韓國心理學會誌 : 臨床
Korean Journal of Clinical Psychology
2000. Vol. 19, No. 1, 161-177

〈특별기고〉

구조 방정식 모형의 적합도 지수 선정기준과 그 근거

홍 세 희†

University of California, Santa Barbara

지난 20여년간 구조 방정식 모형 (structural equation modeling)을 평가하는 많은 적합도 지수가 개발되었다. 그러나, 너무 많은 지수로 인하여 연구자들은 어떤 적합도 지수를 이용해 모형을 평가해야 하는지, 또 어떤 적합도 지수의 값을 보고해야 하는지 결정을 내리는데 오히려 혼란을 겪고 있다. 본 논문에서는 적합도 지수의 선정기준을 제시하고, 요즘 많이 쓰이는 지수들이 이 기준을 만족시키는지 각 지수의 수학적 공식을 이용해 설명하였다. 이를 바탕으로, NNFI, CFI, RMSEA가 바람직한 지수로 추천되었으며, 최근 주목을 받고 있지만 거의 소개되지 않은 RMSEA의 장, 단점에 대하여 자세히 논하였다.

최근들어, 임상 심리학을 비롯한 여러 응용심리 분야의 연구에서 구조 방정식 모형 (SEM: structural equation modeling) 의 적용이 증가 추세에 있다. Tremblay 와 Gardner (1996)의 보고에 의하면, 미국의 심리학 논문을 대상으로 보았을 때, 요인분석, 회귀분석, 변량분석등과 함께 SEM이 가장 인기 있는 통계기법으로 밝혀졌다. 변량분석은 여전히 많이 쓰이는 방법이지만, 그 인기는 약 20년전과 비교하면 하락세에 있다. 많은 학자들이 요인분석을 SEM 의 특별한 한 방법으로 간주함을 감안하면, 현재 미국에서 가장 많이 사용되는 통계기법은 SEM이라고 해도 과언이 아니다. 따라서, SEM은 심리학 연구자에게 필수적인 연구도구라고 할 수 있다.

SEM 이 널리 쓰이는 이유는 여러 가지가 있지만, 그중 대표적인 장점은 다음과 같다. 첫째, SEM에서는 여러개의 측정변수 (measured variable; manifest variable)를 이용해서 추출된 공통변량을 이론변수 (latent variable)로 사용하므로 측정오류 (measurement error)가 통제된다는 점이다. 따라서 SEM을 적용해서 구한 이론변수간의 공변량 계수나 회귀계수 값은 측정변수를 바탕으로 구한 계수 값 보다 정확하다고 할 수 있다. 둘째, 매개변수 (mediator variable)

† 교신저자(Corresponding Author) : 홍 세 희 / Department of Education. University of California, Santa Barbara. USA. / FAX : 805-893-7264 / E-mail : shong@education.UCSB.Edu

의 사용이 용이하다. 매개변수는 특성상 독립변수 및 종속변수의 역할을 동시에 해야 하는데, 회귀분석의 경우, 한 변수는 하나의 역할만을 해야하므로 매개변수의 도입 및 평가가 쉽지 않다. 회귀분석 대신 경로분석 (path analysis)을 사용하면 매개변수를 쉽게 다룰 수 있지만, 경로분석에서는 이론변수가 아닌 측정변수가 사용되므로 측정오류를 제대로 통제할 수 없다. 셋째, 이론 모형에 대한 통계적 평가가 가능하다는 점이다. 즉, 연구자가 개발한 이론 모형이 실제 자료에 얼마나 잘 부합되는 지를 평가하여 이를 바탕으로 연구자는 그 모형을 타당한 모형으로 받아들이거나 수정할 수 있다.

SEM의 분석과정은 대개 (1) 이론모형의 개발, (2) 실제 자료에 대한 모형의 적용, (3) 모형의 평가, (4) 필요시 모형의 수정, (5) 모형 재평가의 순서를 따른다. 첫 번째 단계는 연구자가 문헌연구 등을 통해 이론모형을 개발하는 단계이며, 두 번째 단계에서는 이론모형을 자료에 적용해 AMOS (Arbuckle, 1997), EQS (Bentler, 1995), LISREL (Joreskog & Sorbom, 1996) 등을 이용하여 통계분석을 한다. 세 번째 단계에서는 이론모형이 얼마나 자료에 잘 부합되는지를 평가하고, 적절하게 부합되지 않으면 다음 단계에서 모형을 수정하여 다시 평가를 한다. 모형의 적합도가 나쁘면, 네 번째, 다섯 번째 단계를 반복할 수 있으나, 이 경우 최종 모형을 반드시 다른 자료에 적용하여 교차 타당도 (cross-validation)를 확인하여야 한다.

SEM의 분석과정중, 이론 모형의 평가 단계는 모형을 받아들일 것이냐, 아니면 받아들이지 않을 것이냐를 결정하는 단계이므로 가장 중요한 과정의 하나이다. 이 단계에서는 통계결과를 바탕으로 객관적인 결정을 내리는데 대개의 경우 SEM의 적합도 지수를 이용한다. 요즘 많이 쓰이는 AMOS, EQS, LISREL 등의 프로그램들은 여러 가지 적합도 지수를 제공해 주는데 그 적합도 지수들은 대략 다음과 같다.

AIC: Akaike information criterion (Akaike, 1973),
RMSEA: Root mean square error of approximation (Steiger & Lind, 1980),
NFI: Normed fit index (Bentler & Bonett, 1980),
NNFI: Nonnormed fit index (Bentler & Bonett, 1980),
Hoelter's index (Hoelter, 1983),
GFI: Gooness of fit index (Joreskog & Sorbom, 1984),
AGFI: Adjusted gooness of fit index (Joreskog & Sorbom, 1984),
RFI: Relative fit index (Bollen, 1986),
IFI: Incremental fit index (Bollen, 1989),
ECVI: Expected cross validation index (Browne & Cudeck, 1989),
CFI: Comparative fit index (Bentler, 1990)

그런데, 문제는 많은 경우에 어떤 지수를 보면, 평가하고자 하는 모형의 적합도가 좋은 것으로 나타나고, 또 다른 지수를 보면 적합도가 나쁜 것으로 나타난다는 점이다. 이 때문에 SEM을 적용하는 많은 연구자들은 어떤 적합도 지수를 중심으로 이론모형을 평가해야 하는지 결정하기가 쉽지 않다. 설령, 모든 적합도 지수가 일관되게 모형을 지지한다고 해도, 논문에 어떤 적합도 지수의 값을 보고할 것인가를 결정하는 것도 쉽지 않다. 그렇지만, 각각의 지수는 모형의 적합도를 서로 다른 방식으로 평가하므로, 각 지수의 의미와 수학적 근거를 파악한다면, 어떤 상황에서 어떤 지수를 사용하는 것이 가장 바람직한지 결정하는 것이 보다 용이할 것이다.

많은 연구자들이 Marsh, Balla 와 McDonald (1988)의 연구결과를 바탕으로 적합도 지수를 선택하고 있으나, 그 연구는 80년대 개발된 지수만을 대상으로 하고 있다. 오히려, 90년대 들어 개발된 지수들이 더 많은 장점을 가지고 있음을 고려할 때, Marsh 등의 연구결과를 바탕으로 적합도 지수를 선택를 하는 것은 한계가 있다. 본 논문에서는 흔히 사용되는 적합도 지수의 의미를 수학적 공식을 통해 살펴보고, 각 지수의 장단점을 알아보았다. 이 논문의 주 목적은 응용연구를 하는 연구자에게 적합도 지수의 선택기준을 제시하는 것이므로, 수학적 내용은 가급적 최소한으

로 제한했다.

각 적합도 지수는 모형오류 (model error; model discrepancy), 자유도, 표본크기의 값중, 한 개 또는 그 이상의 값을 다른 방식으로 변환시킨 값이므로 먼저 모형오류와 자유도의 개념에 대해서 알아 보기로 한다.

모형오류와 자유도의 개념

모형오류

모형오류란 모형과 실제 자료와의 차이를 말한다. 즉, 모형을 평가하기 위해서는 모형을 통해 재생된 자료 (reproduced data; expected data) 와 실제 자료 (actual data)를 비교하는데 이 차이가 바로 모형오류이다. 모형오류가 크면, 그 모형은 자료에 잘 부합되지 않은 것이므로 좋은 모형이라고 할 수 없다. 이것은 회귀분석에서 회귀모형을 평가하기 위해 모형을 이용해 예측된 값과 실제값을 비교하여 그 차이가 작아야 회귀모형이 좋다고 하는 것과 같다.

