach, 2003). 이처럼 모형 크기 조건이 다르게 작동하는 이유 중 하나는 RMSEA가 TLI나 CFI와 다르게 추정과정에서 모형과 자료 간 차이를 나타내는 χ^2-df를 df로 나누어주기 때문이다. 이와 같은 과정은 모형의 복잡성에 대한 페널티를 부여하는데, 일반적으로 지표변수의 개수가 증가할수록 자유도도 함께 증가함에 따라(Shi, Lee, &

Terry, 2017) RMSEA는 작은 값을 나타내게 된다. 반대로 자유도가 감소할 경우 RMSEA는 정적으로 편향되어 좋지 않은 적합도를 나타내기 때문에 기존의 RMSEA 가이드라인(예, Browne & Cudeck, 1993)이 과도하게 엄격한 기준이 될 수 있다(Kenny, Kaniskan, & McCoach, 2015). 이와 같은 연구결과들은 모형 크기의 효과로 인하여 적합도 지수의 편향이 발생할 경우 연구자는 일반적인 가이드라인보다 다소 조정된 값을 기준으로 적합도의 효과크기를 해석할 필요가 있음(Moshagen, 2012)을 시사한다. 나아가 현재 범용적으로 사용되고 있는 가이드라인(예, Hu & Bentler, 1999)을 근거로 적합도의 좋고 나쁨을 절대적으로 평가하는 것은 적절한 평가 방식이 아닐 수도 있음을 함의한다.

앞에서 언급된 연구결과들의 경우 애초에 모형 크기의 영향을 받는 T_{ML}을 통해 추정된 적합도 지수 위주로 제시된 반면, SRMR은 잔차 행렬을 이용해 추정되기 때문에 T_{ML}의 편향이 SRMR의 편향으로 이어진다고 보기 어렵다. 그러나 실제로 모형의 df가 감소함에 따라 SRMR도 함께 감소하거나(Taasoobshirazi & Wang, 2016), 작은 표본크기에서 SRMR이 좋지 않은 적합도를 나타내는 경향이 지표변수의 개수가 증가함에 따라 더욱 강해지는(Ximénez, Maydeu-Olivares, Shi, & Revuelta, 2022) 등의 연구결과들을 확인하였을 때, SRMR 역시 모형 크기의 영향으로부터 자유롭지 않음을 알 수 있다.

적합도 효과크기 지수의 연속적 해석

효과크기 가이드라인의 올바른 예시

적합도 효과크기 지수들의 해석을 위한 가이드라인은 그 기준값 자체가 경험적인 배경을 바탕으로 설정됨에 따라 동일한 지수에 대하여 학자마다 조금씩 다른 값을 제안한다. 그럼에도 불구하고 대부분의 가이드라인은 각 효과크기 지수에 대하여 어느 정도 유사한 기준을 지니며, 표 2는 이 가운데 현재 모형 적합도의 보고에 가장 대중적으로 사용되고 있는 기준값들을 제시하고 있다.

적합도 효과크기 지수는 결과를 연속적으로 해석한다는 점에서 χ^2검정과 매우 큰 평가 방법의 차이를 보인다. 이분법적인 해석을 도출하는 통계적 검정과 달리 효과크기 지수는 결과의 해석에 연속성이 반영되어야 한다. 안타깝게도 대다수의 연구 상황에서 적합도의 효과크기는 이분법적으로 해석되고 있는데(Gomer et al., 2019; Lai & Green, 2016), 그 대표적인 원인 중 하나는 가이드라인을 처음 제시하는 과정 자체에서 기준값이 채택 가능한 모형 적합도의 최저 기준, 또는 절단점으로서 제안되기 때문이다. 예를 들어, 적합도 지수를 해석하는 가이드라인으로 가장 많이 사용되고 있는 Browne과 Cudeck(1993)의 연구(RMSEA ≤ .05) 및 Hu와 Bentler(1999)의 연구(CFI≥ .95, RMSEA≤ .06, SRMR≤ .08)의 경우, 모두 기준값을 제시하는 과정에서 적합도 지수가 특정 값 '이상' 또는 '이하' 등의 기준을 제시하였다. 이후, 모형 적합도에 대한 리뷰 연구를 진행한 Schermelleh-Engel 등(2003) 및 Hooper 등(2008) 역시 마찬가지로 RMSEA≤ .10이 적절한 적합도를 나타낸다는 이분법적 기준을 제시하였는데, 이와 같은 가이드라인을 이용해 모형을 평가할 경우 RMSEA=.11과 같이 .10의 기준값을 근소하게(marginally) 달성하지 못하는 모형을 단순히 나쁜 모형으로 해석하는 상황이 발생할 수 있다. 그러나 RMSEA의 가이드

표 2. 대표적으로 사용되고 있는 적합도 효과크기 지수의 가이드라인

적합도 효과크기 지수 추정치	가이드라인		
	기준값	해석	출처
$\widehat{RMSEA} = \sqrt{\frac{\hat{\lambda}}{df(n-1)}} = \sqrt{\frac{\chi^2 - df}{df(n-1)}}$	.10	good fit	Steiger(1989)
	.05	very good fit	
	.01	outstanding fit	
	.06	good fit	Hu와 Bentler(1999)
	.08	reasonable model	Browne과 Cudeck(1993)
	.05	close fit	
$\widehat{SRMR} = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}\left\{(s_{ij} - \hat{\sigma}_{ij})/(s_{ii}s_{jj})^{\frac{1}{2}}\right\}^2}{p(p+1)/2}}$	.08	good fit	Hu와 Bentler(1999)
	.10	acceptable fit	Schermelleh-Engel 등 (2003)
	.05	good fit	
$\widehat{SRMR}_u = \hat{k}^{-1}\sqrt{\frac{\max(e'e - tr(\hat{\Xi}), 0)}{t}}$	$.10 \times \overline{R^2}$ [4]	acceptable fit	Shi 등(2018)
	$.05 \times \overline{R^2}$	close fit	
$CFI = 1 - \frac{Max(\chi^{2_M} - df_M,\ 0)}{Max(\chi^{2_M} - df_M,\ \chi^{2_N} - df_N)}$	.90	.90 이하의 모형은 발전 필요	Bentler와 Bonett(1980)
	.95	good fit	Hu와 Bentler(1999)
	.95	acceptable fit	Schermelleh-Engel 등 (2003)
	.97	good fit	
$\hat{\epsilon} = \left(\frac{T_{ML} - \overline{T}^*_{ML}}{N-1}\right)^{1/2}$	.82	large effect size	Gomer 등(2019)
	.60	medium effect size	
	.42	small effect size	

