Background

Latent Variable Modeling (LVM)

At the core of latent variable modeling is the idea that unobserved latent constructs influence observed variables (Maydeu-Olivares and Millsap 2009). For instance, an individual's underlying math ability can be expected to affect their performance on a math test. The common factor model (Thurstone 1947) posits that the relationship between latent constructs (e.g., math skill) and observed variables (e.g., scores on a math test) can be expressed using a system of equations. We can represent the relationship between J items ($j = 1, \ldots, J$) and M latent constructs ($m = 1, \ldots, M$) in matrix format as the following multiple-group factor analysis model

$$\boldsymbol{y}_{ig} = \boldsymbol{\nu}_g + \boldsymbol{\Lambda}_g \boldsymbol{\eta}_{ig} + \boldsymbol{\epsilon}_{ig}$$

where $\boldsymbol{y}_{ig}$ is a $J \times 1$ vector of observed item scores for individual i from group g, $\boldsymbol{\eta}_{ig}$ indicates a $M \times 1$ vector of latent factor scores, $\boldsymbol{\nu}_g$ is a $J \times 1$ vector of intercepts, $\boldsymbol{\Lambda}_g$ is a $J \times M$ matrix of factor loadings, and $\boldsymbol{\epsilon}_{ig}$ is a $J \times 1$ vector of unique factor variables. The latent factor scores are assumed to be distributed with $M \times 1$ mean vector $\mathrm{E}(\boldsymbol{\eta}) = \boldsymbol{\alpha}$ and $M \times M$ variance-covariance matrix $\mathrm{Cov}(\boldsymbol{\eta}) = \boldsymbol{\Psi}$. The unique factor variables are distributed with mean $\mathrm{E}(\boldsymbol{\epsilon}) = 0$ and variance-covariance matrix $\mathrm{Cov}(\boldsymbol{\epsilon}) = \Theta$. This relationship can also be expressed as the following system of equations at the item level:

$$y_{ijg} = \nu_{jg} + \lambda_{jg}\eta_{ijg} + \epsilon_{ijg}$$

where ϵ_{ijg} and η_{ijg} are independent.

As latent variables lack an inherent or universal scale, such a system suffers from indeterminacy, meaning that there exists no unique solution until some identification constraints are set to provide a scale to the latent variables (Bollen 2014; Maydeu-Olivares and Millsap 2009). For instance, the system can be identified by setting the latent mean and variances to 0 and 1 respectively, allowing for the comparison of factor loadings and intercepts across groups without an anchor variable (Van de Schoot, Lugtig, and Hox 2012). In this approach, the latent variable is standardized and all loadings are estimated freely. Alternatively, a reference indicator may be chosen and its loading and intercept may be constrained to 1 and 0 respectively. This approach sets the scale of the latent variable to the scale of the observed reference variable. Remaining loadings are estimated relative to the reference indicator, and the latent variance is freely estimated.

Once identified, the system of equations can be used to obtain composite sum or average scores or factor scores, which can be used in subsequent analyses investigating group-level differences or as predictor variables in

prediction problems (Maydeu-Olivares and Millsap 2009). Composite sum or average scores may be obtained by simply summing or averaging the raw scores on each item. This simplicity is made possible through several stringent assumptions that are unlikely to hold in practice, such as the unidimensionality of the construct, no measurement error, equal reliability across all items, and a perfectly linear relationship between each item and the latent construct (Maydeu-Olivares and Millsap 2009; DiStefano, Zhu, and Mindrila 2019; Grice 2001). As equal weights are assigned to each item, the sum/average scoring approach assumes that each item contributes equally to the underlying construct. However, this assumption can result in biased or inaccurate scores, particularly when certain items have a stronger relationship with the latent factor than others or when some items are less reliable, thus failing to account for differences in item quality and relevance.

Alternative, more refined (Grice 2001) approaches have been developed to address these limitations for continuous data, such as factor score estimation methods like Bartlett scores (Bartlett 1937), which accounts for the varying contributions of each item by weighting them according to their relationship with the latent construct, providing more accurate and reliable estimates of the underlying factor. Bartlett scores are designed to provide unbiased estimates of the true factor scores that are highly correlated with the common factors by minimizing the error variance (the unique factors) via a least squares approach (Grice 2001; DiStefano, Zhu, and Mindrila 2019). Bartlett factor scores are estimated using the following formula:

$$\widehat{\boldsymbol{F}} = (\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{y}$$

where $\boldsymbol{y}$ is the vector of observed item scores, $\boldsymbol{\Theta}^{-1}$ contains the inverse diagonal of the unique factor variances, and $\boldsymbol{\Lambda}$ is the factor pattern matrix of loadings (Lai et al. 2023).

