Approaches to Studying Measurement Invariance in Multilevel Data with a Level-1 Grouping Variable

by

Heather Gunn

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved April 2016 by the
Graduate Supervisory Committee:

Kevin J. Grimm, Chair
Leona S. Aiken
Hye Won Suk

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

Measurement invariance exists when a scale functions equivalently across people and is therefore essential for making meaningful group comparisons. Often, measurement invariance is examined with independent and identically distributed data; however, there are times when the participants are clustered within units, creating dependency in the data. Researchers have taken different approaches to address this dependency when studying measurement invariance (e.g., Kim, Kwok, & Yoon, 2012; Ryu, 2014; Kim, Yoon, Wen, Luo, & Kwok, 2015), but there are no comparisons of the various approaches. The purpose of this master's thesis was to investigate measurement invariance in multilevel data when the grouping variable was a level-1 variable using five different approaches. Publicly available data from the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K) was used as an illustrative example. The construct of early behavior, which was made up of four teacher-rated behavior scales, was evaluated for measurement invariance in relation to gender. In the specific case of this illustrative example, the statistical conclusions of the five approaches were in agreement (i.e., the loading of the *externalizing* item and the intercept of the *approaches to learning* item were not invariant). Simulation work should be done to investigate in which situations the conclusions of these approaches diverge.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

To have a valid measure, the items that form the scale should function equivalently across different groups of people, such as males and females. If they do not function equivalently, then researchers cannot be certain that the same construct is being measured in all groups, which makes valid comparisons between the groups impossible. For example, the Scholastic Aptitude Test (SAT) should function equivalently for males and females. Because the measure is not perfectly reliable, people who have the same intelligence level would not necessarily have the same observed score. These differences in observed scores for people with the same level of aptitude are assumed to be random. If they were not, and males systematically scored higher than females who have the same level of aptitude, then the test would be biased in relation to gender. In this scenario, the observed scores are influenced by some artifact of the measure. Males and females may respond differently to the items, not because they have different levels of aptitude, but because of unrelated reasons. The test in this case would violate measurement invariance with respect to gender. Measurement invariance, or measurement equivalence, is a property that exists when a test (measure, scale, survey, etc.) functions equivalently across people. It is important to establish measurement invariance prior to making group comparisons to properly interpret the results. If a measure exhibits measurement bias, then differences between groups on the construct of interest may be over- or underestimated. Continuing with the example, if males had a higher mean on the biased

aptitude test than females, it could not be concluded that males were smarter, on average, than females because the observed difference was not solely due to aptitude.

Most measurement invariance methods were developed assuming the participants were sampled independently; however, there are times when individuals are grouped in higher-level units (clusters) and the dependence of the scores needs to be taken into account when examining measurement invariance. The purpose of this master's thesis was to examine whether the same conclusions were reached when investigating measurement invariance for groups nested within clustered units for five different methods. I begin with a formal overview of measurement invariance testing for independently collected data. I then transition to discussing the basics of multilevel modeling and how dependency is typically taken into consideration. Next, I discuss how methods for testing measurement invariance have been extended to deal with dependent data. Finally, I describe an illustrative example and apply the various approaches of testing measurement invariance in multilevel data.

Measurement Invariance

In psychology, there are certain constructs, such as intelligence and personality, which are not measured directly, but measured indirectly by administering a set of items that are thought to be related to the construct. These constructs can be represented as unmeasured or latent variables that influence the responses on a set of observed variables. A measurement model expresses the relationships between the latent variables and the observed variables. One type of measurement model is the linear common factor model. In this model, there are one or more factors that account for the variance in the observed

variables. Once this variance is removed, the observed variables are mutually uncorrelated (Millsap, 2011). These relationships are reflected in the following equation

$$y_i = \tau + \Lambda\eta_i + \varepsilon_i \quad (1)$$

where y_i is a vector of j observed variables for person i, τ is a vector of j intercepts, η_i is a vector of r factor scores for person i, Λ is a $j \times r$ matrix of loadings that relate the factor scores to the observed scores, and ε_i is a vector of j residuals for person i. One of the assumptions of the linear common factor models is that the common factors and unique factors are not correlated. Based on this assumption and Equation 1, the expected covariance structure of y is

$$\Sigma = \Lambda\Psi\Lambda' + \Theta \quad (2)$$

where Σ is a $j \times j$ expected covariance matrix for the observed variables, Ψ is an $r \times r$ matrix of factor variances and covariances, and Θ is a $j \times j$ matrix of unique factor variances and covariances. The covariance matrix Θ is typically diagonal, which illustrates the idea that measured variables are uncorrelated once the common factor(s) has been accounted for.

Researchers in many fields, such as in cross-national consumer research and organizational research, often want to determine if there are differences between groups of individuals in the underlying common factor (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Establishing measurement invariance is essential to making meaningful group comparisons in the underlying common factor. As stated earlier, measurement invariance is a property where the function of a measure does not differ across people. If measurement invariance has been established, then the meaning and

metric of the latent variables are equivalent across the groups tested. This is represented by the following equation

$$P(y|\eta, g) = P(y|\eta) \tag{3}$$

where *y* refers to the observed variables, g refers to group membership, and η refers to the latent variables. If true, the equation states that people who have the same ability on a construct, η, have the same probability of obtaining the same observed score regardless of group membership. The observed score, y, is not related to g, group membership, once η, the latent variable, is taken into consideration. When the parameters in the model (e.g., factor loadings) are equal across groups, those parameters are said to be invariant across group membership. If the model parameters are not the same across groups, then those items with non-invariant parameters are biased or exhibit differential item functioning.

Typically, four levels of measurement invariance are tested to establish factorial invariance (Meredith, 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). The first model tested is the *configural* invariance model (Millsap, 2011; Horn, McArdle, & Mason, 1983). In this model, the common factor model is fit separately in each group, such that

$$y_{ig} = \tau_g + \Lambda_g \eta_{ig} + \varepsilon_{ig} \tag{4}$$

$$\Sigma_g = \Lambda_g \Psi_g \Lambda'_g + \Theta_g \tag{5}$$

where g represents group membership. Equations 4 and 5 are equivalent to Equations 1 and 2, respectively, except group membership is now incorporated into the equations. The groups are constrained to have the same factor structure by constraining the groups to have the same number of factors (i.e., the dimension of Λ is the same across groups)

and the same location of zero loadings in the Λ matrix. To determine if the groups have the same factor structure, researchers examine global fit indices, such as the RMSEA and CFI. If acceptable, then further models can be tested. If, however, the global fit information does not support the viability of the configural invariance model, then testing stops and the researchers conclude that the model is not invariant across groups. Rejection of this model could mean that the common factor model does not fit in one or more of the groups.

The next model is the *metric* invariance model, also known as the *weak* invariance model (Meredith, 1993; Widaman & Reise, 1997). In addition to having the same constraints as the configural invariance model, in the metric invariance model, the factor loadings are constrained to be equal across groups, such that

$$y_{ig} = \tau_g + \Lambda\eta_{ig} + \varepsilon_{ig} \tag{6}$$

$$\Sigma_g = \Lambda\Psi_g\Lambda' + \Theta_g \tag{7}$$

where Equations 6 and 7 are identical to Equations 4 and 5, respectively, except the g subscript is removed from the Λ matrix reflecting that the groups have equal factor loadings. Once again, the global fit information is assessed to determine if the metric invariance model is viable. If acceptable, then a likelihood ratio test can be performed to statistically compare the metric and the configural invariance models because the metric invariance model is nested within the configural invariance model. If the likelihood ratio test is not significant, then there is not a significant loss of fit when the factor loadings are constrained to be equal across groups. If the likelihood ratio test is significant, then local fit information should be assessed to determine where the model misfits. At times, the

lack of invariance may be attributable to one or two factor loadings. In such cases, the researcher can evaluate a partially invariant model, where one or more factor loadings are free to vary across groups. This, however, can lead to interpretation issues because the factor loadings affect the model implied correlations among the measured variables.

If the metric invariance model holds, the next model to test is the *scalar* invariance model, also known as the *strong* invariance model (Meredith, 1993; Widaman & Reise, 1997). This model builds on the metric invariance model and factor intercepts are constrained to be equal across groups, such that

$$y_{ig} = \tau + \Lambda\eta_{ig} + \varepsilon_{ig} \tag{8}$$

$$\Sigma_g = \Lambda\Psi_g\Lambda' + \Theta_g \tag{9}$$

where Equations 8 and 9 are identical to Equations 6 and 7, respectively, except the g subscript is removed from the τ vector, reflecting that the groups have equal intercepts. Similar to before, global fit information is assessed and then a likelihood ratio test is performed between the scalar and metric invariance models, where the scalar invariance model is nested under the metric invariance model. If the test is not significant, then the scalar invariance model is appropriate and any differences among the observed means across groups are attributable to the difference in the means on the factor and not to an artifact of the measure. Achieving strong invariance indicates that the measure being tested is not biased across groups with respect to the observed means. If the test is significant, then the researcher can test for partial scalar invariance where one or more of the intercepts are freed to vary across groups; however, this could lead to interpretation

issues because differences in all observed means would not be strictly due to differences in factor means.

Most researchers end their investigation with the scalar invariance model, but invariance testing should continue with the *strict* invariance model, where unique variances are constrained to be equal across groups (Widaman & Reise, 1997). The strict invariance model can be written as

$$y_{ig} = \tau + \Lambda\eta_{ig} + \varepsilon_i \quad (10)$$

$$\Sigma_g = \Lambda\Psi_g\Lambda' + \Theta \quad (11)$$

where Equations 10 and 11 are identical to Equations 8 and 9, respectively, except the g subscript is removed from the Θ matrix, reflecting that the groups have equal unique variances for each observed variable. Once again, global fit information is assessed and then a likelihood ratio test is performed between the scalar and strict invariance models, where the strict invariance model is nested within the scalar invariance model. If strict invariance holds, then the differences in the means, variances, and covariances of the observed variables across groups are entirely due to differences in the common factors across groups and the measure is not biased across groups.

Multilevel Modeling

One assumption of most statistical models is that the data were collected independently of one another - one score is not influenced by another score. Sometimes data collection schemes measure participants in naturally occurring clusters, such that the scores within a cluster are more related than to scores outside the cluster. Examples of such data collection schemes are when data are collected from students who are nested

within schools, repeated measures that are nested within participants, and children who are nested within families. When independence is violated, the error term decreases because scores within a cluster are not as different from one another compared to scores across clusters. This leads to the test statistic being inflated and the increase of type I error. One way to account for this dependency is to use a multilevel (or hierarchical linear) model.

Multilevel models partition the variance of the outcomes into between- and within-level variance. These partitions are non-overlapping and, when summed together, equal the total variance of the dependent variable, such that

$$\sigma_T^2 = \sigma_B^2 + \sigma_W^2 \tag{12}$$

where σ_T^2 is the total variance, σ_B^2 is the between-cluster variance, and σ_W^2 is the within-cluster variance. The between-level variance, σ_B^2, is the variance of the cluster mean deviations. The within-level variance, σ_W^2, is the variance of the deviations of the raw scores from the cluster means. By modeling the effect of the cluster, the multilevel model takes into account the dependence and the Type I error rate is not inflated.

The intraclass correlation (ICC) is a parameter that calculates the proportion of variance at the between-level compared to the total variance, such that

$$ICC = {\sigma_B^2}\Big/{(\sigma_B^2 + \sigma_W^2)}. \tag{13}$$

If researchers are concerned about independence violations, they typically calculate the ICC. Some researchers argue that dependency only needs to be addressed if the ICC is high, whereas others argue the cluster effect needs to be addressed either by modeling it or controlling for it regardless of the value of the ICC (Nezlek, 2008).

