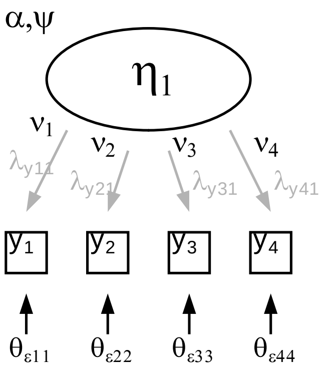
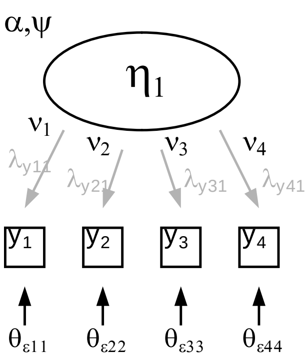
Invariance Tests in Multigroup SEM
Illustration using Meredith's (1993) Terminology

Weak Factorial Invariance

Group A

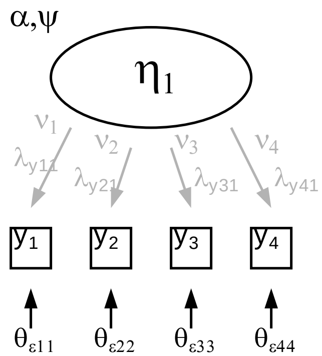


Group B

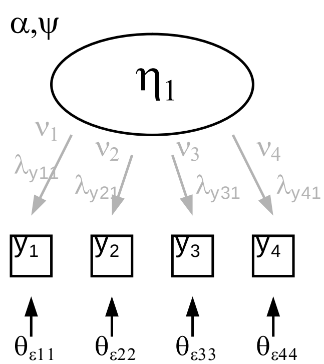


Strong Factorial Invariance

Group A

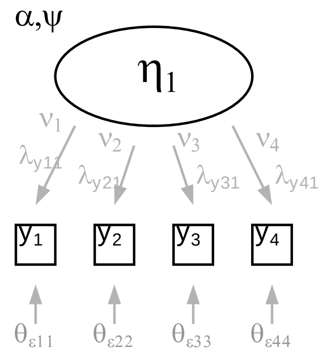


Group B

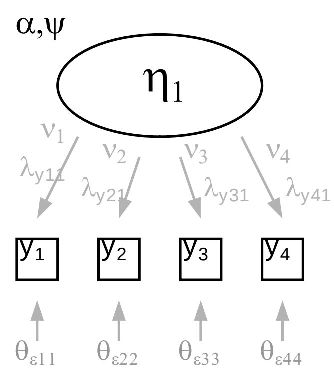


Strict Factorial Invariance

Group A

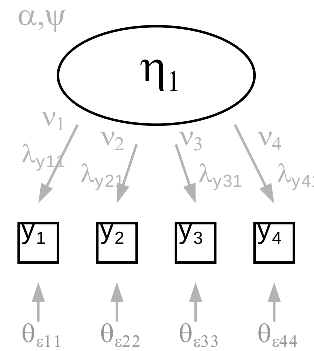


Group B

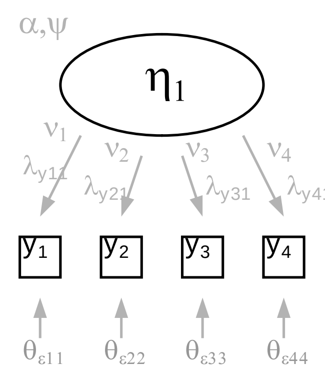


Structural Invariance

Group A

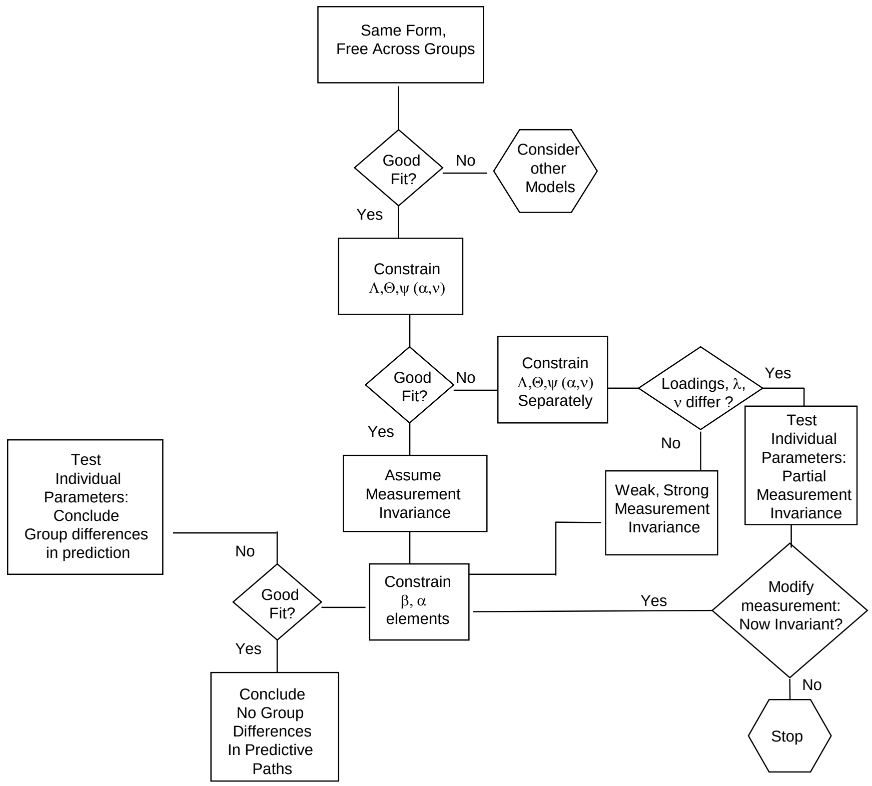


Group B



Note: Grayed elements represent equality constraints across groups. η is the factor, α is the factor mean, ψ is the factor variance, ν is the loading intercept, λ is the factor loading, θ_ε is the measurement residual variance (I use the matrix element symbol instead of just ε in my figure to emphasize that the equality test is a comparison of variances).

Overview of a Suggested Process for Testing for Group Differences in SEM



Comments

The general strategy above follows what Stark and colleagues (Stark, Chernyshenko, & Drasgow, 2006) call a *free baseline approach*. The logic of this approach is that if the model does not fit when there are no cross-group constraints placed on it, rejection or modification of the general model is required. In contrast, one can propose a *constrained baseline approach* in which all parameters are constrained across groups first (which is the same as a single group model). Neither approach is right or wrong per se, but they have different rationales and strengths.

Following the free baseline approach, comparisons should be made between a more constrained model and the baseline model (Bentler, 2000). An omnibus approach is usually used in which classes of parameters (e.g., loadings) are constrained simultaneously. The general overall idea is to establish measurement equivalence before comparing predictive paths across groups to avoid confounding group differences in measurement properties with substantive differences in means or predictive paths across groups. In practice, researchers are often willing to live with weak measurement or sometimes referred to as partial invariance (Byrne, Shavelson, & Muthen, 1989) in which only loadings are equivalent across groups (i.e., strong or weak factorial invariance