Cudeck 과 Henly (1991)에 의하면, 모형오류에는 전집에서의 오류 (discrepancy of approximation), 추정오류 (estimation error), 표본에서의 오류 (sample discrepancy) 가 있다. 본 논문에서는 전집에서의 오류와 표본에서의 오류를 편의상 전집오류와 표본오류라고 하겠다. 전집오류(F_0)란 이론모형이 전집자료 (population data)와 부합되지 않는 정도를 말한다. 추정오류는 한정된 표본을 가지고 미지수를 추정할 경우, 전집을 가지고 미지수를 구하는 경우와 다른 값을 갖게 되는데 이 차이를 추정오류라 한다. 추정오류는 구하고자하는 미지수가 많을수록, 표본크기가 작을수록 커진다. 표본오류(F)는 이론모형이 표본자료와 부합되지 않는 정도를 말한다. 모형은 복잡한 현상을 단순화하여 설명하는 것이므로, 대부분 전집이나 표본에서 모형오류는 존재한다.

표본자료를 이용해 이론모형을 평가할 경우에 얻는 모형오류는 표본오류이며, 전집오류는 표본자료를 이용해서 구할 수 없다. 그러나, 전집오류 값은 표본오류 값을 통해 추정할 수 있다. 즉, 추정된 전집오류 ($\tilde{F_0}$) 는 $\tilde{F_0} = F - \frac{df}{n-1}$의 식을 이용해 구할 수 있다. 여기서, df는 자유도, n은 표본크기를 나타낸다. 위의 식에서 알 수 있듯이, 표본오류는 전집오류에 비해 약 $\frac{df}{n-1}$ 만큼의 오차를 지니고 있으며, 표본크기가 커질수록 그 오차는 작아진다. 표본크기가 무한히 커져 전집에 가까워 지면, 표본오류는 전집오류에 근접하게 된다.

자유도

구조 방정식 모형에서, 자유도는 모형의 간명성 (parsimony)을 나타낸다. 정해진 수의 변수를 이용해 모형을 만들 때, 모형 A의 자유도는 10이고, 모형 B의 자유도는 15라면, 모형 B가 간명하다고 할 수 있다. 자유도의 개념은 다음과 같은 예를 들어 설명할 수 있다. 변수가 4개 있다고 가정하자. 4개의 변수 (A, B, C, D)간에 모두 관계가 있다는 포화모형 (saturated model)을 위해서는 10개의 미지수(free parameter)가 필요하다. 10개의 미지수는 4 변수간의 공변량관계를 나타내는 미지수 6개 (A-B, A-C, A-D, B-C, B-D, C-D간의 공변량)와 4 변수의 변량 4개를 나타낸다 (구조 방정식에서 외적변수 (exogenous variable)의 변량은 구해야 할 미지수에 속한다). 즉, 모형에 4개의 변수가 있다면, 최대 가능한 미지수의 수는 10개가 된다. 이때, 자유도는 최대 가능한 미지수의 수와 이론모형에서 구하고자 하는 미지수간의 차이값을 말한다. 따라서 포화모형의 자유도는 0이다. 만일, A-B, B-C, C-D, D-A간에만 공변량이 있고 A-C와 B-D간의 관계는 없다는 이론모형을 만든다면, 이론모형의 미지수의 수는 8개가 되어, 자유도는 10-8=2가 된다. A-B, B-C, C-D간에만 공변량이 있다는 이론모형을 만든다면, 이론모형의 미지수의 수는 7개가 되어, 자유도는 10-7=3이 된다. 따라서, 변수의 수가 정해져 있을 때, 미지수의 수가 줄어 들수록 모형은 간단해 지며 자유도는 커짐을 알 수 있다. 또한, 미지수의 수와 자유도의 합은 항상 최대 가능한 미지수의 수

임을 알 수 있다. 이 사실들을 수학적으로 표현하면 다음과 같다. 먼저, p는 모형에서 사용되는 측정변수의 수이고 q는 모형에서 구하고자하는 미지수의 수라고 정의한다면, 최대가능한 미지수의 수는 $\frac{p(p+1)}{2}$가 된다. 예를들어, 변수가 4개라면, 최대가능한 미지수의 수는 $\frac{p(p+1)}{2}=\frac{4(5)}{2}=10$이 된다. 따라서 자유도는 $\frac{p(p+1)}{2}-q$라는 공식으로 계산된다.

x^2 검증의 문제점

모형의 적합도를 판단하기 위해 구조 방정식이 개발된 초기에 가장 많이 쓰였던 방법은 x^2 검증이었다. 표본오류 F에 표본크기 (정확히는, 표본크기−1)를 곱해주면, 곱해진 최종값은 x^2 분포를 따르므로 x^2 검증이 가능하다. 즉, $(n-1)F=x^2$ 이며, 이때 자유도는 $\frac{p(p+1)}{2}-q$ 이다. 이 검증에서 영가설은 "전집에서 (in the population) 모형은 변수간의 관계를 완벽히 설명한다" 이다. 영가설을 다르게 표현하면, "전집에서의 모형오류 (F_0)는 0이다." 또는, "모형을 통해 재생된 자료와 전집의 자료간의 차이는 없다." 등이 된다. 이x^2 검증에는 적어도 두가지의 문제점이 존재한다. 첫째, 모형은 복잡한 현상을 효과적이면서도 간명하게 설명하는 것이 기본목적이므로, 모형은 어느 정도 틀리다는 것을 인정해야 한다. 따라서, 연구자는 자신의 모형이 완전하기를 기대할 것이 아니라 모형이 자료를 어느 정도 잘 설명한다면 만족할 수 있는 것이다. 즉, 전집오류가 완전히 없지는 않지만 작다면, 그 모형은 좋은 모형이라고 할 수 있다. x^2 검증에서 영가설의 내용은 너무 엄격하여 모형이 조금만 틀려도 쉽게 기각되며, 또한 연구자의 관심을 반영하지도 못한다 (MacCallum, Browne & Sugawara, 1996). 두 번째 문제점은 x^2값이 n과 F에 의해 결정되므로, x^2 값이 모형오류 뿐아니라 표본크기의 영향도 동시에 반영한다는 점이다. 동일한 모형이 표본크기에 따라 기각될 수도 그렇지 않을 수도 있기 때문에 표본크기가 x^2값을 결정하는 또 하나의 요인이 된다는 점은 바람직하지 않다. 이와같은 문제로 모형을 평가하는데x^2 검증은 더 이상 널리 쓰이지 않으며 대신 적합도 지수(fit index)가 주로 이용된다.

적합도 지수의 선택기준

x^2 검증의 문제점을 해결하기 위해 80년대 초반부터 많은 적합도 지수가 개발되었다. 여러가지 적합도 지수중 어떤 지수값을 이용해 모형을 평가하는 것이 가장 바람직할까? 이 문제에 답하기 위해서는 적합도 지수는 어떤 조건을 충족시켜야 바람직한 적합도 지수라고 할 수 있는가에 먼저 답해야 할 것이다. 모형을 제대로 평가하기 위해서 적합도 지수는 최소한 다음의 두가지 조건을 충족시켜야 한다. 첫째, 적합도 지수는 표본크기에 민감하게 영향을 받지 않아야 한다 (Gerbing & Anderson, 1993; Hu & Bentler, 1995; Marsh, Balla, & Hau, 1996). 표본의 크기에 따라 동일한 모형의 적합도가 달라 진다면 모형을 일관성있게 평가할 수 없을 것이다. 둘째, 적합도 지수는 자료에 잘 부합되면서 동시에 간명한 모델을 선호해야 한다 (Bollen & Long, 1993; Browne & Cudeck, 1989; Cudeck & Browne, 1983; Gerbing & Anderson, 1993; Mulaik, James, Alstine, Bennet, Lind, & Stilwell, 1989; Steiger & Lind, 1980). 모든 적합도 지수는 자료에 잘 부합되는 모형을 선호하지만, 많은 지수들이 모형을 평가할 때 모형의 간명성을 고려하지 않는다. 이러한 지수를 사용하면, 불필요하게 (또는 필요이상으로) 모형을 복잡하게 만들어도 모형이 자료를 충분히 설명하므로 적합도의 값이 좋게 나온다.

회귀분석의 경우를 생각해 보자. 회귀분석의 목적은 소수의 독립변수를 이용해 종속변수의 값을 효과적으로 예측 또는 설명하는 것이다. 종속변수의 값이 효과적으로 예측된다면, R^2의 값이 커질 것이다. 예를 들어, 세 개의 독립변수를 사용했을 경우, $R^2=.70$이고 열 개의 독립변수를 사용했을 경우, $R^2=.73$이라면, 어떤 회귀식을 선택할 것인가? 모형의 적합도만을 고려하여 R^2의 값만 단순히 비교하면, 후자가 되겠지만, 그것은 모형의 간명성을 상실하고 있다. 일곱

개의 변수를 더 사용하였지만, R^2는 겨우 .03만이 증가 되었을 뿐이다. 간명성을 고려하지 않고 모형의 적합도만을 고려한다면 R^2를 높이기 위해 무수한 독립변수가 필요할 것이다. 이러한 문제점을 해결하기 위해 수정된 R^2 (adjusted R^2)가 개발되었으며 회귀모형을 평가하기 위해서는 R^2와 더불어 수정된 R^2도 함께 고려해야 한다. 단지 R^2만을 높이기 위해 별로 중요치 않은 독립변수를 많이 추가하면 수정된 R^2는 오히려 낮아질 수도 있다. 따라서, 수정된 R^2는 모형오류와 간명성을 동시에 고려한 최적의 상태에서 가장 높은 값을 보인다. 마찬가지로, 구조 방정식 모형에서도 적합도 지수는 불필요하게 복잡한 모형을 선호하지 않아야 하며, 모형오류와 간명성을 다 고려한 최적의 상태에서 가장 좋은 값을 보여야 한다.

