1

Measurement Invariance

INTRODUCTION

This primer discusses strategies for testing measurement invariance. The literature on such tests is complex and filled with conflicting advice. Here, I introduce you to strategies for testing invariance using structural equation modeling (SEM). There is a parallel literature on differential item functioning in Item Response Theory, but I do not consider it. I focus on an example where I test for measurement invariance for a mediator in an RET. However, the concepts apply to any measure, be it a mediator, a moderator, an outcome, or a covariate.

The most common formulation of measurement invariance expresses an observed measure (X) as a linear function of (a) a latent variable (LX) presumed to reflect an individual's true standing on the measured construct and (b) random measurement error:

$$X_i = \alpha + \lambda LX_i + \varepsilon_i \qquad [1]$$

where α is the measurement intercept, λ is the path coefficient (factor loading) linking LX to X_i, ε_i is an error term with a mean of 0, and "i" refers to an individual. Figure 1 shows an influence diagram using a measure of depression, D. D is influenced by "true" depression,

indicated in the circle, and random measurement error, ε.

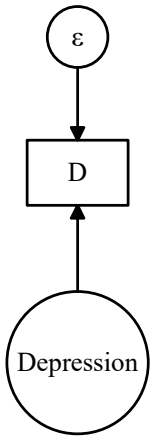

FIGURE 1. Measurement model

Measurement invariance is characterized by equal values of the measurement intercept across groups, equal values of the factor loadings across groups, and/or equal measurement error variances, σ_e^2, across groups. If such equality holds, the measurement parameters are said to be **invariant**. If one or more of the values are not equal, the measurement parameters are said to be **non-invariant**. The terms **loading non-invariance**, **intercept non-invariance**, and **error variance non-invariance** are sometimes used to indicate which type of parameter is problematic. If two or more groups have different values for λ, then this can undermine group comparisons of regression coefficients that regress Y onto LX. If two or more groups have different values for α (or λ), then this can undermine group comparisons of means on LX. I discuss why this is the case and provide examples in my book. I use the term **measurement non-invariance** in this document to refer to the case where either loading or intercept non-invariance occurs.

If one has reason to question the measurement invariance of a measure, it can be helpful in an RET to empirically evaluate its presence, although doing so is not always possible. If not possible, then we must implicitly make assumptions of measurement invariance in our analyses. This assumption is not necessarily damning because perhaps the assumption is reasonable or, even if violated, the violations may not be consequential. In SEM one needs multiple interchangeable indicators of the same variable to evaluate measurement invariance, i.e., we need to work with latent variables.

Psychometricians can explore measurement invariance for items of a scale or at the

level of scale composites (Johnson, Meade & DuVernet, 2009). During scale construction, item level measurement invariance analyses can be used to eliminate items that are ambiguous or that lack psychometric generality. In RETs, concern is usually with measurement invariance at the level of composites. For example, with multiple indicator latent variable models where, say, the CES-D measure of depression, the PH-9 measure of depression, and the Beck depression inventory are used as indicators of depression, one would want to document invariance of the scale composites because, ultimately, composites are the focus of analysis and are what are contained in the model.

Generally speaking, if multiple items that comprise a composite are non-invariant in the same direction (e.g., all of the items show lower measurement intercepts in one group as compared to the other group), this bias will manifest itself in the composite (Steinmetz, 2013). However, it is possible that the non-invariance of a given item in one direction might be offset by non-invariance in an opposite direction of another item, with the opposing non-invariances cancelling when the composite is formed by summing or averaging items. This dynamic is analogous to the case in which items impacted by positive random errors cancel the negative random errors that influence other items when we form item composites. If some items are biased upward by a response set tendency as a function of biological sex but other items are biased downward by a different response set as a function of biological sex, the net effect of the non-invariance when items are summed or averaged can be negligible. By the same token, one or two items that are somewhat non-invariant in the same direction on an inventory with, say, 20 items, may have little impact on substantive conclusions that rely on composites because the non-invariance of these items is swamped by the invariance of the other items when the composite is formed. You will find that most of the measurement invariance research literature focuses on item-level analyses of measurement invariance but this fails to recognize that the majority of social science research works with composite scores and that measurement invariance properties of the composites are where the focus needs to be. Note also that even if the individual items have a binary response metric, when summed or averaged, the metric of the composite is typically treated as being approximately interval-level. Analysis of composite level measurement invariance should use the metric of the composite, not the metric of the individual items, but this rarely is the case in practice.

You also will encounter in the measurement invariance literature a description of different types of invariance and a recommended order in which you should conduct invariance tests. I do not subscribe to this approach. Evaluation of measurement invariance is complicated and cannot be reduced to a rote sequence of steps that are mindlessly followed without taking into account the broader theoretical context. In most RETs, the non-invariance of factor loadings is key as is the non-invariance of measurement intercepts. Non-invariance of measurement error variances usually is of lesser concern because the effects

of such non-invariance often are inconsequential or can be accommodated analytically through SEM.

I will work initially with an example where I have three composite measures of the same construct, depression. All three are self-reports and each is comprised of multiple items. A given scale asks people to complete items on 5 point disagree-agree metrics that are averaged across items. The total scores range from -2 to +2, with 0 representing a neutral point, negative numbers reflecting disagreement with the statements and positive numbers reflecting agreement. Higher scores indicate greater levels of depression. Although the three scales share a common metric (from -2 to +2), this need not be the case in practice. The example is for an elderly population. Depression is a mediator and is assumed to influence the extent to which people adhere to a proscribed weekly exercise protocol. Scores on the outcome range from 0 to 100 and represent percent adherence per week. A person with a score of 50 regularly completes half of the proscribed exercise protocol. A person with a score of 100 regularly completes all of the proscribed protocol. I will assume that each of the three composites are functionally unidimensional and that I have verified this empirically. The study includes both males and females. I ultimately want to evaluate if the effect of depression on exercise adherence varies as a function of biological sex. I also want to explore sex differences in mean levels of depression as a secondary question. I therefore focus on measurement invariance for the depression measures as a function of biological sex. I am going to assume you are reasonably familiar with the basics of SEM and Mplus.

I first provide examples of traditional approaches to the assessment of measurement invariance. These approaches rely on multi-group SEM. I then consider more modern approaches to measurement invariance testing based on alignment analyses, Bayesian methods, equivalence testing, and moderated factor analysis/MIMIC modeling. I also discuss the testing of measurement invariance in longitudinal data and conclude with recommendations for how to test for non-invariance and what to do about it should it occur. Before considering these topics, there are measurement invariance concepts I need to introduce. These include forward and backward analysis, choice of a reference indicator, effect sizes, and omnibus testing.

Forward and Backward Analysis

The traditional method for evaluating measurement invariance uses multi-group SEM. In the depression-exercise adherence study, I might construct a model that conceptualizes the three depression measures as each influenced by a single latent variable of depression and then estimate the factor loadings for the three scales for males and females, separately. I then conduct three pairwise significance tests to see if the factor loading for the first indicator is statistically significantly different for males versus females, if the factor loading

for the second indicator is statistically significantly different for males and females, and if the factor loading for the third indicator is statistically significantly different for males and females when estimates are made from separate model estimations. This is referred to as a **forward strategy** (see Jung & Yoon, 2016). A **backward strategy**, by contrast, first estimates the fit of a model that forces all measurement parameters to be equal across groups, e.g., Mplus is told that when fitting the model to males and females, it must force the factor loadings for the first indicator to be equal across groups, and similarly so for the second and third indicators. I then estimate a second model but I relax one of the equality constraints. For example, the second model might relax the constraint that the factor loading for the first indicator for males must equal the corresponding factor loading for females. I obtain an overall chi square fit index for the fully constrained model and then compare this to the chi square fit index for the relaxed model. If the factor loadings for the first indicator for males and females are equal in the population, we should find that the fit of the constrained and relaxed models are comparable; that the equality constraint does not matter because, after all, the loadings *are* equal in the population. However, if the fit of the relaxed model is much better than the fit of the constrained model, then this suggests the two loadings are not equal in the population. The significance test is formally executed in the form of a chi square difference test, which calculates the difference between the two model chi squares and a p value for the resulting difference. The process is repeated for each factor loading, so I would have three such chi square difference tests.

Whereas the forward strategy begins with a fully unconstrained model and then derives tests of parameter differences in that model, the backwards strategy starts with a fully constrained model and then selectively frees up constraints. Both strategies have strengths and weaknesses. For example, the accuracy of the p values in the backward strategy can be impacted by the degree of ill fit in the relaxed model; if both the constrained and the relaxed model fit the data poorly, then the p value for the chi square difference can be erroneous (Yuan & Bentler, 2004; Yuan & Chan, 2016). I elaborate this dynamic below.

The Choice of a Reference Indicator

A complication with tests of invariance is that the results of the tests can vary depending on the choice of the referent indicator used to define the metric of the latent variable. In order for traditional tests of metric invariance to be valid, the reference indicator must itself have the property of invariance (Cheung & Rensvold, 1999; Johnson, Meade, & DuVernet, 2009). However, in many cases, one does not know *a priori* if the reference indicator is measurement invariant across the target groups. Raykov et al. (2012) suggest an approach that circumvents this problem by defining metrics for the latent variables not through reference indicators but instead by fixing the variance of the latent variable to a value of

1.0. In this case, the choice of a reference indicator becomes moot because there is no reference indicator. The approach of fixing the factor variance to 1.0 is called the **fixed factor variance method**. Another method of scaling the latent variable is called **effects coding** and is described by Little, Slegers and Card (2006). It also does not require choosing a reference indicator but it only is useful when the indicators of the latent variable are on comparable metrics, which often is not the case. For example, if d1 ranges from -2 to +2 with a variance of 1, but d2 ranges from 0 to 100 with a variance of 25, the method can fail. I consider the reference indicator and fixed factor variance methods given their generality.

Non-Invariance Effect Sizes

If measurement non-invariance occurs, this does not mean that it is consequential. Measurement non-invariance can bias regression coefficients, correlations, and factor means but the question is by how much and does it matter? Oberski (2014) and others (e.g., Chen, 2008) suggest examining how parameters of substantive interest change under assumptions of measurement invariance versus non-invariance. As an example, Oberski (2014) examined the effects of metric invariance for a model focused on a substantive path coefficient across 19 countries. He compared the value of the path coefficient when loading equality was enforced across countries for a given loading indicator with the value of the path coefficient when the loadings were freely estimated in the groups. If the change in the substantive path coefficient yielded in the two cases is minimal or functionally zero, then the non-invariance is deemed inconsequential because the path coefficients of primary interest are unaffected by whether or not measurement invariance is imposed.

Independent of the approach advocated by Oberski, methodologists also evaluate non-invariance using standardized effect size indices analogous to Cohen's (1988) effect size approach (Cohen, 1988). For example, Pornprasertmanit, Lee, and Preacher (2014) define indices of standardized effect size for differences in measurement intercepts, loadings, and error variances (see also Pornprasertmanit, 2021). I illustrate both the Obserski and the Pornprasertmanit et al. approaches below.

Omnibus Tests

When testing for measurement invariance, many authors suggest conducting omnibus tests of invariance first and then conditioning exploration of localized tests of invariance on whether the omnibus tests are statistically significant. For example, one might first test for loading invariance across all of the 20 items of a scale considered simultaneously and only if one rejects the null hypothesis of invariance across the 20 items would one pursue invariance tests on an item-by-item basis to locate where the non-invariance resides. I recommend against this practice for several reasons. First, it is possible for the omnibus test

of non-invariance to be statistically non-significant even though a localized, item-specific contrast is meaningfully non-invariant. If the vast majority of items are loading invariant, they can swamp the non-invariance of a single parameter when conducting a global test of non-invariance. Second, the logic of conducting the omnibus test first usually is to protect against inflated Type I error rates that result from conducting multiple contrasts. However, simulations have shown that this strategy typically is deficient for controlling familywise error rates (e.g., Jaccard, 1998). A better strategy is to apply a modified Bonferroni method or the False Discovery Rate (FDR) method to the localized contrasts without an initial "screening" omnibus test. Finally, the two-step strategy undermines the statistical theory for the significance tests of the localized item-specific contrasts. That theory is formulated without taking into account a prior omnibus test as a first step. The omnibus test makes its own assumptions (that may be untenable) and it may lack statistical power. By invoking it, we change the sampling distribution of the localized test, often in unknown ways. As such, we can no longer trust the p values and confidence intervals for the localized tests. Typically, it is better to move directly to localized contrasts and adjust for multiplicity using a modified Bonferroni or FDR method.

MULTIPLE GROUP SEM: TRADITIONAL ANALYSES

I illustrate traditional measurement invariance analyses using the depression-exercise adherence example. I created simulated data from a population in which the first (D1) and third (D3) measures of depression were invariant across males and females in terms of their factor loadings but the second measure (D2) was not. All three measures were invariant across biological sex in terms of their population measurement intercepts and the population latent mean depression scores also were equivalent. When applying multi-group SEM invariance tests, analysts often fix model parameters at convenient values to ensure model identification. It is important to keep in mind the conditional nature of measurement invariance conclusions relative to such decisions (see Steiger, 2002, for elaboration).

Analysis of Factor Loadings

I first illustrate analyses to provide perspectives on the invariance of the factor loadings for depression as a function of biological sex. If, for example, I plan in my RET to evaluate biological sex as a moderator of the effects of depression on exercise adherence, then loading non-invariance of the depression measures across sex is potentially relevant. Given that I have multiple, interchangeable indicators of depression, I am in a position where I can empirically gain perspectives on their loading non-invariance across sex. I first consider a traditional strategy using forward analysis and then a strategy using backward analysis.

A Forward Analysis Strategy

The first set of analyses for the forward strategy seeks to provide an initial sense of loading invariance as well as information to help me choose a reference indicator for depression for later invariance analyses. The relevant Mplus syntax for this step is in Table 1. I number the lines for reference, but Mplus syntax excludes the numbers and the periods after them. The syntax could be more efficient but I sacrifice efficiency in the interest of pedagogy. I assume you have reviewed the basics of Mplus syntax on my website and are familiar with Mplus. This initial analysis assumes that the variance of the target latent variable, depression, is equal or approximately equal in the two groups, males and females. Violations of this assumption might be problematic, so I check its viability in later analyses.

Table 1: Mplus Syntax for Identifying a Reference Indicator

```
1. TITLE: TEST OF LOADING INVARIANCE ;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1 d2 d3 d4 d5 d6 adhere adhere2 dfemale income ethnic ;
5. USEVARIABLES ARE d1 d2 d3 adhere  ;
6. MISSING ARE ALL (-9999) ;
7. GROUPING IS dfemale (0=male 1=female) ;
8. ANALYSIS:
9. ESTIMATOR = MLR ;
10. MODEL:
11. LX BY d1* d2* d3* ;
12. LX@1 ;
13. adhere ON LX ;
14. MODEL female:
15. LX BY d1* d2* d3* (fL1 fL2 fL3) ;
16. LX@1 ;
17. adhere ON LX (fp) ;
18. MODEL male:
19. LX BY d1* d2* d3* (mL1 mL2 mL3) ;
20. LX@1 ;
21. adhere ON LX (mp) ;
22. MODEL CONSTRAINT:
23. NEW(diff1 diff2 diff3);
24. diff1=fL1-mL1 ;
25. diff2=fL2-mL2 ;
26. diff3=fL3-mL3 ;
27. OUTPUT: SAMP RESIDUAL CINTERVAL TECH4 STAND(STDYX) ;
```

Line 1 is the title line. Line 2 tells Mplus where to find the data file. Each line in the data file contains 12 values, space delimited, and contains the scores for a given individual on

the 12 input variables. Line 3 tells Mplus I am going to provide information about the variables that are in the data set. Line 4 provides the names I want to assign to the variables in the order they are encountered in the data file. There are 12 names because there are 12 variables. Line 5 specifies the subset of variables I want to use in the model. Line 6 tells Mplus that if it encounters the value -9999 for any of the variables, it should treat it as missing data. By default, Mplus uses full information maximum likelihood (FIML) for missing data for the single factor CFA conducted here. Line 7 defines the two groups I want to focus on; for the variable called `dfemale`, a score of 0 defines males and a score of 1 defines females. Line 8 tells Mplus I am going to provide information about the type of analysis I want. Line 9 specifies the estimator for the analysis to be robust maximum likelihood, to help deal with non-normality. Lines 10 through 13 provide the general model form to apply to each subgroup, but I will supplement or override aspects of this general model in later commands when I specify a separate model for females and then for males. Line 10 tells Mplus I am going to provide information about the general model. Line 11 specifies a model with a latent variable called LX (I can name this factor anything, but I cannot have a name that exceeds 8 characters) as reflected `BY` 3 observed indicators, variables d1 through d3. These are the three observed depression measures. The * after each variable name tells Mplus to estimate the factor loading for each indicator. Line 12 tells Mplus to fix the variance of the latent variable to 1.0. The @ sign is read as "fix the referenced parameter to a value of….", followed by the value you want to fix the parameter to. In Mplus, listing a variable by name refers to the variance of the variable (or the error variance, if the variable is endogenous). This is why Line 12 fixes the factor variance to 1.0. It can be read as "for the variance of LX, fix it to a value of 1.0." This invokes the fixed factor variance method I described earlier for defining a latent variable metric. The latent variable has a mean of 0 (the Mplus default). Line 13 tells Mplus to regress the adherence outcome variable onto the latent depression variable.