라인에서 제시하는 .10, .08, .05와 같은 값들은 연속선 상의 한 지점에 해당할 뿐이며, 이 값을 가까스로 만족했다고 하여 연구모형이 적합하다고 평가하고 미세하게 만족하지 못했다고 해서 부적합하다고 평가하는 것은 적절하지 않다.

효과크기 지수의 연속성을 직관적으로 이해하기 위하여 그림 2와 같은 가상의 연속선 상에 표시된 각 지수의 가이드라인과 효과의 크기 예시를 제공하였다. 그림 2는 최악의 적합을 의미하는 기저모형부터 완전 적합을 의미하는 포화모형까지 가상의 연속선 상에서 근소하게 기준값을 만족하지 못하는 값(marginal value)의 상대적 위치를 나타낸다. 포화모형은 자료에 대한 모형의 완전 적합을 나타내는 모형으로, 기저모형은 자료를 거의 설명하지 못하는 수준의 적합을 나타내는 모형으로 해석할 수 있다. 적합도 효과크기의 정의에 따라

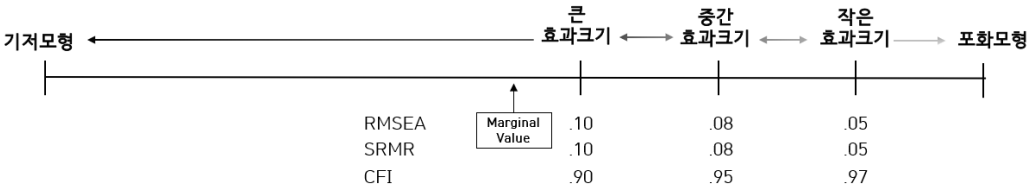

그림 2. 연속선 상에서 근소하게 기준을 만족하지 못하는 적합도 수준의 상대적 위치

효과크기 값이 작을수록 모형이 자료를 잘 설명하고 있음을 고려하였을 때, 적합도의 수준이 완전 적합에 가까워질수록 추정된 효과크기는 작아진다. 만일 연구모형의 RMSEA 값이 .11일 경우, 이는 큰 효과크기의 기준인 .10을 초과하게 된다. 그러나 그림 2를 바탕으로 해당 모형을 해석할 때 모형의 효과크기는 큰 효과크기에서 멀리 떨어져 있는 것도 아니며, .10을 초과했다고 해서 기저모형 쪽에 매우 가까이 위치한 것도 아니다.

이와 같은 연속적인 관점에서의 해석 방식을 가이드라인에 더욱 잘 반영하기 위해서는 가이드라인의 기준값을 적합도의 마지노선으로 인식하는 것이 아닌, 효과크기의 수준을 나타내는 연속선 상의 한 지점으로 이해하고 사용할 필요가 있다. 가이드라인 자체에서 'RMSEA=.10은 큰 효과크기를 나타낸다'고 말하는 해석방식을 제시할 경우 연구자는 .11이라는 RMSEA 값을 가진 모형이 .10의 값을 가진 모형과 설명력의 측면에서 멀리 떨어져 있지 않다고 해석할 수 있다. 반대로 특정 적합도 효과크기 지수가 기준값을 반드시 만족해야 하는 것처럼 가이드라인을 제시할 경우(예, χ^2검정이 기각되고 $SRMR > .08$일 경우 poor fit으로 정의[Asparouhov&Muthén, 2018]), 효과크기의 연속성에 대해 제대로 이해하고 있지 못한 연구자는 이를 절대적 기준으로 받아들이게 되며, 이는 가이드라인의 잘못된 사용방식으로 이어진다(Chen et al., 2008; Kenny, 2015; Kline, 2016; Markland, 2007; West et al., 2012).

적합도 효과크기 지수가 이분법적으로 해석되는 또 다른 원인으로는 기존의 가이드라인들이 기준값에 대한 해석을 하는 과정에서 '좋은 적합도(good fit)', '충분한 적합도(adequate fit)', 또는 '허용 가능한 적합도(acceptable fit)'등과 같은 표현을 이용하기 때문이다. 이와 같은 해석은 적합도의 좋고 나쁨, 혹은 적절성을 질적으로 판단하는 방식으로, 이는 곧 적합도에 대한 이분법적 평가로 이어지게 된다. 하지만 적합도 효과크기 지수들을 이용해 적합도를 평가한다는 것은 모형과 자료 간의 차이를 양적으로 해석하는 것을 의미한다. 일반적으로 대다수의 효과크기 지수들은 연속선 상에서 크기에 대한 정의를 내리기 위해 관습적으로 '작은', '중간 정도의', '큰' 등과 같은 표현들을 이용하였으며(Cohen, 1988), Gomer 등(2019)은 실제로 이와 같은 방식을 이용하여 적합도 효과크기 지수 ϵ를 해석하였다.

'좋은(Good)', 혹은 '적절한(adequate)'과 같은 질적 판단이 아닌 '작은(small)', '중간의(medium)', '큰(large)'과 같은 양적 평가를 기반으로 이루어지는 해석의 경우, 기존의 표현과

4) $\overline{R^2}$=내생변수들의 평균적인 R^2

표 3. 연속성을 반영한 효과크기 지수 가이드라인

적합도 효과크기 지수		효과크기 기준값		
		작은 효과크기	중간 정도의 효과크기	큰 효과크기
RMSEA		.05	.08	.10
SRMR	$\widehat{SRMR}$	.05	.08	.10
	$\widehat{SRMR}_u$	$.05\times\overline{R^2}$	$.10\times\overline{R^2}$	-
ϵ		.42	.60	.82
CFI		.97	.95	.90

주. RMSEA, SRMR, ϵ은 모수이지만 CFI는 추정치에 해당. CFI의 모수 Δ의 경우 다른 지수들과 달리 일반적으로 모수보다 추정치의 형태로 보고되기 때문에 이와 같이 표기함.