다음 장에서는 많은 적합도 지수중, 바람직한 지수를 선택하기 위해, 어떤 지수가 위의 두 조건을 충족시키는지 살펴 보자. AMOS, EQS, LISREL 등의 프로그램의 출력물에 나오는 적합도 지수중, RFI, IFI, Hoelter index 는 현재 거의 쓰이지 않으므로 제외했다. AIC, BIC, CAIC, ECVI는 하나의 모형을 평가하는데 사용되기 보다는 여러 모형을 비교 평가할 때 주로 사용되므로 제외했다. 따라서, 본 논문에서는 현재 많이 사용되는 RMSEA, NFI, NNFI (또는, TLI), CFI, GFI, AGFI만을 다루었다. 이 6개의 지수를 편의상 많이 쓰이는 기준에 따라 상대적 적합도 지수와 절대적 적합도 지수로 나누었다 (Hoyle & Panter, 1995; Marsh, Balla & Hau, 1996) .

적합도 지수의 평가

상대적 적합도 지수 (Relative fit index, incremental fit index)

상대적 적합도 지수는 최악의 모형에 비해 이론모형이 얼마나 자료를 잘 설명하는지를 보여주는 값이다. 최악의 모형은 영 모형 (null model) 또는 기저모형 (base model) 이라고 불린다. 이론모형은 이론에 따라 측정변수들간에 관계를 적절히 설정한 모형인데 반해 기저모형은 측정변수들간에 관계가 없다는 모형이다. 따라서, 기저모형을 최악의 모형이라고 할 수 있다.

기저모형은 각 변수의 변량만 미지수로 계산하고 변수간의 관계를 설명하는 미지수는 계산하지 않으므로 가장 간단한 모형이다. 이에 반해, 포화모형은 모든 가능한 미지수를 다 계산하므로 가장 복잡한 모형이다. 자유도와 모형오류를 통해, 포화모형, 기저모형, 이론모형간의 관계를 보면 다음과 같다. df_s, df_b, df_t (소문자 s, b, t는 각각 포화모형, 기저모형, 이론모형를 가르킨다)를 각각 포화, 기저, 이론모형의 자유도라 할 때, $df_s < df_t < df_b$가 되며, F_s, F_b, F_t를 각각 포화, 기저, 이론모형의 표본오류라 할 때, $F_s < F_t < F_b$가 된다. 기저모형은 극단적으로 간단하여 오류가 너무 큰 모형이고, 포화모형은 극단적으로 복잡하여 간명성을 상실한 모형이다. 여기서, 간단한 (simple) 모형과 간명한 (parsimonious) 모형의 개념을 명확히 구분할 필요가 있다. 간단한 모형은 모형의 적합도에 관계없이 모형에 미지수가 적어서 단순한 모형을 가르키며, 간명한 모형은 적합도를 최대한 좋게 하면서 단순화한 모형이다. 따라서, 기저모형은 간단하지만 간명한 모형은 아니다. 그러나, 이론모형은 적합도와 함께 간명성을 추구해야 한다. 이론모형은 포화모형과 기저모형의 사이에 위치하여, 모형오류와 간명성 두 조건을 만족시키는 최적의 상태에서 결정되어야 한다. 기저모형과 비교하여 이론모형의 적합도를 평가하는 대표적인 상대적 적합도 지수는 NFI, NNFI, CFI이다. 이 세 지수가 위에서 설명된 두 조건을 충족시키는지 알아보자.

NNFI: 이 지수는 가장 오래된 적합도 지수의 하나이다. NNFI는 Bentler와 Bonett (1980)이 Tucker 와 Lewis (1973)가 탐색적 요인분석을 위해 만든 공식을 발전시킨 것이므로 Tucker-Lewis index (TLI)라고도 불린다 (AMOS 출력물에 나오는 TLI 값은 NNFI를 가르킨다). NNFI의 공식은 $\mathrm{NNFI}=\dfrac{(\frac{x^2_b}{df_b}-\frac{x^2_t}{df_t})}{(\frac{x^2_b}{df_b}-1)}$인데 이 공식은 x^2분포의 특성을 이용한 것이다. 완벽한 모형

(전집의 자료를 완벽하게 설명하면서 간명한 모형) 이 존재한다고 가정한다면, 그 모형의 x^2의 기대치는 그 모형의 자유도와 같다. 즉, 모형이 완벽하다면, 그 모형의 x^2값과 자유도간의 관계는 $\frac{x^2}{df}=1$이 된다. 따라서, NNFI 공식에서 분모는 기저모형에 비해 완벽한 모형이 얼마나 좋은지를 나타내며, 분자는 기저모형에 비해 이론모형이 얼마나 좋은지를 나타낸다. 예를 들어, 기저모형에 비해 완벽한 모형이 10 만큼 좋고, 기저모형에 비해 이론모형이 9 만큼 좋다면, NNFI 의 값은 0.9가 된다. NNFI의 값이 높을수록 모형의 적합도는 좋은 것이며, 대략 0.9이상이면 적합도는 좋다고 볼 수 있다.

NNFI가 표본크기에 얼마나 민감한지 그리고 모형을 평가할 때 그 모형의 간명성을 고려하는지 공식을 통해 알아볼 수 있다. 이를위해, 우선, NNFI 공식에 포함된 표본오류 (F) 를 전집오류 (F_o) 로 변환시킨다. 위의 공식에서 $x^2=(n-1)F$ 이다. 표본에서는 전집오류를 알 수 없으므로, 표본오류값을 통해 전집오류를 추정한다. 추정된 전집오류 ($\tilde{F_o}$) 는 $\tilde{F_o}=F-\frac{df}{n-1}$ 의 식을 이용해 구할 수 있다. F값을 $\tilde{F_o}$으로 변환하는 이유는 $\tilde{F_o}$는 전집오류이므로 표본크기의 영향을 받지 않기 때문이다. F값을 $\tilde{F_o}$으로 변환하려면 $F=\tilde{F_o}+\frac{df}{n-1}$의 식을 이용하면 된다. F값을 $\tilde{F_o}$으로 변환하여 여러 적합도 지수를 평가한 예는 McDonald $ 와 Marsh (1990)의 논문에 소개되어 있다.

위의 NNFI 공식에서 F 대신 $\tilde{F_o}+\frac{df}{n-1}$ 을 대입하면, NNFI는 $1-\frac{\frac{\tilde{F_{ot}}}{df_t}}{\frac{\tilde{F_{ob}}}{df_b}}$ 이 된다. 변환된 공식에 사용되는 전집오류와 자유도는 표본크기에 영향을 받지 않는 값들이므로, 표본의 크기가 NNFI의 값에 영향을 미치지 않음을 알 수 있다. 변환된 공식에서 분자는 이론모형과 기저모형의 적합도에 대한 상대적 비교를, 분모는 이론모형과 기저모형의 간명성에 대한 상대적 비교를 나타낸다. 다른 값들이 고정되어 있을 때, 이론모형이 자료를 잘 설명할수록 $\tilde{F_{ot}}$는 작아지고, 따라서 NNFI는 커진다. 또한, 다른 값들이 고정되어 있을 때, 이론모형이 간명할수록, df_t 는 커지고, 따라서 NNFI는 커진다. 변환된 NNFI의 공식을 통해 다음의 사실을 알 수 있다. 첫째, NNFI는 표본의 크기에 별로 영향을 받지 않는다. 둘째, NNFI는 모형오류가 작을수록, 모형이 간명할수록 그 값이 증가한다. 따라서, 모형이 자료를 잘 설명해도, 그 모형이 필요이상으로 복잡하면 NNFI의 값은 높지 않을 수가 있다. NNFI의 값은 대개 0에서 1 사이에 있지만 1이 넘을 수도 있다. NNFI의 값이 0에서 1사이를 벗어날 수 있다는 점을 제외하고는 NNFI는 적합도 지수가 지녀야 할 조건을 잘 충족시키고 있다.

NFI: NNFI와 함께 NFI는 일찍부터 EQS에 포함되어 80년대에 널리 사용되었다. NFI는 NNFI의 값이 0에서 1사이를 벗어날 수 있다는 점을 수정하기 위해 개발되었다. 공식은 $\text{NFI}=\frac{x_b^2-x_t^2}{x_b^2}$이며 그 값은 0에서 1사이를 벗어나지 않는다. NFI의 값이 높을수록 모형의 적합도는 좋은 것이며, 대략 0.9 이상이면 적합도는 좋다고 볼 수 있다.