Measurement Invariance (MI)

Observed test scores can only be considered comparable when measurements are on the same scale and latent construct(s) η are measured equivalently and comparably across groups and/or conditions (g) such that $P(y|\eta, W = g) = P(y|\eta)$, $\forall g$, i.e., when measurement invariance (MI) holds (Mellenbergh 1989; Widaman and Reise 1997). In other words, MI holds when the relationships between test items and the latent construct(s) are identical between groups and/or conditions. Systematic differences in measurement operations across grouping variables indicate measurement noninvariance, which may result in spurious inferences and make it impossible to separate the observed effects from construct-irrelevant attributes (Meredith 1993; Widaman and Reise 1997).

In a factor analytic framework, the parameters of a system relating test items to a latent variable can be

estimated and tested for equivalence using confirmatory factor analysis (Jöreskog 1969). The equivalence or nonequivalence of sets of item parameters determine the level of invariance, which are nested hierarchically (Meredith and Teresi 2006; Meredith 1993). Configural invariance (Horn and McArdle 1992) is said to hold when the factor structure (i.e., the configuration of the items and the factors) is equivalent across groups such that the pattern of zero and nonzero factor loadings is the same across groups. If configural invariance holds, the test measures similar (but not necessarily the same) latent constructs across groups (Widaman and Reise 1997). The factor means and variances are not identified in the configural model, and as different methods of identifying the model can lead to different parameter estimates under the configural model, group differences cannot be tested at this level (Widaman and Reise 1997). Metric (also called pattern or weak) invariance (Horn and McArdle 1992; Meredith and Teresi 2006) holds if loadings are equivalent across groups, suggesting equivalence of the strength and direction of the linear relationship between items and latent constructs across groups. With metric invariance, scores from different groups can be said to have the same unit or interval. Scalar (also called strong) invariance holds if in addition to loadings, intercepts are equivalent across groups. Scalar invariance is necessary for factor means to be comparable, and suggests that the meanings of the items as well as the levels of item responses are invariant across groups (Millsap 2011; Meredith and Teresi 2006; Widaman and Reise 1997; Van De Schoot et al. 2015). Strict invariance holds if all measurement parameters (i.e., loadings, intercepts, and uniqueness) are equivalent across the grouping variables, making any group-level differences in the observed variables entirely attributable to group-level differences in the underlying construct as it is measured identically across groups (Widaman and Reise 1997; Van de Schoot, Lugtig, and Hox 2012). The complete equivalence of measurement parameters necessitated by the strict invariance model is difficult to attain in practice (Van De Schoot et al. 2015), and most often, a partial invariance model (Byrne, Shavelson, and Muthén 1989) with a subset of invariant items is determined to enable group-level comparisons. A minimum of two items with scalar invariance (i.e., equal loadings and intercepts across groups) is needed for the comparison of latent factor means (Byrne, Shavelson, and Muthén 1989).

Testing begins with an unconstrained configural model reflecting the theoretical operationalization of the construct. If the configural model does not have acceptable fit, items may not be measuring the same construct across groups and testing may not proceed. If configural invariance is established, increasingly strict equality constraints are placed on sets of parameters (loadings, intercepts, and uniqueness) across groups to test for metric, scalar, and strict invariance (i.e., from the bottom of the hierarchy of measurement invariance stages to the top) in a sequential (stepwise) specification search where the lower stages need to be cleared at least partially to test for higher levels of invariance. Model fit may be compared with tests such as

the likelihood ratio (LRT) χ^2 test (Cochran 1952). If a particular stage of invariance does not hold (e.g., indicated by a significant χ^2 test), parameters are relaxed one at a time to test for partial invariance at that stage. Fit of the more and less constrained models are compared sequentially, and the parameter leading to the largest χ^2 difference is released. This process of model refitting and comparison continues from the lowest stage of invariance (metric) to the highest (strict) until the LRT is nonsignificant or there are no additional parameters that may be released, arriving at a partial invariance model (Van de Schoot, Lugtig, and Hox 2012). Note if any loadings are released in the partial metric model, the intercepts of noninvariant items are freed in the following models in line with established guidelines (Putnick and Bornstein 2016). Modification indices (Sörbom 1989; Yoon and Millsap 2007; Schmitt and Kuljanin 2008; Yoon and Kim 2014) may also be examined to assess the improvement in the model fit if a parameter was freely estimated vs. constrained.

