Multilevel Measurement Models (for Clustered Item-Level Data)

The Grand Finale!

CFA/IFA/IRT: Single-level Multivariate Measurement Model

+

MLM: Multilevel Model of Univariate Observed Outcome

→

M-SEM: Multilevel Multivariate Measurement Model

The Grand Finale: M-SEM

- **Multilevel structural equation modeling (M-SEM)** is a general term for latent trait measurement models that operate on multiple levels of sampling at once

- It combines the capabilities of:
  - Single-level latent trait measurement models for multivariate item responses with one level of theta(s)
  - Multilevel models for univariate observed outcomes whose variance is partitioned across higher level(s) of sampling

- Now we'll have (at least) **two levels of theta**, operating on:
  - e.g., In clustered data: within-level-1 = person residuals, between-level-2 = cluster random intercepts
    - *We will use this sampling context in our examples*
  - e.g., In longitudinal data: within-level-1 = occasion residuals, between-level-2 = person random intercepts *(assuming a lack of individual differences in change over time)*

Diagram from Pornprasertmanit et al., 2014

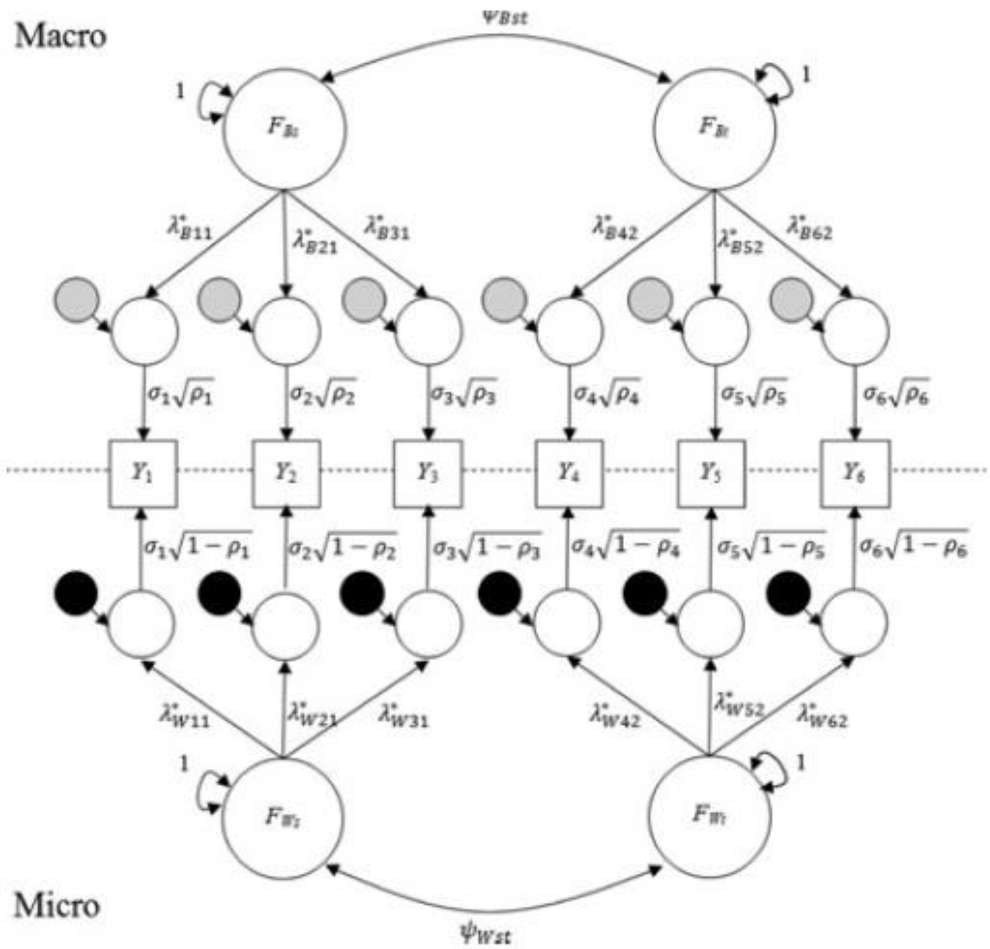

FIGURE 2 The multilevel CFA model. The grey circles represent macro-level unique factors, the variances of which are constrained equal to $1 - \lambda^{*2}_{Brs}$. The black circles represent micro-level unique factors, the variances of which are constrained equal to $1 - \lambda^{*2}_{Wrs}$. σ_r is the standard deviation of indicator r. ρ_r is the intraclass correlation of indicator r.

- Example two-factor M-SEM for 6 responses
  - macro = between-L2
    micro = within-L1
- All is now level-specific:
  - Trait interpretation
  - Assessment of fit, reliability, and validity
- Logical prerequisites:
  - "Enough" item ICCs
  - "Enough" L2*n* and L1*n*

Can I Just Ignore the Clustering?

- Single-level measurement model, ignoring clustering, called:
  - "Individual" (Stapleton et al., 2016)
  - "Disaggregated" (Pornprasertmanit et al., 2014)

- Consequences for measurement model of ignoring clustering
  - **Estimates**: not too much (given level-1 info >>> level-2 info)
  - **Standard errors**: too small (more so with higher ICCs)!
  - **Model fit**: looks too bad (more so with higher ICCs)!
    - Can be remedied somewhat through clustered-sample corrections

- So just use multilevel measurement (M-SEM) instead…
  - Relatively straightforward in application
  - **Intent and interpretation** are a different story!