비교하여 모형과 자료 간 차이를 한층 더 연속적인 관점에서 나타낸다. 표 3은 각 효과크기 지수가 연속선 상에서 작은 효과크기, 중간 효과크기, 큰 효과크기를 나타내는 지점을 제시한 가이드라인으로, 기준값은 표 2에서 언급된 다수의 대표적인 가이드라인을 통합한 결과이다. 앞에서 언급한 바와 같이 완전 적합 검정의 경우 검정 결과에 대한 효과크기 값이 작을수록 높은 수준의 적합도를 의미하며, 각 지수의 적합도에 대한 정의에 따라 높은 수준의 적합도를 가진 모형에 대하여 RMSEA와 SRMR, ϵ는 작은 값을, CFI는 큰 값을 이용해 작은 효과크기를 나타내게 된다. 표 3을 이용하여 적합도의 효과크기 지수를 해석할 경우 연구자는 앞서 언급된 기준값의 절단적 해석 문제를 해결함과 더불어 적합도의 수준을 양적으로 표현하는 것이 가능하다.

근소하게 기준을 만족하지 못하는 적합도(marginal fit)에 대한 평가

효과크기 지수를 해석하는 과정에서 연속성의 중요성은 모형에 대한 실질적인 평가를 내릴 때 더욱 강조된다. 효과크기 지수를 연속적으로 해석하는지 아닌지에 따라 모형의 유용성에 대한 평가가 달라지기 때문이다.

표 4에 제시된 연구들(Heller et al., 2021; Thomsen & Lessing, 2020; Tyler et al., 2020; Yang & McGinley, 2021)은 적합도 평가 과정에서 추정된 효과크기 지수 값이 그 기준을 근소하게 만족하지 못할 경우(marginal fit) 적합도에 대해 질적으로 부정적인 평가를 제시하며 모형을 수정하거나 배제한 사례이다. 이 가운데 Thomsen과 Lessing(2020)의 경우 적합도 효과크기 지수 값을 추정하기 전 이미 가이드라인을 충족하는 모형은 배제할 것을 연구에 명시하였다. 이와 같은 결과들은 적합도 평가 과정에서 가이드라인을 절대적으로 충족해야 하는 이분법적 규칙처럼 사용하는 관행이 계속 이어지고 있음을 가리킨다.

앞서 언급한 바와 같이, 적합도 효과크기 지수가 연속적으로 해석되지 않는 주요 원인 가운데 하나는 기존의 가이드라인들이 제시되는 과정 자체에 연속성이 제대로 반영되지 않았기 때문이다. 만일 연구자가 표 3과 같이

표 4. 근소하게 기준을 충족하지 못한 모형에 대한 해석

출처	효과크기 지수	값	평가에 사용한 가이드라인	해석
Heller 등(2021)	CFI	.97	.95	mixed fit
	RMSEA	.07	.06	
	SRMR	.08	.08	
Thomsen과 Lessing(2020)	CFI	.91	.95	without acceptable fit
	RMSEA	.08	.08	
	SRMR	.09	.08	
Tyler 등(2020)	TLI	.89	.90, .95	mediocre fit
	CFI	.91	.90, .95	
	RMSEA	.06	.05, .08	
Yang과 McGinley(2021)	CFI	.88-.89	.90, .95	poor model fit
	RMSEA	.08	.06, .08	
	SRMR	.08-.11	.08, .10	

주. 회색 영역은 연구자가 이분법적 기준을 바탕으로 문제가 있다고 보고한 효과크기 지수와 값

연속선 상에서 효과크기의 해석을 제시하는 가이드라인을 사용할 경우, CFI=.89의 모형은 CFI=.90의 모형들과 동일하게 단순히 큰 효과크기를 나타내며 비슷한 수준의 유용성을 지닌 모형으로 평가될 수 있다. 기존의 이분법적 가이드라인들이 CFI=.90의 모형은 통과시키면서 CFI=.89의 모형은 적합도의 수준이 낮다는 이유로 배제하였다면, 표 3의 가이드라인은 CFI=.89의 모형과 CFI=.90의 모형을 동일선상에서 사용할 수 있는 근거를 제공한다.

나아가, Tyler 등(2020) 및 Thomsen과 Lessing (2020)의 경우 근소하게 가이드라인을 만족하지 못하는 모형에 대해 오차 간 상관을 포함하거나 변수를 통제하는 등의 기법을 사용하여 적합도를 임의적으로 올렸다. 그러나 효과크기 지수의 가이드라인을 특정 모형을 수정하거나 탈락시키는 유일한 근거로 사용하는 것은 적절하지 않다(Bagozzi & Yi, 1988; Hu & Bentler, 1998; Kenny et al., 2015; Marsh & Balla, 1994; McDonald & Ho, 2002). 반대로 Heller 등(2021) 및 Yang과 McGinley(2021)의 경우 앞의 연구들과 동일하게 가이드라인을 만족하지 못하는 모형을 수정하였으나, 그 과정에서 기준값 외에 모형 자체의 문제(예, Heywood case) 등을 함께 근거로 제시하였다. 이처럼 근소하게 기준값을 만족하지 못하는 효과크기 지수가 산출되었을 때, 해당 모형이 유용하지 않다고 주장하기 위해서는 가이드라인의 기준값 이외에 또 다른 확실한 근거가 필요하다(Marsh & Hau, 1996; Schreiber et al., 2006). 애초에 효과크기의 가이드라인은 모형의 유용성에 대한 절대적 기준이 아니기 때문

이다.