If one is interested in a univariate outcome measure, then multilevel modeling is appropriate. If, however, one is interested in multivariate data, then multilevel structural equation modeling (ML-SEM) is necessary. Expanding Equation 12 for multivariate outcomes results in the following covariance matrix

$$\Sigma_T = \Sigma_B + \Sigma_W \tag{14}$$

where Σ_T is the covariance matrix, Σ_B is the between-level covariance matrix, and Σ_W is the within-level covariance matrix. Structural equation models (SEMs) are specified at each level of the model in ML-SEM (Mehta & Neale, 2005). Furthermore, a measurement model can be fit at each level yielding

$$y_{ik} = (\tau_B + \Lambda_B \eta_{Bk} + \varepsilon_{Bk}) + (\tau_W + \Lambda_W \eta_{Wik} + \varepsilon_{Wik}) \tag{15}$$

$$\Sigma_T = (\Lambda_B \Psi_B \Lambda'_B + \Theta_B) + (\Lambda_W \Psi_W \Lambda'_W + \Theta_W) \tag{16}$$

where y_i is a vector of *j* observed variables for person *i* in cluster *k*, τ_B is a vector of *j* intercepts at the between-level, Λ_B is a matrix of loadings at the between-level that relate the between-level factor scores for cluster *k*, η_{Bk}, to the observed scores, ε_{Bk} is a vector of unique factor scores at the between-level for cluster *k*, τ_W is a vector of intercepts at the within-level, Λ_W is a matrix of loadings at the within-level that relate the within-level factor scores for person *i* in cluster *k*, η_{Wik}, to the observed scores, ε_{Wik} is a vector of unique factor scores at the within-level for person *i* in cluster *k*, Ψ_B is a matrix of factor variances and covariances at the between-level, Θ_B is a matrix of unique variances and covariances at the between-level, Ψ_W is a matrix of factor variances and covariances at the within-level, and Θ_B is a matrix of unique variances and covariances at the within-level.

The between- and within-level models do not need to be equal because there can be different factors across the levels, also known as contextual effects. Additionally, the latent factors at the between-level may not have the same substantive meaning as the latent factors at the within-level (Bovaird & Shaw, 2012).

Multilevel Measurement Invariance

Special considerations need to be taken into account when testing for measurement invariance in the presence of clustered data. If the cluster structure is not controlled for, then the type I error rate inflates (Kim, Kwok, & Yoon, 2012). The groups of interest can be at either level-1 or level-2. To clarify, *group* or grouping variable is used to refer to the sets of people compared in measurement invariance testing. The term *cluster* is used for the structure that is creating dependence in the data. An example of a grouping variable at level-2 is comparing Chinese students to Italian students who are clustered within schools (Wu et al., 2012). The units of interest are students and students are clustered within schools, but each school is homogenous with regard to ethnicity (i.e., of the schools sampled, each school either contains all Chinese students or all Italian students). Beyond controlling for clustering, this scenario does not require analyses different from the analyses of factorial invariance without clustering and all four levels of factorial invariance (i.e., configural, metric, scalar, and strict) can be tested.

Grouping variables can also occur at level-1 while the cluster remains at level-2. An example of group membership occurring at level-1 in clustered data is comparing boys and girls who are clustered within families. Here, gender is the grouping variable and family is the cluster. In the Multi-Court effectiveness trial of the New Beginnings

Program, researchers designed an intervention to improve parenting for divorced parents. One research question they were interested in was if child gender moderated the treatment effect of the program on parental warmth, a construct defined by measures such as the Child Report of Parenting Behavior Inventory (CRPBI). In order to correctly answer that question, measurement invariance analyses were run comparing 559 male and female children who were clustered within 353 families. This scenario (i.e., grouping membership occurring at level-1) is more complicated because the groups and clusters are intertwined. This master's thesis focused on methods that can be used when the grouping variable is at level-1.

Multilevel factor mixture model for known classes approach. A multilevel factor mixture model for known classes can be used to test measurement invariance in a two-level model (Kim, Yoon, Wen, Luo, & Kwok, 2015). Factor mixture modeling is typically used to identify unobserved groups of participants to explain population heterogeneity, but it can also be used for observed classes. This strategy allows for more flexibility in the model set-up than a multiple group model by using a different estimation approach than a multiple group model. The model for the observed scores is

$$[y_{ik}|C_{ik} = c] = [\tau_c + \Lambda_B \eta_{Bk} + \varepsilon_{Bk}] + [\Lambda_{Wc} \eta_{Wick} + \varepsilon_{Wick}] \quad (17)$$

where c refers to the class for participant i in cluster k, y_{ik} is a vector of observed scores for person i in cluster k, τ_c is the vector of intercepts at the between-level for class c, Λ_B is the loading matrix at the between-level, η_{Bk} is the vector of factor scores at the between level for cluster k, ε_{Bk} is the vector of residuals at the between-level for cluster k, Λ_{Wc} is the factor loading matrix at the within-level for class c, η_{Wick} is the vector of

factor scores at the within-level for person i in class c and cluster k, and ε_{Wick} is the vector of residuals at the within-level for person i in class c and cluster k. In this framework, configural and metric invariance can be tested using the Λ_{Wc} matrix. Because scores at the within-level are deviation scores from the cluster means (as seen by the lack of intercepts in the within-level model), scalar invariance can only be tested in the between-level model using the τ_c vector using this approach. Strict invariance can be tested using the residual variances in the within-level model.

Multiple indicator multiple cause approach. Instead of fitting a model separately in each group, a grouping variable can be incorporated into the two-level model using a multiple indicator multiple cause (MIMIC) model (Kim, Yoon, Wen, Luo, & Kwok, 2015; Woods & Grimm, 2011). In MIMIC models, one or more observed variables predict one or more latent variables. As Figure 1 illustrates, to test invariance using MIMIC models for independently and identically distributed data, a factor structure is created where the observed variables load onto the common factor. The grouping variable predicts each person's factor score, which allows the group factor means to differ, and the observed score on a selected set of observed variables, which allows the groups to have different intercepts for these variables. Additionally, the selected set of observed variables is regressed on an interaction term between the latent variable and the grouping variable, which allows for group differences in loadings. This is demonstrated by the following equations

$$\eta_i = \Gamma_\eta x_i + \zeta_i \tag{18}$$

$$y_i = \Lambda\eta_i + \beta_y x_i + \omega_{\eta y}\eta_i x_i + \varepsilon_i \tag{19}$$

where η_i is a vector of person i's factor scores, Γ_η is a matrix of regression coefficients that relate the grouping variable, x_i, to the factor scores, ζ_i is a vector of residuals, y_i is a vector of person i's observed scores, Λ is a matrix of loadings that relate the factor scores to the observed variables, β_y is a vector of regression coefficients that relate the grouping variable to the observed variables, $\omega_{\eta y}$ is a vector of regression coefficients that relate the interaction term between person i's factor scores and the grouping variable to the observed variables, and ε_i is a vector of residuals for person i. In this model metric invariance can be examined by testing each regression coefficient in $\omega_{\eta y}$ to determine if it is significantly different from zero. If there is a significant difference, then the loading to that indicator is not invariant. To evaluate scalar invariance, each regression coefficient in β_y is tested to determine if it is significantly different from zero. If there is a significant difference, then the groups differ on the intercept of the corresponding indicator. Instead of testing each regression coefficient in β_y individually, a model that constrains all of the regression coefficients in β_y to be zero can be compared to a model that that freely estimates those regression coefficients. If the fit of the more constrained model is not significantly worse than the fit of the less constrained model, then scalar invariance holds.

To incorporate a multilevel structure, k clusters need to be included in Equation 19 to form the following equation

$$y_{ik} = (\Lambda_W \eta_{Wik} + \beta_y x_{ik} + \omega_{\eta y} \eta_{Wik} x_{ik} + \varepsilon_{Wik}) + (\tau_k + \Lambda_B \eta_{Bk} + \varepsilon_{Bk}) \qquad (20)$$

where y_{ik} is a vector of observed scores for person i in cluster k, Λ_W is a matrix of loadings that relate within-level factor scores to the observed scores, η_{Wik} is a vector of

factor scores for person i in cluster k, β_y is a vector of regression coefficients that relate the grouping variable, x_{ik}, to the observed variables, $\omega_{\eta y}$ is a vector of regression coefficients that relate the interaction of the within-level latent variables and the grouping variable to the observed variables, ε_{Wik} is a vector of residuals at the within-level, τ_k is a vector of intercepts, Λ_B is a matrix of loadings that relate between-level factor scores, η_{Bk}, to the observed scores, and ε_{Bk} is a vector of residuals at the between-level. As with Equation 19, this model can test differences in loadings between groups (metric invariance) by testing the significance of regression coefficients in $\omega_{\eta y}$. Additionally, the model can test differences in intercepts between groups (scalar invariance) by testing the significance of regression coefficients in β_y. As stated before, a model where the intercepts are constrained to be equal across groups can be compared to a model where the intercepts are not constrained to be equal by fixing the regression coefficients in β_y to zero. The MIMIC approach, however, does not allow for group differences on the unique variances and thus cannot test strict invariance. Because the unique variances are essentially constrained to be equal across groups for all models, this can distort invariance testing of the loadings and intercepts. Additionally, this model does not allow for group differences on the factor variance. One potential benefit to testing invariance using this model is that additional variables (e.g., socioeconomic status, race/ethnicity) can be included in the model that can potentially explain the group differences.

Definition variable approach. A second way to test each loading and intercept individually is to use the definition variable approach (Bauer & Hussong, 2009). A definition variable is not part of the model, but it can be used to constrain values of

parameters in the model (Mehta & Neale, 2005). The constraint allows for group comparisons without incorporating the grouping variable into the model. In this approach, a multilevel SEM is constructed for the entire sample without including the grouping variable. Each parameter of interest (e.g., intercept) is constrained using the MODEL CONSTRAINT command in M*plus* where the grouping variable, a dummy-coded variable, is incorporated into each constraint. For instance, a constraint statement for an intercept would take on the form

$$\tau_j = \gamma_{j0} + \gamma_{j1} x_i \tag{21}$$

where τ_j is the intercept for indicator *j*, γ_{j0} is the intercept when $x_i = 0$, γ_{j1} is the difference in the intercept when $x_i = 1$, and x_i is the grouping variable for person *i* in cluster *k*. If the γ_{j1} parameter is significantly different from zero, then there is a significant difference between the groups on that intercept. All of the loadings, intercepts, and unique variances can be tested in this way. Similar to the MIMIC model approach, potential confounding variables can be included in this model to control for their effects.

Design-based approach. Instead of using a two-level model, a design-based multilevel confirmatory factor analysis (CFA) can be used to compare groups within clustered units (Kim, Kwok, & Yoon, 2012). In this design, TYPE = COMPLEX is used in M*plus*. Rather than decomposing the model into between and within components, this approach specifies a single-level model and uses robust (Huber-White) standard error estimators to correct for dependency. The standard errors for the parameter estimates and the test statistic of the model are adjusted to account for the dependency of scores. If the within- and between-level models are identical, then TYPE = TWOLEVEL and TYPE =

COMPLEX should produce similar results (Wu & Kwok, 2012). In this design, all four levels of invariance testing – configural, metric, scalar, and strict – can be studied when the grouping variable is at level-1.

Muthén's maximum likelihood approach. A fifth approach to study measurement invariance within a multilevel data structure uses a manual set-up of a single-level model to test measurement invariance and Muthén's maximum likelihood (MUML) to estimate the model (Ryu, 2014). To model the dependence within the clusters properly, the level-1 and level-2 components need to be decomposed before separating the data by groups. The two-level model can be written as

$$y_{igk} = [\tau_B + \Lambda_B \eta_{Bk} + \varepsilon_{Bk}] + [\tau_{Wg} + \Lambda_{Wg} \eta_{Wigk} + \varepsilon_{Wigk}] \quad (22)$$

where y_{igk} is the observed score vector for person i in group g and cluster k, τ_B is the between-level intercept, Λ_B is the loading matrix at the between-level, η_{Bk} is the vector of factor scores at the between level for cluster k, ε_{Bk} is the vector of residuals at the between-level for person i in cluster k, τ_{Wg} is the within-level intercept for group g, Λ_{Wg} is the loading matrix at the within-level for group g, η_{Wigk} is the vector of factor scores at the within-level for person i in group g and cluster k, and ε_{Wigk} is the vector of residuals at the within-level for person i in group g and cluster k. The equation above incorporates group structure, but otherwise is equivalent to Equation 15, where a group structure is not specified. In this specification, individuals in the same cluster, regardless of group membership, will have the same between-level model. This is seen by the lack of a subscript g in the level-2 part of the model. The four levels of invariance – configural, metric, scalar, and strict – can be tested in the within-level model.

Because there is no current software program that can estimate the solution associated with the fitting function for Equation 22, Ryu (2014) specified a single-level CFA with two "groups" (hereafter referred to as ^{M}groups) for each level of the grouping variable. So for measurement invariance analyses that compare two groups, such as males and females, there would be a total of four ^{M}groups. Within a group, one ^{M}group defines the within-level model and the other ^{M}group defines the between-level model.