Lines 14 to 17 specify the model for females (be sure to use the same label in the `MODEL` command that you used in Line 7) and lines 18 to 21 specify the model for males. Both of these are identical to the general model with the exception that I added labels to some of the parameters. For females, in line 15 I added the label fL1 to refer to the female loading for the first indicator, fL2 to refer to the female loading for the second indicator, and fL3 to refer to the female loading for the third indicator. The labels have an 8-character maximum; I chose these labels because I felt they were reasonable acronyms. Note that I used the same labels for males in line 19 but used an *m* instead of an *f* as the first letter of the acronym.

Line 22 tells Mplus I want to conduct some contrasts and Line 23 indicates those contrasts will be called diff1, diff2, and diff3. Lines 24 to 26 specify each of the contrasts

using the parameter labels I assigned above. The contrast diff1 is the loading for the first indicator for females minus the loading for the first indicator for males; diff2 is the loading for the first indicator for females minus the loading for the second indicator for males; diff3 is the loading for the third indicator for females minus the loading for the third indicator for males. Finally, line 28 is the output line. I discuss the different options for the output line on the syntax tab of my webpage.

The model yielded reasonable global fit indices. The chi square fit index for the multi-group model was 6.17 with 6 degrees of freedom ($p < 0.98$), the CFI was 1.00, the RMSEA was 0.006 with a 90% confidence interval of 0.000 to 0.048, the p value for close fit was 0.96 and the standardized RMR was 0.006. When you use the `MODEL CONSTRAINT` command, Mplus will not show modification indices. I often first execute the syntax commenting out all of the `MODEL CONSTRAINT` commands so I can examine modification indices to assure a reasonable model fit. Then, I re-execute the program using the full syntax in Table 1. I want to be sure that the model fits well in both groups separately and doing so is often referred to as **configural invariance**, i.e., the basic form of the model reproduces the data well even though the specific values of the parameters can differ across groups. The favorable global fit indices are consistent with configural invariance, but I am also careful to examine fit diagnostics produced by Mplus for each group separately in separate Mplus runs (see my book for details).

Table 2 presents the factor loadings for males and females as taken from the Mplus output, the significance tests of their difference (a * means $p < 0.05$), the 95% confidence intervals for the differences and the margins of error based on the confidence intervals (i.e., the half width of the intervals).

Table 2: Unstandardized Factor Loadings for Forward Analysis

Measure	Female	Male	Difference	CI for Difference	Margin of Error
d1	0.917	0.881	0.036	-0.047 to 0.119	±0.083
d2	0.805	0.906	-0.102*	-0.179 to -0.025	±0.127
d3	0.925	0.874	0.051	-0.028 to 0.130	±0.079

There is a statistically significant difference in the loadings as a function of biological sex for the second depression measure. This suggests that either the first or third depression measure can be used as a reference indicator in later analyses where I move away from the fixed variance approach to a reference indicator approach when defining the metric of the latent variable. Because I performed 3 contrasts comparing males to females (one for each

loading), some researchers argue I should adjust for familywise error by applying a False Discovery Rate correction or a Holm modified Bonferroni correction. I provide programs to do so on the program tab of my website. From the Mplus output, the unadjusted p values for the three contrasts were 0.400 for d1, 0.010 for d2, and 0.206 for d3. When adjusted using FDR logic, they became 0.400, 0.030, and 0.309, respectively; d2 remains non-viable as a reference indicator because of its loading invariance, at least in terms of statistical significance.

Evaluating the magnitude of the loading differences in Table 2 can be challenging because the values of the loadings are impacted by the metrics of both the indicators and the latent variable. In the current case, the standard deviations of d1, d2, and d3 are all close to 1.0, and the variance of the latent variable is fixed at 1.0, so the unstandardized loadings are similar to standardized loadings, but this will not always be the case. Pornprasertmanit (2014) instead uses the effect size framework of Cohen (1988) to convert the raw difference in loadings for a given contrast to a standardized difference using the formula:

$$ES = (\lambda_1 - \lambda_2)\left(\sigma_{LX\text{-}POOLED} / \sigma_{X\text{-}POOLED}\right) \quad [2]$$

where λ_1 is the unstandardized factor loading for group 1 for indicator X, λ_2 is the unstandardized factor loading for group 2 for indicator X, $\sigma_{LX\text{-}POOLED}$ is the pooled standard deviation of the latent variable for the two groups in question, and $\sigma_{X\text{-}POOLED}$ is the pooled standard deviation of indicator X for the two groups (you can use the program for pooled variances and standard deviations on the 'programs' tab of my webpage to calculate a sample estimate of $\sigma_{LX\text{-}POOLED}$ and $\sigma_{X\text{-}POOLED}$ from Mplus output). In the current example for d1, the unstandardized loading for females is 0.917 and for males it is 0.881. The pooled standard deviation for LX is 1.00 (because I fixed the latent variance to be 1.00 in each group) and the pooled standard deviation for d3 (using a female variance = 1.033, female n = 723, male variance = 0.974, male n = 778) is 1.00. The ES is thus (0.917-0.881)(1.00/1.00) = 0.036. A standardized difference of 0.036 seems minor. For d2, the absolute standardized effect size was 0.106 and for d3 it was 0.055.

Based on Table 2, the standardized effect sizes, and substantive considerations, I might tentatively decide to use d1 as the reference indicator. Note that this decision is not based purely on statistical grounds, such as selecting the indicator that produces the largest p value or smallest loading invariance effect size (Thompson, Song, Shi & Liu, 2020). The decision might also be influenced by the non-arbitrariness of the indicator metric, the familiarity of the metric to other scientists/practitioners, and its broader psychometric history. Having made this decision, I next re-do the above analysis but now using d1 as a reference indicator and using bootstrapping instead of robust maximum likelihood for sensitivity purposes. Table 3 presents the relevant Mplus syntax.

Table 3: Mplus Syntax for Loading Invariance with a Reference Indicator

```
1. TITLE: TEST OF LOADING INVARIANCE ;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1 d2 d3 d4 d5 d6 adhere adhere2 dfemale income ethnic ;
5. USEVARIABLES ARE d1 d2 d3 adhere  ;
6. MISSING ARE ALL (-9999) ;
7. GROUPING IS dfemale (0=male 1=female) ;
8. ANALYSIS:
9. ESTIMATOR = ML ; BOOTSTRAP=2000;
10. MODEL:
11. LX BY d1@1 d2* d3* ;
12. LX* ;
13. adhere ON LX ;
14. MODEL female:
15. LX BY d1@1 d2* d3* (fL1 fL2 fL3) ;
16. LX* (fvar) ;
17. adhere ON LX (fp) ;
18. MODEL male:
19. LX BY d1@1 d2* d3* (mL1 mL2 mL3) ;
20. LX* (mvar) ;
21. adhere ON LX (mp) ;
22. MODEL CONSTRAINT:
23. NEW(d2diff d3diff vdiff pdiff);
24. d2diff=fL2-mL2 ;
25. d3diff=fL3-mL3 ;
26. vdiff=fvar-mvar ;
27. pdiff=fp-mp;
28. OUTPUT: SAMP RESIDUAL CINTERVAL(BOOTSTRAP) TECH4 STAND(STDYX) ;
```

The differences in the syntax relative to Table 1 are as follows. Line 9 changes the MLR option to ML (for traditional maximum likelihood) and bootstrapping is invoked with 2000 bootstrap replicates. Lines 11, 15 and 19 each fix the path from LX to the first indicator to the value of 1.0, which functionally passes the metric of d1 to LX, but with adjustments for measurement error. The * symbols tells Mplus to estimate the loadings for d2 and d3, but the * are optional because it is the default, except for the first indicator, whose Mplus default is to fix the indicator at 1.0. Lines 12, 16 and 20 tell Mplus to estimate the variance of LX for males and females, respectively, rather than fixing the LX variance at 1.0. I thus relax the assumption of equal LX variances for males and females from the initial analysis. In addition, I add labels for the two variances (fvar and mvar) for eventual use in the `MODEL CONSTRAINTS` command. In Line 26, I add a contrast (vdiff) between the LX variance for females and males to determine if the variances are statistically significantly different. I also eliminate Line 24 from Table 1, because I am not able to test for differences for the reference

indicators given they were fixed to 1.0. Finally, in Line 28, I add the BOOTSTRAP keyword to the CINTERVAL keyword to obtain bootstrapped confidence intervals. The key results for the factor loadings are in Table 4.

Table 4: Unstandardized Factor Loadings with Reference Indicator

Measure	Female	Male	Difference	CI for Difference	Margin of Error
d1	1.000	1.000	-	-	-
d2	0.877	1.029	-0.151*	-0.225 to -0.077	±0.074
d3	1.009	0.992	0.017	-0.059 to 0.091	±0.076

The magnitudes of the factor loadings are different from those in Table 1 because I am using a different LX metric. In the first analysis I conducted, I fixed the LX variance to 1.0 for both males and females. In the current analysis, I estimated the variances based on the use of d1 as a reference indicator, so the variance of LX becomes that of d1 with an adjustment for measurement error. The variance of LX for females was estimated to be 0.841 and for males it was 0.777, a difference that was not statistically significantly different (based on the contrast in MODEL CONSTRAINTS listed on Line 26, z = 0.85, ns). My presumption in the first analysis of functionally equal LX variances was reasonable.

For the significance tests of loading differences, the contrast for the second loading was again statistically significant indicating loading non-invariance and for the third loading it was again statistically non-significant. When I converted the unstandardized loading differences to standardized effect sizes using Equation 2, the ES for the second loading was -0.141 and for the third loading it was 0.015. The general conclusions about factor loading non-invariance converge with those from the initial analysis.

The case where some loadings exhibit loading invariance but others do not is known as **loading partial invariance**. A common strategy for dealing with it is to conduct a multigroup SEM that constrains the loadings to be equal across groups for the invariant indicators but to allow the non-invariant indicators to be freely estimated across groups. As long as at least one of the indicators other than the reference indicator has the property of loading invariance, a partial invariance model tends to yield valid parameter estimates that properly account for the measurement non-invariance. The syntax for the partial invariance model is identical to that of Table 3 except I change the label for the d3 indicator in Line 19 from mL3 to fL3 to match the label I used for it in Line 15. When two parameters share the same label, Mplus constrains them to be equal when estimating the best fitting values of parameters in the model. (I do not need to do this for d1, because I fixed those loadings at 1.0 so they are, by definition, equal). Thus, for the three indicators, only the loadings for d2

are free to vary across groups. In my analysis of partial invariance, I continue to use bootstrapping but I could also use MLR. In the `MODEL CONSTRAINT` section, I eliminate the diff3 contrast because I have forced the loadings for d3 to be equal. The model yielded good fit: The chi square index was 6.21 with 7 degrees of freedom ($p < 0.52$), the CFI was 1.00, the RMSEA was <0.001 with a 90% confidence interval of 0.000 to 0.042, the p value for close fit was 0.98 and the standardized RMR was 0.007. The factor loadings are in Table 5.

Table 5: Unstandardized Factor Loadings for Partial Invariance Model

Measure	Female	Male	Difference	CI for Difference	Margin of Error
d1	1.000	1.000	-	-	-
d2	0.874	1.033	-0.159	-0.225 to -0.092	±0.074
d3	1.001	1.001	-	-	-

Using the partial invariance model as a reference, I am now in a position to evaluate the implications of the loading non-invariance for d2 using the logic of Oberski (2014). The substantive parameter of interest is the path coefficient for the effect of LX (latent depression) on the percent of exercise adherence. For the partial invariance model, the path coefficient linking latent depression to percent adherence for females was -10.50 (critical ratio (CR) = 16.62, $p < 0.05$, 95% CI = -11.74 to -9.25); for every one unit that depression increases (on the metric of d1 and after adjusting for measurement error), adherence decreases by 10.50 percentage points. For males, the corresponding path coefficient was -11.54 (CR = 16.61, $p < 0.05$, 95% CI = -12.89 to -10.17); for every one unit that depression increases, adherence decreases by 11.54 percentage points. The difference in the estimated effects of depression on adherence for males and females was (-10.50) - (-11.54) = 1.04, which was not statistically significant via the contrast specified in the `MODEL CONSTRAINT` commands in Table 3 on Line 27 (it yielded a critical ratio of 1.11, ns).

Next, I compare these results to a model where I ignore matters of loading invariance by allowing both d2 and d3 to vary per a fully unconstrained multi-group model. That is, I pursue a standard multigroup analysis completely ignoring matters of measurement non-invariance. In this case, the path coefficient linking depression to adherence for females was -10.55 (CR = 16.23, $p < 0.05$, 95% CI = -11.84 to -9.25); for males it was -11.49 (CR = 16.61, $p < 0.05$, 95% CI = -12.89 to -10.13), with significance patterns that were the same as the partial invariance model. Basically, there is not much difference in the values of these coefficients and those from the partial invariance model. The loading non-invariance for d2 as a function of sex does not seem to meaningfully affect the estimate of the substantive

parameter I am interested in no matter how I choose to analyze the data. Oberski would argue based on these results that the measurement non-invariance effect size is trivial.

A Backward Analysis Strategy

As an alternative to the above approach, Raykov, Marcoulides & Millsap (2013) propose a backward analysis strategy. The initial constrained model that forces measurement invariance onto all measurement parameters serves as a basis of comparison for subsequent steps. In the approach by Raykov et al., they allow for unequal group variances on LX and unequal factor means on LX, which minimizes required assumptions. No reference indicator is used for LX; the approach instead relies on the fixed factor variance method where the latent variance is set to 1.0 in one of the groups (in this case, I arbitrarily chose females to fix the LX variance at 1.0). Table 6 presents the Mplus syntax for the initial fully constrained model.[1]

Table 6: Mplus Syntax for Loading Invariance Using Backward Analysis

```
1. TITLE: TEST OF LOADING INVARIANCE ;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1 d2 d3 d4 d5 d6 adhere adhere2 dfemale income ethnic ;
5. USEVARIABLES ARE d1 d2 d3 adhere  ;
6. MISSING ARE ALL (-9999) ;
7. GROUPING IS dfemale (0=male 1=female) ;
8. ANALYSIS:
9. ESTIMATOR = MLR ;
10. MODEL:
11. LX BY d1* d2* d3* ;
12. LX@1 ;
13. [d1] ; [d2]; [d3] ;
14. [LX@0] ;
15. adhere ON LX ;
16. MODEL female:
17. LX BY d1* d2* d3* (fL1 fL2 fL3) ;
18. LX@1 ;
19. [d1] (fi1) ; [d2] (fi2) ; [d3] (fi3) ;
20. [LX@0] ;
21. adhere ON LX (fp) ;
22. MODEL male:
23. LX BY d1* d2* d3* (fL1 fL2 fL3) ;
24. LX ;
25. [d1] (fi1) ; [d2] (fi2) ; [d3] (fi3) ;
```

[1] The incorporation of factor means and intercepts into the model is not critical to evaluating loading invariance. I use it later for tests of intercept invariance. Also, my syntax is inefficient but makes assumptions explicit.

```
26. [LX] ;
27. adhere ON LX (mp) ;
28. OUTPUT: SAMP RESIDUAL CINTERVAL TECH4 STAND(STDYX) ;
```

I focus on code that is noteworthy or novel relative to my discussion of previous Mplus programs. Line 13 in the general model section has three separate commands on it, with each command separated by a semi-colon. This is a space saving device. A variable in brackets tells Mplus to estimate the mean of that variable if it is exogenous and the intercept of the variable if it is endogenous. The variables d1, d2, and d3 are endogenous so these three commands refer to the measurement intercepts for them. They request Mplus to estimate them (which is the default). On lines 19 and 25, the same syntax appears but now for females and for males, separately. Each intercept has a label attached to it in parentheses. Note that the labels for the females are the same as the labels for males. As noted earlier, this tells Mplus to impose an equality constraint; the measurement intercept for males and females for d1 must be equal; the measurement intercepts for d2 for males and females must be equal; and the measurement intercepts for d3 for males and females must be equal.

Lines 17 and 23 specify that the loadings for d1, d2 and d3 should be estimated but given the common labels for males and females, these are constrained to be equal in the groups.

In line 18, I fix the variance of LX for females to 1.0, consistent with the initial forward analysis when I was choosing a reference indicator. In that analysis, I fixed LX to be 1.0 for all of the groups. However, given the other equality constraints in the current model, I can now estimate the variance of LX in the other groups, in this case males, and this is evident on Line 24. If I tried doing so in the forward analysis when choosing a reference indicator, the model would be under-identified. In the current analysis, I do not make the assumption of equal LX variances across groups, which is a strength of the backward analysis strategy. Lines 14, 20 and 26 deal with LX means and I defer discussion of them to the section below on the analysis of measurement intercepts. They are not relevant to tests of loading invariance.