이와 반대로, 실제로 적합도 지수를 연속적으로 해석하여 모형을 평가하는 사례도 다수 존재하는데, 해당 연구들의 경우 기준값이 시뮬레이션으로부터 결정된 값이기 기준값에서 떨어졌다고 해서 무조건 제안된 모형을 기각해서는 안됨을 설명하였으며(Gerpott et al., 2021, Marsh et al., 2004), 이와 같은 기준들을 엄격한 기준이 아닌 가이드라인 정도로 사용할 것을 명시하였다(Williams et al., 2021). 또한, 다양한 연구들에서 적합도 지수를 연속적으로 사용함에 따라 기준값을 근소하게 만족하지 못하는 모형임에도 이를 채택하는 사례들이 제시되었으며(Johnson et al., 2020; Rau et al., 2021), 심지어 RMSEA= .114가 보고되었음에도 다른 요인들을 고려해 해당 모형을 채택하는 등의 (Jansen et al., 2021) 결과도 존재하였다. 특히 CFI =.88 ~ .89 사이의 값을 나타내는 모형임에도 이를 적절한 모형으로 보고하는 연구들도 다수 존재하였는데(Lin et al., 2020; Rojas et al., 2020; Rosen et al., 2020; Thompson & Bergeron, 2020), 특히 Rosen 등(2020)은 .89라는 CFI 값이 다른 지수들과 비교해 상대적으로 조금 작은 것일 뿐이며 West 등(2012) 또는 Williams, O'Boyle과 Yu(2020)이 제안한 바와 같이 적합도 지수를 자동적으로 모형을 기각하는데 적용하는 규칙으로 사용하지 않을 것을 권고하였다.

적합도 효과크기 지수를 연속적으로 해석하는 사례는 국내에서도 적지 않게 확인할 수 있는데, 이들은 CFI 혹은 TLI의 값이 .88에서 .89 사이로 추정됨에도 해당 모형을 적합한 모형으로 보고하였다(김진숙 & 권석만, 2010; 조은영 & 임성문, 2012; 연수진 & 서수균, 2013). 이와 같은 연구 결과들은 현재 경험연구를 수행하는 연구자들 가운데 적합도 지수를 연속적인 지수로 인식해 모형을 평가하는 연구자와 이를 이용해 모형을 이분법적으로 '선정'하는 연구자들이 섞여 있음을 나타낸다. 그러나 연구자가 적합도 효과크기 지수를 바탕으로 선정한 모형이 반드시 가장 좋은 모형인 것은 아니다. 해당 자료를 더 잘 설명하는 모형은 충분히 존재할 수 있으며, 이는 즉 어떠한 모형을 선택하는 과정에서 모형의 옳고 그름에 대한 이분법적 의미를 반영하는 것 자체가 위험한 일임을 의미한다.

적합도 효과크기 지수와 가이드라인의 목적 및 연속성의 특징을 고려하였을 때 연구자는 추정된 효과크기 지수가 가이드라인의 기준값에 근접하게 되면 적합도 자체에 큰 문제가 없음을 명시하고 모형을 사용할 수 있다. 물론, 이와 같은 주장은 heywood case와 같은 모형 자체의 문제가 존재하지 않는다는 조건 하에 성립할 수 있다. 또는, 여전히 적합도 효과크기 지수가 기준값을 만족하지 못하면 해당 모형이 설명력의 측면에서 불완전하다고 생각되어 모형을 수정하고 적합도를 끌어올릴 수도 있다. 하지만 수정 지수를 이용하여 모형에 자유모수를 추가하는 행위에는 확실한 이론적 근거가 바탕이 되어야 한다(Marsh & Hau, 1996; Schreiber et al., 2006). 또한, 일반적으로 적합도를 향상하기 위해 모형을 수정하는 관행이 이미 여러 연구에서 모형의 타당성 및 일반화의 문제 등을 바탕으로 비판(Boomsma, 2000; MacCallum, 1986; MacCallum et al., 1992)받아 왔음을 고려하였을 때, 적합도에 큰 문제가 없음에도 불구하고 오로지 가이드라인을 만족하기 위해 모형을 수정하는 것은 상당히 위험한 행위임을 알 수 있다.

적합도의 효과크기를 확인한다는 것은 적합

도의 실질적 유의성을 확인한다는 의미이며, α와 같은 이분법적 절단점을 이용해 평가하는 통계적 유의성과는 달리 연속성을 바탕으로 효과의 크기를 파악하는 것이다. 표 4에서 제시된 바와 같이, 적합도의 수준을 확인하기 위해 효과크기 지수를 사용함에도 불구하고 이를 이분법적으로 해석하며 가이드라인을 만족하기 위해 모형을 수정하는 것은 효과크기 지수의 본래의 사용 목적에 맞지 않으며, 자료에 대한 충분한 설명력을 지니고 있음에도 불구하고 모형을 배제해 버리는 비효율적인 평가에 해당한다.

결론 및 논의

사회과학 영역에서 구조방정식 모형의 사용이 활발해짐에 따라 대다수 연구자는 가장 대표적인 적합도 평가 도구인 적합도 지수를 중점적으로 활용하여 모형의 유용성을 판단한다. 현재 적합도의 평가 관행은 추정된 적합도 지수가 Hu와 Bentler(1999), Browne과 Cudeck (1993) 등의 가이드라인에서 제시하는 기준값을 만족하면 모형을 통과시키고, 그렇지 못하면 모형을 배제하는 방식이 만연하다(Heene et al., 2012). 심지어 연구모형이 기준값을 조금이라도 만족하지 못할 경우, 오차 간 상관을 임의로 포함하는 등의 기법을 이용하여 어떻게든 그 기준을 충족하고자 한다. 이와 같은 평가 방식은 적합도의 실질적 유의성을 확인하는 과정 자체에 대한 이해가 부족함에 따라 나타나는 문제이다. 본 연구는 이를 해결하기 위하여 효과크기의 관점에서 적합도를 평가하는 다양한 지수들을 소개하고, 연구자들이 적합도를 연속적으로 해석하는데 실질적인 도움이 될 수 있는 효과크기 가이드라인의 예시 및 사용 방법에 대해 논의하였다.