NFI의 공식은 $\frac{F_b-F_t}{F_b}$으로 변환될 수 있으며, NNFI의 경우와 마찬가지로, F 대신 $\tilde{F_o}+\frac{df}{n-1}$ 을 대입하면 $\frac{(\tilde{F_{ob}}-\tilde{F_{ot}})+\frac{df_b+df_t}{n}}{\tilde{F_{ob}}+\frac{df_b}{n}}$ 이 된다. 변환된 NFI를 보면, 표본자료를 바탕으로 계산된 NFI값은 전집을 바탕으로 계산된 NFI에 비해 분모에서는 $\frac{df_b}{n}$ 만큼 분자에서는 $df_b+\frac{df_t}{n}$ 만큼 각각 오차 (biased error)가 있음을 알 수 있다. 이 오차는 표본의 크기가 커질수록 작아진다. 즉, 표본의 크기가 무한대로 커져 분모와 분자에 있는 오차가 0에 가까워져야만, 표본자료를 바탕으로 해서 전집에서의 이론모형의 적합도를 잘 추정할 수 있다. 따라서, 표본크기가 전집에 가까워질 정도로 크지 않는 한 NFI 값은 오차를 포함한다. 이로인해, NFI가 여전히 널리

사용되는 지수이지만, Bollen과 Long (1993), Gerbing과 Anderson (1993) 등은 NFI는 가급적 사용되지 않아야 된다고 주장하였다.

CFI: 이 지수는 90년대에 가장 인기있는 지수의 하나이며, NFI가 표본크기에 영향받는 점을 보완하기 위해 개발 되었다. 앞에서 소개된 두 지수가 x^2 분포를 바탕으로 개발되었지만, CFI는 비중심적 x^2 분포 (noncentral x^2 distribution) 를 바탕으로 개발되었다. 일반적으로 알려진 x^2 분포는 중심적 x^2 분포 (central x^2 distribution) 이며, 이는 비중심적 x^2 분포의 특별한 한 형태이다. 본 논문에서는 별도의 지칭이 없으면, x^2 분포는 중심적x^2 분포를 가르킨다.

90년대에 등장한 새로운 적합도 지수들이 x^2 분포를 바탕으로 개발되지 않고, 비중심적 x^2 분포를 바탕으로 개발된 이유는 $(n-1)F$ 가 x^2 분포를 따른다는 것이 "비 현실적인" 가정을 기초로 성립되기 때문이다 (MacCallum, Browne & Sugawara, 1996). 즉, "이론모형은 완벽하게 맞다"는 가정하에서 (즉, 영가설이 맞다는 가정하에서), $(n-1)F$ 는 x^2 분포를 따른다. 모형은 복잡한 현상을 간명하게 잘 설명하는 것이 기본 목적이므로, 모형은 어느 정도 틀릴 수밖에 없으며 따라서 그 가정은 비 현실적이다. 보다 현실적인 가정은 "이론모형은 어느 정도 틀리다." 이며, 이 가정하에서는 $(n-1)F$ 는 비중심적 x^2 분포를 따른다. 중심적 x^2 분포와 비중심적 x^2 분포의 차이는 비중심 모수 (noncentrality parameter)인 λ에 의해 결정된다. 이론모형이 틀릴수록, λ는 커져서 $(n-1)F$의 분포는 중심적 x^2 분포에서 벗어나 비중심적 x^2 분포가 된다. 따라서, λ는 전집오류의 크기를 반영한다 (Bentler, 1990; McDonald & Marsh, 1990). 전집에서 모형오류가 없다면 λ는 0이 되어 $(n-1)F$ 는 중심적 x^2 분포를 따르지만, 모형오류가 존재한다면 λ는 0이 아니므로 $(n-1)F$ 는 비중심적 x^2 분포를 따른다. CFI는 이런 비중심 모수의 특성을 이용해 개발되었다. 즉, CFI의 공식은 $\frac{\lambda_b-\lambda_t}{\lambda_b}$ 이다. 비중심 모수와 전집오류는 $\lambda=(n-1)F_o$ 의 관계를 가지므로 위의 공식은 $\frac{F_{ob}-F_{ot}}{F_{ob}}$로 변환될 수 있다. CFI를 계산할 때, 전집오류값을 구하는 것은 불가능하므로 표본오류값을 이용하여 추정한다. 표본자료를 이용해 추정된 CFI의 공식은 $CFI=\frac{\tilde{F}_{ob}-\tilde{F}_{ot}}{\tilde{F}_{ob}}$ 이 된다. CFI값은 0과 1.0 사이이며, 어떤 이론모형에 대한 CFI값이 대략 0.9이상이면 그 모형의 적합도는 좋은 것으로 간주한다.

CFI의 공식은 NFI의 공식과 매우 유사한데 그 차이점은 CFI는 전집오류를 바탕으로 계산되는 값이고, CFI는 표본오류를 바탕으로 계산되는 값이라는 점이다. NFI에 비해, CFI의 강점은 다음과 같다. 첫째, 보다 현실적인 가정에 기초한 비중심적 x^2 분포를 바탕으로 개발되었다. 둘째, 전집오류를 바탕으로 계산되므로 표본의 크기에 영향을 받지 않는다 (McDonald & Marsh, 1990; Bandalos, 1997). 그러나, CFI의 문제점은 모형의 간명성을 고려하지 않는다는 점이다. 따라서, 아무리 불필요한 미지수를 모형에 포함해도 모형의 적합도는 좋아지기만 할 뿐, 나빠지지는 않는다. CFI는 모형의 간명성은 고려하지 않지만, 표본크기에 영향을 받지 않고 모형오류를 측정하므로, 모형의 적합도를 평가할 때 유용한 지수이다. 그러나, 모형이 간명성을 상실한채 불필요하게 복잡해도 CFI는 좋아질 수 있으므로 CFI값 하나만으로 모형을 평가하는 것은 바람직하지 않다. CFI는 모형의 간명성을 고려하는 다른 지수 (예, NNFI, RMSEA)와 함께 고려되어야 한다. 예를들어, CFI값은 좋은 반면, NNFI와 뒤에서 설명될 RMSEA값은 나쁘면, 모형은 자료를 잘 설명하지만, 불 필요하게 복잡해져 있어 간명성을 상실하고 있다고 해석할 수 있다.

절대적 적합도 지수 (Absolute fit index)

절대적 적합도 지수는 이론모형의 적합도를 다른 모형 (예, 기저모형)의 적합도와 비교해 상대적으로 평가하지 않고, 이론모형이 자료와 얼마나 잘 부합되는지를 절대적으로 평가한다. 대표적인 절대적 적합도 지수는 GFI, AGFI, RMSEA이다.

GFI와 AGFI: 이 두 지수는 LISREL에 가장 먼저

포함되어 LISREL 사용자들에 의해 80년대에 널리 사용 되었다. 먼저, GFI의 공식을 살펴보면, GFI= $1-\frac{(S-\Sigma)'\Lambda(S-\Sigma)}{S'\Lambda S}$ 이다. 여기서 S는 표본 공변량 행렬로서, 표본의 자료를 뜻한다 (구조 방정식 모형에서 쓰이는 기본 자료의 형태는 표본 공변량 행렬이다). Σ은 이론모형을 통해 재생된 공변량 행렬로서, 재생된 자료를 뜻한다. 모형의 적합도를 평가하기 위해서는 S와 Σ를 비교하는데, 그 차이가 작을수록 모형은 자료에 잘 부합되는 것이다. 즉, 두 행렬을 비교하는 것은 회귀분석에서 원 자료와 회귀식을 통해 예측된 자료를 비교해 회귀식을 평가하는 것과 같은 이치이다. Λ는 가중치를 두는 행렬로서 추정방법 (예, maximum likelihood, weight least square 등) 에 따라 그 값이 다르다. GFI의 식에서 분모는 모형이 없어서 자료가 전혀 설명되지 않는 상태를 말하며 분자는 이론모형에 의해 자료가 어느 정도 설명된 상태를 말한다. 이론모형에 의해 자료가 많이 설명될수록 분자는 작아져서, GFI는 커지게 된다. 따라서, GFI는 모형이 존재하지 않는 상태에 비해 이론모형을 도입하면 자료가 얼마나 많이 설명되는지를 보여준다. 여기서 모형이 존재하지 않는 상태와 기저모형을 사용하는 상황은 확실히 구분될 필요가 있다. 상대적 적합도 계산에 사용되는 기저모형은 최악의 모형이지만 엄연히 존재하는 모형이다.

GFI는 공식이 의미하는 바가 회귀분석의 R^2와 매우 유사하다. 회귀분석에서 $R^2=1-\frac{SS_{residual}}{SS_{tot}}$ 이므로, GFI와 마찬가지로, 분자는 회귀식이 자료를 설명하는 정도를 나타내며, 분모는 회귀식이 없어서 자료가 전혀 설명되지 않은 상태를 나타낸다. R^2처럼 GFI도 모형이 불 필요하게 복잡해지는 것에 제약을 가하지 않으므로, 이를 보완하기 위해 AGFI가 개발되었다. 구조방정식 모형을 복잡하게 만들수록 그 모형은 자료를 조금이라도 더 잘 설명하므로 GFI의 값은 증가하지만, 별로 중요치 않은 변수간의 관계까지 모형에서 설명하면 모형은 불필요하게 복잡해지므로 AGFI의 값은 오히려 낮아질 수 있다. 따라서, AGFI는 수정된R^2와 유사하다. GFI와 AGFI값은 0과 1.0 사이이며, 어떤 이론모형에 대한 GFI값이 대략 0.9이상이면 그 모형의 적합도는 좋은 것으로 간주한다. AGFI값은 GFI값보다 대부분의 경우 낮게 나오며, AGFI값이 대략 0.85이상이면 그 모형의 적합도는 좋은 것으로 간주할 수 있다.

