Nested Models, Model Modifications, and Correlated Errors

Nested Models

It is often recommended that researchers compare the fit of their model to alternative models. A chi-square difference test can be conducted using chi-square values and degrees of freedom from any two *nested models*. A nested model is a model that uses the same variables (and cases!) as another model but specifies at least one additional parameter to be estimated. The model with fewer restrictions or more free parameters (i.e., fewer degrees of freedom), which could be called a *reduced model*, is nested within the more restricted model, which could be called the *full model*. Some examples include comparing a one-factor to a two-factor model with the same variables, comparing a model which imposes equality constraints (two parameters are set to be equal) to a model that does not, or comparing a model that sets one or more parameters to particular values compared a model that allows them to be freely estimated. There are several ways this can be done. Whatever the hypothesis being tested, two models that differ in one or more parameters are compared.

The chi-square test, referred to as a "likelihood ratio test", is simply the difference between the full model and the reduced model (symbolized as $\Delta\chi^2$, using the difference in degrees of freedom as the degrees of freedom for the test (Δdf). Tests are typically computed by hand and compared to a standard chi-square table to determine significance.[1]

$$\begin{aligned}\Delta\chi^2 &= \chi_{Fewer}{}^2 - \chi^2_{More}\\ \Delta df &= df_{Fewer} - df_{More}\end{aligned}$$

Not all nested models are obvious and not all obvious model comparisons are nested, so researchers should exercise some caution (Rigdon, 1999). Tests are also sensitive to sample size. With large sample sizes, even small magnitude differences in fit will be significant. Just as with other significance tests, researchers must decide whether differences are of practical magnitude or not. There are several ways of quantifying the magnitude of difference in chi-square. One that I like to use is based on Cohen's effect size measure w, where $w = \sqrt{\Delta\chi^2 / \left(N \cdot \Delta df\right)}$, where $\Delta\chi^2$ and Δdf are the difference in chi-square and df of the two models (Dziak, Lanza, & Tan, 2014). The values can be roughly interpreted in terms of Cohen's suggested standards for a small (.1), medium (.3), and large (.5) effect. A more precise approach, but less easily interpreted, uses relative fit indices, such as the TLI or McDonald's Noncentrality Index (Cheung & Rensvold, 2002; Fan & Sivo, 2010).[2] Work by Fan and Sivo suggests that a difference larger than .02 in the McDonald index represents a substantial magnitude difference and works well for a variety of circumstances. Keep in mind that these are magnitude measures not significance tests. The chi-square difference test is the statistical significance test.

Modification Indices

Modification indices, which involve a series of simple nested model tests, can be requested from most computer packages. A modification index is a one degree of freedom chi-square test of the addition of a new parameter or the deletion of a parameter. Each modification index represents the change in the overall chi-square for the fit of the model if that particular parameter is changed. Thus, a significant chi-square value (greater than 3.84, because this is the chi-square critical value for $df = 1$) will significantly improve the fit of the model. This type of chi-square difference test, known as a likelihood ratio test, is the type of modification index used in Mplus, LISREL, and Amos. EQS provides two slightly different types of modification indices, which are similar and asymptotically equivalent tests, called the Lagrange multiplier for adding parameters to the model and Wald tests for eliminating parameters from the model. (The Lagrange multiplier test is sometimes called the Score test elsewhere). Some packages also will print the expected change in a parameter, representing what the unstandardized or standardized value of the added path

[1] To be consistent with West et al. (2023), I am using the terminology of "Fewer" and "More," which refer to the number of parameters estimated, instead other terminology that is used such as M0 and M1 or M1 and M2 (used by Kline, 2023). In any event, the first model will always have a worse or equal fit to the second model resulting in a chi-square difference that is 0 or greater and a df difference that is positive.

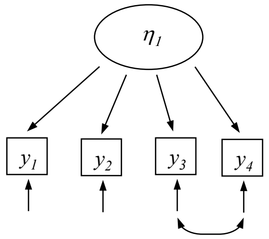
[2] I have created a spreadsheet for computing the chi-square difference test, Cohen's w, and the difference in the McDonald Centrality Index, which can be found here: http://web.pdx.edu/~newsomj/semclass/NCI.xlsx.

would be. Another way to roughly gauge the size of the modification index is to compare the chi-square difference relative to the overall chi-square value (i.e., by what percentage will the chi-square be improved by the addition of the new parameter to the model). I'll note that I have noticed that the estimated standardized parameter changes do not always correspond exactly to the values obtained after the changes in made in the model.