When I executed the syntax, most indices pointed towards a reasonable model fit, with the exception of the chi square statistic: The chi square was 29.24 with 8 degrees of freedom ($p < 0.001$), the CFI was 0.99, the RMSEA was <0.001 with a 90% confidence interval of 0.037 to 0.083, the p value for close fit was 0.22, and the standardized RMR was 0.026.[2]

Next, I relaxed the loading equality constraint for d1 and allowed that loading to vary between males and females by changing the label fL1 to mL1 on Line 23; now the label is no longer the same as the corresponding label on Line 17. The resulting chi square for this

[2] The primary source of the ill fit is that the model forced the loadings for d2 for males and females to be equal when, in fact, they are not.

relaxed model was 24.81 with 7 degrees of freedom. I performed a chi square difference test adjusting for scaling correction factors (see my book). I used the program for doing so on the programs tab of my website ("Scaled chi square difference test"). The chi square difference was 4.33 with 1 degree of freedom and $p < 0.0372$. I then repeated the chi square difference testing process for d2 after restoring the across group equality constraint for the d1 loading but relaxing it for d2. I found the chi square difference between the constrained and relaxed model was 22.25 with 1 degree of freedom, $p < 0.001$. For d3, the corresponding test yielded a chi square difference of 7.47 with 1 degree of freedom, $p < 0.006$. I then adjusted the p values for multiple contrasts (k=3) using the False Discovery Rate, as recommended by Raykov et al. (2013). (The program for FDR corrections also is on the program tab on my website). The adjusted p values for the contrasts were <0.001, 0.009, and 0.037 for d2, d3, and d1, respectively. Thus, contrary to the forward analysis, all three loading differences are declared statistically significant. Note that when I created the population data via simulation, the loadings for both d1 and d3 were invariant across sex in the population and only the loadings for d2 were non-invariant. The backward method, in this case, led to erroneous conclusions; the prior forward analysis led to correct conclusions.

One reason the backward solution may have failed is that both of the chi square values used to form the chi square difference were large and indicative of poor model fit. Statisticians have shown that the chi square difference test can perform poorly in such cases (e.g., Yuan & Bentler, 2004; Yuan & Chan, 2016). Another weakness of the backward analysis is that it only yields a test of statistical significance; it does not yield confidence intervals for parameter differences nor is it amenable to effect size analysis when both chi squares forming the chi square difference test suggest each model is ill fitting. To be sure, there are work arounds for effect size analysis, but consideration of them is beyond the scope of this primer.

Analysis of Measurement Intercepts

To test for measurement intercept invariance, researchers use similar approaches to those described above for the analysis of loading invariance. However, the tests are less straightforward and they often require stronger assumptions.

A Backward Analysis Strategy

Traditional analyses of localized intercept invariance tend to rely on backward analyses. There are many variations, most of which have non-trivial limitations. One of the more reasonable approaches is that suggested by Raykov, Marcoulides & Millsap (2013), but it suffers from the same limitations described above when applied to tests of loading invariance, i.e., it only provides a significance test and it can yield incorrect conclusions if

the two models comprising the chi square difference test have large and statistically significant chi squares indicative of poor model fit. In this approach to intercept testing, the baseline model for purposes of chi square differencing is the same as the factor loading baseline model presented in Table 6. I refer here to that table to highlight features of the syntax relevant to intercept invariance tests for the depression and exercise adherence example.

Lines 20 and 26 in Table 6 refer to the mean of LX. I fix it at 0 for females (Line 20) and estimate it for males (Line 26). The choice of which group to fix it for is arbitrary, but the underlying mathematics are such that the modeling allows for the LX means to vary for females and males. Several existing backward analytic strategies for intercept testing require the assumption of no mean LX differences across groups. The Raykov et al. (2013) approach is an exception. Line 18 fixes the variance of LX to 1.0 for females but estimates the variance of LX for males. The choice of which group to fix the variance to 1.0 also is arbitrary. This allows the LX variances to differ for groups. Lines 19 and 25 indicate that the model is to estimate the measurement intercepts for each indicator (d1, d2, and d3) for males and females. However, by assigning common labels to the respective intercepts, I force them to be equal for females and males during model estimation. In sum, the model is characterized by across group equality constraints for all loadings and all measurement intercepts but not for LX means or LX variances. The model does not need a reference indicator for defining the metric of LX because the variance of it is set in Line 18 for females which then implies a metric for the males by virtue of all the other equality constraints in the model.

The intercept testing strategy is to fit the above model and note the chi square fit index for it. Then, to test the null hypothesis of no difference between female and male population measurement intercepts for d1, relax the equality constraint for the d1 measurement intercepts by changing the label for the male intercept on Line 25 from fi1 to mi1. Now the label differs from the label for the corresponding intercept for females on Line 19. The chi square for this "relaxed" model is calculated and differenced from the baseline model chi square using the "Scaled chi square difference test" on the programs tab of my website. This process is then repeated, separately, for the intercepts for d2 and d3, but each time using the baseline model with all of the intercepts set to being equal across groups.

For d1, the model chi squares were 29.24 (df=8) and 28.83 (df=7) and the corrected chi square difference was 0.503, df=1, $p < 0.48$. For d2, the model chi squares were 29.24 and 29.23 and the corrected chi square difference was 0.11, df=1, $p < 0.73$. For d3, the model chi squares were 29.24 and 28.83 and the corrected chi square difference was 1.14, df=1, $p < 0.29$. The FDR adjusted p values for d1, d2 and d3 were 0.72, 0.73, and 0.72, respectively. None of the contrasts were statistically significant.

ALIGNMENT TESTS

Asparouhov and Muthén (2014) have suggested an alternative approach to measurement invariance analysis using multi-group SEM that eschews the traditional strategies. It uses what they call **alignment testing**. Alignment testing seeks to identify a good fitting model that allows for underlying measurement non-invariance but that yields estimates of group means and variances in the presence of that non-invariance using what they call an a priori defined **simplicity function**. The first step of the approach is to estimate a model in which all factor means and factor variances in all groups are constrained to equal 0 and 1, respectively, but with the factor loadings and measurement intercepts freely estimated in each group. Modifications to the resulting estimates during the iteration process are introduced at each iteration using rules dictated by the simplicity function. The simplicity function uses an algorithm that loosens the initial restrictions on the group means and variances in ways that yield better estimates of their differences across groups, all while also yielding estimates of measurement non-invariance across groups based on the initial loading and intercept estimates. The approach draws on the logic of rotation criteria used with exploratory factor analysis. Essentially, a non-identified model where factor means and factor variances vary in conjunction with factor loadings and factor variances is made identified by invoking the simplicity requirement. The default simplicity function used by Mplus works best for cases where there are a few large non-invariant measurement parameters rather than many medium-sized non-invariant measurement parameters. The mathematics of the approach are complex and I do not delve into them here. See Asparouhov and Muthén (2014) for details and Byrne and van de Vijver (2017) for a reasonably accessible description of the approach.

Alignment tests use the notion of **approximate measurement invariance**. They assume one can obtain reasonable estimates of group latent mean differences and variances even if some measurement non-invariance is present rather than assuming strict non-invariance, the latter of which is probably unrealistic. Multi-group alignment analysis typically is used for two purposes, (1) to evaluate measurement invariance properties of latent variable indicators, and (2) to evaluate latent mean differences between groups. My focus here is primarily on the former.

Alignment Implementation

I implement the approach in Mplus using the `ALIGNMENT` subcommand in conjunction with mixture modeling. Mixture modeling is a specialized form of multiple group SEM. Alignment testing works best when there are many groups, but it also can perform well with a small number of groups, such as the depression and exercise example that I consider here

which has only two groups. The first step is to affirm the reasonableness of the overall configural model, which in the depression-adherence example, is a single factor model with three indicators. It turns out that this model is just-identified, so it perfectly fits the data within each group. The question of model fit and a viable common model in the groups is therefore moot.

In Mplus syntax, the `ALIGNMENT` option has two settings: `FIXED` and `FREE`. For the `FIXED` option, a factor mean is fixed to be zero in a reference group chosen by the researcher (e.g., females) but the factor means are estimated in all other groups. For the `FREE` option, all factor means are estimated. The `FREE` option is more general than the `FIXED` option, but the latter usually performs better when there is a small number of groups. I will use it here.

The relevant syntax appears in Table 7. I again use an inefficient programming strategy but one that makes key defaults explicit. I comment on syntax that is new.

Table 7: Alignment Approach

```
1. TITLE: ALIGNMENT TEST ;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1 d2 d3 d4 d5 d6 adhere adhere2 dfemale income ethnic ;
5. USEVARIABLES ARE d1 d2 d3 ;
6. MISSING ARE ALL (-9999) ;
7. CLASSES = c(2) ; !number of classes
8. KNOWNCLASS = c(dfemale = 0 1) ; !variable  values for groups
9. ANALYSIS:
10. TYPE=MIXTURE ;
11. ESTIMATOR=MLR ;
12. ALIGNMENT=FREE;
13. !ALIGNMENT=FIXED(0);
14. MODEL:
15. %OVERALL%
16. LX BY d1* d2* d3* ;
17. [d1] ; [d2] ; [d3] ;
18. %c#1%
19. LX BY d1* d2* d3* (L1_1 L1_2 L1_3) ;
20. [d1] (i1_1) ; [d2] (i1_2) ; [d3] (i1_3);
21. %c#2%
22. LX BY d1* d2* d3* (L2_1 L2_2 L2_3) ;
23. [d1] (i2_1) ; [d2] (i2_2) ; [d3] (i2_3);
24. OUTPUT: ALIGN CINTERVAL SAMP RESIDUAL TECH4 TECH8 ;
```

In Line 10, the type of analysis is identified as a mixture analysis, which is necessary to implement alignment testing. Mixture models are like multi-group SEM but where group membership can be parameterized as being either known or unknown. In the current

example, group membership is known, namely the groups are defined by the variable reflecting biological sex, `dfemale`. In alignment testing, the groups are referred to as **classes** rather than groups. Line 7 defines the label used to refer to each class (in this case, I use the letter c as a label), after which I specify the number of classes within the parentheses. On Line 8, I indicate I will use known classes and then in parentheses I indicate the variable that identifies the classes (`dfemale`) and the values on the variable that define the classes (0 = males, 1 = females). The first value listed will be referred to by Mplus as class 1, the second value listed will be referred to as class 2, and so on if you have more than two classes. Line 11 indicates I use robust maximum likelihood estimation and Line 12 indicates I will use the `FREE` option. I later change this to `FIXED`, but on this first run, I use `FREE` for reasons that will be apparent shortly. (I commented out Line 13 and explain it later).

The `MODEL` command beginning with Line 14 has the same structure as multigroup SEM, but the generalized model across classes is specified under `%OVERALL%`. The separate class subcommands occur under the `%c#1%` and `%c#2%` labels. The first class, `c#1`, refers to males and the second class, `c#2`, refers to females, per Line 8. The model is simple: the latent variable LX is indicated by d1, d2 and d3 and I estimate the three measurement intercepts. There is no reference indicator as the `ALIGNMENT` option assigns metrics to LX internally vis-à-vis the alignment algorithm. By default, the measurement error variances and a number of other parameters are estimated, but I do not specify them here in the interest of space. They will automatically appear on the Mplus output.

I assign labels to each of the factor loadings and measurement intercepts but doing so is optional. I introduce a labeling system in this example that comes in handy for mixture modeling applications and that I make use of later, so I recommend you use it. Consider the loading labels for males, which are (`L1_1 L1_2 L1_3`). I choose an arbitrary letter as the first part of the label, in this case, the letter L for "loading." I follow that with a number to represent the "class" or group the loading refers to. Males are class 1, so I use the number 1 for them and females are class 2, so I use the number 2 for them. I then insert an underline and follow it by a 1 for loading 1 (d1), a 2 for loading 2 (d2), and a 3 for loading 3 (d3). I repeat the same process for the intercepts. Again, you do not have to use this approach, nor do you have to use labels at all. However, I take advantage of the strategy below when I describe Bayesian approaches.

On the output line (Line 24), I remove the usual request for standardized estimates and modification indices because these are not allowed for alignment testing. I add the keyword `ALIGN`, which produces specialized output for alignment analysis. Importantly, alignment analyses in Mplus currently do not permit me to include covariates or outcomes in my model. I therefore restrict the modeling of the current example to the single latent variable and its three indicators. Alignment modeling can be applied to multifactor scenarios but

doing so is beyond the scope of this primer.

When I executed the syntax in Mplus, I obtained the following message:

```
STANDARD ERROR COMPARISON INDICATES THAT THE FREE ALIGNMENT MODEL MAY BE POORLY
IDENTIFIED.USING THE FIXED ALIGNMENT OPTION MAY RESOLVE THIS PROBLEM.
```

This message occurred because I have few groups. I need to use `FIXED` alignment rather than `FREE`. I comment out Line 12 in Table 7 and uncomment Line 13. Line 13 specifies a `FIXED` strategy and the number in parentheses refers to the value on the variable that defines the classes, in this case `dfemale`, that I want to use as the reference group. With many groups/classes, Mplus often makes a recommendation on the above output about which group should be used as the reference group. Usually, it is the group whose estimated latent mean is closest to 0. In this case I use males (after inspecting the output labeled "Means" under `MODEL RESULTS` in the `FREE` analysis) because it was closest to zero.

I now re-run the syntax. The output of interest in the `FIXED` analysis results is in the section called `ALIGNMENT OUTPUT` under the subsection `INVARIANCE ANALYSIS`. The output for the case of 2 groups appears in a somewhat different format than that for 3 or more groups, although the core information is the same. In the Appendix, I present an alignment analysis for four ethnic groups and walk you through that output. It covers more concepts and technical details than what I consider here for the two-group case.

The key output for the factor loading for d1 is:

```
Loadings for D1
 Group      Group       Value       Value      Difference  SE        P-value
     1          0       0.906       0.885       0.021      0.015       0.168
 Approximate Measurement Invariance Holds For Groups: 0 1
```

The loading for females (Group 1) was 0.906 and for males, it was 0.885, a difference of 0.021. The estimated standard error of the difference was 0.015. The p value for the contrast reflecting the loading difference was statistically non-significant ($p<0.168$). An approximate 95% confidence interval for the difference is obtained by multiplying the estimated standard error by 1.96 and then subtracting (lower limit) or adding (upper limit) the result to the difference. It equals -0.008 to 0.050 A verbal conclusion is provided by Mplus to indicate whether approximate invariance holds for the groups with respect to the loading. The statement is based on an alpha level of 0.001 rather than the traditional alpha of 0.05. I explain the reasons for this practice in the Appendix, but the choice is somewhat ad hoc. If you want to use a different criterion, you can ignore the Mplus generated verbal description.

I can convert the loading difference to a standardized effect size using Equation 2 and

the program for a pooled SD on my website (with the relevant variance values needed appearing in the MODEL RESULTS section of the output). The standardized effect size difference was 0.021, which in this case, is the same as the unstandardized difference.

Here is the output for the d2 and d3 loadings:

```
Loadings for D2
Group     Group      Value      Value     Difference  SE       P-value
    1         0      0.797      0.902     -0.105      0.029      0.000
Approximate Invariance Was Not Found For This Parameter.

Loadings for D3
Group     Group      Value      Value     Difference  SE       P-value
    1         0      0.915      0.875      0.040      0.021      0.051
Approximate Measurement Invariance Holds For Groups: 0 1
```

For d2, the loading for females (Group 1) was 0.797 and for males, it was 0.902, with a difference of -0.105 and an estimated standard error of the difference of 0.029. The p value for the contrast was statistically significant, which conforms to our previous forward analysis using traditional methods. The standardized effect size for the d2 difference was -0.11. For d3, the loading for females was 0.915 and for males, it was 0.875, with a difference of 0.04. The p value for the contrast was not statistically significant. The standardized effect size for the d3 difference was 0.040. When I apply the FDR correction for multiple contrasts (using the program on my website), the p values for the three loading differences adjust from 0.168, <0.001, and 0.051 to 0.168, < 0.001, and 0.077, respectively.

Here are the results for the three measurement intercepts:

```
Intercept for D1
 Group     Group      Value      Value     Difference  SE       P-value
     1         0      0.040      0.027      0.014      0.019      0.463
Approximate Measurement Invariance Holds For Groups: 0 1


Intercept for D2
 Group     Group      Value      Value     Difference  SE       P-value
     1         0      0.026      0.020      0.006      0.019      0.742
Approximate Measurement Invariance Holds For Groups: 0 1


Intercept for D3
 Group     Group      Value      Value     Difference  SE       P-value
     1         0      0.033      0.053     -0.020      0.019      0.306
Approximate Measurement Invariance Holds For Groups: 0 1
```

For intercept difference effect sizes, Pornprasertmanit (2021) suggests converting the unstandardized differences to a form of Cohen's d using the following formula:

$$ES_{ij} = (\alpha_i - \alpha_j) / \sigma_{X\text{-}POOLED}) \quad [5]$$

For d1, d2, and d3, the effect sizes are 0.014, 0.007, and -0.022, which seem small. None of the differences are statistically significant either using traditional p values or FDR adjusted p values.

Asparouhov and Muthén (2014) suggest that alignment studies should be interpreted cautiously if more than 20% of the parameter estimates are non-invariant. In our example, one of the 6 parameters was non-invariant (the loading for d3), which is 16.7%. Byrne and van de Vijver (2017) suggest the guideline may be too simplistic. For example, they found the method worked well when *all* of the parameters were non-invariant. More research is needed in this regard.