CHAPTER 2

METHOD

Illustrative Example

To illustrate how the five methods work with real data, I analyzed publicly available data from the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K). The ECLS-K is a longitudinal study of 21,260 students who began kindergarten in 1998; however, for these analyses, only 17,809 students (49.28% female) were included because listwise deletion was used for participants who had incomplete data on any of the variables used in the analysis. To analyze data using the MUML approach, there has to be complete data. Listwise deletion was used rather than multiple imputation because common imputation routines do not account for clustering and current imputation routines for multilevel data are still in development (Enders, Mistler, & Keller, in press). Additionally, if any student had missing data on the gender variable, they were removed from the analysis. The 17,809 students with complete data were clustered in 943 schools. The ECLS-K used a multistage random sampling approach. In the first stage, schools were randomly sampled and in the second stage, children were randomly selected from those schools using a list of all kindergartners in the school. Participating students were representative of the class of entering kindergarten children in 1998 in the United States (US Department of Education, National Center for Education Statistics, 2009). These data were chosen as the analytical example because the children were naturally clustered within schools and their gender, which is a level-1 grouping variable, was the focus of my illustration regarding measurement invariance.

In the fall of kindergarten, the teachers were interviewed about their children's behavior and direct assessments were administered to the students to measure academic and non-academic skills. The four measures that are the focus of my investigation were the "Internalizing Problem Behavior Scale", the "Externalizing Problem Behaviors Scale", the "Interpersonal Skills Scale", and the "Approaches to Learning Scale". These four measures were assessed through rating scales completed by the students' teachers on a 1-4 scale. Internalizing behavior scores were the average of four items assessing *anxiety*, *loneliness*, *low self-esteem*, and *sadness*. Externalizing behavior scores were the average of five items that measured behaviors such as *fighting*, *arguing*, and *impulsivity*. The interpersonal skills scale score was the average of five items that measured skills such as *the ability to make friends*, *expressing feelings*, and *exhibiting empathy*. Finally, the approaches to learning scale score was the average of six items that measured behaviors that can be disruptive or conducive to learning in the school setting such as *attentiveness*, *learning independence*, and *eagerness to learn*. The four measures were used as indicators of a single factor that represented early childhood behavior within a school context. The *internalizing* and *externalizing* behavior scales were negatively valenced (i.e., higher scores indicate more problems) whereas the *interpersonal skills* scale and *approaches to learning* scale were positively valenced. For four of the approaches, the scales were analyzed as such; but, for the MUML approach, the *interpersonal* and *approaches to learning* were recoded to be negatively valenced because the model did not converge otherwise.

Planned Analyses

The five approaches discussed above were used to investigate measurement invariance of the factor structure for early childhood behavior with respect to gender. To fit these models, the M*plus* v. 7.3 (Muthén & Muthén, 1998-2012) software was used. M*plus* is a general latent variable modeling program with various estimation approaches that were conducive to this project. For example, the MUML estimation routine is available, definition variables can be specified, and latent variable interactions can be specified using the XWITH command to evaluate metric invariance (Woods & Grimm, 2011). The conclusions of the five approaches could not be statistically compared, but they were assessed for degree of agreement. As with all statistical analyses, a set of assumptions were made to use each approach. If one or more of the assumptions was violated for an approach, then that could be an explanation for any divergence of the conclusions. Table 1 highlights the assumptions of each approach as related to measurement invariance testing. The conclusions also may diverge because not all approaches can test all levels of invariance. Additionally, different approaches test scalar invariance at difference levels of the model. Table 2 summarizes at which level invariance can be tested at for each model.

Definition variable approach. The M*plus* v. 7.3 (Muthén & Muthén, 1998-2012) software does not allow for certain constraints when using the analysis TYPE = TWOLEVEL. Specifically, a definition variable cannot be incorporated into a two-level model. To implement the definition variable approach, I performed a two-step procedure. First, two data sets were created. One data set consisted of scores on the items that were centered within context (CWC). The total sample size for this data set was 17,809 - the

number of individuals. The second data set consisted of cluster means, which were calculated by averaging the individual scores within each school. The total sample for this data set was 943 - the number of clusters.

Second, I estimated a multiple group model for two groups where one group defined the between-level portion of the model and the other group defined the within-level portion of the model. There were no constraints across the two groups. The within-level portion was a single-level analysis using item scores that were CWC. The between-level portion was essentially an aggregated analysis, or a single-level analysis of the cluster means. Chan (1998) refers to this as an additive composition model. One of the assumptions of this analysis is that the cluster means were equally reliable (i.e., the cluster sizes were equal), which they were not in this case because the schools did not have the same number of students sampled. The average cluster size was 18.89 students per school with a minimum and maximum cluster size of 1 and 27, respectively. Additionally, not all of the students within a school were sampled, so the cluster means should ideally be treated as unobserved variables (Ludtke, Marsh, Robitzsch, Trautwein, Asparouhov, & Muthén, 2008). Given the limitations of using definition variables in multilevel SEMs in M*plus*, the cluster means were treated as observed variables, rather than as latent variables, so the results may be biased when using the definition variable approach.

MUML approach. To estimate the models using the MUML approach, I used the SAS macro and modified the example M*plus* syntax file provided by Ryu (2014). The

SAS macro created the input data file and computed the statistics needed in the invariance model for MUML estimation.

Fit evaluation. The invariance models were tested sequentially, taking global fit statistics (e.g., RMSEA, SRMR) into account in addition to comparative fit indices (e.g., AIC, BIC) and likelihood ratio tests. Because the sample size was large, the likelihood ratio tests were overpowered and could not solely be relied upon to determine invariance. Likelihood ratio tests were calculated for the MUML, definition variable, and MIMIC approaches, whose models were estimated using maximum likelihood. To estimate the models for the design-based approach and the multilevel factor mixture model for known classes approach, I used robust maximum likelihood (MLR). When models are estimated with MLR, the Satorra-Bentler likelihood ratio (SB LR) test is recommended over the likelihood ratio test for testing nested models (Satorra & Bentler, 2010).

Three comparative fit indices - Akaike's information criterion (AIC), the Bayesian information criterion (BIC), and the sample-size adjusted BIC (SABIC) - and the log-likelihood were calculated for all models. For three approaches (multilevel factor mixture model for known classes, definition variable, and the MIMIC), these log-likelihood based fit statistics were the only fit statistics available. Because of this, it was not possible to assess global fit for a single model using these approaches; however, model comparisons were possible. The χ^2, RMSEA, CFI, TLI, SRMR, and local fit statistics were only able to be calculated for the design-based and MUML approaches.

Partial invariance. Partial invariance was examined in different ways depending on the approach used to study measurement invariance. Partial invariance for the design-

based approach and MUML approach was determined by looking at local fit statistics, such as the modification indices.

In the case of partial invariance for the mixture modeling approach, multiple models were fit where each item parameter (e.g., loading, intercept, or unique variance) was freed individually to determine which item parameter should be freed to vary across groups. Whichever model had the lowest AIC, BIC, SABIC, and SB LR test was chosen as the partial invariance model.

Even though local fit indices were not available for the MIMIC and definition variable approaches, there was a way to identify which item should be freed on a local level. If a more constrained invariance model (e.g., scalar invariance model) had poor fit compared to a less constrained invariance model (e.g., metric invariance model), the results from the less constrained invariance model were reexamined to determine the parameter or parameters that were dependent on the grouping variable. Because of the set-up of the two approaches, there were statistical tests that compared the groups on the parameters of interest.

In the case of the configural invariance model using the MIMIC approach, there were three regression coefficients that captured the relationship between the product of the latent variable and grouping variable with the three items that were not the reference variable (see Figure 1). In the metric invariance model, these three regression coefficients were fixed to zero. If the metric invariance model had significantly worse fit than the configural invariance model, then the item that was associated with the regression coefficient with the highest *t*-value from the three *t*-tests was chosen to vary across

groups in the partial metric invariance model. For example, if the *t*-test that tested if the regression coefficient from the interaction to the *interpersonal* item was significantly different from zero had the highest absolute *t*-value compared to the other two *t*-tests, then that relationship was added back into the metric invariance model.

In the configural invariance model using the definition variable approach, constraint statements were incorporated into the model that constrained the parameters (e.g., loadings and intercepts) to be dependent on the definition variable - the grouping variable (see Equation 20). Three of the constraint statements constrained the loadings for the three non-reference variables to be dependent on the grouping variable. Each of these statements contains a parameter, γ_{j1}, that tests if the group difference on the loading is significantly different from zero. To estimate the metric invariance model, these three constraint statements were removed from the analysis, effectively forcing boys and girls to have the same loading. If the metric invariance model had significantly worse fit than the configural invariance model, then the results from the configural invariance analysis were reexamined. The item associated with the highest absolute *t*- value, which tested the γ_{j1} parameter, had its loading constraint statement added back into the metric invariance model, creating a partial metric invariance model.

If a partial invariance model was more appropriate than a full invariance model, then any further parameters associated with the biased item or items were not constrained to be equal across groups in further models. For instance, if the loading for the *externalizing* item was freed to vary in a partial metric model, then the intercept and the

unique variance for the *externalizing* item were not be constrained to be equal across groups when testing for scalar invariance and strict invariance.

Identification. To be consistent across models, I used similar identification constraints for all approaches. For all models, the same reference variable was chosen to have a loading equal to one in both groups at the within-level and at the between-level if one existed and to have an invariant intercept. The rest of the identification constraints are provided for each approach.

Multilevel factor mixture model for known classes. In addition to the above constraints, the mean of the within-level factor was fixed to zero for boys and freed to vary for girls. The mean of the between-level factor was fixed to zero for both genders. The within-level factor variances were not constrained to be equal across gender. The between-level model and between-level factor distribution (mean and variance) were constrained to be equal across groups. The between-level unique variances were not estimated. The syntax for the configural model and the final model for the multilevel mixture model approach is provided in Appendix C.

MIMIC. Similar to the mixture model, the mean of the between-level factor was fixed to zero for both genders and the mean of the within-level factor for boys was fixed to zero and freed to vary for girls. It was impossible to separately estimate the variances for boys and girls on either the within-level factor or the between-level factor. Additionally, the unique variances at the between-level were not specified, which matches the analyses from Kim, Yoon, Wen, Luo, & Kwok (2015). The syntax for the

configural model and the final model for the MIMIC approach is provided in Appendix D.

Definition variable. The mean of the within-level factor was fixed to zero for boys and freely estimated for girls. The variance of the within-level factor was free to vary across groups. The between-level model and the between-level factor distribution were estimated for the whole group, not within each gender. The syntax for the configural model and the final model for the definition variable approach is provided in Appendix E.

Design-based. The factor mean was fixed to zero for boys and freely estimated for girls. The factor variances were freely estimated for both genders. The syntax for the configural model and the final model for the design-based approach is provided in Appendix F.

MUML. The mean of the between-level factor was constrained to be zero for both genders. The mean of the within-level factor was fixed to zero for boys and freely estimated for girls. The school-level model was constrained to be equal across the genders. The student-level model was constrained to be equal across the two ^{M}groups within a gender. The syntax for the configural model and the final model for the MUML approach is provided in Appendix G.

CHAPTER 3

RESULTS

Intraclass Correlations

To determine the relative magnitude of the between-level variance to the total variance and examine whether a multilevel structure was necessary to validly model the data, the ICCs were calculated for each item using an unconditional model. The ICC for the *internalizing* item was .11, indicating that approximately 11% of the variance in the *internalizing* item was at level-2. The ICCs for the *externalizing* item, *approaches to learning* item, and *interpersonal* item were .07, .12, and .15, respectively. Because these ICCs were not negligible, the clustering of observations within schools needed to be taken into consideration.

Reference Indicator

To identify these models, one item had to be chosen to be invariant (in the loading and intercept) across the groups (in addition to other constraints). If a biased item was chosen to be invariant, it could distort invariance testing for the other items (Cheung & Rensvold, 1999). To determine which item should be invariant, I tested four MIMIC models where all items except the reference variable were regressed on the grouping variable and on the interaction between the grouping variable and the latent variable (see Figure 1). All four models were equal to each other except the reference variable was different for each model (e.g., the *externalizing* item was the reference variable for one model and the *approaches to learning* item was the reference variable for another model). For each model, three regression coefficients, ω, captured the relationship between the

interaction and three of the items and another three regression coefficients, β, captured the relationship between the grouping variable and the same three items. The β and ω for the fourth item were fixed to zero to designate it as the reference variable. If ω was not significantly different from zero, then the product term was not significantly related to the item response. In other words, there was not a significant difference between the groups on the factor loading for that item. For the three models where the *internalizing* item was not the reference variable, the regression coefficient, ω, associated with the *internalizing* item was not significant (see Table 3). This indicated that regardless of the reference variable, the *internalizing* item had an invariant loading. Because none of the other three items had this distinction, the *internalizing* item was chosen to be the reference variable for all models. Thus, the loading of the *internalizing* item was fixed to one, which caused the factor to be negatively valenced with higher factor scores were indicative of worse early school behavior.