두 지수의 공식에는 표본크기가 포함되어 있지 않으므로 외견상으로는 표본크기의 효과가 없어 보인다. 그러나 Bollen (1990)에 의하면, 표본크기의 효과는 다음의 두가지로 규정된다. 첫째는 표본크기가 적합도 지수 공식에 포함되어 있는 경우이다. 이 경우, 표본크기는 직접적으로 지수값에 영향을 미친다. 앞에서 설명된x^2 검증이나 NFI가 이 경우에 속한다. 둘째는 적합도 지수의 표집분포 (sampling distribution)의 평균값이 표본크기와 관련이 있는 경우이다. 두 번째 경우는 모의실험 연구 (Monte Carlo simulation)를 통해 알 수 있다. 모의실험 연구에서는 우선 특정 이론모형에 맞는 표본을 생성한다. 작은 표본 크기부터 큰 표본 크기까지 (예, 100, 300, 500, 1000), 각각의 표본크기의 자료를 수 없이 (예, 10000번) 생성한다. 그리고, 생성된 각각의 표본 자료에 모형을 적용해 적합도 지수를 얻어낸다. 각각의 표본크기에 10000번씩 모형을 평가하므로 적합도 지수의 값은 하나의 분포를 이룰 것이며, 이 분포의 평균값을 구할 수 있다. 이 평균값이 표본크기를 늘릴수록 체계적으로 커지거나 작아지면 그 지수는 두 번째 경우의 표본크기의 영향을 받는 것이다. 모의실험 연구 결과에 의하면 GFI와 AGFI는 두 번째 경우에 속한다 (Bollen, 1990; Gerbing & Anderson, 1993).

요약하면, GFI 는 모형을 평가할 때, 모형의 간명성을 고려하지 않으며, 표본크기의 영향을 받는다. AGFI 는 모형의 간명성은 고려하지만 역시 표본크기의 영향을 받는다. 표본크기의 영향을 받는다는 점으로 인하여 GFI와 AGFI의 최근 사용빈도는 80년대에 비해 많이 줄었다.

RMSEA: 이 지수는 다른 적합도 지수에 비해 일찍 개발되었으나 그 동안 거의 사용되지 않다가 90년대 들어와 새로이 주목을 받게 되었다. RMSEA도 CFI처럼 비중심적 x^2 분포를 바탕으로 개발되었다. 비

중심 모수인 λ는 전집오류의 크기를 반영하지만 $\lambda=(n-1)F_o$ 의 관계에서 보듯이, λ는 모형오류 뿐만 아니라 표본크기에 의해서도 동시에 영향을 받으므로, RMSEA는 λ 대신 F_o을 바탕으로 개발된 지수이다. F_o이 표본크기에 영향을 받지않고 전집에서의 모형오류를 나타내는 지수지만, F_o을 이용해 모형을 평가하는데는 다음의 두 가지 문제가 있다. 첫째, F_o는 모형의 적합도만 고려하지 모형의 간명성은 고려하지 않는다. 둘째, F_o는 제곱값의 형태이므로 해석이 쉽지 않다. F_o가 제곱값인 이유를 간단히 설명하면, 회귀식에서 모형오류를 $SS_{residual}$이 나타내는데 이것이 제곱의 형태로 된 것과 같은 이치이다. 이 두 문제를 해결하기 위한 간단한 해결책은 첫째, F_o을 자유도로 나누고, 둘째, 그 값의 제곱근을 구하면 된다. 이 과정을 거쳐 개발된 지수가 RMSEA인데, 그 공식은 RMSEA= $\sqrt{\frac{F_o}{df}}$ 이다. 표본자료로부터 RMSEA 값을 추정할 때는 RMSEA=$\sqrt{\frac{\tilde{F_o}}{df}}$의 식을 이용한다. RMSEA는 표본크기에 영향을 받지 않는 전집오류를 이용해 구해지므로 표본크기의 영향을 받지 않는다 (McDonald & Marsh, 1990; Bandalos, 1997). 또한, RMSEA의 공식에 자유도도 포함되므로, RMSEA는 모형을 평가할 때, 모형오류와 간명성을 동시에 고려한다. 다른 적합도의 경우 그 값이 대개 0에서 1.0 사이에 결정되며 값이 1.0에 가까울수록 좋은 적합도를 나타내지만, RMSEA의 경우 그 값의 하한선은 0이지만 (RMSEA값이 음수로 나오면 0으로 간주한다) 상한선은 제한되지 않으며, 값이 작을수록 좋은 적합도를 나타낸다. 대략적인 기준으로, RMSEA 〈 .05이면 좋은 적합도 (close fit), RMSEA 〈 .08이면 괜찮은 적합도 (reasonable fit), RMSEA 〈 .10이면 보통 적합도 (mediocre fit), RMSEA 〉 .10 이면 나쁜 적합도 (unacceptable fit)를 각각 나타낸다 (Browne & Cudeck, 1993).

앞에서 본 바와 같이, RMSEA는 바람직한 지수가 충족시켜야할 여러 조건을 만족시키고 있다. 이외에도 이 지수는 여러 가지 장점을 갖고 있는데도 불구하고, 그 장점들이 아직 많이 알려지지 않아서 널리 쓰이고 있지 않으므로, 다음장에서는 RMSEA의 장점과 단점에 대해 자세히 다루었다.

RMSEA의 장점

신뢰구간 설정 가능: RMSEA의 분포가 알려져 있으므로 신뢰구간 (주로 90% 구간)의 설정이 가능하다 (Browne & Cudeck, 1993). AMOS, EQS, LISREL에서 구해 주는 RMSEA 값은 단지 표본 자료를 바탕으로 추정된 값에 지나지 않으므로 이 값에는 오차가 있다. 추정된 RMSEA 값만을 통해서는 그 오차의 정도를 알 수 없으며 신뢰구간을 통해서 알 수 있다. 신뢰구간의 간격이 넓으면 추정된 RMSEA 값에 오차가 많으며 그 만큼 신뢰할 수 없음을 뜻한다. 반대로, 신뢰구간의 간격이 좁으면 추정된 RMSEA 값은 정확하며, 연구자는 추정된 RMSEA값을 바탕으로 보다 자신있게 이론모형을 평가할 수 있음을 뜻한다.

예를들어, 추정된 RMSEA 값이 .05라면, 앞에서 제시된 기준에 의하면 이론모형의 적합도는 좋다고 볼 수 있다. 그러나, 추정된 RMSEA 값의 정확도를 모르므로 90% 신뢰구간도 동시에 고려할 필요가 있다. 만일 90% 신뢰구간이 (.01, .10)이라면, 다른 유사한 표본을 이용해 같은 이론모형을 수없이 반복적으로 평가한다고 가정했을 때, 90%의 경우는 추정된 RMSEA 값이 .01에서 .10사이의 속한다는 의미이므로 표본에 따라 동일한 이론모형이 거의 완벽한 모형일 수도 있고, 좋은 모형일 수도 있고, 또 나쁜 모형일 수도 있다는 것이다. 이런 상황이라면, 추정된 RMSEA 값이 .05라도 모형의 적합도가 좋다고 결론 짓는 것은 매우 성급한 일이다. 반대의 경우도 생각해 볼 수 있다. 추정된 RMSEA 값이 .05이고, 90% 신뢰구간이 (.048, .052)라면, 다른 유사한 표본을 이용해 같은 이론모형을 평가한다 해도 대개의 경우 (약 90%의 경우) 추정된 RMSEA 값은 .05정도일 것이므로, 이론모형의 적합도는 좋다고 결론을 내려도 무방할 것이다.

그럼, 90% 신뢰구간이 넓게 나와 모형의 적합도를 결정하기가 어려울 때는 어떻게 할 것인가? 신뢰도 구간의 넓이는 표본크기와 자유도에 의해 결정되므로

수학적으로 볼 때, 이중 하나 또는 둘의 값을 크게 하면 신뢰도 구간을 좁힐 수 있다. 그러나, 대개의 경우 신뢰도 구간을 좁히기 위해 자유도를 높이는 방법은 바람직하지 않다. 자유도를 높인다는 의미는 이론모형을 억지로 간단하게 하는 것이므로, 이 경우 모형에서 중요한 변수간의 관계를 설명하지 않음으로 인해 모형의 적합도가 나빠질 수 있기 때문이다. 대신, 표본크기를 늘리는 것이 합리적이다. RMSEA의 신뢰도 구간이 넓은 경우에는 표본을 늘려서 보다 정확한 추정값을 얻을 수 있을 때까지 모형의 적합도에 대한 평가를 보류하는 것이 현명하다.