모형 평가의 전반적인 과정에 대한 이해를 돕기 위하여 본 연구에서는 우선 적합도 평가의 첫 번째 단계인 χ^2검정을 간략하게 소개하고, 실제 연구에서 빈번하게 기각되는 χ^2검정 결과가 어떠한 의미를 갖는지에 대해 논하였다. 표본크기를 비롯한 몇몇 한계점들로 인하여 χ^2검정 결과는 현재 형식적으로만 보고되고 있으나 그 형식 자체도 제대로 지켜지고 있지 않으며, 검정의 결과를 완전 적합 영가설에 대하여 해석하는 사례는 거의 찾을 수 없다. 본문에서도 강조했듯이 χ^2검정은 적합도의 통계적 유의성을 평가하는 거의 유일한 도구로서 매우 중요한 의의를 지닌다. 검정의 p값조차 제대로 보고하지 않고 표본크기를 근거로 χ^2검정 자체를 배제하기보다, 검정이 기각됨에 따라 모형이 자료에 완벽하게 적합하지는 않으며 완전 적합에서 얼마나 떨어져 있는지에 대해서는 실질적 유의성을 통해 파악한다고 해석하는 것이 적절한 χ^2검정 결과의 해석이라 볼 수 있다.

다음으로, 본 연구에서는 적합도의 실질적 유의성을 평가할 수 있는 다양한 지수에 대한 소개가 이루어졌다. 과거부터 오랜 기간 사용되고 있는 전통적 지수부터 최근 새롭게 발전한 지수까지 다양하고 핵심적인 종류의 평가지수를 이용해 적합도의 실질적 유의성을 확인하는 것이 가능하다. 이와 같은 지수들을 효과크기로 이용하는 과정에서 첫 번째로 주의를 요하는 개념은 일반적인 검정 결과에 대한 효과크기와 달리 적합도에 대한 효과크기의 경우 완전 적합검정 결과에 대한 것이기 때문에 효과크기가 작을수록 모형과 자료가

서로 합치함을 나타낸다는 것이다. 적합도의 수준은 효과크기와 부적 관계를 이루고 있으며, 효과크기 지수 값이 작을수록 연구자는 모형 적합도의 수준이 높다고 주장할 수 있다.

적합도의 실질적 유의성을 평가하는 과정에서 주의해야 하는 두 번째 요점이자 적합도 효과크기 지수의 목적을 달성하기 위하여 가장 중점적으로 고려해야 하는 요소는 추정된 지수의 해석과정에서 연속성을 반영하는 것이다. 실질적 유의성의 경우 연속선 상에서 모형과 자료 간의 차이를 확인하는 것이기 때문에 추정된 적합도 효과크기 지수 값은 연속적으로 해석되어야 하며(Hu & Bentler, 1998), 이는 곧 효과크기 지수의 가이드라인에서 제시되는 기준값들을 데드라인, 또는 절단 값이 아닌 말 그대로의 가이드라인 정도로 사용해야 함을 의미한다(Marsh, Hau & Wen, 2004). 본 연구는 기존의 효과크기 지수 가이드라인들을 정리하여 연속성이 반영된 새로운 가이드라인의 예시를 제공하였으며, 이를 바탕으로 가이드라인의 기준값을 근소하게 만족하지 못하는 모형을 사용하는 것에 논리적으로 문제가 없음을 설명하기 위해 노력하였다.

본 연구는 모형 적합도를 평가하고 해석하는 과정에서 어려움을 겪는 내용 영역 연구자들에게 실용적으로 도움이 되고자 하는 목적 아래 적합도를 평가할 수 있는 다양한 지수를 재소개하고 연속적인 해석을 위한 가이드라인의 사용 방식을 제안하였다. 그럼에도 불구하고 실제로 적합도 효과크기 지수를 사용하는 과정에는 여러 종류의 문제들이 복합적으로 존재한다. 본 연구에서도 설명하였듯이 적합도 효과크기 지수는 모형 조건의 영향을 받으며, 추정된 지수 값이 기준을 만족하지 못하는 것이 실제로 모형 설정 과정에서의 심각한 문제에 해당하는지 아닌지에 대한 명확한 이유를 알기 위해서는 잔차 행렬 등을 통해 모형을 복합적으로 진단하는 과정이 요구된다(McDonald & Ho, 2002). 나아가, 효과크기를 해석하는 과정에는 효과크기 모수에 대한 구간 추정치인 신뢰구간을 함께 확인하여 효과크기의 정확성을 평가하는 단계도 필요하다(Maydeu-Olivares, 2017; Maydeu-Olivares & Shi, 2017). RMSEA의 경우 현재 대부분의 통계 프로그램에서 신뢰구간을 함께 제공하고 있으나, 그 외의 지수들은 구간 추정치의 정보가 디폴트로 제공되지 않는다. 이에 따라, 본 연구에서 제시한 다양한 효과크기 지수를 신뢰구간의 관점에서 어떻게 추정하고 사용할 수 있는가에 대한 확장된 논의가 필요할 수 있다.

적합도 지수가 처음 발전된 이래 이를 이용하여 모형을 평가하는 행위 자체에 대한 근본적인 한계에 대한 연구들 역시 지속적으로 제기되어왔다. 기본적으로 적합도 지수는 점추정치 형식으로 제공되는데, 그에 따른 표집오차의 문제는 함께 수반될 수 밖에 없으며 그에 따라 동일한 모형에 대해 표본이 바뀔 때마다 추정치가 다른 값을 제공하게 된다(Kline, 2016). 또한, 적합도 지수는 모형을 평가하는 다양한 요소 중 하나일 뿐이며(Marsh & Balla, 1994), 이외에도 모형의 타당성을 주장하기 위해서는 추정치의 해석 가능성이나 모형의 복잡성 등 다양한 요인이 고려되어야 한다(Hu & Bentler, 1998).