Multilevel Factor Mixture Model for Known Classes Approach

Table 4 lists the fit statistics for all models tested using this approach. The configural invariance model converged, but because global fit statistics (e.g., RSMEA, CFI) were not available, it was difficult to determine if the fit of this model was good; however, comparative fit indices were provided, AIC = 130,310, BIC = 130,567, SABIC = 130,462. The metric invariance model converged and had greater information criteria (AIC, BIC, and SABIC) than the configural model in addition to a significant SB LR test, scaled χ^2 (3) = 85.57, $p < .001$, suggesting that the metric invariance model fit worse than the configural invariance model. To investigate further, three partial metric invariance

models were tested where the factor loading for each of the non-reference variables was freed to vary across groups. The partial metric invariance model that freed the *externalizing* item had the lowest AIC, BIC, SABIC, and scaled χ^2 of the three models so that model was chosen to be the partial metric invariance model. Comparing this model to the configural model, it had a lower BIC and non-significant SB LR test, scaled χ^2 (2) = 3.16, p = .207, indicating that the partial metric invariance model fit better than the configural invariance model. Because of this, I chose the partial metric invariance model over the metric invariance model with the loading for *externalizing* behavior separately estimated for boys and girls. To test for partial scalar invariance, the intercepts of the *interpersonal skills* and *approaches to learning* indicators were fixed to be equal across groups. (The intercept for *internalizing* was already fixed to invariance for identification purposes and the parameters associated with the *externalizing* indicator were not constrained to be equal across groups in subsequent models.) The partial scalar invariance model had a higher AIC, BIC, and SABIC compared to the partial metric invariance model in addition to a significant SB LR test, scaled χ^2 (2) = 130.26, p < .001. Two revised partial scalar invariance models were analyzed to determine which intercept or intercepts to free. The model that freed the intercept for the *approaches to learning* item fit better than the model that freed to intercept for the *interpersonal* item. The former model had a significant SB LR test, scaled χ^2 (1) = 57.56, p < .001, but because the test was overpowered due to the large sample size, this model was accepted and no further intercepts were freed to vary across groups. Finally, the partial strict invariance model was equal to the revised partial scalar invariance model except the unique

variances for the *interpersonal* and *internalizing* items were constrained to be equal across groups. This model had a non-significant SB LR test, scaled χ^2 (2) = 3.46 p = .178, and lower BIC and SABIC values compared to the partial scalar invariance model, indicating a better fit. In sum, a partial strict invariance model was the final model where the loading, intercept, and unique variance of the e*xternalizing* item and the intercept and unique variance of the *approaches to learning* item were freed to vary across the genders.

In the final model, the standard deviation of the latent factor for males and females were 0.21 and 0.20, respectively. Females had a mean on the latent variable at the within level that was 0.09 units, or 0.42 standard deviations, lower than males, indicating females had better early school behavior because the factor was negatively valenced. The mean difference was 0.42 standard deviations, a small to medium effect size according to Cohen (1988). The loading for the *externalizing* item for females was 1.68 and 2.14 for males, denoting that the relationship between the *externalizing* item and the factor was stronger for boys than for girls. The standard errors for these loadings were 0.055 for females and 0.062 for males. The intercept for the *approaches to learning* item for boys was 2.85 and 2.91 for girls. For a boy and a girl who had the same latent factor score, the boy was more likely to be rated in the lower categories of the *approaches to learning* item, indicating worse early school behavior. Table 5 provides the estimates for all other parameters in the measurement model.

Multiple Indicator Multiple Cause Approach

Table 6 contains the fit statistics for all the models analyzed using the multiple indicator multiple cause approach. Because unique variances cannot vary across groups

in MIMIC models, I refer to the model where the loadings and intercepts are freed to vary across groups as the configural* model. Similarly, I will use the terms metric* and scalar* to refer to the model where the loadings are constrained to be equal across groups and to the model where the loadings and intercepts are constrained to be equal across groups, respectively.

In the configural* model, all items except the reference variable were regressed on the grouping variable and on the interaction between the grouping variable and the latent variable. The global fit of the model could not be determined because global fit statistics were not able to be calculated; however, comparative fit indices were provided, AIC = 107,321, BIC = 107,500, SABIC = 107,427. The metric* model, where the interaction between the latent variable and the grouping variable was eliminated from the model, restricting the genders to have equal loadings, fit significantly worse than the configural* model, χ^2 (3) = 195.70, $p < .001$. Additionally, the AIC, BIC, and SABIC were higher for the metric* model than the configural* model. To determine which loading(s) should be freed to vary across groups for the partial metric invariance* model, I looked at the results from the configural* model and examined the *t*-statistic for the three regression coefficients between the three non-reference items and the interaction. The item with the largest (positive or negative) *t*-statistic was the item that had the biggest loading difference between the two genders. The regression coefficient for the *externalizing* item had the highest *t*-statistic (-14.387) so that path was added back to the model. The partial metric* model fit significantly worse than the configural* model, χ^2

(2) = 18.65, $p < .001$, but had a lower BIC so was chosen as final metric invariance model.

The next model tested was the partial scalar* invariance model. The intercepts of the *approaches to learning* and *interpersonal* indicators were constrained to be equal across groups. This model had significantly worse fit than the partial metric* model, χ^2 (2) = 148.13, $p < .001$. To determine which intercept to free, I referred back to the results from the partial metric* model which tested for group differences on the intercepts of the non-reference items. Excluding the *externalizing* item (because it has a different loading between groups), the *approaches to learning* item had the highest *t*-statistic (12.14), indicating that girls had a significantly greater intercept than boys (because boys were coded 0 and girls were coded 1 in the data set). Even though the fit for this revised partial scalar* model was significantly worse than the partial metric* model, χ^2 (1) = 77.74, $p < .001$, this model was chosen as the final invariance model.

The residual variance of the within-level factor for the final model was 0.05. The coefficient for the regression of the latent variable on gender was significant, $\gamma = -0.09$, $p < .001$, indicating that girls had better early childhood behavior in a school context than boys. In terms of effect size, boys had a factor mean 0.41 standard deviations greater than the factor mean for girls. As seen in Table 7, boys and girls differed significantly on the loading for the *externalizing* item, $\omega = -0.52$, $p < .001$, and on the intercept for the *approaches to learning* item, $\beta = 0.06$, $p < .001$. A one-point increase on the latent factor was associated with a 2.08 increase on the teacher ratings on the *externalizing* scale for boys, but only a 1.56 increase for girls. To understand the effect of non-invariance in the

intercepts for the *approaches to learning* item, I divided the gender difference in the intercepts by the gender difference in the observed means. The difference in the intercepts between the genders on the *approaches to learning* scale was 0.06 and the difference in observed means was 0.27. Roughly 22% of the observed mean difference can be explained by the gender difference in intercepts.

Definition Variable Approach

Table 8 lists the fit statistics for all models tested using the definition variable approach. The global fit of the configural invariance model could not be determined because global statistics were not able to be calculated, but comparative fit indices were provided, AIC = 96,478, BIC = 96,760, SABIC = 96,646. To create the metric invariance model, the three constraint statements that allowed the genders to differ on the loadings were removed from the model. The metric invariance model had higher comparative fit indices and significantly worse fit compared to the configural invariance model, χ^2 (3) = 100.30, $p < .001$. To determine which loading or loadings were causing the misfit, I examined the three *t*-tests from the configural model that individually tested if the group difference on the three loadings was significantly different from zero. The *t*-test for the *externalizing* item was the only one of the three *t*-tests that was significant, $t = -5.08$, $p < .001$. A constraint statement was added back to the metric invariance model to allow for group differences on the loading for the *externalizing* item. Partial metric invariance held, χ^2 (2) = 4.18, $p = .124$. Additionally, the BIC and SABIC were lower for the partial metric invariance model than for the configural invariance model. The partial scalar invariance model, which constrained the intercepts of the *interpersonal* and *approaches*

to learning indicators to be equal across groups, fit significantly worse than the partial metric invariance model, χ^2 (2) = 135.13, $p < .001$. To determine which intercept or intercepts were causing the misfit, I examined the two *t*-tests from the partial metric invariance model that individually tested if the group difference on the intercepts for the *approaches to learning* and *interpersonal* indicators was significantly different from zero. The *t*-statistic for the *approaches to learning* item, $t = 10.95$, $p < .001$, was larger than the *t*- statistic for the *interpersonal* item, $t = 7.94$, $p < .001$, so a constraint statement was added to the partial scalar invariance model that constrained the intercept of the *approaches to learning* item to be dependent on the grouping variable, effectively creating a revised partial scalar invariance model. The revised partial scalar invariance model fit significantly worse than the partial metric invariance model, χ^2 (1) = 60.00, $p < .001$, but because the test was overpowered due to the high sample size, this model was accepted. Strict invariance held, χ^2 (2) = 5.09, $p = .078$. Additionally, the strict invariance model had lower BIC and SABIC values than the revised partial scalar invariance model.

As shown in Table 9, girls had a loading on the *externalizing* item that was 0.43 units lower than boys, indicating that the relationship between the factor and that item was stronger for boys. Additionally, girls had an intercept on the *approaches to learning* item that was 0.06 units higher than the intercept for boys. So for a boy and a girl who had the same factor score, the girl, on average, would have a higher rating on the *approaches to learning* item than the boy. The pooled within-level factor standard deviation was 0.19. Boys were constrained to have a within-level factor mean of zero and

the group difference on the within-level factor mean was -0.08, indicating that girls were 0.41 standard deviations lower on the factor than boys.

Design-Based Approach

Table 10 contains the global fit indices for every model tested using the design-based approach. The configural invariance model had good fit, RMSEA = .035, CFI = .997. The metric invariance model also had good fit, RMSEA = .043, CFI = .991; however, the SB LR test was significant, scaled χ^2 (3) = 83.44, $p < .001$, suggesting that the metric invariance model fit significantly worse than the configural invariance model. To investigate further, I tested a partial metric invariance model where the loading for the *externalizing* variable was freed to vary across groups. This loading was chosen because it had the highest modification index (*MI* = 64.25). The global fit of the model was good (RMSEA = .030, CFI = .996) and the SB LR test was non-significant, scaled χ^2 (2) = 2.39, $p = .303$. Additionally, the AIC, BIC, and SABIC were lower for this model than they were for the configural model. Because of these results, I moved forward with the partial metric invariance model and tested a partial scalar invariance model where all of the intercepts except the intercept for the *externalizing* item were constrained to be equal across groups. The global fit of this model was good (RMSEA = .043, CFI = .989), but the SB LR test was significant, scaled χ^2 (2) = 104.76, $p < .001$. To test which intercept or intercepts to free, I examined the modification indices. The intercept for the *approaches to learning* item was associated with a higher modification index (*MI* = 41.68) than the intercept for the *interpersonal* item (*MI* = 11.62). (For identification purposes, the intercept for the *internalizing* item was constrained to be invariant.) So the intercept for

the *approaches to learning* item was freed to vary across groups in addition to the loading and intercept for the *externalizing* item in the revised partial scalar invariance model. This model also had a significant χ^2-difference test, χ^2 (1) = 49.59, $p < .001$; however, because the sample size was so large, trivial differences can yield a significant result (Tucker & Lewis, 1973). Finally, a partial strict invariance model was tested (only the unique variances for the *internalizing* and *interpersonal* items were constrained to be equal across groups) and strict invariance held, RMSEA = .037, CFI = .989, scaled χ^2 (2) = 3.15, $p = .207$. In conclusion, the final invariance model using the design-based approach was a partial strict invariance model where the factor loading, intercept, and unique variance of the *externalizing* item and the intercept and unique variance of the *approaches to learning* item were freed to vary across the genders.

According to the final model, females had a factor variance of 0.05 and males had a factor variance of 0.045. Additionally, females had a mean on the latent variable that was 0.08 units (or 0.39 standard deviations) lower than males, consistent with the other approaches. As seen in Table 11, the loading for the *externalizing* item was 1.99 for males and 1.58 for females. The intercept for the *approaches to learning* item was 2.85 for males and 2.91 for females, a difference of 0.06. The difference in observed means on this item was 0.27, meaning that roughly a quarter of the difference in observed means on this item was due to the difference in the intercept between the genders. (The remaining difference is due to the gender differences on the latent factor.)