Including Other Variables after Alignment Analysis

In the traditional multigroup SEM analyses for the depression-adherence example, I included exercise adherence as an outcome in the model when evaluating measurement invariance. I did so for two reasons. First, it allowed me to evaluate the practical implications of non-invariance for parameters of substantive interest using the logic of Oberski (2014). Second, it provided additional information for estimating the factor loadings for d1, d2, and d3. By linking a fourth variable to LX (namely exercise adherence), the model takes into account the relationships of LX, d1, d2, d3 *and* adherence when estimating the loadings, i.e., it uses more information.

Is it possible to include a broader set of covariates and outcomes in alignment analyses? Strictly speaking, no it is not and this is a limitation of the method. Marsh et al. (2018) suggest an approach for adding covariates and outcomes to alignment tests that preserves the results of an initial alignment analysis without covariates and outcomes and that integrates the results into a larger SEM model. In their strategy, a subset of the values of the loadings and measurement intercepts from the initial alignment analysis are used as fixed values in the expanded SEM model of interest. I think this approach is suboptimal for several reasons. First, the strategy presumes the loading and measurement intercept estimates in the initial alignment model apply to the expanded model when, in fact, this may not be the case. This is because the expanded analysis used by Marsh et al. does not fully take into account the relationships of the indicators to the outcome/covariates when estimating the loadings. Second, the Marsh et al. (2018) approach requires one to use an arbitrary metric for the latent LX variable which I personally prefer to avoid. My preference is to use metrics that are meaningful. Given these limitations, I illustrate here instead a less formal approach than that of Marsh et al. that is in the same spirit as Marsh et al. but that addresses these limitations. This is not to say that my method is preferable to that of Marsh

et al. Both approaches have strengths and weaknesses.

In my approach, I use a traditional multigroup SEM that includes the adherence outcome. I first conduct a traditional alignment analysis with no covariates or outcomes. Then, I run a second model that relies on the general conclusions I arrive at in the initial alignment analysis. These conclusions were that the measurement intercepts of d1, d2 and d3 are functionally equivalent across biological sex as are the factor loadings for d1 and d3. I use a reference indicator of my choice (in this case, d1) to define the metric of LX. The reference indicator, of course, should be reasonably measurement invariant across groups. Table 8 presents the relevant syntax for the second step model.

Table 8: Partial Invariance Expanded SEM Model after Alignment Analysis

```
1. TITLE: EXPANDED MODEL ;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1 d2 d3 d4 d5 d6 adhere adhere2 dfemale income ethnic ;
5. USEVARIABLES ARE d1 d2 d3 adhere  ;
6. MISSING ARE ALL (-9999) ;
7. GROUPING IS dfemale (0=male 1=female) ;
8. ANALYSIS:
9. ESTIMATOR = MLR ;
10. MODEL:
11. LX BY d1@1 d2* d3* ;
12. LX* ;
13. [d1] ; [d2] ; [d3] ;
14. adhere ON LX ;
15. MODEL female:
16. LX BY d1@1 d2* d3* (fL1 fL2 fL3) ;
17. LX* (fvar) ;
18. [LX@0] ;
19. [d1] (fi1) ; [d2] (fi2) ; [d3] (fi3) ;
20. adhere ON LX (fp) ;
21. MODEL male:
22. LX BY d1@1 d2* d3* (fL1 mL2 fL3) ;
23. LX* (mvar) ;
24. [LX*] ;
25. [d1] (fi1) ; [d2] (fi2) ; [d3] (fi3) ;
26. adhere ON LX (mp) ;
27. OUTPUT: SAMP RESIDUAL MOD(ALL 4) CINTERVAL TECH4 STAND(STDYX) ;
```

All of the syntax is self-explanatory with a few exceptions. First, note that I impose equality constraints on the measurement parameters in accord with the conclusions of the initial alignment analysis. Second, I set the latent mean for females to zero (Line 18) but I estimate the mean of LX for males (Line 24). This is a programming trick often used in

multigroup SEM when at least one across-group equality constraint holds for the loadings. The mathematics are such that the LX mean that is estimated (in this case, for males) will equal the mean LX difference between that group and the group whose LX mean is fixed to zero, in this case females. The mean difference is in units of the metric of the reference indicator, d1, after adjusting for measurement error. If I have more than two groups, then I only fix one of the LX means to zero (called the reference group, a group of my choosing) and I then estimate the mean LX in the remaining groups. Each LX mean is the mean LX difference between that group and the reference group.

If my presumption about the viability of the across-group equality conclusions from the initial alignment analysis is wrong, this will be evident in a poor fitting model and large modification indices for these parameters. The model did, in fact, produce a good model fit with no notable modification indices.

Suppose that one parameter I am interested in substantively is the magnitude of the mean difference in LX for males versus females. For comparative purposes in the spirit of Oberski and to compare my conclusions when I ignore measurement non-invariance, I fit a multi-group model where I relaxed all of the across sex equality constraints except one; I constrained the measurement intercept for d1 to be equal across males and females. Without this constraint, the model is under-identified. Table 9 shows the coefficients for the two models, with the results for the substantive parameter that is of interest in red.

Table 9: Comparison of Three Models

Parameter	Partial Invariance Model	Minimal Partial Invariance Model
LX → d1	1.00/1.00	1.00/1.00
LX → d2	1.033*/0.874*	1.029*/0.877*
LX → d3	1.001*/1.001*	0.992*/1.009*
d1 intercept	0.024/0.024	0.028/0.028
d2 intercept	0.013/0.013	0.021/0.015
d3 intercept	0.034/0.034	0.054/0.021
$M_{LXmales}$-$M_{LXfemales}$	0.009 ($p < 0.99$)	0.776 ($p < 0.99$)

(notes: $M_{LXmales}$ = mean LX for males; $M_{LXfemales}$ =mean LX for females; for the partial invariance models with two entries, male parameter value is listed first, then female value; * $p<0.05$)

Per the last row of Table 9, the non-invariance in d2 loadings that the alignment analyses identified had trivial effects on inferences and characterizations for the parameter of substantive interest, in this case the mean LX difference between males and females. This also was true when I compared the results focused on the path coefficient from depression to adherence. Via the logic of Oberski (2014), whether I adjust for non-invariance does not seem to have substantive implications.

Bayesian Alignment Analysis

In this section, I apply Bayesian modeling to alignment analyses. I assume you are familiar with Bayesian approaches to SEM. If not, see my book for an introduction to them before reading this section (or you can skip this section).

As noted, alignment analysis allows for approximate measurement invariance rather than strict invariance but it does so through the invocation of a simplicity function. A more informed approach to approximate measurement invariance uses Bayesian SEM (Muthén & Asparouhov, 2012). Muthén & Asparouhov (2012) argue that traditional models of measurement invariance that invoke strict invariance of loadings and/or measurement intercepts assume an unrealistic prior distribution for hyperparameter differences of measurement parameters. They argue that it may be more reasonable to use an informative prior distribution for measurement parameter value differences such that they are seen as being *approximately* zero by using a small-variance, informative prior distribution for them vis-a-vis an approach they call BSEM. Using such approximate invariance finds an analytic solution in which the measurement non-invariance across groups is allowed to be small based on the presumed prior distribution of the parameter differences. This approach to measurement invariance is not the same as simply minimizing a simplicity criterion but it reflects the spirit of alignment analysis in that it allows for small amounts of non-invariance. It does so by bringing to bear a prior distribution of parameter differences.

Mplus offers the option of using a combined alignment and BSEM strategy. The choice of variance hyperparameters for the prior distribution of loading or measurement intercept differences is important; parameter estimates can be significantly affected by it. Both BSEM and its use in alignment analysis has been criticized because of difficulties in justifying the values for the prior distributions (Byrne & van de Vijver, 2017). Future research needs to provide researchers with better practical guidance on the choice of hyperparameter prior parameters. In the current example, I choose below what seem to be reasonable values but I am the first to acknowledge their arbitrariness. Relevant syntax is in Table 10, followed by commentary on it.

Table 10: Alignment Analysis with BSEM

```
1. TITLE: ALIGNMENT WITH BSEM ;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1 d2 d3 d4 d5 d6 adhere adhere2 dfemale income;
5. USEVARIABLES ARE d1 d2 d3 ;
6. MISSING ARE ALL (-9999) ;
7. CLASSES = c(2) ; !number of groups
8. KNOWNCLASS = c(dfemale = 0 1) ; !give id numbers of groups
9. ANALYSIS:
10. TYPE=MIXTURE ;
11. ESTIMATOR = BAYES; BITERATIONS=100000(50000); BCONVERGENCE =.01 ;
12. ALIGNMENT=FIXED(0 BSEM);
13. MODEL:
14. %OVERALL%
15. LX BY d1* d2* d3* ;
16. [d1] ; [d2] ; [d3] ;
17. %c#1%
18. LX BY d1* d2* d3* (L1_1 L1_2 L1_3) ;
19. [d1] (i1_1) ; [d2] (i1_2) ; [d3] (i1_3);
20. %c#2%
21. LX BY d1* d2* d3* (L2_1 L2_2 L2_3) ;
22. [d1] (i2_1) ; [d2] (i2_2) ; [d3] (i2_3);
23. MODEL PRIORS:
24. DO (1,3) DIFFERENCE (L1_#-L2_#)~N(0,0.01);
25. DO (1,3) DIFFERENCE (i1_#-i2_#)~N(0,0.01);
26. OUTPUT: ALIGN CINTERVAL RESIDUAL TECH4 TECH8 ;
```

The syntax follows closely the format of Table 8. One difference is in Line 11, where the estimator is specified as `BAYES` instead of `MLR`. I also override the Mplus default number of iterations and convergence criteria for Bayesian analysis so as to produce a more stable solution. In Line 12, I again indicate a `FIXED` solution with the reference group being males, but now I add `BSEM`, which tells Mplus to use the BSEM approach in conjunction with alignment. Given the use of BSEM, I need to specify the prior distributions of key parameters. I use the Mplus defaults of uninformative priors except to specify informative priors for the group differences between the loadings (Line 24) and the group differences between the intercepts (Lines 24). Line 23 indicates I will provide information about certain prior distributions.

Line 24 uses a shorthand notation to specify the three loading differences, namely the d1 difference for males versus females, the d2 difference for males versus females, and the d3 difference for males versus females. The word `DO` creates a sequential loop ranging from the first number in the parentheses to the last number in parentheses, in this case 1 and 3. The loop integers, in this case the number 1, then 2, then 3. The active number in the loop

will be substituted for the # sign in the notation to the right. The first time through the loop, the specification is L1_1-L2_1, which are the labels I used for the first loading for class 1 minus the first loading for class 2. The second time through the loop, the specification is L1_2-L2_2, which are the labels I used for the second loading for class 1 minus the second loading for class 2. The third time through the loop, the specification is L1_3-L2_3, which are the labels I used for the third loading for class 1 minus the third loading for class 2. The same format is used for the measurement intercepts in Line 25.

The right most terms in Lines 24 and 25 state that the prior difference distribution should be assumed to be normally distributed (the letter N) with a mean of 0 (the expected value of the loading difference) and a variance of 0.01 (which represents a standard deviation of 0.10). It is the specification of the variance of the prior distribution that puts in motion the Bayesian basis for approximate measurement invariance. In the current example, the observed measures of d1, d2, and d3 each have a variance of approximately 1.0 and this also is true of the latent variable, LX. This means that I can think of the loadings much like standardized factor loadings and a typical disparity of 0.10 units or less between loadings (which translates into a variance of 0.01, per Line 24) is often considered to be small. In practice, the observed standard deviations of the indicators may differ from 1.0 and their metric must therefore be taken into account when specifying hyperparameters. For example, if d1 has a standard deviation of 10, this will affect the magnitude of the loading and specification of the variance hyperparameter must account for this. As noted, there are few guidelines for making such choices, which is a non-trivial limitation of the BSEM method. I do not delve here into the complex issues that must be considered because my goal is to illustrate the general logic of the approach.

The model yielded a reasonable model fit but I do not review fit statistics for Bayesian models here in the interest of space; see my book for details. Here is the key alignment output for the loadings for d1, d2 and d3:

```
Loadings for D1
 Group  Group   Value   Value Difference  SE  P-value   Lower 2.5%   Upper 2.5%
     1      0   0.909   0.887   0.022   0.016   0.042     -0.002       0.060
 Approximate Invariance (Noninvariance) Holds For Groups: 0 1


Loadings for D2
 Group  Group   Value   Value Difference  SE  P-value   Lower 2.5%   Upper 2.5%
     1      0   0.804   0.900  -0.096   0.027   0.000     -0.153      -0.047
 Approximate Invariance Was Not Found For This Parameter.
```

```
Loadings for D3
 Group  Group   Value   Value Difference  SE  P-value   Lower 2.5%   Upper 2.5%
     1      0   0.917   0.877   0.040   0.021   0.006      0.007      0.087
 Approximate Invariance (Noninvariance) Holds For Groups: 0 1
```

The columns `Lower 2.5%` and `Upper 2.5%` represent 95% credible intervals. Bayesian analyses do not use z tests, so the p values are *not* based on the loading differences divided by their standard errors (indeed, the standard errors have a different interpretation in Bayesian frameworks as opposed to null hypothesis testing frameworks). All of the conclusions that Mplus produces in this analysis are similar to those of the MLR alignment analysis but note that this is only true if one uses the 0.001 alpha level that Mplus relies on. For example, the p value for d3 is 0.006 but because it is not less than 0.001, approximate invariance is said to hold.

Here are the results for the measurement intercepts:

```
Intercept for D1
 Group  Group   Value   Value Difference  SE  P-value   Lower 2.5%   Upper 2.5%
     1      0   0.039   0.026   0.013   0.018   0.237     -0.022      0.050
 Approximate Measurement Invariance Holds For Groups: 0 1


Intercept for D2
 Group  Group   Value   Value Difference  SE  P-value   Lower 2.5%   Upper 2.5%
     1      0   0.025   0.019   0.006   0.018   0.372     -0.030      0.043
 Approximate Measurement Invariance Holds For Groups: 0 1


Intercept for D3
 Group  Group   Value   Value Difference  SE  P-value   Lower 2.5%   Upper 2.5%
     1      0   0.033   0.052  -0.019   0.019   0.153     -0.058      0.017
 Approximate Measurement Invariance Holds For Groups: 0 1
```

These results also comport with the conclusions of the MLR analysis.

In my opinion, the BSEM approach is promising but because results can be dependent on the hyperparameters in the prior distributions and because we have few guidelines about setting them, it probably is better to use the MLR method, unless context dictates otherwise.

Concluding Comments on Alignment Analysis

Alignment analysis is an interesting approach to evaluating the presence of loading and measurement intercept invariance. The strategy is relatively new. For statistical details of the approach, see Asparouhov & Muthén (2014). For applications, see Lomazzi (2017), Munck, Barber & Torney-Purta (2017), Flake & McCoach (2017), and Marsh et al., (2018). Most applications have dealt with scenarios where there are a large number of groups to

compare (I used only two groups, males and females). Alignment analysis has been extended to categorical (binary or ordinal) indicators in Mplus. Note also that BSEM can be used either with or without alignment analyses. For an example of measurement invariance analysis that uses BSEM without alignment, see Shi et al. (2019). In general, when working with BSEM, you should consider conducting sensitivity analyses for different hyperparameter values. As well, Asparouhov & Muthén (2014) suggest following up alignment analyses with a localized simulation to assure solution stability (see Rudnev, 2020, for how to program such simulations in Mplus). Parenthetically, alignment analyses can include correlated errors among indicators, if desired.

EQUIVALENCE TESTING AND MEASUREMENT INVARIANCE

A goal of tests of measurement invariance is to evaluate group equivalence of measurement intercepts and measurement focused factor loadings. Many methodologists argue that traditional null hypothesis testing is ill-suited to the task. To illustrate the logic using a non-measurement example with means, suppose a researcher wants to test if there are sex differences in the mean starting salaries for new Assistant Professors at major universities in the United Sates. The traditional null hypothesis is that the difference between the two population means is zero:

H_0: $\mu_M - \mu_F = 0$

The alternative hypothesis is that the difference between the two populations is not zero:

H_1: $\mu_M - \mu_F \neq 0$

If we collect sample data, analyze it, and reject the null hypothesis ($p < 0.05$), then we confidently conclude that the population sex difference in mean salaries is not exactly zero, i.e., we reject the null hypothesis. If the statistical test yields a statistically non-significant result, then it is not the case that we can accept the null hypothesis (H_0) and conclude that there is no sex difference in average salaries. The null hypothesis specifies that the difference in salaries for males and females is exactly zero and there is no way to know with any reasonable degree of certainty that the difference in salaries equals a single, exact value, such as zero. We find ourselves in an uncomfortable position of having to suspend judgment: We can't say that the sex difference in salaries is not zero because the p value is larger than 0.05; but we also can't say that there is not a sex difference. In statistics, we are taught to say we "failed to reject the null hypothesis" rather than "there is no difference" or that "the two groups are the same."

The concept of approximate measurement invariance was introduced to the

measurement field vis-à-vis alignment analysis and BSEM. However, there is a branch of statistics called **equivalence testing** that also embraces approximate effects in the sense that one seeks to make statements of *functional* equivalence between groups not *exact* equivalence (Chow & Liu, 2000; Lakens, Scheel & Isager, 2018; Welleck, 2010). The logic is somewhat different from alignment analysis and BSEM, so I elaborate it here.