depending on the interest in intercepts). As long as SEM is used to assess differences in prediction across groups, weak invariance across groups should be sufficient, because the differences in measurement residuals across groups should not affect relations among latent variables as long as the measurement residual variances are allowed to freely vary across groups. (And, in fact, constraining measurement residuals to be equal across groups when they are truly not equal will lead to biases in the predictive paths). On the other hand, strict measurement invariance (as opposed to weak or strong invariance) will be required if the goal is to compare groups in subsequent analyses using a composite index of the items, because group differences in the amount of measurement error across groups can impact the results (Millsap & Kwok, 2004).

It is important to realize that testing only a subset of loadings or intercepts for measurement invariance can be problematic, because there is an interdependence of loadings (intercepts) and factor variances (factor means). This makes sense if you remember that the factor variance is altered when a different indicators is chosen as the referent. For testing factorial invariance on only a subset of loadings (or other parameters) for a factor, special procedures are needed to be ensure that the choice of referent indicator does not obscure the correct identification of the specific indicators that differ (see Cheung & Lau, 2012; Yoon & Millsap, 2007).

Mean comparisons (of factors, α, or intercepts, ν) may not always be of interest, but when they are important, factorial invariance of the measurement intercepts should be examined. Two cases in which intercept invariance should be established: 1) bias may be introduced because groups are combined or assumed equivalent in later analyses, and 2) mean differences between groups are of substantive interest (analogous to t-test or ANOVA comparisons). In other cases, in which the researcher is interested in examining predictive differences between groups, one would not necessarily assume that group differences on the mean of an independent or dependent variable would affect associations with other variables within a group. Loadings should always be constrained equal across groups when comparing means, because factor means are a function of the measurement intercept and the loading (see Newsom, 2015, Chapter 1).