Muthén's Maximum Likelihood Approach

Table 12 contains the fit statistics for all models analyzed using the MUML approach. In order to get the model to converge, the *interpersonal* and *approaches to learning* items were recoded to be similarly valenced to the *internalizing* and *externalizing* items. The fit of the configural model was acceptable, RMSEA = .058, CFI = .987, SRMR = .027. Additionally, fit of the metric model was good, RMSEA = .061, CFI = .983, SRMR = .035, but the χ^2-difference test was significant, χ^2 (3) = 85.49, p < .001. The modification indices indicated that freeing the loading for the *externalizing* item would lead to better fit. This partial metric invariance model had good fit, RMSEA = .056, CFI = .987, SRMR = .027, and the fit was not significantly different from the fit of the configural invariance model, χ^2 (2) = 3.26, p = .196. Partial scalar invariance did not hold, χ^2 (2) = 109.18, p < .001, so revised partial scalar invariance models were considered. The intercept for the *approaches to learning* item was freed to vary across groups because it had the highest modification index (*MI* = 61.68). The fit of this revised partial scalar invariance model was significantly different from the fit of the partial metric invariance model, χ^2 (1) = 49.62, p < .001, but was accepted due to the test being overpowered. Strict invariance held, χ^2 (2) = 4.66, p = .097. The final invariance model using the MUML approach was a partial strict invariance model where the loading, intercept, and unique variance of the *externalizing* item and the intercept and unique variance of the *approaches to learning* item were freed to vary across boys and girls.

Using the results from the partial strict invariance model, the factor variance for girls was 0.04 and for boys was 0.04. The pooled standard deviation for these two variances was 0.20. Girls had a factor mean 0.08 units or 0.39 standard deviations lower

than boys. As Table 13 illustrates, boys had a larger within-level factor loading for the *externalizing* item ($\lambda = 2.23$) than girls ($\lambda = 1.82$). Additionally, boys had a higher intercept on the *approaches to learning* item ($\tau = 0.12$) than girls ($\tau = 0.07$), indicating that for a boy and girl who had the same factor score, the girl was more likely to have been rated with a higher score than the boy on that item.

Summary

The conclusions of the five approaches converged; the loading for the *externalizing* indicator and the intercept for the *approaches to learning* indicator were not invariant across gender. Determining invariance was difficult due to the large sample size and limited fit information. While the specific fit statistics were different across approaches, the same overall pattern emerged. For instance, the scaled χ^2 for testing the metric invariance model using the multilevel mixture factor model for known classes was 78.01 with three degrees of freedom and was 109.79 with two degrees of freedom for testing the partial scalar invariance model. If I rejected the first model, then I would have to reject the second model given that the scaled χ^2 was larger and there were fewer degrees of freedom. There was a similar pattern for the design-based approach where the scaled χ^2 for testing the metric invariance model was 83.44 with three degrees of freedom and was 104.76 with two degrees of freedom for testing the partial scalar invariance model.

The values of the parameters were similar across approaches. For instance, in the partial strict invariance model using the design-based approach the loadings for the *internalizing*, *approaches to learning*, and *interpersonal* indicators were 1.00, -2.45, and

-2.48, respectively. Boys had a loading of 1.99 on the *externalizing* scale and girls had a loading of 1.58. In the partial strict invariance model using the definition variable approach, the loadings for the *internalizing*, *approaches to learning*, and *interpersonal* indicators were 1.00, -2.43, and -2.51, respectively. Boys had a loading of 2.15 on the *externalizing* scale and girls had a calculated loading of 1.72. Even when different parameterizations were used, similar results emerged. A within-level intercept and a between-level intercept were estimated in the models that used the MUML approach whereas one overall intercept was estimated using the multilevel factor mixture model for known classes (refer to Equation 22 and Equation 17, respectively). In the partial strict invariance model using the MUML approach, the within-level intercept for the *internalizing* scale was $\tau = 0.038$ and the between-level intercept was $\tau = 1.544$. The sum of these two intercepts (1.582) is roughly equal to the intercept estimated in the partial strict invariance model using the multilevel mixture approach ($\tau = 1.584$). So while there were different parameterizations, overall, the approaches converged on similar parameter estimates.

One of the divergences of conclusions across approaches was the different estimates of the between-level loadings, especially for the *interpersonal* and *approaches to learning* indicators. The design-based approach does not estimate the between-level loadings. The multilevel mixture model and definition variable approaches had similar estimates. The between-level loading estimates for the *interpersonal* and *approaches to learning* indicators using the MUML approach were different because those two indicators were reverse coded in order to estimate the model. The between-level loading

estimates using the MIMIC approach were different than the other approaches; they were positive and greater in value. This is probably because the between-level unique variances were not estimated in this approach, which affected the between-level model.

CHAPTER 4

DISCUSSION

Measurement invariance testing is essential for making valid group comparisons on scale scores. Typically in measurement invariance testing, the measurement model is estimated separately for each group, where parameters can be constrained to be equal across the groups to test for different levels of invariance. However, if the data have a hierarchical structure and the grouping variable is at level-1 (e.g., comparing boys and girls who are clustered within schools), then the dependence of the scores needs to be taken into consideration. The goal of this master's thesis was to compare five different approaches to testing measurement invariance in multilevel data structures when the grouping variable was at level-1.

The five approaches I used to study measurement invariance test for invariance in different ways. I, therefore, focused on the substantive conclusions garnered from each approach. The statistical conclusions of these approaches (e.g., whether a metric invariance model fit better than a scalar invariance model and if there was a significant difference in a loading) were compared to assess their degree of agreement.

The substantive conclusions of all five approaches were the same – the factor loading of the *externalizing* item and the intercept of the *approaches to learning* item were not invariant across genders. The sample size for my illustrative example was large, making likelihood ratio tests overpowered. This resulted in having no clear cut-off criteria for determining when invariance held. Having no clear criteria, I followed a consistent approach to model comparison and determined that each approach, while

providing different information, yielded the same conclusions regarding the level of measurement invariance.

One of the difficulties of comparing the five different approaches was that the type of information provided by each approach differed. To be consistent across approaches, I relied on the comparative fit indices (i.e., AIC, BIC, SABIC) and the likelihood ratio tests for models estimated using maximum likelihood and the comparative fit indices and the Satorra-Bentler likelihood ratio tests for models estimated using robust maximum likelihood. Even though global fit statistics were calculated for the design-based and MUML approaches, I did not rely on this information in order to be consistent across approaches. This information, however, can help in the decision-making process of measurement invariance testing. Chen (2007) developed a set of criteria based on change in global fit statistics that could be used; however, the set of criteria were developed for single-level data and sample sizes less than or equal to 1,000. It would be important to validate his results for testing measurement invariance in multilevel data before relying on them for measurement invariance testing in hierarchical data structures. In addition to fit information, the approaches differed in how they estimated parameters. The multilevel mixture model for known classes, design-based, and MUML approaches estimate a model separately for each group and constrain parameters to be equal - the standard way of conducting measurement invariance analyses. In contrast, the MIMIC and definition variable approaches estimate one model for the sample and specific parameters are predicted (to allow differences) or not predicted (to constrain the parameter to be equal across groups) by a grouping variable.

Limitations

One of the limitations of my thesis was that my analyses were data driven. I did not have substantive theory about which indicator may be biased across gender and may have capitalized on chance when determining which loading, intercept, or unique variance to free across groups. If researchers do not have substantive theory to guide decisions, they can use cross-validation. One method to cross-validate the model is to split the data set into a calibration sample and a validation sample (Bentler, 1980). Rather than running measurement invariance tests on the full sample, researchers can run tests on just the calibration sample and make modifications (e.g., free a parameter to vary across groups) to improve model fit. The final model with empirical modifications is then tested using the validation sample. If the model has good fit in the validation sample, then the modifications were appropriate and the model is generalizable.

The generalizability of the results of my master's thesis is limited because the analyses were based on one illustrative example. For this illustrative example, the conclusions of the five approaches were the same; however, this does not necessarily mean that the conclusions will be the same in all situations. If the factor model becomes more complex (e.g., more items, more factors), then the conclusions of the approaches may diverge. For instance, as Table 1 explicates, one of the assumptions of the design-based approach is that the within-level model is equal to the between-level model. Wu and Kwok (2012) found that when the within-level model was complex or the between-level model was more complex than the within-level model, the factor loading estimates were biased. In my illustrative example where I had one factor and four indicators, the

assumption of the equivalency of the within-level and between-level models was likely to hold. But if the number of items of a scale increases (e.g., 20 items), then the possibility for complexity at both levels of the model increases. In this scenario, the within-level and between-level models may not be equal and the estimates of the design-based approach may be biased and diverge from the estimates of the other approaches.

Another situation where the conclusions of the approaches may diverge is when the factor distributions differ substantially between the groups. In the MIMIC approach, the factor variance is estimated for the whole sample and the groups are not able to differ on that variance. It may not always be substantively appropriate to assume that the factor variances across groups should be equal. For instance, boys tend to have a higher variability than girls on tests of math and reading (Machin & Pekkarinen, 2008). If there is a group difference in the factor variance and that group difference is not modeled, then the tests of invariance can be distorted.

Additionally, the conclusions of the approaches may diverge when the unique variances differ substantially across groups. In the MIMIC approach, a group difference on the unique variances is not able to be incorporated into the model. If the groups do differ on the unique variances, this can distort invariance testing of the loadings and the intercepts. In this scenario, I would expect the conclusions of the MIMIC approach to differ from the conclusions of the other four approaches.

Future Directions

Simulation work should be conducted that investigates situations in which the conclusions of the approaches diverge (e.g., the within-level and between-level models

are not equal, the factor variance is substantially different between two groups), how sensitive the analyses are to violations of those assumptions, the sample size requirements for each approach, and the power differences of the approaches.

To accurately and precisely estimate the fit and the parameters of these models, an adequate sample size is needed. There are many factors that influence the minimum sample size required such as the level of communality of the variables and how equal the sample sizes are across groups (MacCallum, Widaman, Zhang, & Hong, 1999). Simulation studies found that sample sizes as low as 100 participants per group can have enough power to detect measurement invariance in independent data (Meade & Bauer, 2007). To test measurement invariance in clustered data, an adequate sample size is also needed at level-2. The accuracy of MUML approximation depends on the sample size within clusters and the sample size at level-2 (Yuan & Hayashi, 2005). A previous simulation study investigated multilevel measurement invariance for a sample size of 1600 with 200 clusters (Ryu, 2015). Kim et al. (2015) investigated sample sizes between 600 and 3200 participants with the number of clusters varying between 60, 100, or 160 clusters. They found that smaller sample sizes did not have enough power to detect noninvariance. They recommended a sample size of at least 2,000 participants using a balanced design. Future research should expand on these results and determine the minimum sample size required to detect invariance for all approaches. This may guide which approach should be chosen when investigating measurement invariance in multilevel data.

In this illustrative example, the sample sizes of the two groups were roughly equal in magnitude. There are situations where the sample sizes are unequal such as investigating measurement equivalence across ethnicity. If there is an imbalance in sample sizes between the groups, then power to detect factor mean differences is lower compared to situations where the sample size is balanced (Kaplan & George, 1995). For independent data, the MIMIC approach has more power to detect group differences than a two-group item response theory (IRT) model (Woods, 2009). It is reasonable to assume that the definition variable and MIMIC approaches may have more power to detect noninvariance in multilevel data than the other approaches because a model is not estimated separately in each group. Future research is needed to confirm this assumption.

Conclusions

While the conclusions of the five approaches converged in this study, there are benefits and limitations to each approach. I would not recommend using the MIMIC approach for invariance testing because a group difference on the factor variance is not able to be modeled and because the unique variances are not able to differ between groups. This can distort invariance testing of the other parameters (i.e., loadings and intercepts). Additionally, incorporating the between-level unique variances into the model leads to estimation and convergence issues. But without incorporating them, the between-level model is distorted. One weakness of the definition variable approach is that it ignores the original data structure. Rather than using the uncentered values of the indicators to estimate a two-level model, the within-level and between-level models are estimated separately using the deviations from the cluster means and the cluster means of

the indicators, respectively. Measurement invariance is tested in the within-level model. Because of this set-up, uncertainty in the cluster means is not able to be taken into account because the cluster means, as well as the deviations from the cluster means, are treated as observed variables in the model. This can cause the model to be incorrectly estimated. The MUML approach is a good approach if there are no missing data. Otherwise, this approach requires listwise deletion, which can distort the results if the data are not missing completely at random. If the factor structure is not complex (e.g., small number of items), the design-based approach is a good approach to use and has the least computational demand. Overall, I would recommend that researchers use the multilevel mixture model for known classes approach to test for measurement invariance in multilevel data structures. This approach appears to have the least assumptions that can distort invariance testing though further research needs to be done to support this.