The first step in applying the equivalence testing framework is to specify what is called an **equivalence limit**. An equivalence limit is the value of a parameter difference that separates trivial differences from meaningful differences. It is specified a priori by the researcher. For example, suppose that the true average annual salary of males and females in the population of Assistant Professors is $1. Technically, the null hypothesis of no sex difference in mean salaries is not true, but the difference is so small that the two groups can be said to be functionally equivalent. What if the true mean difference was $10? How about $100? How about $1,000? At what point does the magnitude of the difference become meaningful? That point is called the equivalence limit. If two groups differ by an amount less than the (absolute value of the) equivalence limit, the groups are said to be functionally equivalent on the parameter of interest. As applied to measurement invariance, even though two groups may differ on their factor loadings or measurement intercepts, an equivalence limit specifies a difference magnitude that allows us to say the groups are "functionally equivalent" on the parameters.

One variant of equivalence testing relies on the use of confidence intervals to make statements of functional equivalence. Suppose for the sex difference salary example, I set the equivalence limit to be $5,000. This means if the true population difference between males and female means is between -$5000 and +$5,000, I will consider the groups to be functionally equivalent on their average salaries. The interval -$5,000 to $5,000 is called the **equivalence interval.** I collect sample data, calculate the sample mean difference and the 95% confidence interval (CI) for that difference. Suppose the CI is -$2,000 to +$2,000. Given the 95% CI for the mean difference is fully contained within the equivalence interval of -$5,000 to +$5,000, I can confidently declare the groups to be "functionally" equivalent.

I can apply the equivalence testing approach to measurement invariance independent of the use of alignment or BSEM. To do so, one must first define an equivalence limit for each factor loading and for each measurement intercept for the latent variable indicators. As with BSEM, this can be challenging and must take into account the metrics of the variables. Some researchers suggest defining standards based on standardized effect sizes using Equations 2 and 5, but in the final analysis, such standards can be arbitrary. For both loadings and intercepts, standardized effect sizes less than 0.10 are often said to be "ignorable," which would yield an equivalence interval of -0.10 to 0.10 for both of them.

To apply this strategy, one needs to convert raw loading or intercept differences to

standardized effect sizes via Equations 2 and 5 and then bootstrap 95% confidence intervals for them to determine if they are within the (standardized effect size) equivalence interval of -0.10 to 0.10. I illustrate here the equivalence testing framework using an equivalence limit of 0.10 or an equivalence interval of -0.10 to 0.10 as applied to intercept differences and loading differences for the depression-adherence example. The relevant Mplus syntax is in Table 11 (which is similar to the syntax in Table 3) in which d1 is used as the reference indictor (i.e., I assume it has invariant properties).

Table 11: Mplus Syntax for Standardized Effect Size

```
1. TITLE: EQUIVALENCE TEST LOADING DIFFERENCES ;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1 d2 d3 d4 d5 d6 adhere adhere2 dfemale income ethnic ;
5. USEVARIABLES ARE d1 d2 d3 adhere  ;
6. MISSING ARE ALL (-9999) ;
7. GROUPING IS dfemale (0=male 1=female) ;
8. ANALYSIS:
9. ESTIMATOR = ML ; BOOTSTRAP=2000;
10. MODEL:
11. LX BY d1@1 d2* d3* ;
12. d2 ; d3 ;
13. LX* ;
14. adhere ON LX ;
15. MODEL female:
16. LX BY d1@1 d2* d3* (fL1 fL2 fL3) ;
17. d2 (fe2);  d3 (fe3) ;
18. LX* (fvarLX) ;
19. adhere ON LX (fp) ;
20. MODEL male:
21. LX BY d1@1 d2* d3* (mL1 mL2 mL3) ;
22. d2 (me2);  d3 (me3) ;
23. LX* (mvarLX) ;
24. adhere ON LX (mp) ;
25. MODEL CONSTRAINT:
26. NEW(fvard2 fvard3 mvard2 mvard3 pooledd2 pooledd3 pooledLX esd2 esd3);
27. fvard2=fvarLX*fL2*fL2 + fe2 ; !female d2 observed var
28. fvard3=fvarLX*fL3*fL3 + fe3 ; !female d3 observed var
29. mvard2=mvarLX*mL2*mL2 + me2 ; !male d2 observed var
30. mvard3=mvarLX*mL3*mL3 + me3 ; !male d3 observed var
31. pooledd2=sqrt(((723-1)*fvard2+(778-1)*mvard2)/(723+778-2));
32. pooledd3=sqrt(((723-1)*fvard3+(778-1)*mvard3)/(723+778-1));
33. pooledLX=sqrt(((723-1)*fvarLX+(778-1)*mvarLX)/(723 +778-2));
34. esd2=(fL2-mL2)*(pooledLX/pooledd2) ;
35. esd3=(fL3-mL3)*(pooledLX/pooledd3) ;
36. OUTPUT: SAMP RESIDUAL CINTERVAL(BOOTSTRAP) TECH4 STAND(STDYX) ;
```

I use bootstrapping (Lines 9 and 36) because the sampling distributions of standardized effect size statistics often are non-normal. I add statements to estimate the measurement errors (Lines 12, 17, 22) so I can label them and make use of the labels in the `MODEL CONSTRAINT` commands. I provide labels to the variances of LX (Lines 18 and 23) also so I can reference them in the `MODEL CONSTRAINT` commands.

ln addition to Equation 2, I make use of the following formulae in my programming:

variance of observed measure $d_k = var(LX)*L_k^2 + var(e_k)$

where var(LX) is the variance of the latent variable X, L_k is the factor loading from LX to d_k, and var(e) is the error variance of d_k, and

pooled standard deviation = $\sqrt{((n_1-1)\,var(x_1)+(n_2-1)\,var(x_2))/(n_1+n_2-2)}$

where n_1 is the sample size for group 1, n_2 is the sample size for group 1, var(x_1) is the variance of the variable, X, for group 1, and var(x_2) is the variance of the variable X for group 2. The first formula is executed in Lines 27 to 30 of Table 11 and the second formula is executed in Lines 31-33. Equation 2 is executed in Lines 34 and 35.

The loading difference standardized effect size for females minus males for d2 was -.14 with a 95% confidence interval of -0.21 to -0.07. The loading difference effect size for females minus males for d3 was 0.015 with a 95% confidence interval of -0.052 to 0.081. To make a firm conclusion, the confidence interval must be fully contained within the equivalence interval or it must be fully outside the equivalence interval. For d3, the confidence interval was fully contained within the equivalence interval so I conclude that the d3 loadings are functionally equivalent for males and females. For d2, the lower limit of the 95% confidence interval was outside the equivalence interval but the upper limit was within it. This means I cannot make a strong conclusion one way or the other because the confidence interval is too wide. I must "suspend judgment." I need a larger sample size to reduce the confidence interval width to make a definitive statement.

To evaluate the functional equivalence for the d1 loading difference, I can change the reference indicator for LX to d3 (which also is loading invariant) and make corresponding changes throughout the program. These changes are in red in Table 12.

Table 12: Mplus Syntax for d1 Standardized Effect Size

```
1. TITLE: EQUIVALENCE TEST LOADING DIFFERENCES ;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1 d2 d3 d4 d5 d6 adhere adhere2 dfemale income ethnic ;
```

```
5. USEVARIABLES ARE d1 d2 d3 adhere  ;
6. MISSING ARE ALL (-9999) ;
7. GROUPING IS dfemale (0=male 1=female) ;
8. ANALYSIS:
9. ESTIMATOR = ML ; BOOTSTRAP=2000;
10. MODEL:
11. LX BY d1* d2* d3@1 ;
12. d2 ; d3 ;
13. LX* ;
14. adhere ON LX ;
15. MODEL female:
16. LX BY d1* d2* d3@1 (fL1 fL2 fL3) ;
17. d2 (fe2);  d1 (fe1) ;
18. LX* (fvarLX) ;
19. adhere ON LX (fp) ;
20. MODEL male:
21. LX BY d1* d2* d3@1 (mL1 mL2 mL3) ;
22. d2 (me2);  d1 (me1) ;
23. LX* (mvarLX) ;
24. MODEL CONSTRAINT:
25. NEW(fvard2 fvard1 mvard2 mvard1 pooledd2 pooledd1 pooledLX esd2 esd1);
26. fvard2=fvarLX*fL2*fL2 + fe2 ; !female d2 observed var
27. fvard1=fvarLX*fL1*fL1 + fe1 ; !female d1 observed var
28. mvard2=mvarLX*mL2*mL2 + me2 ; !male d2 observed var
29. mvard1=mvarLX*mL1*mL1 + me1 ; !male d1 observed var
30. pooledd2=sqrt(((723-1)*fvard2+(778-1)*mvard2)/(723+778-2));
31. pooledd1=sqrt(((723-1)*fvard1+(778-1)*mvard1)/(723+778-1));
32. pooledLX=sqrt(((723-1)*fvarLX+(778-1)*mvarLX)/(723 +778-2));
33. esd2=(fL2-mL2)*(pooledLX/pooledd2) ;
34. esd1=(fL1-mL1)*(pooledLX/pooledd1) ;
35. OUTPUT: SAMP RESIDUAL CINTERVAL(BOOTSTRAP) TECH4 STAND(STDYX) ;
```

The standardized effect size for d1 was -0.015 with a 95% confidence interval of -0.08 to 0.05. I conclude that the d1 loading for females is functionally equivalent to that for males because the confidence limit is fully contained in the equivalence interval of -0.10 to 0.10.[3]

It is beyond the scope of this primer to delve into the application of equivalence testing to measurement invariance in depth. However, I encourage you to explore this framework and consider its applicability to measurement invariance paradigms. Equivalence testing applications to measurement invariance exist in the literature, but most rely on equivalence limits defined using RMSEAs. This is problematic because RMSEAs are difficult to interpret substantively and have challenging statistical properties when applied to measurement invariance (Yuan & Chan, 2016; Shi, Maydeu-Olivares, DiStefano, 2018; Edwards, 2013; Saris, Satorra, & van der Veld, 2009). For an interesting application of

[3] By changing the reference indicator, the results will not perfectly generalize to the original analysis due to the conditional nature of the modeling, but I would expect them to be reasonably close.

equivalence testing to measurement invariance that relies on BSEM, see Shi et al. (2019).

LONGITUDINAL MEASUREMENT NON-INVARIANCE

To this point, I have only considered tests of measurement invariance across groups. As discussed in my book, also relevant is measurement invariance across time. If you study children in grade 6 and then again in grade 8 and you wish to make statements about longitudinal influence or changes in means, then it is important that the basic psychometric properties of your measure do not change over time. In this section, I extend the depression and exercise adherence example to a two-wave design with an 18-month interval between waves. Given an elderly population, it is possible that loadings or measurement intercepts for depression could change across time due to maturation. It is important that we know if this is so to make valid inferences. Figure 2 shows the model I use.

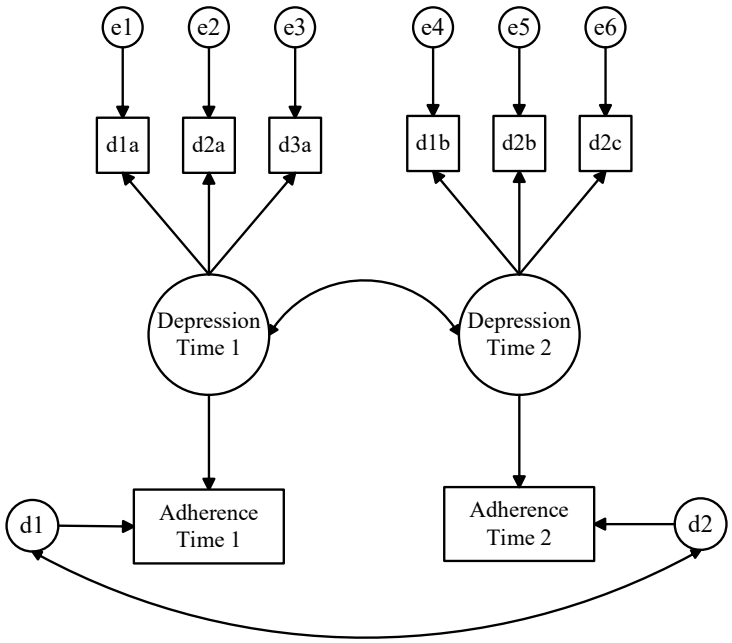

FIGURE 2. Longitudinal model

Frequently, longitudinal models include autoregressive effects, which means I would place a causal arrow from the latent depression variable at time 1 to the latent depression variable at time 2 in Figure 2. The idea is that the correlation between LX at the two time points is due to this causal effect. A different dynamic might be that there are stable

variables (e.g., gender, SES) that impact depression at both times and these unmeasured common causes create the correlation between LX at times 1 and 2. It is possible, of course, that both dynamics operate. By simply correlating the two latent variables, per Figure 2, the correlation between the latent variables reflects both dynamics, without teasing out the relative contributions of each separate dynamic. In the present case, I decide to simply correlate the latent variables to simultaneously take into account both dynamics, a decision that is inconsequential for my primary purpose of evaluating measurement non-invariance.

The same issue applies to the two endogenous variables for exercise adherence. I accommodate the issue by correlating the disturbances in Figure 2, d1 and d2, to capture all of the sources of correlation between them, both autoregressive and common cause.

Analysis of Factor Loadings

I first use a forward analysis strategy to find a reasonable reference indicator for the latent depression measure, much like I did with the traditional multiple group SEM. This strategy again makes the assumption that the variance of the latent variable is the same at both time points, which may or may not be the case. Data I present later suggests the assumption is not unreasonable. Table 13 presents relevant Mplus syntax (note: I change the names for the variables in Line 4 so that they conform better to Figure 2).

Table 13: Mplus Syntax to Find Reference Indicator for Loadings

```
1. TITLE: LOCATE REFERENCE VARIABLE - LONGITUDINAL ;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1a d2a d3a d1b d2b d3b adhere1 adhere2 dfemale income ethnic;
5. USEVARIABLES ARE d1a d2a d3a d1b d2b d3b adhere1 adhere2  ;
6. MISSING ARE ALL (-9999) ;
7. ANALYSIS:
8. ESTIMATOR = MLR;
9. MODEL:
10. LX1 BY d1a* d2a* d3a* (Ld1a Ld2a Ld3a) ;
11. LX1@1 (varLX1) ;
12. [d1a] (id1a) ;  [d2a] (id2a) ; [d3a] (id3a) ;
13. adhere1 ON LX1 ;
14. LX2 BY d1b* d2b* d3b* (Ld1b Ld2b Ld3b) ;
15. LX2@1 (varLX2) ;
16. [d1b] (id1b) ;  [d2b] (id2b) ; [d3b] (id3b) ;
17. adhere2 ON LX2 ;
18. LX1 WITH LX2;
19. adhere1 WITH adhere2 ;
20. MODEL CONSTRAINT:
21. NEW(Ldiff1 Ldiff2 Ldiff3);
```

```
22. Ldiffd1=Ld1a-Ld1b ;
23. Ldiffd2=Ld2a-Ld2b ;
24. Ldiffd3=Ld3a-Ld3b ;
25. OUTPUT: SAMP RESIDUAL CINTERVAL TECH4 STAND(STDYX) ;
```

All of the syntax should be familiar, except for Lines 18 and 19. These lines use the `WITH` subcommand and tell Mplus to correlate the two exogenous latent depression measures and the two disturbances for the endogenous adherence variables. I use the fixed factor variance method for scaling the two latent variables, implicitly assuming the variances of the two latent variables are comparable (I relax the assumption shortly).

The model yielded reasonable global fit indices and no notable modification indices when I ran it without the `MODEL CONSTRAINTS` commands (Lines 21 to 24). The chi square fit index was 18.03 with 18 degrees of freedom ($p < 0.46$), the CFI was 1.00, the RMSEA was 0.001 with a 90% confidence interval of 0.000 to 0.023, the p value for close fit was 0.99 and the standardized RMR was 0.011. Here is the output of the loading differences for time 1 minus time 2 based on the `MODEL CONSTRAINT` subcommands:

```
New/Additional Parameters
                                                    Two-Tailed
                  Estimate       S.E.    Est./S.E.    p-Value
    LDIFFD1          0.035      0.026      1.371      0.171
    LDIFFD2          0.006      0.025      0.247      0.805
    LDIFFD3          0.023      0.026      0.875      0.382
```

None of the contrasts were statistically significant, suggesting that any of the three indicators are viable candidates to be the reference indicator. If I convert the raw loading differences to standardized effect sizes using Equation 2, they are 0.036, 0.006, and 0.023, respectively. I might decide for substantive reasons that I prefer d1 as the reference indicator).

I next re-run the analysis using d1a and d1b as reference indicators but with bootstrapping (for sensitivity) and to evaluate latent variable variance differences as well. The syntax appears in Table 14.