대부분의 구조방정식 프로그램은 RMSEA의 추정값과 그 값의 90% 신뢰구간을 제공한다. RMSEA의 추정 값만 제공하고 90% 신뢰구간이 제공되지 않는 경우에는 FITMOD (Browne, 1992) 라는 간단한 DOS 프로그램을 이용해 구할 수 있다. FITMOD는 비 상업용 프로그램이며, 저자의 홈페이지 (http://education.ucsb.edu/~shong/)에서 구할 수 있다. FITMOD의 사용법 및 결과해석에 대한 예는 홍세희 (1999)에 소개되어 있다.

모형의 적합도에 대한 가설검증 가능: 앞에서 살펴본 바와 같이, x^2 검증 방법은 그 영가설의 내용이 너무 엄격하다는 문제를 지니고 있다. x^2 가설검증 방법에서 영가설의 내용은 "이론모형은 전집자료에 완벽하게 부합된다." 이므로, 이 방법은 "완벽한 적합도에 대한 가설검증 (test of perfect fit)" 이라고도 불린다. 모형은 복잡한 현상을 가급적 단순하게 설명하는 것이므로, 모형이 완벽하게 자료를 설명하기를 기대하는 것은 무리이다. 대신, 모형이 완벽하지는 않지만 어느정도 자료를 잘 설명할 수 있다면 모형은 유용하다고 볼 수 있다. 따라서, 영가설의 내용을 보다 현실적으로 바꿔 가설검증을 한다면 가설검증의 결과가 훨씬 의미있게 해석될 수 있을 것이다. x^2 가설검증 방법의 비 현실적인 영가설의 내용을 수정하기 위해, Browne 과 Cudeck (1992)은 RMSEA를 이용한 가설검증 방법을 제안했다. Browne과 Cudeck에 의하면 RMSEA=.05 는 완벽한 적합도는 아니지만 좋은 적합도를 나타내는 값이므로, 이 값을 이용해 "좋은 적합도에 대한 가설검증 (test of close fit)"을 할 수 있다. 이 경우, 영가설은 "RMSEA≤.05" 이며, 이 영가설이 기각되지 않으면 모형의 적합도는 좋다고 할 수 있다. RMSEA를 이용해, 완벽한 적합도에 대한 가설검증의 영가설을 표현하면, "RMSEA=0" 이다. 대부분의 구조방정식 프로그램은 RMSEA를 이용한 가설검증의 결과를 제공한다. 그렇지 않은 경우에는 FITMOD (Browne, 1992)를 이용해 구할 수 있다.

모든 가설검증 방법이 표본크기의 영향을 받듯이, 좋은 적합도에 대한 가설검증도 표본크기의 영향을 받는다. 즉, 표본크기가 커질수록, 검증력 (power)이 커져서 조그만 차이값에도 민감하여 영가설을 기각할 가능성이 커진다. t 검증을 예로들면, 두 집단간의 평균차이가 작아도, 표본크기가 크면, 검증력이 커져서 조그만 차이값에도 민감하여 영가설을 기각할 가능성이 커진다. 따라서 검증력이 너무 작거나 크면, 가설검증의 결과를 해석할 때, 주의를 요한다.

좋은 적합도에 대한 가설검증에서는 표본크기가 클수록 그리고 자유도가 클수록, 검증력이 커진다 (MacCallum, Browne, & Sugawara, 1996; MacCallum & Hong, 1997). Cohen (1965; 1992)에 의하면 α =.05일 때, 적절한 검증력 정도는 .8이며, 검증력이 너무 높거나 낮으면, 가설검증의 결과를 너무 신뢰하지 않는 것이 좋다. 예를들어, 표본의 크기가 너무 커서 검증력이 .99 정도가 나온다면, 모형이 좋건 나쁘건 상관없이 모형이 기각되므로, 이 경우에는 CFI, NNFI, RMSEA 값을 이용해 모형의 적합도를 해석하는게 바람직하다. 반대로, 검증력이 너무 낮은 경우에는, 모형의 적합도가 나빠도 모형이 기각되지 않을 가능성이 크므로 RMSEA의 신뢰구간을 살펴보고 모형의 적합도를 해석하는게 바람직하다. 즉, 신뢰구간이 너무 넓다면, 표본의 크기를 늘려서 모형의 적합도에 관한 보다 정확한 추정치를 얻을 때까지, 모형의 적합도에 대한 판단을 유보해야 한다.

구조 방정식 모형에 대한 검증력 계산방법에 대해서는 MacCallum, Browne과 Sugawara (1996) 의 논문과 MacCallum 과 Hong (1997)의 논문을 참고하기 바란다. 검증력 계산은 SEMPOWER (Hong, 1999) 라

는 프로그램을 이용해 계산할 수 있으며, 저자의 홈페이지 (주소: http://education.ucsb.edu/~shong/)에서 구할 수 있다.

RMSEA의 단점

위에서 본 바와 같이 RMSEA는 많은 장점을 지니고 있지만 단점 역시 지니고 있다. RMSEA의 공식을 보면, RMSEA 값은 모형오류와 자유도에 의해서 결정된다. 즉, 모형오류가 크거나 자유도가 작으면 RMSEA는 커지는데, 문제는 자유도가 작은 경우이다. 자유도는 모형이 복잡한 경우에도 작아지지만, 모형에서 사용되는 변수의 수가 너무 작을 때도 작아진다. 예를들어, 변수가 3개인 경우, 최대가능한 미지수는 $\frac{p(p+1)}{2}=\frac{3(3+1)}{2}=6$에 불과하므로, 이 값에서 미지수의 수를 빼면, 자유도는 더욱 작아진다. 자유도가 가장 큰 기저모형의 경우에도 자유도는 3에 불과하며, 이론모형의 경우에는 기저모형에 미지수를 추가하므로, 자유도는 3보다 작아지게 된다. 이처럼, 모형에 사용되는 변수의 수가 너무 작은 경우에는, 모형오류가 작아도 자유도가 너무 작아 RMSEA는 크게 나올 수 있다. 따라서, 모형에 사용되는 변수의 수가 너무 작은 경우, RMSEA값은 아주 나쁜 적합도를 보이고 (RMSEA>.10), CFI와 NNFI는 좋은 적합도를 보인다면 (CFI, NNFI>.90), RMSEA값이 작은 변수의 수에 영향을 받는 것을 암시하므로, CFI와 NNFI위주로 모형의 적합도를 평가하는 것이 바람직하다.

적합도 지수의 해석 예

적합도 지수의 실제 해석 예를 다루기 위해 확인적 요인분석 (confirmatory factor analysis)의 예를 보자. 확인적 요인분석의 예를 위해 "자율 지향성 (Autonomy)"을 측정하는 3개의 문항과 "대인관계 지향성 (Sociotropy)"을 측정하는 3개의 문항을 사용하였다. 이 두 성격특질이 지나치게 강하면 우울증으로 발전될 가능성이 높다고 여러 학자에 의해 제안되었다 (Arieti & Bemporad, 1980; Beck, 1983; Blatt & Zuroff, 1992). 이 예에서 사용된 6개의 문항은 두 성격특질을 측정하기 위해 개발된 Personal Style Inventory (Robins, Ladd, Welkowitz, Blaney, Diaz, & Kutcher, 1994)에서 선택된 것이다. Personal Style Inventory에서는 각 성격특질을 측정하기 위해 24개의 문항이 사용되지만, 이 예에서는 편의상 3개의 문항만을 각각 사용하였다. 자율 지향성을 측정하는 3개의 문항의 내용은 다음과 같다.

A1. 다른 사람이나 환경으로 인해 내 계획에 차질이 생겼을 때 나는 몹시 화가 난다.

A2. 나는 내 사 생활을 침해하는 사람을 좋아하지 않는다.

A3. 다른 사람이 내 행동이나 일에 대해 지시를 하면 불쾌해진다.

대인관계 지향성을 측정하는 3개의 문항의 내용은 다음과 같다.

S1. 나를 불행하게 하는 관계라도 그 관계를 쉽게 끝내지 못한다.

S2. 나는 다른 사람들을 기쁘게 해주려고 너무 애를 쓰는 편이다.

S3. 사람들이 내게 어떻게 반응할 지에 대해 신경을 많이 쓴다.

이 경우, A1, A2, A3, 세 문항은 자율 지향성이라는 공통된 개념을 측정하므로 세 문항간에 상관이 높을 것이고, 역시 S1, S2, S3, 세 문항도 대인관계 지향성이라는 공통된 개념을 측정하므로 세 문항간에 상관이 높을 것이다. 또한, 자신의 일에 대해서 지나치게 신경쓰는 사람은 남에게도 지나치게 신경쓸 가능성이 높으므로, 자율 지향성을 측정하는 세 문항들과 대인관계 지향성을 측정하는 세 문항들간의 상관 역시 낮지는 않을 것이다. 여섯 변수간의 복잡한 관계를 간명하게 2요인 구조로 설명하기 위해 그림 1과 같은 이론모형이 개발되었다. 그림 1에서 점선으로 표시된 미지수 C는 일단 무시하자.