REFERENCES

Bauer, D. J., & Hussong, A. (2009). Psychometrical approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods, 14,* 101-125.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, *31*, 419-456. Reprinted in C. Fornell (Ed.), *A second generation of multivariate analysis, Vol. 1*. New York: Praeger (1982).

Bovaird, J. A., & Shaw, L. H. (2012). Multilevel structural equation modeling. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 501-518). New York, NY: Guilford Press.

Chan (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*, 234-246.

Chen, F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 464-504.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1-27.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum.

Enders, C. K., Mistler, S. A., and Keller, B. T. (in press). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *The Southern Psychologist*, *1,* 179-188.

Kaplan & George (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 2,* 101-118.

Kim, E., Kwok, O., & Yoon, M. (2012). Testing factorial invariance in multilevel data: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal, 19*, 250-267.

Kim, E., Yoon, M., Wen, Y., Luo, W., & Kwok, O. (2015). Within-level group factorial invariance with multilevel data: Multilevel factor mixture and multilevel MIMIC models. *Structural Equation Modeling, 22*, 603-616.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203-229.

MacCallum, R., Widaman, K., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84-99.

Machin, S., & Pekkarinen, T. (2008). Global sex differences in test score variability. *Science, 322*, 1331-1332.

Meade, A., & Bauer, D. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 611-635.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10,* 259-284.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58,* 525-543.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, & N. A. Card (Eds.), *Handbook of developmental research methods* (pp. 109-126). New York, NY: Guilford Press.

Millsap, R. E., & Everson, H. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research, 26*, 479-497.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass, 2*, 842-860.

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology, 67*, 172-194.

Ryu, E. (2015). Multiple group analysis in multilevel structural equation model across level 1 groups. *Multivariate Behavioral Research, 50,* 300-315.

Satorra, A., & Bentler, P.M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, 75, 243-248.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross national consumer research. *Journal of Consumer Research, 25,* 78-90.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

US Department of Education, National Center for Education Statistics. (2009). Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) Kindergarten through Eighth Grade Full Sample Public-Use Data and Documentation (DVD). (NCES 2009-005). Washington, DC: Author.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance abuse domain. In K. J. Bryant, *Alcohol and substance use research* (pp.281-324). Washington, D.C.: American Psychological Association.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*, 1-27.

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

Wu, J., & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multi-level approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 19,* 16-35.

Wu, W., Lu, Y., Tan, F., Yao, S., Steca, P., Abela, J., & Hankin, B. (2012). Assessing measurement invariance of the Children's Depression Inventory in Chinese and Italian primary school student samples. *Assessment*, *19*, 506-516.

Yoon, M., & Millsap, R. E., (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, *14,* 435-463.

Yuan, K.-H., & Hayashi, K. (2005). On Muthén's maximum likelihood for two-level covariance structure models. *Psychometrika*, *70*, 147-167.

APPENDIX A

TABLES

Table 1

Assumptions of each approach as related to measurement invariance

Approach	Assumptions/Constraints
ML Mixture	Between-level matrices are equal across groups; intercept estimated at the between-level
MIMIC	Factor variances are equal across groups; within-level unique variances are equal across groups; between-level unique variances are not estimated; between-level matrices are equal across groups; intercept estimated at the between-level, but the group difference estimated at the within-level
Definition Variable	Cluster means are equally reliable; intercept estimated at the between-level and at the within-level
Design-based	Between-level and within-level matrices are equal; intercept estimated at the within-level
MUML	Between-level matrices are equal across groups; listwise deletion; intercept estimated at the between-level and at the within-level

Table 2

Levels of invariance that can be tested by each approach

Approach	Configural	Metric	Scalar	Strict
ML Mixture	x[1]	x[1]	x[2]	x[1]
MIMIC	x[1]	x[1]	x[3]	
Definition Variable	x[1]	x[1]	x[1]	x[1]
Design-based	x[4]	x[4]	x[4]	x[4]
MUML	x[1]	x[1]	x[1]	x[1]

Notes: x indicates that the level of invariance can be tested using that approach, [1]The level of invariance is tested at level-1, [2]The level of invariance is tested at level-2, [3]The intercepts are estimated at level-2, but invariance is calculated at level-1, [4]There is only one level of invariance

Table 3

Testing MIMIC models to identify the reference variable

Reference Variable	Item tested	Test statistic[1]	*p*-value
Internalizing	Externalizing	-14.387	<.001
	Interpersonal	1.902	.057
	Approaches to Learning	4.257	<.001
Externalizing	Internalizing	1.124	.261
	Interpersonal	-3.404	.001
	Approaches to Learning	0.073	.942
Interpersonal	Internalizing	0.538	.591
	Externalizing	14.913	<.001
	Approaches to Learning	-3.853	<.001
Approaches to Learning	Internalizing	0.140	.889
	Externalizing	14.048	<.001
	Interpersonal	-0.026	.980

Note: [1]This is the test statistic for the regression coefficient that captures the relationship between the item tested and the interaction of the grouping variable and the latent variable, ω

Table 4

Fit statistics for multilevel factor mixture model for known classes approach

Model	AIC	BIC	SABIC
Configural	130,310	130,567	130,462
Metric	130,413	130,646	130,551
Partial Metric[1]	130,310	130,551	130,453
Partial Metric[2]	130,377	130,618	130,520
Partial Metric[3]	130,411	130,652	130,553
Partial Scalar[4]	130,443	130,669	130,577
Partial Scalar[5]	130,370	130,604	130,508
Partial Scalar[6]	130,423	130,657	130,562
Partial Strict[7]	130,372	130,590	130,501

Notes: [1]Externalizing loading freed to vary across groups, [2]Interpersonal loading freed to vary across groups, [3]Approaches to Learning loading freed to vary across groups, [4]Externalizing loading and intercept freed to vary across groups, [5]Approaches to Learning intercept freed to vary across groups in addition to the Externalizing loading and intercept, [6]Interpersonal intercept freed to vary across groups in addition to the Externalizing loading and intercept, [7]Externalizing loading, intercept, and unique variance and Approaches to Learning intercept and unique variance freed to vary across groups

Table 5

Parameter estimates for the partial strict invariance model using the multilevel factor mixture model for known classes approach

	Within-level				Between-level			
	Loadings		Unique Variances		Loadings	Intercepts		Unique Variances
Item	Males	Females	Males	Females		Males	Females	
Internalizing	1	1	0.207	0.207	1	1.584	1.584	0.019
Externalizing	2.140	1.677	0.250	0.184	1.318	1.764	1.647	0.018
Interpersonal	-2.473	-2.473	0.081	0.081	-3.436	2.857	2.857	0.017
Approaches to Learning	-2.397	-2.397	0.155	0.146	-3.617	2.849	2.914	0.006

Table 6

Fit statistics for the models using the MIMIC approach

Model	AIC	BIC	SABIC
Configural	107,321	107,500	107,427
Metric	107,511	107,667	107,603
Partial Metric[1]	107,336	107,499	107,433
Partial Scalar[2]	107,480	107,628	107,568
Partial Scalar[3]	107,469	107,625	107,561
Partial Scalar[4]	107,412	107,567	107,504

Notes: [1]Externalizing loading freed to vary across groups, [2]Externalizing loading and intercept freed to vary across groups, [3]Interpersonal intercept freed to vary across groups in addition to the Externalizing loading and intercept, [4]Approaches to Learning intercept freed to vary across groups in addition to the Externalizing loading and intercept

Table 7

Parameter estimates for the partial scalar invariance model using the MIMIC approach

	Loadings			Intercepts		Unique Variances
Item	λ_w	ω_w	λ_b	τ_b	β_w	θ_w
Internalizing	1	0	1	1.590	0	0.220
Externalizing	2.077	-0.520	1.020	1.765	-0.116	0.217
Interpersonal	-2.255	0	8.523	2.861	0	0.082
Approaches to Learning	-2.187	0	5.723	2.848	0.064	0.169

Note: Parameters with no decimal places were fixed to that number

Table 8

Fit statistics for the models using the definition variable approach

Model	AIC	BIC	SABIC
Configural	96,478	96,760	96,646
Metric	96,573	96,831	96,726
Partial Metric[1]	96,478	96,745	96,637
Partial Scalar[2]	96,610	96,860	96,759
Partial Scalar[3]	96,536	96,795	96,690
Partial Strict[4]	96,537	96,781	96,682

Notes: [1]Externalizing item loading freed to vary across groups, [2]Externalizing item loading and intercept freed to vary across groups, [3]Externalizing item loading and intercept and Approaches to Learning intercept freed to vary across groups, [4]Externalizing item loading, intercept, and unique variance and Approaches to Learning intercept and unique variance freed to vary across groups

Table 9

Parameter estimates for the partial strict invariance model using the definition variable approach

	Within-Level						Between-Level		
	Loadings		Intercepts		Unique Variances		Loadings	Intercepts	Unique Variances
Item	Overall Model	Group Difference[1]	Overall Model	Group Difference[1]	Overall Model	Group Difference[1]			
Internalizing	1	N/A	0.039	N/A	0.194	N/A	1	1.541	0.037
Externalizing	2.147	-0.427	0.120	-0.108	0.233	-0.058	1.324	1.638	0.039
Interpersonal	-2.509	N/A	-0.098	N/A	0.075	N/A	-2.693	2.968	0.025
Approaches to Learning	-2.429	N/A	-0.125	0.062	0.145	-0.006	-2.693	2.983	0.022

Notes: Parameters with no decimal places were fixed to that number, [1]Girls were coded 1 and boys were coded 0, so a negative difference indicates a lower value for girls on that parameter

Table 10

Fit statistics for the models using the design-based approach

Model	χ^2 (*df*)	RMSEA	CFI	SRMR	AIC	BIC	SABIC
Configural	46.72 (4)	.035	.997	.009	109,574	109,761	109,685
Metric	119.95 (7)	.043	.991	.034	109,670	109,834	109,834
Partial Metric[1]	54.97 (6)	.030	.996	.011	109,573	109,744	109,674
Partial Scalar[2]	137.85 (8)	.043	.989	.023	109,683	109,838	109,775
Partial Scalar[3]	93.35 (7)	.037	.993	.017	109,625	109,788	109,722
Partial Strict[4]	95.61 (9)	.033	.993	.023	109,626	109,774	109,713

Notes: [1]Externalizing item loading freed to vary across groups, [2]Externalizing item loading and intercept freed to vary across groups, [3]Externalizing item loading and intercept and Approaches to Learning intercept freed to vary across groups, [4]Externalizing item loading, intercept, and unique variance and Approaches to Learning intercept and unique variance freed to vary across groups

Table 11

Parameter estimates for the partial strict invariance model using the design-based approach

	Loadings		Intercepts		Unique Variances	
Item	Males	Females	Males	Females	Males	Females
Internalizing	1	1	1.586	1.586	0.226	0.226
Externalizing	1.994	1.578	1.765	1.638	0.278	0.205
Interpersonal	-2.481	-2.481	2.859	2.859	0.101	0.101
Approaches to Learning	-2.453	-2.453	2.847	2.907	0.160	0.150

Table 12

Fit statistics for the models using the MUML approach

Model	χ^2 (*df*)	RMSEA	CFI	SRMR	AIC	BIC	SABIC
Configural	322.74 (20)	0.058	0.987	0.027	106,619	106,899	106,785
Metric	408.23 (23)	0.061	0.983	0.035	106,698	106,955	106,850
Partial Metric[1]	326.00 (22)	0.056	0.987	0.027	106,618	106,883	106,775
Partial Scalar[2]	435.18 (24)	0.062	0.982	0.031	106,723	106,972	106,871
Partial Scalar[3]	375.61 (23)	0.059	0.984	0.030	106,666	106,923	106,818
Partial Strict[4]	380.28 (25)	0.056	0.984	0.030	106,666	106,908	106,809

Notes: [1]Externalizing item loading freed to vary across groups, [2]Externalizing item loading and intercept freed to vary across groups, [3]Externalizing item loading and intercept and Approaches to Learning intercept freed to vary across groups, [4]Externalizing item loading, intercept, and unique variance and Approaches to Learning intercept and unique variance freed to vary across groups

Table 13

Parameter estimates for the partial strict invariance model using the MUML approach