Table 14: Mplus Syntax using Reference Indicator for Longitudinal Analysis

```
1. TITLE: LOADING INVARIANCE - LONGITUDINAL ;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1a d2a d3a d1b d2b d3b adhere1 adhere2 dfemale income ethnic;
5. USEVARIABLES ARE d1a d2a d3a d1b d2b d3b adhere1 adhere2  ;
6. MISSING ARE ALL (-9999) ;
7. ANALYSIS:
```

```
8. ESTIMATOR = ML; BOOTSTRAP=2000;
9. MODEL:
10. LX1 BY d1a@1.0 d2a* d3a* (Ld1a Ld2a Ld3a) ;
11. LX1* (varLX1) ;
12. [d1a] (id1a) ;  [d2a] (id2a) ; [d3a] (id3a) ;
13. adhere1 ON LX1 ;
14. LX2 BY d1b@1.0 d2b* d3b* (Ld1b Ld2b Ld3b) ;
15. LX2* (varLX2) ;
16. [d1b] (id1b) ;  [d2b] (id2b) ; [d3b] (id3b) ;
17. adhere2 ON LX2 ;
18. LX1 WITH LX2;
19. adhere1 WITH adhere2 ;
20. MODEL CONSTRAINT:
21. NEW(Ldiff2 Ldiff3 Lvardiff);
22. Ldiff2=Ld2a-Ld2b ;
23. Ldiff3=Ld3a-Ld3b ;
24. Lvardiff=varLX1-varLX2;
25. OUTPUT: SAMP RESIDUAL CINTERVAL(BOOTSTRAP) TECH4 STAND(STDYX) ;
```

The syntax speaks for itself. Here are the results for the three contrasts in the `MODEL CONSTRAINTS` command:

```
New/Additional Parameters
                                                    Two-Tailed
                    Estimate       S.E.    Est./S.E.    p-Value
    LDIFF2            -0.032      0.029     -1.106      0.269
    LDIFF3            -0.015      0.029     -0.516      0.606
    LVARDIFF           0.062      0.045      1.385      0.166
```

The difference in the variance of LX1 and LX2 was not statistically significant (see the last row), giving me more confidence in the initial analysis I ran to identify a reference indicator. The loading differences across time for d2 and d3 were again not statistically significantly different, affirming the initial results from the first analysis. The standardized effect size for the loading differences across time for d2 and d3 were -0.030 and -0.013.

To gain further perspective on loading non-invariance vis-à-vis the logic of Oberski, I compared parameter estimates for two models, (a) a fully constrained model that constrained the factor loadings to be equal across time as well constraining the factor variances to be equal, with (b) a model that imposes no constraints, thereby ignoring all non-invariance no matter how large. I used d1 as the reference indicator at both time points in each analysis coupled with robust maximum likelihood estimation. The models fit the data reasonably well. The first model yielded a chi square fit index of 20.34 with 21 degrees of freedom ($p < 0.50$), the CFI was 1.00, the RMSEA was <0.001 with a 90% confidence interval of 0.000 to 0.021, the p value for close fit was 0.99 and the standardized RMR was 0.014. The second model yielded a chi square of 18.03 with 18 degrees of freedom ($p <$

0.46), the CFI was 1.00, the RMSEA was 0.001 with a 90% confidence interval of 0.000 to 0.023, the p value for close fit was 0.99 and the standardized RMR was 0.011.

The four parameters of most interest are (1) the path coefficient from latent depression to exercise adherence at time 1, (2) the path coefficient from latent depression to exercise adherence at time 2, (3) the correlation between latent depression at time 1 and latent depression at time 2, and (4) the correlation between the adherence disturbance terms at time 1 and time 2. Table 15 presents the results for the two models. The differences in results for the two models is trivial. I can ignore non-invariance matters in this case in the sense that how I analyze the data (taking it into account or ignoring it) does not seem to matter.

Table 15: Comparison of Two Models

Parameter	Model 1 Value	Model 2 Value
LX time 1 → adherence time 1	-11.14*	-11.02*
LX time 2 → adherence time 2	-10.74*	-10.86
Correlation LX1 with LX2	0.535*	0.535*
Correlation adhere time 1 and adhere time 2	0.012	0.012

(notes: * $p < 0.05$)

Analysis of Measurement Intercepts

To analyze longitudinal measurement intercept invariance using forward analysis, I first need to identify an indicator that I am reasonably confident is measurement intercept invariant across time. Statisticians differ in how they approach this task. For example, the backward analysis strategy described earlier by Raykov et al. (2013) can be adapted to gain initial perspectives on the matter as long as the fits of the component models of the chi square difference tests are not both poor. The syntax for the initial constrained model that I will subsequently relax is in Table 16.

Table 16: Mplus Syntax for Longitudinal Intercept Invariance Using Backward Analysis

```
1. TITLE: TEST OF LONGITUDINAL INTERCEPT INVARIANCE ;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1a d2a d3a d1b d2b d3b adhere1 adhere2 dfemale income ethnic;
5. USEVARIABLES ARE d1a d2a d3a d1b d2b d3b adhere1 adhere2  ;
6. MISSING ARE ALL (-9999) ;
```

```
7. ANALYSIS:
8. ESTIMATOR = MLR;
9. MODEL:
10. LX1 BY d1a* d2a* d3a* (Ld1a Ld2a Ld3a) ;
11. LX1@1 ;
12. [LX1@0] ;
13. [d1a] (id1a) ;  [d2a] (id2a) ; [d3a] (id3a) ;
14. adhere1 on LX1 ;
15. LX2 BY d1b* d2b* d3b* (Ld1a Ld2a Ld3a) ;
16. LX2* ;
17. [LX2*] ;
18. [d1b] (id1a) ;  [d2b] (id2a) ; [d3b] (id3a) ;
19. adhere2 on LX2 ;
20. LX1 WITH LX2 ;
21. adhere1 WITH adhere2 ;
19. OUTPUT: SAMP RESIDUAL CINTERVAL TECH4 MOD(ALL 4) STAND(STDYX) ;
```

All of the syntax should be familiar. By using common labels, I force all of the loadings and all of the measurement intercepts to be equal across time. I fix the variance of LX to 1.0 at time 1 (Line 11) but allow it to be estimated at subsequent time points (Line 16). Similarly, I fix the LX mean to be 0 at time 1 (Line 12) but allow it to be estimated at subsequent time points (Line17). The fit of this model was reasonable. The chi square fit index was 23.54 with 22 degrees of freedom ($p < 0.68$), the CFI was 1.00, the RMSEA was 0.007 with a 90% confidence interval of 0.000 to 0.023, the p value for close fit was 1.00 and the standardized RMR was 0.012.

I conduct a contrast by making one change to the syntax in Table 16, then re-run the syntax, and form a chi square difference test. To test the difference of the intercept for d1, I change the label for it in Line 18 from `(id1a)` to `(id1b)` so that the intercepts are no longer constrained to be equal. The chi square for this model was 21.34 with 21 degrees of freedom. The difference between it and the constrained model is 23.54-21.34 = 2.00 and if I scale this difference by the relevant correction factors given the use of MLR (see the programs on my website), I obtain a chi square difference of 1.18, $p < 0.276$. If I restore the `(id1b)` label back to `(id1a)` and then repeat this process for d2, the scaled chi square difference was 3.89, $p < 0.049$ and for d3 it was 0.23, $p < 0.632$. Applying the FDR method to these p values yielded adjusted p values of 0.414, 0.147 and 0.632, respectively.

I next apply a forward analysis based on Jung and Yoon (2016) with d1 as my reference indicator because it is both loading invariant and measurement intercept invariant. Table 17 presents the relevant syntax.

Table 17: Mplus Syntax for Longitudinal Intercept Invariance Using Forward Analysis

```
1. TITLE: TEST OF LONGITUDINAL INTERCEPT INVARIANCE USING FORWARD STRATEGY;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1a d2a d3a d1b d2b d3b adhere1 adhere2 dfemale income ethnic;
5. USEVARIABLES ARE d1a d2a d3a d1b d2b d3b ;
6. MISSING ARE ALL (-9999) ;
7. ANALYSIS:
8. ESTIMATOR = MLR;
9. MODEL:
10. LX1 BY d1a@1 d2a* d3a* (Ld1a Ld2a Ld3a) ;
11. LX1* ;
12. [LX1@0] ;
13. [d1a] (id1a) ;  [d2a] (id2a) ; [d3a] (id3a) ;
14. adhere1 on LX1 ;
15. LX2 BY d1b@1 d2b* d3b* (Ld1a Ld2a Ld3a) ;
16. LX2* ;
17. [LX2*] ;
18. [d1b] (id1a) ;  [d2b] (id2b) ; [d3b] (id3b) ;
19. adhere2 on LX2 ;
20. LX1 WITH LX2 ;
21. adhere1 WITH adhere2 ;
22. MODEL CONSTRAINT:
23. NEW(idiff2 idiff3);
24. idiff2=id2a-id2b ;
25. idiff3=id3a-id3b ;
26. OUTPUT: SAMP RESIDUAL CINTERVAL TECH4 STAND(STDYX) ;
```

All the syntax should be familiar. I again fix the mean of LX to 0 at time 1 but estimate it at time 2. Here is the output for the measurement intercept differences :

```
New/Additional Parameters
                                                    Two-Tailed
                   Estimate       S.E. Est./S.E.      p-Value
    IDIFF2           -0.045      0.023     -1.989      0.047
    IDIFF3           -0.013      0.023     -0.570      0.569
```

The results are comparable to the backward analysis and when I adjust for multiplicity using the FDR method, the two p values are statistically non-significant; they are 0.094 and 0.569, respectively. I can document the trivialness of the differences vis-à-vis the logic of Oberski, but in the interest of space, I leave this as an exercise for you.

In sum, the analyses suggest the three indicators are longitudinally loading invariant

and measurement intercept invariant.[4]

MODERATED FACTOR ANALYSIS AND MIMIC ANALYSIS

The multi-group strategy dominates the measurement invariance literature but it suffers from several limitations. First, it cannot easily accommodate assessments of measurement invariance as a function of many-valued quantitative variables. Suppose I want to know if a measure is loading and/or intercept invariant across levels of income. One strategy might be to subdivide income into three or four groups and to conduct multi-group analyses on these groups, but, as discussed in my book, such false dichotomization or trichotomization often is unsatisfactory. As well, the sample size demands for each group can become unwieldy. A second limitation of multi-group invariance analysis is that it is difficult to analyze measurement invariance as a function of multiple variables at the same time or to examine the effects of one variable on measurement invariance while holding other variables constant. For example, I might ask if a measure of depression is invariant across biological sex, income, and age, all while holding constant ethnicity.

Measurement invariant methods based on MIMC (Multiple Indicators Multiple Causes Modeling) approaches and moderated factor analysis have been proposed to circumvent these limitations (Bauer, 2017; Bauer, Belzak & Cole, 2020; Curran et al., 2014; Woods & Grimm, 2011). An advantage of these methods is that they can accommodate both nominal and quantitative variables in measurement invariance modeling, hence they are more general than the multigroup framework. In addition, they are executed using single group SEM analytic frameworks, hence they often are less sample size demanding.

Analysis of Measurement Intercepts

The MIMIC approach for analyzing measurement invariance is most straightforward when analyzing measurement intercept invariance, so I consider it first. I will conduct an analysis of measurement intercept invariance across biological sex for the depression and exercise adherence example in order to compare its results with the multi-group approach. I will bring to bear continuous variables as possible sources of measurement invariance. The influence diagram is in Figure 3 (disturbance terms are indicated by the letter r).

[4] For another example of longitudinal measurement non-invariance testing, see the example for Chapter 11.

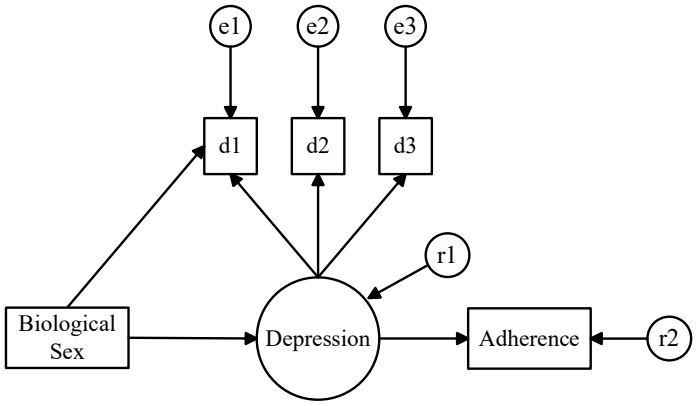

FIGURE 3. MIMIC analysis of measurement invariance

This is a single group analysis with the most notable feature being a causal arrow from biological sex directly to one of the indicators, in this case, d1. If this path is statistically significant, it suggests there are differences in the mean of d1 as a function of biological sex despite holding the latent depression variable constant (which is accomplished in the path from latent depression to d1). This is indicative of measurement intercept invariance. Note that this analysis assumes all of the factor loadings are invariant because there is only one group. Table 18 presents the relevant Mplus syntax.

Table 18: Mplus Syntax for MIMIC Intercept Invariance

```
1. TITLE: TEST OF INTERCEPT INVARIANCE USING MIMIC MODEL FOR SEX;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1 d2 d3 d4 d5 d6 adhere adhere2 dfemale income ethnic;
5. USEVARIABLES ARE d1 d2 d3 adhere dfemale  ;
6. MISSING ARE ALL (-9999) ;
7. ANALYSIS:
8. ESTIMATOR = MLR ;
9. MODEL:
10. LX BY d1@1 d2* d3* (L1 L2 L3) ;
11. LX ON dfemale  ;
12. LX*
13. [LX@0] ;
14. [d1] (I1) ;  [d2] (I2) ;  [d3] (I3) ;
15. d1 ON dfemale  ;
16. adhere ON LX ;
17. OUTPUT: SAMP RESIDUAL CINTERVAL MOD(ALL 4) TECH4 STAND(STDYX) ;
```

All the syntax should be familiar. I fix the mean of LX to 0 (which is the Mplus default) and use d1 as the reference indicator. Line 15 regresses d1 onto biological sex, which is the test of measurement intercept invariance for d1.

The global fit indices for the model suggested a good model fit. The chi square index was 2.09 with 4 degrees of freedom ($p < 0.72$), the CFI was 1.00, the RMSEA was <0.001 with a 90% confidence interval of 0.000 to 0.029, the p value for close fit was 0.99 and the standardized RMR was 0.004. There were no meaningful modification indices greater than 4. Note that despite a mis-specified model (where the loading for d2 is non-invariant across biological sex), none of the traditional model diagnostics suggest a problem. SEM has its limitations.

The output of primary interest is the regression of d1 onto dfemale:

```
MODEL RESULTS
                                                    Two-Tailed
                    Estimate       S.E.    Est./S.E.    P-Value
D1       ON
    DFEMALE            0.018      0.028      0.660      0.509
```

The path coefficient was statistically non-significant (critical ratio (CR) = 0.66, *ns*), suggesting little support for intercept invariance. Because it is binary, I can convert the estimate of 0.018 to a standardized effect size index using the program "z to t or semi-part r" on the "Programs" tab of my webpage. It was <0.001.

To evaluate the measurement intercept invariance of each indicator, d1, d2 and d3, one might be tempted to add arrows from biological sex (`dfemale`) to each indicator. However, this will produce an under-identified model. There have been several suggestions for dealing with the problem, but a commonly used one is to re-run the model separately for each indicator where non-invariance is hypothesized but change the arrow from the covariate (in this case, `dfemale`) to the target indicator. FDR or Holm modified Bonferroni corrections for multiplicity can be invoked across the contrasts, as desired. In the present case, none of the contrasts were statistically significant nor large in magnitude. The conclusion is consistent with what I found using the multi-group paradigm.

Table 19 shows the syntax where I predict measurement intercept invariance as a function of two variables, biological sex and annual income (measured in units of thousands of dollars).

Table 19: Mplus Syntax for MIMIC Intercept Invariance with Sex and Income

```
1   TITLE: TEST OF INTERCEPT INVARIANCE USING MIMIC MODEL FOR SEX AND INCOME;
2. DATA: FILE IS invariance.dat ;
```

```
3. DEFINE:
4. CENTER income (GRANDMEAN) ;
5. VARIABLE:
6. NAMES ARE id d1 d2 d3 d4 d5 d6 adhere adhere2 dfemale income ethnic;
7. USEVARIABLES ARE d1 d2 d3 adhere dfemale income  ;
8. MISSING ARE ALL (-9999) ;
9. ANALYSIS:
10. ESTIMATOR = MLR ;
11. MODEL:
12. LX BY d1@1 d2* d3* (L1 L2 L3) ;
13. LX ON dfemale income  ;
14. LX* ;
15. [LX@0] ;
16. [d1] (I1) ;  [d2] (I2) ;  [d3] (I3) ;
17. d1 ON dfemale income  ;
18. adhere ON LX ;
19. OUTPUT: SAMP RESIDUAL CINTERVAL MOD(ALL 4) TECH4 STAND(STDYX) ;
```

Line 3 with the `DEFINE` command tells Mplus I want to transform one or more of my variables. Line 4 tells Mplus to mean center the income variable. A score of zero on the transformed variable will now equal the mean income. The key word `(GRANDMEAN)` instructs Mplus to calculate the mean across all individuals in the sample. Between the words `CENTER` and `(GRANDMEAN)` are the names of variables to mean center. Line 7 adds the income variable to the model. Line 13 regresses the latent depression variable onto both biological sex and (the mean centered) income. Line 17 regresses d1 onto `dfemale` and `income`. Each predictor will have an unstandardized coefficient associated with it and each coefficient is interpreted per traditional multiple regression with multiple predictors. This line tests for measurement intercept non-invariance as a function of sex and income because LX is held constant relative to d1.