이론모형에 대한 적합도를 평가하기 위해 AMOS를 이용하여 표 1과 같은 결과를 얻었다고 가정하자.

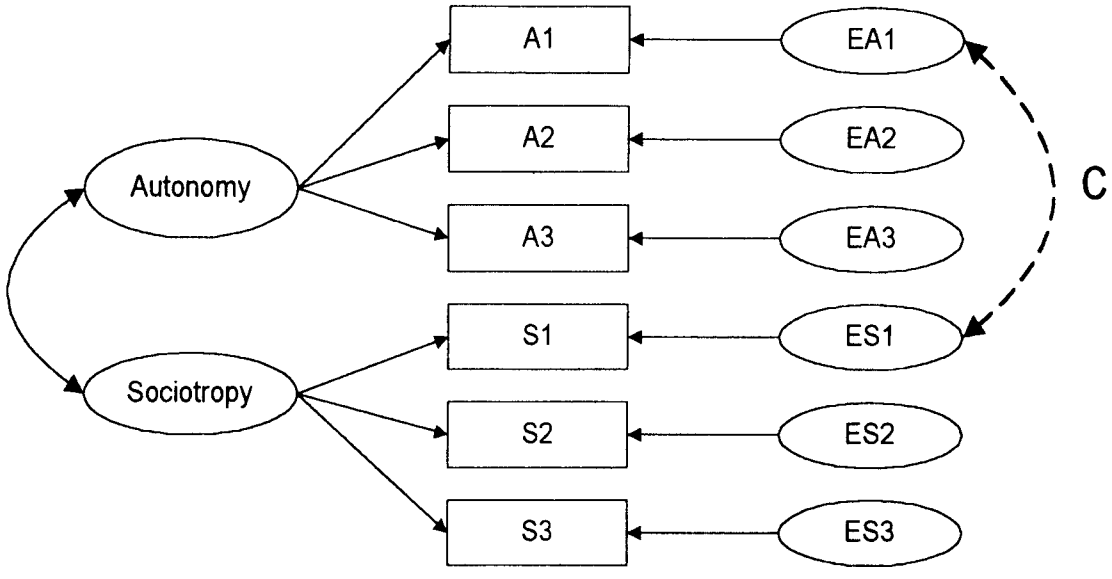

그림 1. 모형 B (EA1과 ES1간의 공변량은 모형 C에만 포함)

표 1. AMOS 출력물의 예

Model	NPAR	CMIN	DF	*P*	TLI	CFI
Your_model	13	8.948	8	0.347	0.989	0.992
Saturated model	21	0.000	0			1.000
Independence model	6	171.876	15	0.000	0.000	0.000

Model	RMSEA	LO 90	HI 90	PCLOSE
Your_model	0.041	0.000	0.148	0.477
Independence model	0.381	0.331	0.433	0.000

표 1의 출력물은 모형의 평가에 있어서 중요한 부분만 요약한 것이다.

출력물에서 Your_model은 이론모형, Saturated model은 포화모형, Independence model은 기저모형을 각각 가르킨다. NPAR는 모형에서 구하고자하는 미지수의 수, CMIN는 x^2 값, DF는 자유도, P는 x^2 검증에서의 p값을 각각 나타낸다.

이론모형, 포화모형, 기저모형 각각에 대하여 자유도와 미지수의 수의 합은 6개의 측정변수에 대한 최대가능한 미지수의 수, 즉 $\frac{p(p+1)}{2}=\frac{6(6+1)}{2}=21$이다. 출력물에서 보면 각 모형에 대해 NPAR와 DF의 합은 21이다. 포화모형에서는 가능한 모든 미지수를 구하므로 미지수의 수는 21이며 따라서 자유도는 0이다. 포화모형에서는 변수간의 모든 관계를 설명하므로 x^2값은 항상 0이다. 기저모형은 변수간의 관계는 존재하지 않는다는 모형이므로 각 변수의 변량만이 미지수이다. 따라서 기저모형에서 미지수의 수는 모형의 측정변수의 수와 같다. 출력물을 보면 기저모형의 미지수는 6인데 이는 측정변수의 수와 같다. 기저모형에 대한 x^2값과 자유도는 상대적 적합도를 구하는데 사용된다.

출력물에서 P값이 .05 이하면 α =.05일 때 "완벽한 적합도에 대한 가설검증 (test of perfect fit)"에 대한 영가설이 기각된 것이다. 즉, x^2 검증에서 영가설이 기각된 것이다. 이 경우에는 "이론모형이 전집자료에 완벽하게 부합된다."는 영가설이 기각되므로 모

형의 적합도는 완벽하다고 할 수 없다. 그러나, 대개의 경우에 x^2 검증에서 영가설은 기각되므로, 영가설이 기각되어도 이론모형을 기각할 것이 아니라 다른 적합도 지수를 고려할 필요가 있다. x^2 검증 결과를 중요시하지 않는 이유는 x^2값이 모형 적합도뿐만 아니라 표본크기에 의해서도 영향을 받기 때문이다. 그러나, 이 예의 경우 P값이 0.347이므로 영가설은 기각되지 않았다. 따라서, 이론모형이 전집자료에 거의 완벽하게 부합된다고 해석할 수 있다.

x^2 검증결과 영가설이 기각되지 않았으므로, 적합도 지수값들도 역시 매우 좋을 것이다. 출력물에서 중요시해 보아야할 적합도 지수값들은 CFI, NNFI (AMOS에서는 TLI)와 RMSEA값이다. CFI는 .992, TLI는 .989, RMSEA는 .041이므로 이론모형의 적합도 역시 매우 좋다. 출력물에서 포화모형에 대한 TLI값이 계산되지 않은 이유는 TLI값은 1보다 클 수 있으며 상한가가 제한 되어 있지 않기 때문이다. 또한, "좋은 적합도에 대한 가설검증 (test of close fit)"의 결과를 보면, 영가설 (RMSEA $\leq$.05) 이 기각되지 않았으므로 모형은 측정변수간의 관계를 효과적으로 설명한다고 볼 수 있다. AMOS에서 좋은 적합도에 대한 가설검증의 결과는 RMSEA값 옆에 나와있는 PCLOSE 값을 보면 된다. 이 값이 .05보다 크면, α =.05일 때 영가설이 기각되지 않은 것이다.

그러면, 이론모형의 적합도는 좋다고 결론내릴 수 있을까? 결론을 내리기 전에 RMSEA값에 추정오류가 어느 정도인지 알아보기 위해 신뢰구간을 살펴 볼 필요가 있다. AMOS에서 RMSEA의 신뢰구간은 RMSEA값 옆에 나와있는 LO 90과 HI 90 값을 보면 된다. LO 90과 HI 90는 90% 신뢰구간의 하한값과 상한값을 각각 의미한다. 이 경우, 90% 신뢰구간은 (.000, .148)이므로 간격이 너무 넓어서 추정된 RMSEA값에 오차가 많으며 그 만큼 신뢰할 수 없다. 즉, 다른 유사한 표본을 이용해 같은 이론모형을 수없이 반복적으로 평가한다고 가정했을 때, 90%의 경우는 추정된 RMSEA 값이 .000에서 .148사이의 속한다는 의미이므로 표본에 따라 동일한 이론모형이 거의 완벽한 모형일 수도 있고, 좋은 모형일 수도 있고, 또 나쁜 모형일 수도 있다는 것이다. 이런 상황에서, 신뢰구간을 고려하지 않고 추정된 RMSEA 값만을 통해 모형의 적합도가 좋다고 결론 짓는 것은 매우 성급한 일이다. 이 예에서 표본크기는 73명에 불과하므로 표본크기를 늘릴 때 까지는 모형의 적합도에 대해 명확한 결론을 내리기가 어렵다. 작은 표본크기로 인하여, 위에서 설명한 x^2 검증의 결과도 역시 신뢰하기 어렵다. 즉, 영가설이 기각되지 않은 것이 모형의 적합도가 완벽하기 때문인지 아니면 표본크기가 작기 때문인지 불명확하다.

다음으로, 적합도와 간명성이 다른 여러 모형이 있을 때, CFI, NNFI, RMSEA값이 어떤 식으로 변하는지를 알아보자. 앞에서 사용된 모형을 편의상 모형 B라 하자. 중요한 미지수를 없애면 세 적합도 지수가 얼마나 나빠지는 지를 알아보기 위해 모형 B에서 요인간의 상관을 구하는 미지수를 없앴다. 요인간의 상관이 무시할 수 없을 정도로 크므로 (이 경우, 상관값은 .5), 상관이 없는 모형 (모형 A) 의 적합도는 나빠져야 한다. 반대로, 불필요한 미지수를 포함시켰을 때, 세 적합도 지수가 어떻게 변하는지 알아보기 위해 모형 B에 EA1과 ES1간에 상관을 구하는 미지수를 추가했다. 이렇게 불필요하게 복잡해진 모형을 모형 C라 하자. 그림 1에 점선으로 표시된 C가 모형 C에 추가된 것이다.