본 연구의 주요 목적은 현재 구조방정식 모형의 평가 과정에 사용되는 주요 적합도 지수를 이용해 적합도의 실질적 유의성을 해석한다는 것이 모형 평가의 측면에서 어떠한 의미

를 지니는지에 대해 논의하는 것이었다. 이와 같은 논의를 바탕으로 실제로 모형 적합도를 평가하는 과정에 있는 연구자, 특히 적합도 지수 값이 Hu와 Bentler(1999)나 Browne과 Cudeck(1993) 등과 같은 가이드라인에서 제공하는 기준에 근접한 결과를 가진 연구자들이 해당 모형을 배제하지 않을 수 있는 근거를 제시할 수 있을 것으로 기대된다. 연구자는 적합도 평가의 가장 큰 목적이 가이드라인에서 제시하는 기준을 충족하는 것이 아님을 인지하고, 이를 바탕으로 모형의 유용성에 대해 넓은 관점에서 효율적인 판단을 내려야 한다.

참고문헌

김진숙. & 권석만. (2010). 인지행동적 요인과 부부 불만족도 사이의 관계. 한국심리학회지: 일반, 29(2), 265-288.

연수진. & 서수균. (2013). 이성관계에서 안정애착이 갈등해결전략과 관계만족도에 미치는 영향: 자기효과와 상대방효과. 한국심리학회지: 일반, 32(2), 411-428.

조은영. & 임성문. (2012). 자아해석과 주관적 안녕감 및 우울간의 관계: 인지적 유연성, 자기개념 명확성의 매개효과와 자기복잡성의 조절효과. 한국심리학회지: 일반, 31(2), 493-519.

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, 49(2), 155-173. https://doi.org/10.1007/BF02294170

Asparouhov, T., & Muthén, B. (2018). SRMR in Mplus.

Bagozzi, R., & Yi, Y. (1988). On the Evaluation of Structural Equation Models. *Journal of the Academy of Marketing Sciences*, 16(1), 74-94. http://dx.doi.org/10.1007/BF02723327

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, 42(5), 815-824. https://doi.org/10.1016/j.paid.2006.09.018

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88(3), 588-606. https://doi.org/10.1037/0033-2909.88.3.588

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

Bollen, K. A. (1989). *Structural equations with latent variables.* John Wiley & Sons. https://doi.org/10.1002/9781118619179

Boomsma, A. (2000). Reporting Analyses of Covariance Structure. *Structural Equation Modeling: A Multidisciplinary Journal,* 7(3), 461-483. https://doi.org/10.1207/S15328007SEM0703_6

Brosseau-Liard, P. E., & Savalei, V. (2014). Adjusting Incremental Fit Indices for Nonnormality. *Multivariate Behavioral Research,* 49(5), 460-470. https://doi.org/10.1080/00273171.2014.933697

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park,

CA: Sage.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An Empirical Evaluation of the Use of Fixed Cut-Off Points in RMSEA Test Statistic in Structural Equation Models. *Sociological Methods and Research,* 36(4), 462-494. https://doi.org/10.1177/0049124108314720

Cheng, C., & Wu, H. (2017). Confidence Intervals of Fit Indexes by Inverting a Bootstrap Test. *Structural Equation Modeling: A Multidisciplinary Journal,* 24(6), 870-880. https://doi.org/10.1080/10705511.2017.1333432

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. B. (2003). Finite Sampling Properties of the Point Estimates and Confidence Intervals of the RMSEA. *Sociological Methods & Research*, 32(2), 208-252. https://doi.org/10.1177/0049124103256130

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of Estimation Methods, Number of Indicators per Factor, and Improper Solutions on Structural Equation Modeling Fit Indices. *Structural Equation Modeling: A Multidisciplinary Journal,* 2(2), 119-143. https://doi.org/10.1080/10705519509540000

Fan, X. (2001). Statistical Significance and Effect Size in Education Research: Two Sides of a Coin. *The Journal of Educational Research,* 94(5), 275-282. http://www.jstor.org/stable/27542335

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal,* 6(1), 56-83. https://doi.org/10.1080/10705519909540119

Gerpott, F. H., Rivkin, W., & Unger, D. (2021). Stop and Go, Where is My Flow? How and When Daily Aversive Morning Commutes are Negatively Related to Employees' Motivational States and Behavior at Work. *Journal of Applied Psychology*. Advance online publication. https://doi.org/10.1037/apl0000899

Goffin, R. D. (1993). A comparison of two new indices for the assessment of fit of structural equation models. *Multivariate Behavioral Research,* 28(2), 205-214. https://doi.org/10.1207/s15327906mbr2802_3

Goffin, R. D. (2007). Assessing the adequacy of structural equation models: Golden rules and editorial policies. *Personality and Individual Differences,* 42(5), 831-839. https://doi.org/10.1016/j.paid.2006.09.019

Gomer, B., Jiang, G., & Yuan, K.-H. (2019). New effect size measures for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(3), 371-389. https://doi.org/10.1080/10705511.2018.1545231

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Johns Hopkins University Press.

Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S.

(2007). Testing! Testing! One, two, three-Testing the theory in structural equation models! *Personality and Individual Differences,* 42(5), 841-850. https://doi.org/10.1016/j.paid.2006.10.001

Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indexes with respect to violations of uncorrelated errors. *Structural Equation Modeling: A Multidisciplinary Journal,* 19(1), 36-50. https://doi.org/10.1080/10705511.2012.634710

Heller, A. S., Stamatis, C. A., Puccetti, N. A., & Timpano, K. R. (2021). The distribution of daily affect distinguishes internalizing and externalizing spectra and subfactors. *Journal of Abnormal Psychology,* 130(4), 319-332. https://doi.org/10.1037/abn0000670

Helmstadter, G. C. (1964). Principles of psychological measurement. Appleton-Century-Crofts.

Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling: A Multidisciplinary Journal,* 14(3), 361-390. https://doi.org/10.1080/10705510701301602

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *The Electronic Journal of Business Research Methods,* 6(1), 53-60. https://doi.org/10.21427/D7CF7R

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Sage Publications, Inc.