	Within-level						Between-level		
	Loadings		Intercepts		Unique Variances		Loadings	Intercepts	Unique Variances
Item	Males	Females	Males	Females	Males	Females			
Internalizing	1	1	0.038	0.038	0.206	0.206	1	1.544	0.023
Externalizing	2.228	1.823	0.118	0.018	0.246	0.187	1.262	1.638	0.022
Interpersonal	2.558	2.558	0.097	0.097	0.080	0.080	3.078	2.039	0.020
Approaches to Learning	2.512	2.512	0.124	0.066	0.153	0.147	3.257	2.023	0.007

APPENDIX B

FIGURES

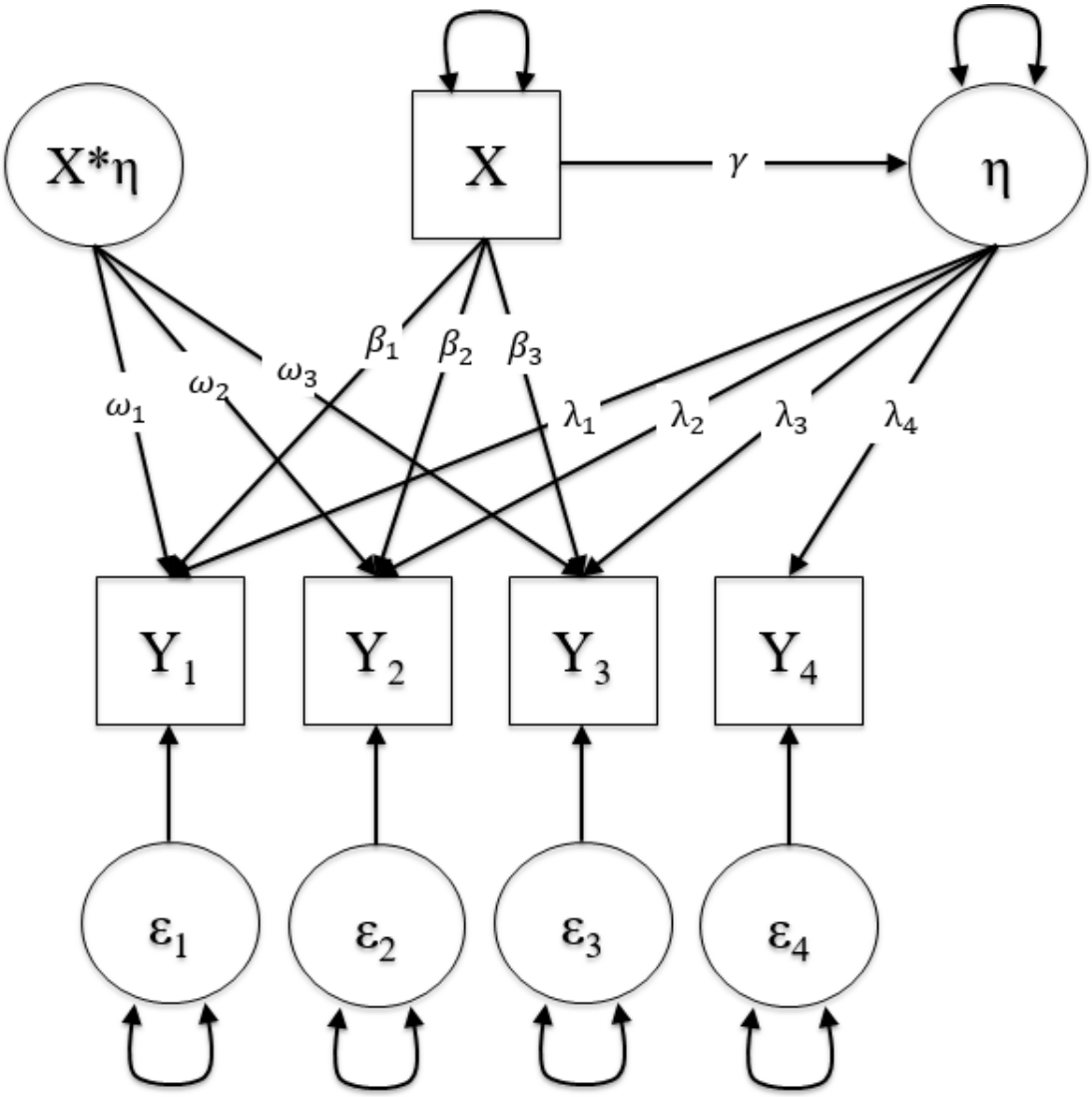

Figure 1. Path diagram for the configural invariance MIMIC model with item Y_4 designated as the reference variable.

APPENDIX C

MPLUS 7.3 INPUT FILES FOR CONFIGURAL INVARIANCE AND PARTIAL STRICT INVARIANCE MODELS – MULTILEVEL MIXTURE MODEL FOR KNOWN CLASSES APPROACH

```
TITLE: Configural Invariance Model

DATA:
FILE = eclsk_listwise_deletion.txt;

VARIABLE:
NAMES = s1_id cfemale t1learn t1interp t1intern t1extern;
USEVARIABLES = t1learn t1interp t1extern t1intern;
CLUSTER = s1_id;
CLASSES = class (2);
KNOWNCLASS = class (cfemale=0 cfemale=1);
! 0 = Male, 1 = Female;

ANALYSIS:
TYPE = TWOLEVEL MIXTURE;
ESTIMATOR = MLR;
PROCESSORS = 5;
INTEGRATION = MONTECARLO;

MODEL:
%WITHIN%
   %OVERALL%
   fw1 BY t1learn* t1interp t1extern t1intern@1;

   %class#1%
   fw1 BY t1learn*-1 t1interp*-1 t1extern*1 t1intern@1;
   t1learn t1interp t1extern t1intern;
   [fw1@0];
   fw1;

   %class#2%
   fw1 BY t1learn*-1 t1interp*-1 t1extern*1 t1intern@1;
   t1learn t1interp t1extern t1intern;
   [fw1];
   fw1;

%BETWEEN%
   %OVERALL%
   fb1 BY t1learn*-1 t1interp*-1 t1extern*5 t1intern@1;
   [fb1@0];
   fb1;
   t1learn t1interp t1extern t1intern;

   %class#1%
   [t1intern] (i1);
```

[t1learn t1interp t1extern];

%class#2%
[t1intern] (i1);
[t1learn t1interp t1extern];

```
TITLE: Partial Strict Invariance Model

DATA:
FILE = eclsk_listwise_deletion.txt;

VARIABLE:
NAMES = s1_id cfemale t1learn t1interp t1intern t1extern;
USEVARIABLES = t1learn t1interp t1extern t1intern;
CLUSTER = s1_id;
CLASSES = class (2);
KNOWNCLASS = class (cfemale=0 cfemale=1);
! 0 = Male, 1 = Female;

ANALYSIS:
TYPE = TWOLEVEL MIXTURE;
ESTIMATOR = MLR;
PROCESSORS = 5;
INTEGRATION = MONTECARLO;

MODEL:
%WITHIN%
   %OVERALL%
   fw1 BY t1learn* t1interp t1extern t1intern@1;

   %class#1%
   fw1 BY t1learn*-1 t1interp*-1 t1intern@1 (l1-l3);
   fw1 BY t1extern*1;
   t1intern t1interp (r1-r2);
   t1learn t1extern;
   [fw1@0];
   fw1;

   %class#2%
   fw1 BY t1learn*-1 t1interp*-1 t1intern@1 (l1-l3);
   fw1 BY t1extern*1;
   t1intern t1interp (r1-r2);
   t1learn t1extern;
   [fw1];
   fw1;

%BETWEEN%
   %OVERALL%
   fb1 BY t1learn*-1 t1interp*-1 t1extern*5 t1intern@1;
   [fb1@0];
   fb1;
```

```
t1learn t1interp t1extern t1intern;

%class#1%
[t1intern t1interp] (i1-i2);
[t1extern t1learn];

%class#2%
[t1intern t1interp] (i1-i2);
        [t1extern t1learn];
```

APPENDIX D

MPLUS 7.3 INPUT FILES FOR CONFIGURAL INVARIANCE AND PARTIAL SCALAR INVARIANCE MODELS – MIMIC APPROACH

```
DATA:
FILE IS eclsk_listwise_deletion.txt;

VARIABLE:
NAMES = s1_id cfemale t1learn t1interp t1intern t1extern;
USEVARIABLES = t1learn t1interp t1extern t1intern cfemale;
MISSING = .;
WITHIN = cfemale; !(0=Male 1=Female);
CLUSTER = s1_id;


ANALYSIS:
TYPE = TWOLEVEL RANDOM;
ESTIMATOR = ML;
ALGORITHM = INTEGRATION;
PROCESSORS = 5;

MODEL:
%WITHIN%
   FW1 BY t1learn*-1 t1interp*-1 t1extern*2 t1intern@1;

   Inter | FW1 XWITH cfemale; ! creating an interaction

   t1learn ON Inter; !testing invariance of loading
   t1interp ON Inter;
   t1extern ON Inter;

   t1learn ON cfemale; !testing invariance of intercept
   t1interp ON cfemale;
   t1extern ON cfemale;

   t1learn t1interp t1extern t1intern;

   [FW1@0]; ! fix within factor mean (for boys) to 0
   FW1 ON cfemale; ! group difference in a within-level factor

%BETWEEN%
   FB1 BY t1learn* t1interp t1extern t1intern@1;
```

```
DATA:
FILE IS eclsk_listwise_deletion.txt;

VARIABLE:
NAMES = s1_id cfemale t1learn t1interp t1intern t1extern;
USEVARIABLES = t1learn t1interp t1extern t1intern cfemale;
MISSING = .;
WITHIN = cfemale; !(0=Male 1=Female);
CLUSTER = s1_id;


ANALYSIS:
TYPE = TWOLEVEL RANDOM;
ESTIMATOR = ML;
ALGORITHM = INTEGRATION;
PROCESSORS = 5;

MODEL:
%WITHIN%
   FW1 BY t1learn*-1 t1interp*-1 t1extern*2 t1intern@1;

   Inter | FW1 XWITH cfemale;

   !t1learn ON Inter;
   !t1interp ON Inter;
   t1extern ON Inter;

   t1learn ON cfemale;
   !t1interp ON cfemale;
   t1extern ON cfemale;

   t1learn t1interp t1extern t1intern;

   [FW1@0];
   FW1 ON cfemale;

%BETWEEN%
          FB1 BY t1learn* t1interp t1extern t1intern@1; 75
```

