Simulation Study Brainstorming

**Purpose (overall study)**:

# Clarify and demonstrate common issues related to the interpretation and reliability of subscale scores derived from hierarchical/bi-factor/two-tier IRT models.

# Propose and evaluate a method for obtaining more reliable and meaningful subscale scores from the hierarchical/bi-factor/two-tier models.

**Empirical analysis**:

* Purpose/goal – (I.) above
* Summary: The empirical analysis will demonstrate the implications of naively interpreting subscores resulting from a hierarchical multidimensional IRT model (e.g., bifactor/two-tier) using real data from the SF-36 health outcomes instrument. In particular, we’ll show how (i) naïve interpretation of specific factors is incorrect when the goal of the assessment is to produce subscale scores that represent ability on the specific factors entirely (as opposed to the specific factors after controlling for the general factor), and (ii) these estimates are marred by poor reliability that compounds the problem with interpreting them as subscores. We’ll then demonstrate how these problems may be mitigated by a corrective procedure.
* Method
  + Demonstrate 3 model fittings, each representing a different goal for the overall assessment
    - Model fitting 1:
      * Interest lies in controlling for nuisance factors, which in our case might be positively/negatively worded items
    - Model fitting 2:
      * Interest lies in obtaining subscale scores using a bi-factor model (single-tier)
    - Model fitting 3:
      * Interest lies in obtaining subscale scores using a two-tier model (2 general factors)
    - For each:
      * Discuss how the specific factors would be interpreted and highlight issues when these interpretations don’t align with the goals of the assessment
      * Demonstrate the poor reliability of the subscale scores:
        + Reliability indices
        + Test information/standard error of measurement
  + Demonstrate the corrective procedure
    - Demonstrate the corrective procedure on Model fitting 2 and Model fitting 3
    - Once again, discuss the resulting interpretation of subscale scores & show reliability/test information/etc. for the newly obtain scores. Contrast these with the previous models/results.

**Simulation study**:

* Purpose/goal – (II.) above
* Summary: The simulation study is used to assess the performance of the corrective procedure in terms of its ability to produce more reliable and accurate subscale scores. We’ll manipulate factors related to the structure of the simulated instrument that we think might influence the procedure’s performance. We’re interested in identifying scenarios in which use of this corrective procedure is (and is not) particularly important in terms of improving subscale score accuracy/reliability
* **Possible evaluation criteria**:
* Reliability measures
* Test information
* Residual correlations (Reckase, 1997; Segall, 2001)?
* True/Estimated
  + Correlation between
  + RMSE
  + Bias
* **Exhaustive list of factors to consider**:
* Number of specific dimensions
* Number of items per dimension
  + If so, keep fixed across subscales?
* Magnitude of alphas on specific dimensions
  + As the magnitude gets smaller relative to the general dimension, then there’s less and less reason to report subscale scores. That is, in a real application, such a situation would mean that the user should just report general scores and not subscale scores at all.
  + Despite this, stakeholders sometimes still require people to report subscale scores
* Ratio of alphas, specific/general
  + This is another way of looking at the magnitude of the alphas on the specific dimensions. As the ratio of alpha\_specific / alpha\_general gets smaller, it means that the specific dimension gets less and less important.
* DBM
  + Manipulate the DBM, i.e., manipulate how far away from the general dimension the subscale is.
  + This one is a given
  + Difficulty: inextricably linked to magnitude of alphas/ratio of alphas/heterogeneity among angles in subscale
* Heterogeneity among angles in a subscale
  + I.e., How dispersed are the item-specific angles of best measurement
  + Homogenous vs. Heterogenous
* Correlation between general dimensions (for two-tier)
* Fitted model
  + For nuisance factors; bi-factor; two-tier

**Random notes**:

“If loadings are high on the general factor but low on the specific factor, it might be most defensible to report the general factor scores alone.” [because most of the reliable portion of the subscale was due to the general factor and thus the subscale scores were largely redundant.]

“In summary, when loadings on the specific factors are high, factor score estimates for the specific scores can be meaningful as long as score users are carefully educated about the fact that each subscale reflects information above and beyond the skill or trait reflected in the general score.

Reliability measures

* Omega Hierarchical
  + Model-based reliability index
  + Estimates the proportion of variance in total scores that can be attributed to a single general factor, thereby, treating variability in scores due to group factors as measurement error
  + E.g. OmegaH = 0.93 would mean that 93% of the variance of unit-weighted total scores can be attributed to the individual differences on the general factor. The square root is the correlation between the general factor and the observed total scores.
* Omega Hierarchical Subscale
  + Extended to subscales by computing the unique variance associated with each group factor once portioning out variance associated with a general factor
  + Index reflects the reliability of a subscale score after controlling for the variance due to the general factor (Reise, Bonifay et al., 2013)
* deMars
  + 1 – (s^2\_e / s^2\_{\theta}) [Thissen & Orlando, 2001, p. 118; Wainer, Bradlow & Wang, 2007, p. 76]