Multilevel Intent: Word Salad

2 distinct types of cluster-level constructs:	**Shared experience**	**Aggregation of disparate individual responses**
Difference in item types:	Rate *your school* (common is the purpose)	Rate *yourself* (common is unintended)
Stapleton et al. (2016); Kozlowski & Klein (2000)	"shared"	"configural"
Marsh et al. (2012)	"climate"	"contextual"
Lüdtke et al. (2011)	"reflective"	"formative"
Within-level-1 person variance reflects:	Unreliability or disagreement	Expected and targeted variation
Between-level-2 cluster variance reflects:	Differences in actual construct of interest	Similarity for multiple and unknown reasons

- Flavors of multilevel measurement models:
  - "Multilevel CFA/SEM" predicts "continu-ish" item responses
  - "Multilevel IRT" predicts categorical item responses

Measurement Invariance across Clusters

- Measurement non-invariance by cluster [(Jak et al., 2013)]
  - **Uniform**: only intercepts/thresholds differ (main effect)
  - **Non-uniform**: at least loadings differ (≈interaction with factor)

- Levels of measurement invariance ("lack of cluster bias")
  - **Configural**: same factor structure form within-L1 and between-L2
  - **Weak**: same factor loadings within-L1 and between-L2
    - Otherwise, we can't meaningfully consider the analysis to yield the "between" and "within" parts of the same factor (are different traits)
  - **Strong**: no leftover random intercept variance in item responses after prediction by between-L2 factor(s)
    - Otherwise, some other cluster-level variable (beyond the between-L2 factor) is affecting the expected response for each cluster
    - Often done to fix model non-convergence or NPD solutions
- Structural invariance refers to factor relations across levels

Should I Use Across-Level Constraints?

- Constraining factor loadings equal across within-L1 and between-L2 ("weak cluster invariance") is often recommended:
  - For the **between traits** to be interpreted as the **cluster aggregates of the within traits** (i.e., as is the case for random intercepts in MLMs)
  - To improve parsimony and aid in model convergence (Jak, 2019)
  - If all loadings are estimated, within-level trait variances are fixed to 1 for **shared identification**; between-level trait variances are then estimated
    - Alternatively, use a marker item and estimate trait variance at both levels
  - Is assumed in three-level MLMs (Rasch-type items in people in clusters)

- Otherwise, trait variances must be separately identified at each level (although trait means are only parameters at between-L2)
  - Within-L1 traits and between-L2 traits capture **conceptually different constructs** (and can't be easily put back together again)
  - It would not make sense to constrain leftover random intercept variances to 0 in this case (see also Geldhof et al., 2014)
  - It would also not make sense to use the same marker item at both levels

Complications: CFA vs. IFA/IRT

- The prior (prototypical) references all dealt with the "CFA" case of the measurement model in M-SEM (now estimated by FIML):
  - **CFA**: Continuous, normally-distributed person responses to items (or other outcomes) are predicted linearly by person latent traits

- So the total amount of variance in each response can be partitioned into model-estimated orthogonal components → within-L1 and between-L2 covariance matrices
  - Within-L1: individual deviations around cluster means
  - Between-L2: cluster mean deviations around sample mean
  - ≈ "Latent centering" version of cluster-mean-centering

- Because a saturated model (of all possible variances and covariances) is then possible *at each level*:
  - Get usual indices of model fit (and modification indices to fix it), although overall fit indices mostly address the within-L1 model (see Ryu & West, 2009; Hsu et al., 2015)
  - Flexible range of models (e.g., nothing, no covariances, saturated)

Complications: CFA vs. IFA/IRT

- Switching to a **generalized version** of M-SEM (i.e., multilevel IRT) then implies:
  - Intercept/threshold measurement model parameters are then "**unit-specific**": conditional on their corresponding random effects = 0 (not a distinction in CFA as a general-type model)
  - Because level-1 residual variance is not estimated, there is no easy "saturated model" for within-level covariances unless you resort to limited information estimation (i.e., via polychoric correlations)

- Potential ambiguity about interpretation of between-L2 factor loadings: **between** or **contextual**?
  - **Between** = all level-2; **contextual** = level-2 after controlling for level-1 (whenever level-1 variable still has level-2 variance in it)
  - Given that the latent traits are uncorrelated across levels, we believe the between-L2 loadings are indeed **between**
  - How else to verify? In non-Bayes estimation, remove the within-L1 traits—if level-2 loadings change a lot, they must be contextual

Example Models (as 03_modelxx.stan)

1. Empty (non-measurement) two-level model with correlated random item intercepts
2. Within-school (WS) measurement model with correlated random item intercepts and within-school discriminations fixed=1
3. WS measurement model with correlated random item intercepts and estimated within-school discriminations using standardized theta
4. WS measurement model with correlated random item intercepts and estimated within-school discriminations using item1=marker
5. WS and between-school (BS) measurement model with uncorrelated random item intercepts and estimated level-specific WS (item1=marker) and BS (item10=marker) discriminations
6. WS and BS measurement model with uncorrelated random item intercepts and estimated level-constrained WS (item1=marker) and BS (item1=marker) discriminations
7. WS and BS measurement model without random item intercepts and with estimated level-constrained WS (item1=marker) and BS (item1=marker) discriminations
8. WS and BS measurement model with uncorrelated random item intercepts and free/reduced lunch MLM predictor and estimated level-constrained WS (item1=marker) and BS (item1=marker) discriminations