APPENDIX E

MPLUS 7.3 INPUT FILES FOR CONFIGURAL INVARIANCE AND PARTIAL STRICT INVARIANCE MODELS – DEFINITION VARIABLE APPROACH

```
DATA:
FILE (within) = Deviations.txt;
FILE (between) = Clustermeans.txt;

VARIABLE:
NAMES = s1_id cfemale t1learn t1interp t1intern t1extern;
USEVARIABLES = t1learn t1interp t1extern t1intern;
CONSTRAINT = cfemale;

ANALYSIS:
TYPE = MEANSTRUCTURE;
ESTIMATOR = ML;
ITERATIONS = 1000000;
PROCESSORS = 5;

MODEL:
BF BY t1learn* t1interp t1extern t1intern@1;
WF BY t1learn* t1interp t1extern t1intern@1;

MODEL within:
WF BY t1learn* t1interp t1extern t1intern@1 (l1 l2 l3 l4);
[WF] (alpha); !alpha
WF (var);

BF BY t1learn@0 t1interp@0 t1extern@0 t1intern@0;
[BF@0];
BF@0;
WF WITH BF@0;

[t1learn t1interp t1extern t1intern] (tau1 tau2 tau3 tau4);

t1learn t1interp t1extern t1intern (epsilon1 epsilon2 epsilon3 epsilon4);


MODEL between:
BF BY t1learn*-1 t1interp*-1 t1extern*2 t1intern@1;
[BF@0];
BF;

WF BY t1learn@0 t1interp@0 t1extern@0 t1intern@0;
[WF@0];
WF@0;
WF WITH BF@0;

t1learn t1interp t1extern t1intern;
```

```
MODEL CONSTRAINT:
NEW(gammal10*-2.5 gammal11*.04 gammal20*-2.5 gammal21*.04 gammal30*1
gammal31*.04
   gammai10*2.8 gammai11*.28 gammai20*2.8 gammai21*.29 gammai30*1.8
gammai31*-.27
   gammae10*.14 gammae11*0 gammae20*.07 gammae21*.01 gammae30*.24
gammae31*-.07
   gammae40*.2 gammae41*-.01 gammam*.035 varm*.05 gammav*.01);

l1 = gammal10 + gammal11*cfemale;
l2 = gammal20 + gammal21*cfemale;
l3 = gammal30 + gammal31*cfemale;

tau1 = gammai10 + gammai11*cfemale;
tau2 = gammai20 + gammai21*cfemale;
tau3 = gammai30 + gammai31*cfemale;

epsilon1 = gammae10 + gammae11*cfemale;
epsilon2 = gammae20 + gammae21*cfemale;
epsilon3 = gammae30 + gammae31*cfemale;
epsilon4 = gammae40 + gammae41*cfemale;

alpha = 0 + gammam*cfemale;
var = varm + gammav*cfemale;
```

```
DATA:
FILE (within) = Deviations.txt;
FILE (between) = Clustermeans.txt;

VARIABLE:
NAMES = s1_id cfemale t1learn t1interp t1intern t1extern;
USEVARIABLES = t1learn t1interp t1extern t1intern;
CONSTRAINT = cfemale;

ANALYSIS:
TYPE = MEANSTRUCTURE;
ESTIMATOR = ML;
ITERATIONS = 1000000;
PROCESSORS = 5;

MODEL:
BF BY t1learn* t1interp t1extern t1intern@1;
WF BY t1learn* t1interp t1extern t1intern@1;

MODEL within:
WF BY t1learn* t1interp t1extern t1intern@1 (l1 l2 l3 l4);
[WF] (alpha); !alpha
WF (var);

BF BY t1learn@0 t1interp@0 t1extern@0 t1intern@0;
[BF@0];
BF@0;
WF WITH BF@0;

[t1learn t1interp t1extern t1intern] (tau1 tau2 tau3 tau4);

t1learn t1interp t1extern t1intern (epsilon1 epsilon2 epsilon3 epsilon4);


MODEL between:
BF BY t1learn*-1 t1interp*-1 t1extern*2 t1intern@1;
[BF@0];
BF;

WF BY t1learn@0 t1interp@0 t1extern@0 t1intern@0;
[WF@0];
WF@0;
WF WITH BF@0;

![t1learn t1interp t1extern t1intern] (tau1 tau2 tau3 tau4);
```

```
t1learn t1interp t1extern t1intern;


MODEL CONSTRAINT:
NEW(gammal30*1 gammal31*.04
   gammai10*2.8 gammai11*.28 gammai30*1.8 gammai31*-.27
   gammae10*.14 gammae11*0 gammae30*.24 gammae31*-.07
   gammam*.035 varm*.05 gammav*.01);

l3 = gammal30 + gammal31*cfemale;
tau1 = gammai10 + gammai11*cfemale;
tau3 = gammai30 + gammai31*cfemale;
epsilon1 = gammae10 + gammae11*cfemale;
epsilon3 = gammae30 + gammae31*cfemale;

alpha = 0 + gammam*cfemale;
var = varm + gammav*cfemale;
```

APPENDIX F

MPLUS 7.3 INPUT FILES FOR CONFIGURAL INVARIANCE AND PARTIAL STRICT INVARIANCE MODELS – DESIGN-BASED APPROACH

```
DATA:
FILE IS eclsk_listwise_deletion.txt;

VARIABLE:
NAMES = s1_id cfemale t1learn t1interp t1intern t1extern;
USEVARIABLES = t1learn t1interp t1extern t1intern;
GROUPING = cfemale (0=Male 1=Female);
CLUSTER = s1_id;

ANALYSIS:
TYPE = COMPLEX;
ESTIMATOR = MLR;

MODEL:
F1 BY t1learn* t1interp t1extern t1intern@1;

MODEL Male:
F1 BY t1learn*-1 t1interp*-1 t1extern*2 t1intern@1;
[F1@0];
F1;

[t1intern] (i1);
[t1learn t1interp t1extern];

t1learn t1interp t1extern t1intern;

MODEL Female:
F1 BY t1learn* t1interp t1extern t1intern@1;
[F1];
F1;

[t1intern] (i1);
[t1learn t1interp t1extern];

t1learn t1interp t1extern t1intern;

OUTPUT:
MODINDICES;
```

```
DATA:
FILE IS eclsk_listwise_deletion.txt;

VARIABLE:
NAMES = s1_id cfemale t1learn t1interp t1intern t1extern;
USEVARIABLES = t1learn t1interp t1extern t1intern;
GROUPING = cfemale (0=Male 1=Female);
CLUSTER = s1_id;

ANALYSIS:
TYPE = COMPLEX;
ESTIMATOR = MLR;

MODEL:
F1 BY t1learn* t1interp t1extern t1intern@1;

MODEL Male:
F1 BY t1learn*-1 t1interp*-1 t1intern@1 (l1-l3);
F1 BY t1extern*2;
[F1@0];
F1 (var1);

[t1intern] (i1);
[t1interp] (i3);
[t1learn t1extern];

t1interp t1intern (r1-r2);
t1extern t1learn;

MODEL Female:
F1 BY t1learn*-1 t1interp*-1 t1intern@1 (l1-l3);
F1 BY t1extern*2;
[F1] (mean);
F1 (var2);

[t1intern] (i1);
[t1interp] (i3);
[t1learn t1extern];

t1interp t1intern (r1-r2);
t1extern t1learn;

OUTPUT:
MODINDICES;
```

APPENDIX G

MPLUS 7.3 INPUT FILES FOR CONFIGURAL INVARIANCE AND PARTIAL STRICT INVARIANCE MODELS – MUML APPROACH

```
TITLE: Configural Invariance Model

DATA:
FILE = mumlinput.dat;
TYPE = means fullcov;
NGROUPS = 4;
NOBSERVATIONS = 478.304 8554.696 464.696 8312.304;

VARIABLE:
NAMES = intern extern interp learn;
USEVARIABLES = intern extern interp learn;

MODEL:
intern_b BY intern@4.34768;
extern_b BY extern@4.34768;
learn_b BY learn@4.34768;
interp_b BY interp@4.34768;

intern_w BY intern@1;
extern_w BY extern@1;
interp_w BY interp@1;
learn_w BY learn@1;

[intern@0 extern@0 interp@0 learn@0];
intern@0 extern@0 interp@0 learn@0;

bw BY intern_b@1 extern_b interp_b learn_b;
wi BY intern_w@1 extern_w interp_w learn_w;

bw WITH wi@0;

MODEL g1:
bw BY intern_b@1
     extern_b (1)
     interp_b (2)
     learn_b (3);
[bw@0]; bw (4);
[intern_b] (5); [extern_b] (6); [interp_b] (7); [learn_b] (8);
intern_b (9); extern_b (10); interp_b (11); learn_b (12);

wi BY intern_w@1
     extern_w (13)
     interp_w (14)
     learn_w (15);
[wi@0]; wi (16);
```

```
[intern_w] (17); [extern_w] (18); [interp_w] (19); [learn_w] (20);
intern_w (21); extern_w (22); interp_w (23); learn_w (24);

MODEL g2:
intern_b BY intern@0;
extern_b BY extern@0;
interp_b BY interp@0;
learn_b BY learn@0;

bw BY intern_b@0 extern_b@0 interp_b@0 learn_b@0;
[bw@0]; bw@0;
[intern_b@0 extern_b@0 interp_b@0 learn_b@0];
intern_b@0 extern_b@0 interp_b@0 learn_b@0;

wi BY intern_w@1
     extern_w (13)
     interp_w (14)
     learn_w (15);
[wi@0]; wi (16);
[intern_w] (17); [extern_w] (18); [interp_w] (19); [learn_w] (20);
intern_w (21); extern_w (22); interp_w (23); learn_w (24);


MODEL g3:
intern_b BY intern@4.34783;
extern_b BY extern@4.34783;
learn_b BY learn@4.34783;
interp_b BY interp@4.34783;

bw BY intern_b@1
     extern_b (1)
     interp_b (2)
     learn_b (3);
[bw@0]; bw (4);
[intern_b] (5); [extern_b] (6); [interp_b] (7); [learn_b] (8);
intern_b (9); extern_b (10); interp_b (11); learn_b (12);

wi BY intern_w@1
     extern_w (25)
     interp_w (26)
     learn_w (27);
[wi] (28); wi (29);
[intern_w] (17); [extern_w] (30); [interp_w] (31); [learn_w] (32);
intern_w (33); extern_w (34); interp_w (35); learn_w (36);
```

```
MODEL g4:
intern_b BY intern@0;
extern_b BY extern@0;
interp_b BY interp@0;
learn_b BY learn@0;

bw BY intern_b@0 extern_b@0 interp_b@0 learn_b@0;
[bw@0]; bw@0;
[intern_b@0 extern_b@0 interp_b@0 learn_b@0];
intern_b@0 extern_b@0 interp_b@0 learn_b@0;

wi BY intern_w@1
     extern_w (25)
     interp_w (26)
     learn_w (27);
[wi] (28); wi (29);
[intern_w] (17); [extern_w] (30); [interp_w] (31); [learn_w] (32);
intern_w (33); extern_w (34); interp_w (35); learn_w (36);

OUTPUT:
SAMPSTAT RESIDUAL MOD;
```

```
TITLE: Partial Strict Invariance Model

DATA:
FILE = mumlinput.dat;
TYPE = means fullcov;
NGROUPS = 4;
NOBSERVATIONS = 478.304 8554.696 464.696 8312.304;

VARIABLE:
NAMES = intern extern interp learn;
USEVARIABLES = intern extern interp learn;

MODEL:
intern_b BY intern@4.34768;
extern_b BY extern@4.34768;
learn_b BY learn@4.34768;
interp_b BY interp@4.34768;

intern_w BY intern@1;
extern_w BY extern@1;
interp_w BY interp@1;
learn_w BY learn@1;

[intern@0 extern@0 interp@0 learn@0];
intern@0 extern@0 interp@0 learn@0;

bw BY intern_b@1 extern_b interp_b learn_b;
wi BY intern_w@1 extern_w interp_w learn_w;

bw WITH wi@0;

MODEL g1:
bw BY intern_b@1
     extern_b (1)
     interp_b (2)
     learn_b (3);
[bw@0]; bw (4);
[intern_b] (5); [extern_b] (6); [interp_b] (7); [learn_b] (8);
intern_b (9); extern_b (10); interp_b (11); learn_b (12);

wi BY intern_w@1
     extern_w (13)
     interp_w (14)
     learn_w (15);
[wi@0]; wi (16);
```

```
[intern_w] (17); [extern_w] (18); [interp_w] (19); [learn_w] (20);
intern_w (21); extern_w (22); interp_w (23); learn_w (24);

MODEL g2:
intern_b BY intern@0;
extern_b BY extern@0;
interp_b BY interp@0;
learn_b BY learn@0;

bw BY intern_b@0 extern_b@0 interp_b@0 learn_b@0;
[bw@0]; bw@0;
[intern_b@0 extern_b@0 interp_b@0 learn_b@0];
intern_b@0 extern_b@0 interp_b@0 learn_b@0;

wi BY intern_w@1
     extern_w (13)
     interp_w (14)
     learn_w (15);
[wi@0]; wi (16);
[intern_w] (17); [extern_w] (18); [interp_w] (19); [learn_w] (20);
intern_w (21); extern_w (22); interp_w (23); learn_w (24);


MODEL g3:
intern_b BY intern@4.34783;
extern_b BY extern@4.34783;
learn_b BY learn@4.34783;
interp_b BY interp@4.34783;

bw BY intern_b@1
     extern_b (1)
     interp_b (2)
     learn_b (3);
[bw@0]; bw (4);
[intern_b] (5); [extern_b] (6); [interp_b] (7); [learn_b] (8);
intern_b (9); extern_b (10); interp_b (11); learn_b (12);

wi BY intern_w@1
     extern_w (25)
     interp_w (14)
     learn_w (15);
[wi] (28); wi (29);
[intern_w] (17); [extern_w] (30); [interp_w] (19); [learn_w] (32);
intern_w (21); extern_w (34); interp_w (23); learn_w (36);
```

```
MODEL g4:
intern_b BY intern@0;
extern_b BY extern@0;
interp_b BY interp@0;
learn_b BY learn@0;

bw BY intern_b@0 extern_b@0 interp_b@0 learn_b@0;
[bw@0]; bw@0;
[intern_b@0 extern_b@0 interp_b@0 learn_b@0];
intern_b@0 extern_b@0 interp_b@0 learn_b@0;

wi BY intern_w@1
     extern_w (25)
     interp_w (14)
     learn_w (15);
[wi] (28); wi (29);
[intern_w] (17); [extern_w] (30); [interp_w] (19); [learn_w] (32);
intern_w (21); extern_w (34); interp_w (23); learn_w (36);

OUTPUT:
SAMPSTAT RESIDUAL MOD;
```