While MI testing is commonly performed in the context of grouping variables such as ethnicity, gender, or time, it may also be of interest to test the equivalence of measurement operations across different studies. When pooling data from multiple studies, establishing MI within each study does not necessarily translate to achieving MI across the studies, or that the measure(s) administered in the studies are comparable (Curran and Hussong 2009). As such, MI testing may be performed considering study membership as a grouping variable.

References

Bartlett, Maurice S. 1937. "The Statistical Conception of Mental Factors." *British Journal of Psychology* 28 (1): 97.

Bollen, Kenneth A. 2014. *Structural Equations with Latent Variables*. John Wiley & Sons.

Byrne, Barbara M, Richard J Shavelson, and Bengt Muthén. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance." *Psychological Bulletin* 105 (3): 456.

Cochran, William G. 1952. "The χ2 Test of Goodness of Fit." *The Annals of Mathematical Statistics*, 315–45.

Curran, P. J., and A. M. Hussong. 2009. "Integrative Data Analysis: The Simultaneous Analysis of Multiple Data Sets." *Psychological Methods* 14 (2): 81–100. https://doi.org/10.1037/a0015914.

DiStefano, Christine, Min Zhu, and Diana Mindrila. 2019. "Understanding and Using Factor Scores: Considerations for the Applied Researcher." *Practical Assessment, Research, and Evaluation* 14 (1): 20.

Grice, James W. 2001. "Computing and Evaluating Factor Scores." *Psychological Methods* 6 (4): 430.

Horn, John L, and J Jack McArdle. 1992. "A Practical and Theoretical Guide to Measurement Invariance in Aging Research." *Experimental Aging Research* 18 (3): 117–44.

Jöreskog, K. G. 1969. "A General Approach to Confirmatory Maximum Likelihood Factor Analysis." *Psychometrika* 34(2): 183–202. https://doi.org/10.1007/BF02289343.

Lai, Mark HC, Winnie Wing-Yee Tse, Gengrui Zhang, Yixiao Li, and Yu-Yu Hsiao. 2023. "Correcting for Unreliability and Partial Invariance: A Two-Stage Path Analysis Approach." *Structural Equation Modeling: A Multidisciplinary Journal* 30 (2): 258–71.

Maydeu-Olivares, Alberto, and Roger E Millsap. 2009. "The SAGE Handbook of Quantitative Methods in Psychology."

Mellenbergh, Gideon J. 1989. "Item Bias and Item Response Theory." *International Journal of Educational Research* 13 (2): 127–43.

Meredith, William. 1993. "Measurement Invariance, Factor Analysis and Factorial Invariance." *Psychometrika* 58: 525–43.

Meredith, William, and Jeanne A Teresi. 2006. "An Essay on Measurement and Factorial Invariance." *Medical Care* 44 (11): S69–77.

Millsap, Roger E. 2011. *Statistical Approaches to Measurement Invariance*. Routledge. https://doi.org/10.4324/9780203821961.

Putnick, Diane L, and Marc H Bornstein. 2016. "Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research." *Developmental Review* 41: 71–90.

Schmitt, Neal, and Goran Kuljanin. 2008. "Measurement Invariance: Review of Practice and Implications." *Human Resource Management Review* 18 (4): 210–22.

Sörbom, Dag. 1989. "Model Modification." *Psychometrika* 54 (3): 371–84.

Thurstone, L. L. 1947. *Multiple Factor Analysis*. University of Chicago Press.

Van de Schoot, Rens, Peter Lugtig, and Joop Hox. 2012. "A Checklist for Testing Measurement Invariance." *European Journal of Developmental Psychology* 9 (4): 486–92.

Van De Schoot, Rens, Peter Schmidt, Alain De Beuckelaer, Kimberley Lek, and Marielle Zondervan-Zwijnenburg. 2015. "Measurement Invariance." *Frontiers in Psychology*. Frontiers Media SA.

Widaman, Keith F, and Steven P Reise. 1997. "Exploring the Measurement Invariance of Psychological Instruments: Applications in the Substance Use Domain."

Yoon, Myeongsun, and Eun Sook Kim. 2014. "A Comparison of Sequential and Nonsequential Specification Searches in Testing Factorial Invariance." *Behavior Research Methods* 46: 1199–1206.

Yoon, Myeongsun, and Roger E Millsap. 2007. "Detecting Violations of Factorial Invariance Using Data-Based Specification Searches: A Monte Carlo Study." *Structural Equation Modeling: A Multidisciplinary Journal* 14 (3): 435–63.