The global fit indices for the model suggested good model fit. The chi square index was 2.68 with 6 df ($p < 0.85$), the CFI was 1.00, the RMSEA was <0.001 with a 90% confidence interval of 0.000 to 0.019, the p value for close fit was 1.00 and the standardized RMR was 0.005. There were no meaningful modification indices greater than 4.

The output of primary interest is the regression of d1 onto dfemale and income:

```
MODEL RESULTS
                                                    Two-Tailed
                    Estimate       S.E. Est./S.E. P-Value
D1       ON
    DFEMALE            0.018      0.028      0.645      0.519
    INCOME             0.003      0.003      0.917      0.359
```

Both of the path coefficients are statistically non-significant. For income, for every $1,000

that income increases, the intercept for d1 is predicted to increase 0.003 units, holding constant biological sex and latent depression. As before, I can repeat the analysis for each of the other indicators, d2 and d3 (separately). Neither predictor was statistically significant relative to measurement intercept invariance for d1, d2 and d3. Note that if I wanted to, I could introduce interaction terms and polynomials in the analysis, just as I would for any regression analysis.

Analysis of Factor Loadings

I illustrate the analysis of loading invariance from a MIMIC perspective using a relatively new feature in Mplus. In the previous section, I modeled a measurement intercept as a linear function of biological sex and income in the form of the following equation:

$$\alpha = \alpha_1 + \beta_1 \text{ Sex} + \beta_2 \text{ Income} \quad [6]$$

In this next example, I repeat this analysis but also model a factor loading as a linear function of the same predictors:

$$\lambda = \alpha_2 + \beta_3 \text{ Sex} + \beta_4 \text{ Income} \quad [7]$$

For purposes of illustration, I will model the loading for the indicator d2. The relevant syntax appears in Table 20. I can model any measurement parameter I want, but I use these particular contrasts to show the general logic of the syntax.

Table 20: Mplus Syntax for MIMIC Loading Invariance

```
1  TITLE: TEST OF LOADING INVARIANCE;
2. DATA: FILE IS invariance.dat ;
3. DEFINE:
4. CENTER income (GRANDMEAN) ;
5. VARIABLE:
6. NAMES ARE id d1 d2 d3 d4 d5 d6 adhere adhere2 dfemale income ethnic;
7. USEVARIABLES ARE d1 d2 d3 adhere dfemale income  ;
8. MISSING ARE ALL (-9999) ;
9. CONSTRAINT = dfemale income ;
10. ANALYSIS:
12. ESTIMATOR = MLR ;
12. MODEL:
13. LX BY d1@1 d2* d3* (L1 L2 L3) ;
14. LX ON dfemale income  ;
15. LX* ;
16. [LX@0] ;
17. [d1] (I1) ;  [d2] (I2) ;  [d3] (I3) ;
```

```
18. d1 ON dfemale income  ;
19. adhere ON LX ;
20. MODEL CONSTRAINT:
21. NEW (ad2 b1d2 b2d2);
22. L2=ad2+b1d2*dfemale+b2d2*income ;
23. OUTPUT: SAMP CINTERVAL ;
```

With a few exceptions, all of the syntax should be familiar. On Line 9, I add a `CONSTRAINT` subcommand under the `VARIABLES` command to tell Mplus the names of input variables from the data set that I will be referencing in the `MODEL CONSTRAINT` commands (starting at Line 20). On Line 18, I indicate I am going to model the measurement intercept for d1 as a function of dfemale and income, just as I did in the previous example. However, as I will show shortly, I am doing so in conjunction with modeling the d2 factor loading as a function of these two variables. This is an improvement over the prior model where the loading for d2 was treated as invariant (which actually is a misspecified model given how I created the simulated data).

On Line 21, I declare three new parameters that will be derived in the `MODEL CONSTRAINT` command, `ad2` for the intercept of Equation 7, `b1d2` for the regression coefficient for `dfemale` predicting λ_{d2} in Equation 7 and `b2d2` for the regression coefficient for `income` predicting λ_{d2} in Equation 7. On Line 22, I tell Mplus to impose a constraint on the loading labeled L2 (see Line 13) when deriving parameter estimates. In this case, the constraint is to make L2 be a linear function of `dfemale` and `income` with reference to the designations for the intercept (`ad2`) and the two regression coefficients (`b1d2` and `b2d2`), which Mplus will estimate.

After executing the syntax, I re-examine the measurement intercept results that I analyzed earlier but now it carries the additional assumption/possibility that the loading for d2 is non-invariant. Here is the output:

```
MODEL RESULTS
                                                    Two-Tailed
                    Estimate       S.E. Est./S.E. P-Value
D1       ON
    DFEMALE            0.019      0.028      0.675      0.500
    INCOME             0.003      0.003      0.893      0.372
```

Both coefficients are statistically non-significant, consistent with my prior analysis; there is not convincing support for measurement intercept non-invariance for d1.

Before characterizing the results for d2 loading invariance, let us first look at the results that Mplus provides for the loading estimates for d1, d2, and d3. Here is the output:

```
MODEL RESULTS
                                                    Two-Tailed
                  Estimate       S.E. Est./S.E. P-Value

 LX       BY
    D1               1.000      0.000    999.000    999.000
    D2             999.000      0.000    999.000    999.000
    D3               0.999      0.019     51.390      0.000
```

The estimate for d1 is 1.00 because it was the reference indictor. The loading for d3 was 0.999. The loading for d2 is not provided; Mplus prints a 999 when it cannot report a valid value for a parameter. It did so in this case because the value of the loading depends upon `dfemale` and `income;` there is no single value for it. To gain a sense of the loading values, we need to examine the output that regressed L2 onto `dfemale` and `income`. Here is the relevant output:

```
 New/Additional Parameters
                                                    Two-Tailed
                  Estimate       S.E.    Est./S.E.   p-Value
    AD2              1.012      0.023     43.402      0.000
    B1D2            -0.125      0.030     -4.142      0.000
    B2D2             0.000      0.003      0.109      0.913
```

For Equation 7 that expressed the loading for d2 as a linear function of `dfemale` and `income`, the estimates were:

λ_{d2} = 1.012 + -0.125 `dfemale` + 0.000 `income`

The intercept in the above equation is the predicted value of the outcome when all of the predictors equal zero. A score of 0 on `dfemale` represents males. Because I mean centered `income`, a score of 0 on `income` represents the average income in the sample. Given this, the loading for males who have a "typical " income (as reflected by the mean income) is 1.012. The loading for females who have a "typical" income is -0.125 units lower than 1.012, as indicated by the coefficient for `dfemale` (which was statistically significant, critical ratio = -4.142, p <0.01). It equals 1.012 – 0.125 = 0.887. Because the coefficient of -0.125 is statistically significant, there is evidence for loading non-invariance across gender.

It is instructive to compare this result for the multi-group loading invariance analysis that used d1 as a reference group. From Table 4, the loading for males was 1.029 and for females it was 0.88, yielding a difference on -0.151, which was statistically significant (p < 0.01). This is close to what I found above. To be sure, the current analysis is the sex difference in d2 loadings holding constant income at its mean whereas the multi-group analysis did not control for income. Despite this, the conclusions were similar.

After conducting the analysis for d2, I can, repeat the analysis for d3 and then for d1, although for the latter I need to define a new reference indicator. For a discussion of strategies for combining the separate analyses, see Bauer (2017).

General Comments on Moderated Factor Analysis and MIMIC Models

The MIMIC and moderated factor analytic approaches represent a flexible strategy for measurement invariance analysis. The most recent incarnation of the method is called moderated non-linear factor analysis (MNFA) and is described in Bauer (2017) and Curran et al. (2014). Like other methods of measurement invariance analysis, the approach makes assumptions that are sometimes questionable, but it can accommodate a much wider range of scenarios than traditional multi-group SEM. Some simulation work suggests that the loading invariance portion of the approach produces inflated Type I error rates (Finch, 2005, Woods, 2009; Millsap, 2011). Both Bauer (2017) and Currant et al. (2014) suggest possible solutions to this problem, one of which is using FDR or Holm modified Bonferroni adjustments. Another weakness is the occasional need to use stepped strategies rather than simultaneous multivariate strategies due to identification or convergence problems. Bauer (2017) extends the method to modeling variances of latent variables, not just loadings and measurement intercepts.

CONCLUDING COMMENTS

There is no one best approach to test for measurement non-invariance. Before conducting an RET, I make explicit the substantive questions I seek to answer and the type of measurement invariance assumptions required to answer them. I then choose measures for the RET with one of the selection criteria being the likely measurement invariance of the measures relative to the questions I seek to answer. I conduct measurement invariance analyses for a measure in my RET if I can build an a priori case for the possible presence of consequential non-invariance. However, by the time I reach the stage of conducting the RET, I hope to have my measurement house in order so that I need not worry about issues like reliability, validity, and measurement non-invariance. It is hard enough to analyze RET data without having also to confront messy measurement.

In the RET proper, issues of measurement invariance lurk in the background whenever I compare means for two or more groups or whenever I explore moderation. All RETs conduct two group comparisons for the treatment versus control conditions, so you need to think about measurement invariance for such contrasts. Given that people are randomized to condition, there is little reason to expect those in the treatment condition, as a whole, to orient differently to measures as compared to people in the control condition. Measurement

invariance is probably a non-issue in this case. The one exception might be for a post-treatment measure whose psychometrics are affected by the treatment per se. For example, after undergoing treatment for anxiety, people might interpret items differently on an anxiety scale than they did at baseline. Such treatment-based re-interpretations might be strong enough that the construct the anxiety test purportedly measures ends up being different for the treatment and control groups.

For moderator analyses in RETs, measurement non-invariance as a function of the moderator variable is relevant because it can create false moderation or it can mask true moderation. For example, if a measure is loading non-invariant across the values of the moderator variable, then the scale is calibrated differently for the groups defined by the moderator. These different calibrations can make an effect artificially appear stronger or weaker in one group compared to another. It would be like quantifying income in one group using U.S. dollars and in the other group using pesos but not taking into account the exchange rate. It is helpful to rule out the operation of meaningful measurement non-invariance if such non-invariance is suspected.

In this primer, I have emphasized the evaluation of measurement non-invariance for composites. This is not to imply that I do not sometimes explore measurement invariance at the item level for a given scale. I might find, for example, that a composite exhibits measurement invariance across subgroups but that half the items on the scale exhibit non-invariance in one direction and the other half exhibit non-invariance in the opposite direction, with the non-invariance cancelling at the level of the composites. Such a result would certainly give me pause about the psychometric utility of a scale whose items operate so differently in different subgroups. Again, I try to screen out such scales before I conduct my RET. Nevertheless, I sometimes find myself in a position where item-level measurement invariance analyses are necessary (see below). All of the principles and methods discussed in this primer readily generalize to item level analyses; there are just many more input variables and larger covariance matrices to contend with. Sometimes such item level analysis require using methods for ordinal level or binary measures. Mplus can accommodate such models, although I have not discussed them in this primer.

Recommendations for Testing Measurement Non-Invariance

To test measurement invariance, I usually embrace a sensitivity approach such that I analyze the data from multiple perspectives, hoping that conclusions converge. If they do not, then I use my best judgment based on the empirics, theory, past research, and common sense to make inferences about the operative invariance dynamics.

If the sample size is sufficient and if the moderator/contrast variable has few values, then for tests of loading invariance, I lean towards using a forward multigroup SEM analysis

with a carefully selected reference indicator. I use alignment analysis for sensitivity purposes. The sample size demands for forward analysis of multigroup SEM typically needs to be at least 100 per group, but this can vary upwards or downwards depending on model complexity, non-invariance effect sizes, and variable distributions. I like to execute analyses first using robust maximum likelihood and then using bootstrapping.[5] I evaluate effect sizes, where possible, using the logic of Oberski. I also am sensitive to the fact that one cannot rely on traditional null hypothesis testing to assert invariance because one can never accept the null hypothesis. Thus, in addition to statistical non-significance and effect size evaluation, I apply or at least keep in mind equivalence testing logic, even if I do so in more informal ways.

Testing invariance for measurement intercepts in a multi-group scenario is challenging because approach viability depends, in part, on empirics. For example, the Raykov et al. (2013) backward analysis is reasonably effective for null hypothesis tests of measurement intercept non-invariance as long as the chi square indices that comprise the chi square difference are not both indicative of poor model fit, i.e., you do not contrast one bad fitting model with another bad fitting model. The alignment method is more straightforward as a test of measurement intercept non-invariance as is the test that uses MIMIC modeling, so I lean towards these methods when evaluating intercept non-invariance.

For cases where the different "groups" derive from continuous or many-valued quantitative variables, the MIMIC and moderated factor analysis approaches (including MLNFA) are the primary analytic methods of choice.

You will encounter other approaches than those I have discussed in this primer, but many of them have non-trivial limitations. For example, some methodologists suggest multi-group SEM but use changes in the comparative fit index (CFI) rather than chi square difference testing when comparing constrained versus relaxed models to make conclusions. The recommended rule of thumbs or cutoff values for changes in the CFI to use are somewhat ad hoc and recommendations for the values of such cutoffs often vary from one simulation study to the next. Part of the reason for such variation is that the simulations sometimes use different SEM software. It turns out that how the CFI is defined varies for different software, such as how the independence model that feeds into the calculation of the CFI is defined. As such, the cutoff values you use will differ depending on the SEM software you use.

I offer the above recommendations with some trepidation because none of the methods are perfect and assumption free. This is an area that still needs development. My treatment of invariance modeling has assumed linearity between continuous latent variables and the continuous latent variable indicators but the methods can be readily adapted to work with

[5] Although I did not discuss it, one can use bootstrapping in alignment analyses.

non-linear relationships as well. Consideration of such applications is beyond the scope of this introductory primer.

When working at the level of composites in your RET, evaluations of measurement invariance are best pursued in the presence of interchangeable indicators of the same construct, per the depression and exercise adherence example. If you have only a single multi-item composite, you can pursue item level measurement non-invariance analyses with the idea that such analyses might provide you with clues about measurement non-invariance of the composite. For example, if all of the items are functionally measurement invariant, then the composite likely will be so as well. If a small minority of the items are measurement invariant in one direction (e.g., males have a stronger factor loading than females) and another small minority of items are measurement invariant in the opposite direction (e.g., females have a stronger factor loading than males), then the composite might be measurement invariant vis-à-vis cancellation when the composite is formed.

Recommendations for What to Do Given Measurement Non-Invariance

When the Analysis Targets Items of a Single Scale

If one finds meaningful non-invariance at the item level for a single scale, recommendations in the literature usually are to revise the offending items, eliminate the items, or to conduct partial invariance analyses. For RETs, item revision usually is not possible because the data have already been collected. Eliminating items is a two-edged sword. On the one hand, it addresses the documented non-invariance. On the other hand, it might change the content universe of the original scale and, as such, undermine concept coverage. The strategy also limits the applicability of previously documented scale norms and psychometrics. This approach should be used with care.

Partial invariance modeling using item level data are problematic because they require RET modeling at the item level. If a scale has 15 items and the RET has a baseline, a posttest, and a follow-up assessment of it, this adds 45 variables to the covariance matrix. The sample size and ensuing model complexity, coupled with low reliability of individual items, can undermine effective SEM analysis, especially if non-invariance on multiple constructs and multiple scales needs to be dealt with. Thus, most recommendations to use partial invariance analysis at the item level are not practical.

When the Analysis Targets Composites

When composites of interchangeable indicators show meaningful non-invariance, one strategy for dealing with it is to drop the offending scale. One does not undermine concept coverage with this strategy because, after all, the various composite indicators are thought

to be interchangeable. In general, using more indicators of a construct tends to increase statistical power and precision, at least up to a point (e.g., Marsh, Hau, Balla, & Grayson, 1998; Cole & Preacher, 2014). You can use the power analysis program (called Power: CFA) on my website to explore the ramifications of dropping an indicator in terms of statistical power. For example, a power analysis on a population correlation of 0.25 between a mediator and an outcome in which both constructs have four indictors whose reliability are 0.75 and a sample size of 125 yields power for a two tailed test of about 0.76. If I drop one of the indicators for the mediator, the power is trivially affected; it remains at about 0.76. Dropping an indicator in this case is non-consequential, at least in terms of statistical power.

My general point is that I might find that dropping an indicator is not consequential compared to leaving in an indicator that is a poor measure vis-à-vis, say, loading non-invariance. If dropping an indicator reduces you to having two indicators of the latent construct in question, then you need to be careful of possible empirical under-identification and convergence instability (see the video for the Power: SEM program on my website). If dropping an indicator reduces your model to having a single indictor, you can still address measurement error using strategies described in the other primer for Chapter 3 focused on measurement error.