Multifactor models present additional complications. Structural relations between a set of predictors and an outcome will depend on correlations among the predictors, for example. So, in order to meaningfully interpret differences in prediction across groups, one would normally want to assume equivalence in the correlations among the predictors. Finally, with large sample sizes, significant differences may be found for very small magnitude differences, and the researcher needs to decide which differences are important. Calculating magnitude of the difference is encouraged (Cheung & Rensvold, 2002; Fan & Sivo, 2010). As outlined in the handout *Nested Models, Model Modifications, and Correlated Errors*, one can use relative fit indices such as the CFI or McDonald's noncentrality index or Cohen's w to evaluate how large the chi-square difference is. As a separate matter, the mean differences across groups can be evaluated in terms of effects size as well, using Cohen's standardized effect size d, for instance (Hancock, 2001).

The issues discussed in this handout only scratch the surface of the relevant issues for multigroup SEM. Comparisons with categorical indicators (and threshold constraints), comparisons with nonnormal and missing data, and comparisons across more than two groups are not discussed because of time limitations in the course. Some of these issues have not been thoroughly considered, but many of them are discussed in Roger Millsap's book (2011).

References and Recommended Readings

Bentler, P.M. (2000). Rites, wrongs, and gold in model testing. *Structural Equation Modeling, 7*, 82-91.
Bontempo, D. E., & Hofer, S. M. (2007). Assessing factorial invariance in cross-sectional and longitudinal studies. In A.D. Ong & M. van Dulmen (Eds.), *Handbook of methods in positive psychology* (pp. 153-175). Oxford University Press.
Byrne, B.M., Shavelson, R.J., & Muthen, B. (1989). Testing for the equivalence of factorial covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.
Cheung, G. W. & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. Organizational
Cheung, G.W. & Rensvold, R.B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. Journal of Management, 25, 1-27.
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.
Choi, J., Fan, W., & Hancock, G. R. (2009). A note on confidence intervals for two-group latent mean effect size measures. *Multivariate behavioral research, 44(*3), 396-406.
Fan, X., & Sivo, S. (2009). Using _goodness-of-fit indices in assessing mean structure Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 54–69.
Gabler. Millsap, R.E. (2011). Statistical Approaches to Measurement Invariance. New York, NY: Routledge Academic.
Hancock, G. R. (200 l). Effect size, power, and sample size determination for structured means modeling and mimic approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66*, 373-388.
Kim, E.S., & Yoon, M. (2011). Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. Structural Equation Modeling, 18(2), 212-228.
Kim, J. O., & Ferree, G. D., Jr. (1981). Standardization in causal analysis. *Sociological Methods and Research, 10*, 187-210.
Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525-543.
Millsap, R. E, & Yun-Tein, J. (2004). Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behavioral Research., 39, 479-515.*
Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33, 403-424.*
Millsap, R. E., & Kwok, O.M. (2004). Evaluating the impact of partial measurement invariance on selection in two populations. Psychological Methods, 9, 93-115.
Millsap, R.E. (2011). *Statistical Approaches to Measurement Invariance.* New York: Routledge. Research Methods, 15(2), 167-198
Newsom, J.T. (2015). *Longitudinal Structural Equation Modeling: A Comprehensive Introduction*. New York: Routledge.
Sass, D.A. (2011). Testing Measurement Invariance and Comparing Latent Factor Means Within a Confirmatory Factor Analysis Framework. Journal of Psychoeducational Assessment, 29(4), 347-363.
Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292.
Temme, D. (2006). Assessing measurement invariance of ordinal indicators in cross-national research. In S. Diehl and R.
Terlutter (Eds.), International Advertising and Communication: Current Insights and Empirical Findings (pp. 455-472).
Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4-69.
Werts, C.E., Rock, D.A., Linn, R.L., & Joreskog, K.G. (1977). Validating psychometric assumptions within and between several populations. *Educational and Psychological Measurement, 37, 863-872.*
Yoon, M., & Millsap, R. E. (2007). Detecting Violoations of Factorial Invariance Using Data-9 Based Specification Searches: A Monte Carlo Study. *Structural Equation Modeling:A Multidisciplinary Journal, 143, 435–463.*