I suggest that you make one modification at a time, because each change may affect other parameters in the model and may, therefore, alter the relative importance of modification indices in a revised model. Keep in mind that many modifications to a model are considered by reviewers to be exploratory. Modifications can lead to development of incorrect models (MacCallum, 1986). Because a large number of tests are conducted when modification indices are requested, the group of tests are subject to familywise error (alpha inflation). I suggest that when changes are made they are reported and that you make sure they can be theoretically justified.

Chi-square Difference Tests with Non-normal Data
I would like to briefly mention here that, with robust chi-square adjustments for non-normal data, it is not appropriate to use a simple difference in chi-square values, and a weighted correction is required (Satorra & Bentler, 1999; Satorra & Bentler, 2010). Mplus allows the user to specify difference tests (DIFFTEST) and saves out information from the first model and compares it to the second model. I also have an Excel sheet that I use to compute correct difference test if another package is used. I will return to this issue when we discuss remedies for non-normal data later in the course.

Correlated Errors
One nested model modification that is fairly commonly mentioned by researchers, is an addition to the measurement model that allows measurement errors (or, better, "measurement residuals") to be correlated.[3] Typically, researchers begin testing a confirmatory factor model by assuming that measurement residuals are independent of one another (known sometimes as "local independence"). This assumption implies that the variance of a particular item not caused by the factor has a source that is unique to that particular variable. This assumption is not always valid, so researchers may incorporate correlated errors in a factor model. Inclusion of the correlation will decrease the loadings for the items involved if the correlation is positive. The following figure graphically illustrates a correlation between errors for items 3 and 4.



Correlated errors are often due to similar item wording or content (e.g., questions "I feel blue whenever my spouse is around" and "I feel sad whenever my spouse is around" from a life satisfaction measure). In this example, there is systematic variance shared by these two items (perhaps due to the phrasing or feelings specific to the spouse) that is not due to the factor. The basis for including a correlation between measurement residuals can be data driven (e.g., suggested by modification indices) or theoretically driven. A common, and theoretically motivated, application is in longitudinal research in which the measurement residual for the same variable is correlated over time (Newsom, 2024). Data driven correlated errors will likely appear exploratory to reviewers, but my experience is that one or two conceptually justified correlated errors added post hoc are not met with too much resistance. Correlated errors also may be used to account for method effects (e.g., telephone interviews and face-to-face interviews, observation vs. self-report). One

[3] It is most common to describe the errors associated with indicators as "measurement errors" when they really represent other sources of error, something better termed a *unique error* or a *unique component*, and so I prefer the term measurement residual, which does not necessarily imply random measurement error. Essentially the unique error is any source of variation in the indicator that is not accounted for by the latent variable and may include systematic and meaningful causes as well as random error. That is, we should think about unique errors as any variation in the indicator due to unknown sources (i.e., not accounted for by the latent variable). It is the unique systematic variance that is correlated, not random error.

common type of method effect that you may encounter has to do with items that are positively or negatively worded (e.g., "I find statistics dreadful" vs. "I find statistics most interesting"; see Tomás & Oliver, 1999, for detailed discussion). In such cases, model misfit is indicated for correlations among measurement residuals for items worded in the same direction. Modifications to the model involving correlated errors are subject to the same precautions as mentioned above.

References

Cheung, G W., & Resnvold, R. B (2002). Evaluating Goodness of fit indexes for testing measurement invariance. *Structural Equation Modelling: A Multidisciplinary Journal, 9*, 233–255.

Dziak, J. J., Lanza, S. T., & Tan, X. (2014). Effect size, statistical power, and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 534-552.

Fan, X., & Sivo, S. (2009). Using goodness-of-fit indices in assessing mean structure invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 54–69.

Kline, R.B. (2023). *Principles and practice of structural equation modeling, fifth edition.* New York: Guilford.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100*, 107-120.

Newsom, J.T. (2024). *Longitudinal structural equation modeling: A comprehensive introduction, second edition*. New York: Routledge.

Rigdon, E.E. (1999). Using the Friedman method of ranks for model comparison in structural equation modeling. *Structural Equation Modeling A Multidisciplinary Journal, 6*, 219-232

Satorra, S., & Bentler, P.M. (1999). *A scaled difference chi-square test statistic for moment structure analysis.* Unpublished technical report, Universitat Pompeu Fabra, Barcelona, Spain.

Satorra, A. & Bentler, P.M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. Psychometrika 75: 243. doi:10.1007/s11336-009-9135-y

Tomás, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling*, 6, 84–98.

West, S.G., Wu, W., McNeish, D., & Savord, A. (2023). Model Fit in Structural Equation Modeling. In R.H. Hoyle, *Handbook of structural equation modeling*, *second edition* (pp. 185-205). Guilford.