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

Jansen, M., Becker, M., & Neumann, M. (2021). Dimensional comparison effects on (gendered) educational choices. *Journal of Educational Psychology,* 113(2), 330-350. http://dx.doi.org/10.1037/edu0000524

Johnson, A., Nelson, J. M., Tomaso, C. C., James, T., Espy, K. A., & Nelson, T. D. (2020). Preschool executive control predicts social information processing in early elementary school. *Journal of Applied Developmental Psychology.* 71, Article 101195. https://doi.org/10.1016/j.appdev.2020.101195

Jöreskog, K. G., & Sörbom, D. (1981). LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods. Chicago: International Educational Services

Jöreskog, K. G., & Sörbom, D. (1984). *Advances in factor analysis and structural equation models.* Lanham: Rowman & Littlefield Publishers.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika,* 34(2, Pt.1), 183-202.

https://doi.org/10.1007/BF02289343

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal,* 10(3), 333-351. https://doi.org/10.1207/S15328007SEM1003_1

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

Kenny, D. A. (2020). Measuring model fit. https://davidakenny.net/cm/fit.htm

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement,* 56(5), 746-759. https://doi.org/10.1177/0013164496056005002

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling* (4th ed.). New York, NY: The Guilford Press.

van Laar, S., & Braeken, J. (2022). Caught off Base: A Note on the Interpretation of Incremental Fit Indices. *Structural Equation Modeling: A Multidisciplinary Journal,* 29(6), 935-943. https://doi.org/10.1080/10705511.2022.2050730"

Lai, K. (2019). A simple analytic confidence interval for CFI given nonnormal data. *Structural Equation Modeling: A Multidisciplinary Journal,* 26(5), 757-777. https://doi.org/10.1080/10705511.2018.1562351

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research,* 51(2-3), 220-239. https://doi.org/10.1080/00273171.2015.1134306

Lin, S.-H. (J.), Chang, C.-H. (D.), Lee, H. W., & Johnson, R. E. (2021). Positive family events facilitate effective leader behaviors at work: A within-individual investigation of family-work enrichment. *Journal of Applied Psychology,* 106(9), 1412-1434. https://doi.org/10.1037/apl0000827

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.

MacCallum, R. (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

Maiti, S. S., & Mukherjee, B. N. (1991). Two new goodness-of-fit indices for covariance matrices with linear structures. *British Journal of Mathematical and Statistical Psychology,* 44(1), 153-180. https://doi.org/10.1111/j.2044-8317.1991.tb00953.x

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting

model fit in structural equation modelling. *Personality and Individual Differences,* 42(5), 851-858. https://doi.org/10.1016/j.paid.2006.09.023"

Marsh, H. W., & Balla, J. R. (1994). Goodness of fit in confirmatory factor analysis: The effects of sample size and model parsimony. *Quality and Quantity,* 28(2), 185-217. https://doi.org/10.1007/BF01102761

Marsh, H. W., & Hau, K.-T. (1996). Assessing Goodness of Fit: Is Parsimony Always Desirable? *The Journal of Experimental Education,* 64(4), 364-390. https://doi.org/10.1080/00220973.1996.10806604

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin,* 103(3), 391-410. https://doi.org/10.1037/0033-2909.103.3.391

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research,* 33(2), 181-220. https://doi.org/10.1207/s15327906mbr3302_1

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

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of Fit in Structural Equation Models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275-340). Lawrence Erlbaum Associates Publishers.

Maydeu-Olivares, A., & Shi, D. (2017). Effect sizes of model misfit in structural equation models: Standardized residual covariances and residual correlations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences,* 13(1), 23-30. https://doi.org/10.1027/1614-2241/a000129

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika,* 82(3), 533-558. https://doi.org/10.1007/s11336-016-9552-7"

Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2018). Assessing fit in structural equation models: a Monte-Carlo evaluation of RMSEA versus SRMR confidence intervals and tests of close fit. *Structural Equation Modeling: A Multidisciplinary Journal,* 25(3), 389-402. https://doi.org/10.1080/10705511.2017.1389611

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods,* 7(1), 64-82. https://doi.org/10.1037/1082-989X.7.1.64

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin,* 107(2), 247-255. https://doi.org/10.1037/0033-2909.107.2.247

McDonald, R. P. (1989). An index of

goodness-of-fit based on noncentrality. *Journal of Classification*, 6(1), 97-103. https://doi.org/10.1007/BF01908590

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: *A Methodological Paradox. Philosophy of Science,* 34(2), 103-115. https://doi.org/10.1086/288135

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal,* 19(1), 86-98. https://doi.org/10.1080/10705511.2012.634724

Pavlov, G., Maydeu-Olivares, A., & Shi, D. (2021). Using the Standardized Root Mean Squared Residual (SRMR) to Assess Exact Fit in Structural Equation Models. *Educational and Psychological Measurement,* 81(1), 110-130. https://doi.org/10.1177/0013164420926231

Rau, R., Carlson, E. N., Back, M. D., Barranti, M., Gebauer, J. E., Human, L. J., ... & Nestler, S. (2021). What is the structure of perceiver effects? On the importance of global positivity and trait-specificity across personality domains and judgment contexts. *Journal of Personality and Social Psychology,* 120(3), 745-764. http://dx.doi.org/10.1037/pspp0000278

Rojas, N. M., Yoshikawa, H., & Melzi, G. (2020). Preschool teachers' use of discourse practices with Spanish-speaking dual language learners. *Journal of Applied Developmental Psychology.* 69, Article 101158. https://doi.org/10.1016/j.appdev.2020.101158

Rosen C.C., Dimotakis N., Cole M.S., Taylor S.G., Simon L.S., Smith T.A., Reina C.S. (2020). When challenges hinder: An investigation of when and how challenge stressors impact employee outcomes. *Journal of Applied Psychology.* 105(10), 1181-1206. https://doi: 10.1037/apl0000483.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research*, 8(2), 23-74.