An alternative to dropping an indicator is to leave the indicator in the model but to use partial invariance modeling to compensate for the measurement non-invariance that is operating. The spirit of such analyses is that among the multiple indicators you have, there are a sufficient number of invariant ones to compensate for the non-invariance of the other indicators; you can still take advantage of the information that the non-invariant indicators provide as long as there are a sufficient number of invariant indicators. Guidelines have been proposed for the minimum number of invariant indicators you need for partial invariance to effectively capture true population means and regression coefficients. However, many of these guidelines are subjective and not based in empirics.

There have been a handful of simulation studies that have sought to evaluate the performance of partial invariance modeling (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2013; Byrne, Shavelson & Muthén, 1989; Chiorri et al., 2014; De Beuckelaer & Swinnen, 2018; Donahue, 2006; Flake & McCoach, 2017; Guenole & Brown, 2014; Hsiao & Lai, 2018; Pokropek, Davidov & Schmidt, 2019; Steinmetz, 2013, 2018; van de Schoot et al., 2013). The simulation conditions and types of invariance explored vary across studies, as do the results, so it is difficult to offer general conclusions. In my view, the simulations suggest that (a) the more invariant indicators there are, the better,[6] (b) factor

[6] Having at least half the indicators be invariant seems to work well, but simulation studies also have found that fewer than this can yield good estimates in certain contexts.

loading invariance is generally handled better by partial invariance modeling than measurement intercept invariance, although the latter also can be satisfactorily addressed with partial invariance modeling, (c) it usually is better to err on the side of declaring an indicator non-invariant than invariant when applying partial invariance models, and (d) more indicators of a latent variable tend to produce better substantive estimates using partial invariance modeling than few indicators. The traditional multigroup SEM partial invariance strategy I described using robust maximum likelihood tends to fare well for cases of loading invariance. The alignment robust maximum likelihood method tends to fare well for cases of measurement intercept invariance. However, there is some evidence for downwardly biased standard errors for these modeling strategies in some scenarios. Work is needed to provide better guidance to applied researchers about the suitability of partial invariance modeling.

There are so many forms of non-invariance, so many degrees and patterns of non-invariance, so many approaches to diagnosing and addressing non-invariance, and so many design differences (e.g., sample size, number of indicators, types of moderators, number of groups), that addressing measurement invariance can be intimidating. As research advances in this domain, the task should become more manageable.

REFERENCES

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling*, 21, 495-508.

Bauer, D.J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, 22, 507–526.

Bauer, D.J., Belzak, W. & Cole, V. (2020). Simplifying the assessment of measurement invariance over multiple background variables: using regularized moderated nonlinear factor analysis to detect differential item functioning, *Structural Equation Modeling*, 27, 43-55.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456-466

Byrne, B & van de Vijver, F. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: A paradigmatic cross-cultural application. *Psicothema*, 29, 539-551.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95, 1005–1018.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, 25, 1-27.

Chow, S-C. & Liu, J-P. (2000). *Design and analysis of bioavailability and bioequivalence studies*. Marcel Dekker Press.

Chiorri, C., Day, T., & Malmberg, L. E. (2014). An approximate measurement invariance approach to within-couple relationship quality. *Frontiers in Psychology*, 5, 983.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Curran, P., McGinley, J., Bauer, D., Hussong, A., Burns, A., Chassin, L., Sher, K. & Zucker, R. (2014). A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. *Multivariate Behavioral Research*, 49, 214–223.

De Beuckelaer, A., & Swinnen, G. (2018). Biased latent variable mean comparisons due to measurement noninvariance: A simulation study. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Methods and applications in cross-cultural analysis* (2nd ed.) (pp. 127-156). Routledge.

Donahue, B. H. (2006). The effect of partial measurement invariance on prediction. Doctoral dissertation thesis, University of Georgia, Athens.

Edwards, M. C. (2013). Purple unicorns, true models, and other things I've never seen. *Measurement: Interdisciplinary Research & Perspective*, 11, 107–111.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, 29, 278-295.

Flake, J.K. & McCoach, D. B. (2017). An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. *Structural Equation Modeling*, 25, 212-231.

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, 5, 980.

Hsiao, Y. & Lai, M. (2018). The impact of partial measurement invariance on testing moderation for single and multi-level data. *Frontiers in Psychology*, 9, 740.

Jaccard, J. (1998). *Interaction effects in factorial analysis of variance*. Sage.

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika*, 71, 173–191.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling*, 16, 642-657.

Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: forward, backward, and factor-ratio tests. *Structural Equation Modeling*, 23, 567-584.

Lakens, D., Scheel, A. & Isager, P. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1, 259-269.

Little, T. (2013). *Longitudinal structural equation modeling*. Guilford.

Lomazzi, V. (2017). Using alignment optimization to test the measurement invariance of gender role attitudes in 59 countries. *Methods, Data, and Analyses*, 12, 1-27.

Marsh, H., Guo, J., Parker, P., Nagengast, B., Asparouhov, T., Muthén, B. & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods*, 23, 524-545.

Millsap, R. (2011). *Statistical approaches to measurement invariance*. Routledge.

Munck, I., Barber, C., & Torney-Purta, J. (2017). Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009: The alignment method applied to IEA CIVED and ICCS. *Sociological Methods and Research*, 47, 687-728.

Muthén, B. & Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods*, 17, 313-335.

Muthén, B., & Asparouhov, T. (2013). BSEM measurement invariance analysis. Mplus Web Notes: No. 17. Retrieved from https://www.statmodel.com/examples/webnotes/webnote17.pdf

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, 22, 45-60.

Pornprasertmanit, S. (2021). A note on effect size for measurement invariance. Downloaded from https://cran.r-project.org/web/packages/semTools/vignettes/partialInvariance.pdf

Pokropek, A., Davidov, E. & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling*, 26, 724-744.

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research*, 49, 518-543.

Raykov, T., Marcoulides, G. A., & Li, C.-H. (2012). Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement*, 72, 954-974.

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Factorial invariance in multiple populations: A multiple testing procedure. *Educational and Psychological Measurement*, 73, 713-727.

Rudnev, M. (2020). Alignment method for measurement invariance: Tutorial. Downloaded from https://maksimrudnev.com/2019/05/01/alignment-tutorial/.

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, 16, 561–582.

Shi, D., Maydeu-Olivares, A. & Distefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivariate Behavioral Research*, 53, 210-222.

Shi, D., Song, H., DiStefano, C., Maydeu-Olivares, A., McDaniel, H. & Jiang, Z. (2019). Evaluating factorial invariance: An interval estimation approach using Bayesian structural equation modeling. *Multivariate Behavioral Research*, 4, 224-245.

Steiger, J. H. (2002). When constraints interact: A caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. *Psychological Methods*, 7, 210–227.

Steinmetz, H. (2018). Estimation and comparison of latent means across cultures. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural analysis: Methods and applications* (2nd ed.) (pp. 95-126). Routledge.

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 9, 1–12.

Thompson, Y., Song, H., Shi, D. & Liu, Z. (2020). It matters: Reference indicator selection in measurement invariance tests. *Educational and Psychological Measurement*, 81, 5-38.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, 4, 1-15.

Welleck, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. Chapman and Hall

Woods, C. & Grimm K. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, 35, 339–361.

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, 64, 737-757.

Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods*, 21, 405-422.

APPENDIX: FOUR GROUP ALIGNMENT TEST

In this Appendix, I review output for a four-group alignment problem using the depression and exercise adherence example but now focusing on four ethnic groups rather than males and females. I use robust maximum likelihood (MLR) as my estimation algorithm. Table A.1 presents the relevant Mplus syntax.

Table A.1: Four Group Alignment Approach

```
1. TITLE: ALIGNMENT TEST ;
2. DATA: FILE IS invariance.dat ;
3. VARIABLE:
4. NAMES ARE id d1 d2 d3 d4 d5 d6 adhere adhere2 dfemale income ethnic ;
5. USEVARIABLES ARE d1 d2 d3  ;
6. MISSING ARE ALL (-9999) ;
7. CLASSES = c(4) ; !number of classes
8. KNOWNCLASS = c(ethnic = 1 2 3 4) ; !variable  values for groups
9. ANALYSIS:
10. TYPE=MIXTURE ;
11. ESTIMATOR=MLR ;
12. ALIGNMENT=FREE;
13. !ALIGNMENT=FIXED(3);
14. MODEL:
15. %OVERALL%
16. LX BY d1* d2* d3* ;
17. [d1] ; [d2] ; [d3] ;
18. %c#1%
19. LX BY d1* d2* d3* (L1_1 L1_2 L1_3) ;
20. [d1] (i1_1) ; [d2] (i1_2) ; [d3] (i1_3);
21. %c#2%
22. LX BY d1* d2* d3* (L2_1 L2_2 L2_3) ;
23. [d1] (i2_1) ; [d2] (i2_2) ; [d3] (i2_3);
24. %c#3%
25. LX BY d1* d2* d3* (L3_1 L3_2 L3_3) ;
26. [d1] (i3_1) ; [d2] (i3_2) ; [d3] (i3_3);
27. %c#4%
28. LX BY d1* d2* d3* (L4_1 L4_2 L4_3) ;
29. [d1] (i4_1) ; [d2] (i4_2) ; [d3] (i4_3);
30. OUTPUT: ALIGN CINTERVAL SAMP RESIDUAL TECH4 TECH8 ;
```

In Line 7, I indicate there are four groups. In Line 8, I list the numbers that identify each group for the variable called ethnic. Using Line 12, I first run the `FREE` model, but then had to change to the `FIXED` model (Line 13) based on the warning message. I added Lines 24 to 29 to accommodate the extra two classes.

The first part of the alignment output summarizes the approximate invariance patterns:

```
APPROXIMATE MEASUREMENT INVARIANCE (NONINVARIANCE) FOR GROUPS

 Intercepts/Thresholds
   D1          1 2 3 4
   D2          1 2 3 4
   D3          1 2 3 4

 Loadings for LX
   D1          1 2 3 4
   D2          1 2 3 4
   D3          1 2 3 4
```

For each intercept and each loading, there was approximate invariance across group. Non-invariance is not an issue.

Mplus the lists the estimated factor means for each group and summarizes the significance test of their differences:

```
FACTOR MEAN COMPARISON AT THE 5% SIGNIFICANCE LEVEL IN DESCENDING ORDER

 Results for Factor LX

           Latent    Group      Factor
 Ranking    Class    Value       Mean     Groups With Significant Smaller Mean
     1         4         4       0.101
     2         1         1       0.020
     3         2         2       0.012
     4         3         3       0.000
```

The highest mean was for Group 4 (0.101) and the lowest mean was for Group 3 and none of the means were statistically significant different from the overall grand mean.

Here is the detailed analysis for the loading for d1:

```
Loadings for D1
 Group     Group      Value      Value     Difference  SE       P-value
     2         1      0.912      0.904      0.008      0.033      0.816
     3         1      0.903      0.904     -0.001      0.033      0.974
     3         2      0.903      0.912     -0.009      0.024      0.715
     4         1      0.916      0.904      0.012      0.040      0.759
     4         2      0.916      0.912      0.004      0.033      0.893
     4         3      0.916      0.903      0.013      0.032      0.682
 Approximate Measurement Invariance Holds For Groups:
 1 2 3 4
 Weighted Average Value Across Invariant Groups:       0.908
 R-square/Explained variance/Invariance index:       0.978
```

The upper table reports the loadings and pairwise comparisons for the different groups. None of the loading differences were statistically significant. Beneath the table, Mplus gives the verbal description, in essence telling us that loading invariance holds across all groups. Next, it reports the average loading of the groups (adjusted for sample size) of 0.908. Finally, it reports an omnibus R square index that I personally do not find that helpful so I do not consider it here; see Asparouhov & Muthén (2014) for details.

Next for the d1 loading, Mplus provides the following table:

```
Invariant Group Values, Difference to Average and Significance
 Group      Value Difference         SE    P-value
     1      0.904     -0.004      0.025      0.869
     2      0.912      0.004      0.015      0.811
     3      0.903     -0.005      0.014      0.707
     4      0.916      0.008      0.025      0.745
```

This table reflects an analysis that Mplus conducts for each group to determine if the loading for that group is approximately invariant. Here is what Mplus does: In the initial stage, Mplus conducts a significance test of the loading difference for all possible pairs of groups and "connects" two groups if the p value for the comparison is larger than 0.01. The algorithm then determines the largest connected group; in the present case it was four groups because none of the pairwise loading differences had $p < 0.0$. This is referred to as the "invariant set." Mplus then tests if each member's loading is statistically significantly different from the average loading in the "invariant set." If the p value for the test is less than 0.001, the loading for the group is declared to be non-invariant. The above table provides for each group its loading, the difference between the loading and the average loading of the invariant set, the standard error of the difference, and the p value for the test against the average.

Here is the same information for the other loadings and the measurement intercepts:

```
Loadings for D2
 Group     Group      Value      Value     Difference  SE       P-value
     2         1      0.862      0.862      0.000      0.031      1.000
     3         1      0.856      0.862     -0.006      0.032      0.842
     3         2      0.856      0.862     -0.006      0.025      0.800
     4         1      0.871      0.862      0.009      0.034      0.791
     4         2      0.871      0.862      0.009      0.028      0.748
     4         3      0.871      0.856      0.015      0.029      0.598
 Approximate Measurement Invariance Holds For Groups:
 1 2 3 4
 Weighted Average Value Across Invariant Groups:       0.862
 R-square/Explained variance/Invariance index:       0.971
```

```
Invariant Group Values, Difference to Average and Significance
 Group      Value Difference         SE    P-value
     1      0.862      0.001      0.023      0.979
     2      0.862      0.001      0.015      0.966
     3      0.856     -0.006      0.014      0.692
     4      0.871      0.010      0.021      0.643

 Loadings for D3
 Group     Group      Value      Value     Difference  SE       P-value
     2         1      0.904      0.912     -0.008      0.034      0.817
     3         1      0.919      0.912      0.008      0.033      0.816
     3         2      0.919      0.904      0.016      0.026      0.545
     4         1      0.888      0.912     -0.023      0.039      0.554
     4         2      0.888      0.904     -0.015      0.033      0.644
     4         3      0.888      0.919     -0.031      0.032      0.333
 Approximate Measurement Invariance Holds For Groups:
 1 2 3 4
 Weighted Average Value Across Invariant Groups:       0.908
 R-square/Explained variance/Invariance index:       0.901

 Invariant Group Values, Difference to Average and Significance
 Group      Value Difference         SE    P-value
     1      0.912      0.004      0.025      0.882
     2      0.904     -0.004      0.016      0.795
     3      0.919      0.011      0.014      0.426
     4      0.888     -0.019      0.024      0.422

 Average Invariance index:       0.846

Intercepts/Thresholds
 Intercept for D1
 Group     Group      Value      Value     Difference  SE       P-value
     2         1      0.007      0.018     -0.011      0.029      0.697
     3         1     -0.008      0.018     -0.026      0.029      0.365
     3         2     -0.008      0.007     -0.015      0.022      0.497
     4         1      0.011      0.018     -0.007      0.032      0.830
     4         2      0.011      0.007      0.004      0.026      0.872
     4         3      0.011     -0.008      0.019      0.027      0.468
 Approximate Measurement Invariance Holds For Groups:
 1 2 3 4
 Weighted Average Value Across Invariant Groups:       0.005
 R-square/Explained variance/Invariance index:       0.945

 Invariant Group Values, Difference to Average and Significance
 Group      Value Difference         SE    P-value
     1      0.018      0.014      0.022      0.525
     2      0.007      0.003      0.013      0.844
     3     -0.008     -0.013      0.013      0.328
     4      0.011      0.007      0.020      0.726
```

Intercept for D2

Group	Group	Value	Value	Difference	SE	P-value
2	1	-0.018	-0.006	-0.012	0.027	0.649
3	1	0.000	-0.006	0.006	0.026	0.825
3	2	0.000	-0.018	0.018	0.023	0.431
4	1	0.015	-0.006	0.021	0.030	0.496
4	2	0.015	-0.018	0.033	0.027	0.221
4	3	0.015	0.000	0.015	0.026	0.569

Approximate Measurement Invariance Holds For Groups:
1 2 3 4
Weighted Average Value Across Invariant Groups: -0.004
R-square/Explained variance/Invariance index: 0.925

Invariant Group Values, Difference to Average and Significance

Group	Value	Difference	SE	P-value
1	-0.006	-0.002	0.020	0.926
2	-0.018	-0.014	0.014	0.308
3	0.000	0.004	0.013	0.748
4	0.015	0.019	0.019	0.329

Intercept for D3

Group	Group	Value	Value	Difference	SE	P-value
2	1	0.028	0.004	0.025	0.029	0.401
3	1	0.025	0.004	0.021	0.028	0.452
3	2	0.025	0.028	-0.003	0.024	0.885
4	1	-0.013	0.004	-0.017	0.034	0.616
4	2	-0.013	0.028	-0.042	0.031	0.182
4	3	-0.013	0.025	-0.038	0.030	0.209

Approximate Measurement Invariance Holds For Groups:
1 2 3 4
Weighted Average Value Across Invariant Groups: 0.016
R-square/Explained variance/Invariance index: 0.353

Invariant Group Values, Difference to Average and Significance

Group	Value	Difference	SE	P-value
1	0.004	-0.012	0.021	0.568
2	0.028	0.013	0.015	0.393
3	0.025	0.009	0.013	0.490
4	-0.013	-0.029	0.023	0.204