표 2에서 보듯이, 중요한 미지수를 없앴을 때 (모형 A), 세 적합도 지수는 그렇지 않았을 때 (모형 B) 보다 현격히 나빠졌다.

불필요한 미지수를 추가했을 경우, 모형의 간명성은 나빠지지만 모형의 적합도는 좋아지므로, 모형 C의 x^2 값은 모형 B의 x^2 값에 비해 낮아졌다. CFI도 모형의 간명성을 고려하지 않으므로, 근소하지만 더

표 2. 모형 A, B, C의 적합도

모형	χ^2	df	CFI	NNFI	RMSEA
A	21.070	9	.883	.872	.136
B	8.948	8	.992	.989	.041
C	8.270	7	.994	.983	.050

높아졌다. 그러나, NNFI와 RMSEA값은 불필요한 미지수를 추가함으로 해서 오히려 나빠졌다. 포함된 미지수가 모형의 간명성을 해치면서 적합도를 좋게한 정도가 아주 미미하므로 불필요하게 복잡한 모형 C의 적합도는 모형 B의 적합도에 비해 오히려 나빠지는 것이다. 따라서, 모형을 평가할 때는 모형의 적합도와 간명성을 동시에 고려하는 지수를 최소한 하나는 포함시켜야 한다.

결론

80년대와 90년대 초반에 구조 방정식 모형을 평가하는 많은 적합도 지수가 개발되었다. Bollen (1989), Marsh, Balla, McDonald (1988) 등은 아직 어떤 적합도 지수가 가장 나은지에 대해서 의견이 분분하므로 가급적 여러개의 지수를 이용해 모형을 평가하고, 여러개의 지수를 보고할 필요가 있다고 주장하였다. 이로 인해, 대부분의 연구자들은 별다른 기준없이 3개에서 5개 정도의 적합도 지수를 선정해 모형을 평가하고 보고해 왔다. 그러나, 각 적합도 지수는 각각 다른 특성을 지니고 있으므로, 연구자들은 이론모형을 평가하기 위해서는 우선 적절한 기준을 가지고 그에 맞는 적합도 지수를 선택해야 한다. 그러기 위해서는 주요 적합도 지수가 어떤 방식으로 모형을 평가하는지를 이해해야 한다. 즉, 각 지수의 공식이 의미하는 바를 이해할 필요가 있다.

본 논문에서 적합도 지수를 평가한 바에 의하면 NNFI, CFI, RMSEA가 다른 지수에 비해 바람직한 적합도의 기준을 대체로 만족시킨다고 볼 수 있다. 그러나, 물론 이 지수들도 문제점이 없는 것은 아니다. 모든 상대적 적합도는 두 가지 문제를 기본적으로 갖고 있다. 첫째, 상대적 지수는 절대적 지수에 비해 추정방법 (예, maximum likelihood estimation, generalized least square, asymptotic distribution free 등)에 더 민감하다 (Sugawara & MacCallum, 1993). 즉, 추정방법을 달리함에 따라 상대적 지수의 값이 크게 변할 수 있다. 둘째, 상대적 지수는 기저모형과 이론모형의 적합도 지수를 바탕으로 계산되므로 상대적 지수의 분포는 이변량 (bivariate) 분포를 따른다. 이변량 분포의 수학적 복잡성으로 인해, 상대적 지수의 분포는 아직 불명확하다. 이로인해, NNFI, CFI의 경우 신뢰구간의 설정이 불가능하며 추정된 값에 대한 정확도도 알 수 없다.

이에 비해 RMSEA는 많은 장점을 지니고 있으므로 NNFI, CFI와 함께 많이 사용되기를 기대한다. RMSEA의 장점은 여러 가지가 있으나, 가장 중요한 장점은 신뢰구간의 설정이 가능하다는 것이다. RMSEA를 이용해 모형을 평가할 때는 RMSEA의 추정치를 이용하기 전에 신뢰구간을 통해 그 추정치의 정확도에 대해 먼저 평가하기를 권한다. 위에서 강조한 대로 RMSEA를 비롯해 NNFI, CFI가 바람직한 지수로 제시되었으나, 그 지수들 역시 어느 상황에서나 적당한 지수는 아니므로 연구자는 상황에 따라 가장 적합한 지수를 선택해야 할 것이다.

참고문헌

홍세희 (1999). 문항반응 이론과 요인분석을 이용한 척도개발 및 타당화. 임상심리학회 3차 워크샵 교재. 한국 임상심리학회.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N., & Csaki, F. (Eds.), *Proceedings of the 2nd International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267-281.

Arbuckle, J. L. (1997). *AMOS users' guide version 3.6*. Chicago, IL: SmallWaters Corporation.

Arieti,S., & Bemporad, J. (1980). The psychological organization of depression. *American Journal of Psychiatry, 136,* 1365-1369.

Bandalos, D. L. (1997). Assessing sources of error in structural equation models: The effects of sample size, reliability, and model misspecification. *Structural Equation Modeling, 4,* 177-192.

Beck, A. T. (1983). Cognitive therapy of depression: New perspectives. In P. J. Clayton & J. E. Barrett (Eds.), *Treatment of depression: old controversies and new approaches* (pp. 265-290). New York: Raven Press.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17,* 303-316.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Bentler, P. M. (1990). *EQS for windows user's guide.* Encino, CA: Maultivariate Software, Inc.

Blatt, S. J., & Zuroff, D. C. (1992). Interpersonal relatedness and self-definition: Two prototypes for depression. *Clinical Psychology Review, 12,* 527-562.

Bollen, K. A. (1986). Sample size and Bentler and Bonett's nonnormed fit index. *Psychometrika, 51,* 375-377.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods and Research, 17,* 303-316.

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107,* 256-259.

Bollen, K. A., & Long, J. S. (1993). Introduction. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 1-9). Newbury Park, CA: Sage.

Browne, M. W. (1992). *FITMOD: Point and interval estimates of measures of fit of a model.* Department of Psychology, Ohio State University.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24,* 445-455.

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.

Cohen, J. (1965) Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of Clinical Psychology.* New York: McGraw-Hill.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147-167.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the " problem" of sample size: A clarification. *Psychological Bulletin, 109,* 512-519.

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 40-65). Newbury Park, CA: Sage.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research, 11,* 325-344.

Hoyle, R. H. & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Eds.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158-176). Newbury Park, CA: Sage.

Hu, L-. T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Eds.), *Structural equation modeling: Concepts, issues, and applications*

(pp. 76-99). Newbury Park, CA: Sage.

Joreskog, K. G., Sorbom, D. (1984). *LISREL-VI user's guide* (3rd ed.). Mooresville, IN: Scientific Software, Inc.

Joreskog, K. G., Sorbom, D. (1996). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago, IL: Scientific Software International.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance sturucture modeling. *Psychological Methods, 1*, 130-149.

MacCallum, R. & Hong, S. (1997). Power analysis for covariance structure models using GFI and AGFI. *Multivariate Behavioral Research, 32*, 193-210.

Marsh, H. W., Balla, J. R., & Hau, K-. T. (1996). An evaluation of incremental fit indices: a clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 315-353). Lawrence Erlbaum Associates.

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*, 391-410.

McDonald, R. P. & Marsh, H. W. (1990). Choosing a multivariate model: noncentrality and goodness of fit. *Psychological Bulletin, 107*, 247-255.

Mulaik, S. A., James, L. R., Alstine, J. V., Bennet, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*, 430-445.

Robins, C. J., Ladd, J., Welkowitz, J., Blaney, P. H. Diaz, R., & Kutcher, G. (1994). The Personal Style Inventory: Preliminary validation studies of new measures of sociotropy and autonomy. *Journal of Psychopathology and Behavioral Assessment, 16*, 277-301.

Steiger, J. H., & Lind, J. M. (1980, June). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Sugawara, H. M., & MacCallum, R. C. (1993). An effect of estimation method on incremental fit indices for covariance structure models. *Applied Psychological Measurement, 17*, 365-377.

Tremblay, P. F., & Gardner, R. C. (1996). On the growth of structural equation modeling in psychological journals. *Structural Equation Modeling, 3*, 93-104.

Tucker, L. R., & Lewis, C. A. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

원고접수일 1999. 12. 1
수정원고접수일 2000. 1. 13
게재결정일 2000. 1. 26 ■

The Criteria for Selecting Appropriate Fit Indices in Structural Equation Modeling and Their Rationales

Sehee Hong

University of California, Santa Barbara

Over the past two decades, many indices for evaluating the fit of structural equation models have been developed. The existence of many indices, however, causes confusion among researchers about which indices to use to evaluate their models and about which indices to report. In the present study, two criteria for selecting appropriate indices were proposed and several popular indices were evaluated in terms of the criteria. Based on the evaluation, NNFI, CFI, and RMSEA were recommended. Finally, the advantages and disadvatages of RMSEA were discussed in detail because this index has only recently been recognized as one of the most informative indices and thus it is not widely known and used.