Schreiber, J. B., Stage, F. K., King, J., Nora, A., & Barlow, E. A. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research,* 99(6), 323-337. https://doi.org/10.3200/JOER.99.6.323-338

Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The Relationship Between the Standardized Root Mean Square Residual and Model Misspecification in Factor Analysis Models. *Multivariate Behavior Research,* 53(5), 676-694. https://doi.org/10.1080/00273171.2018.1476221.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the Model Size Effect on SEM Fit Indices. *Educational and Psychological Measurement,* 79(2), 310-334. https://doi.org/10.1177/0013164418783530

Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivariate Behavioral Research,* 53(5),

676-694.
https://doi.org/10.1080/00273171.2018.1476221

Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. *Journal of Experimental Education,* 74(3), 267-288.
https://doi.org/10.3200/JEXE.74.3.267-288

Steiger, J. H. (1989). EzPATH: A supplementary module for SYSTAT and SYGRAPH. Systat, Inc.

Taasoobshirazi, G., & Wang, S. (2016). The performance of the SRMR, RMSEA, CFI, and TLI: An examination of sample size, path size, and degrees of freedom. *Journal of Applied Quantitative Methods,* 11(3), 31-40.

Tanaka, J. S., & Huba, G. J. (1989). A general coefficient of determination for covariance structure models under arbitrary GLS estimation. *British Journal of Mathematical and Statistical Psychology,* 42(2), 233-239.
https://doi.org/10.1111/j.2044-8317.1989.tb00912.x

Tanaka, J. S. (1993). Multifaceted Conceptions of Fit in Structural Equation Models. In K. A. Bollen, & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 10-39). Newbury Park, CA: Sage.

Thomsen, T., & Lessing, N. (2020). Children's emotion regulation repertoire and problem behavior: A latent cross-lagged panel study. *Journal of Applied Developmental Psychology,* 71, 101198.
https://doi.org/10.1016/j.appdev.2020.101198

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher,* 25(2), 26-30.
https://doi.org/10.2307/1176337

Thompson P. S., Bergeron D.M., Bolino M.C. (2020) No obligation? How gender influences the relationship between perceived organizational support and organizational citizenship behavior. *Journal of Applied Psychology*. 105(11):1338-1350.
https://doi: 10.1037/apl0000481

Tukey, J. W. (1991). The Philosophy of Multiple Comparisons. *Statistical Science*, 6(1), 100-116.

Tyler, C. P., Olsen, S. G., Geldhof, G. J., & Bowers, E. P. (2020). Critical consciousness in late adolescence: Understanding if, how, and why youth act. *Journal of Applied Developmental Psychology,* 70, 101165.
https://doi.org/10.1016/j.appdev.2020.101165

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.

Williams, A. L., Craske, M.G., Mineka, S., Zinbarg, R.E. (2021). Neuroticism and the longitudinal trajectories of anxiety and depressive symptoms in older adolescents. *Journal of Abnormal Psychology*. 130(2), 126-140.
https://doi.org/10.1037/abn0000638.

Williams, L. J., O'Boyle, E., & Yu, J. (2020). Condition 9 and 10 tests of model confirmation: A review of James, Mulaik, and Brett (1982) and contemporary alternatives.

Organizational Research Methods. 23, 6-29. http://dx.doi.org/10.1177/1094428117736137

Wilkerson, M., & Olson, M. R. (1997). Misconceptions about sample size, statistical significance, and treatment effect. *The Journal of Psychology: Interdisciplinary and Applied,* 131(6), 627-631. https://doi.org/10.1080/00223989709603844

Ximénez, C., Maydeu-Olivares, A., Shi, D., & Revuelta, J. (2022). Assessing Cutoff Values of SEM Fit Indices: Advantages of the Unbiased SRMR Index and Its Cutoff Criterion Based on Communality. *Structural Equation Modeling: A Multidisciplinary Journal,* 29(3), 368-380. https://doi.org/10.1080/10705511.2021.1992596

Yang, P. J., & McGinley, M. (2021). Commonalities and specificities of positive youth development in the US and Taiwan. *Journal of Applied Developmental Psychology,* 73(1), 101251. https://doi.org/10.1016/j.appdev.2021.101251

Yuan, K. H., & Marshall, L. L. (2004). A new measure of misfit forcovariance structure models. *Behaviormetrika,* 31(1), 7-90. https://doi.org/10.2333/bhmk.31.6

Zhang, X., & Savalei, V. (2016). Bootstrapping confidence intervals for fit indexes in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal,* 23(3), 392-408. https://doi.org/10.1080/10705511.2015.1118692

1차원고접수 : 2024. 03. 20
2차원고접수 : 2024. 07. 01
최종게재결정 : 2024. 07. 28

한국심리학회지: 일반
Korean Journal of Psychology: General
2024, Vol. 43, No. 3, 199-230
http://dx.doi.org/10.22257/kjp.2024.9.43.3.199

Overall evaluation of structural equation models and reflection on effect size and continuity

So-Hyun Yoo **Su-Young Kim**

Department of Psychology, Ewha Womans University

Structural equation model, which is widely used to describe the relationship between latent variables, can be judged by its goodness of fit. The χ^2 test for statistical significance and the effect size index for practical significance of model fit use dichotomous and continuous interpretation approach to evaluate the usefulness of the model, respectively. However, despite the fact that the level of fit is represented on a continuum for practical significance, the calculated effect size index is interpreted dichotomously by using the guideline as an absolute standard. The present study discusses the process of assessing the practical significance of fit in terms of the effect size index and the correct use of guidelines so that researchers evaluating the fit of a model can interpret the level of fit on a continuum. We begin with a brief discussion of the importance of assessing statistical significance using χ^2 test, and then define the concept of effect size in the context of structural equation models. We then introduce the different types of goodness of fit effect size indices and describe the characteristics of the guidelines used to interpret them. Finally, we provide examples of appropriate guidelines for interpreting calculated effect size index values on a continuum and discuss examples of incorrect model evaluation when continuity is not reflected, as well as the correct interpretation of models with marginal fit.

Key words : Structural equation models, interpretation of goodness of fit, effect sizes, continuity, use of guidelines.