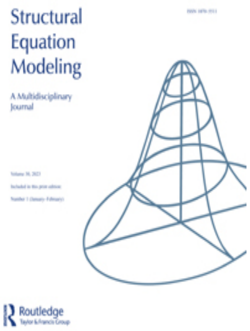

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hsem20

Utilizing Moderated Non-linear Factor Analysis Models for Integrative Data Analysis: A Tutorial

Joseph M. Kush, Katherine E. Masyn, Masoumeh Amin-Esmaeili, Ryoko Susukida, Holly C. Wilcox & Rashelle J. Musci

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2023, VOL. 30, NO. 1, 149–164
https://doi.org/10.1080/10705511.2022.2070753

Utilizing Moderated Non-linear Factor Analysis Models for Integrative Data Analysis: A Tutorial

Joseph M. Kush[a], Katherine E. Masyn[b], Masoumeh Amin-Esmaeili[a], Ryoko Susukida[a], Holly C. Wilcox[a] and Rashelle J. Musci[a]

[a]Johns Hopkins Bloomberg School of Public Health; [b]Georgia State University School of Public Health

ABSTRACT
Integrative data analysis (IDA) is an analytic tool that allows researchers to combine raw data across multiple, independent studies, providing an improved measurement of latent constructs as compared to single study analysis or meta-analyses. This is often achieved through the implementation of moderated non-linear factor analysis (MNLFA), an advanced modeling approach that allows for covariate moderation of item and factor parameters. The current paper provides an overview of this modeling technique, highlighting distinct advantages most apt for IDA. We further illustrate the complex model building process involved in MNLFA by providing a tutorial using empirical data from five separate prevention trials. The code and data used for analyses are also provided.

KEYWORDS
Factor scores; integrative data analysis; measurement invariance; moderated non-linear factor analysis

The practice of combining information across multiple independent studies, commonly referred to as integrative data analysis (IDA), is becoming increasingly popular in the social and medical sciences (see Graham et al., 2017; Gross et al., 2018, Rose et al., 2018; Sibley & Coxe, 2020). IDA may be simply defined as "fitting models to data that have been pooled across multiple studies" (Curran & Hussong, 2009, p. 82). IDA provides distinct advantages over single study analysis or even meta-analyses, including increased power through larger combined sample sizes, the ability to formulate and explore previously unachievable research questions (e.g., combining multiple cross-sectional studies measured at different timepoints to allow for longitudinal data analysis), improved measurement of latent constructs through robust psychometric instruments and obtaining increased frequencies of low base-rate behaviors. More recently, researchers have begun incorporating moderated non-linear factor analysis (MNLFA) into IDA studies as a novel modeling approach, allowing researchers to test whether a measure is invariant across multiple covariates simultaneously, such as race, sex, and study membership.

Although the current methodological development of MNLFA has been quite extensive (Bauer, 2017; Bauer & Hussong, 2009; Curran et al., 2014; Curran & Hussong, 2009), there is a growing need among applied researchers for guidance on implementing these models in practice within the larger IDA framework. There are several challenges and decisions researchers must consider when conducting these analyses, including determining which combinations of predictors should be included as moderators of the factor mean and variance, as well as considering differential item functioning among certain item intercepts and loadings. The overall aim of this paper is to provide a detailed tutorial on the process of MNLFA model building and implementation for broader IDA studies.

1. Integrative Data Analysis

Incorporating data from multiple sources has the ability to yield greater insight into scientific questions than any single study by providing researchers with more heterogeneous populations, larger sample sizes, and greater precision in parameter estimates, for example. In reality, the practice of combining and analyzing quantitative data is largely varied. One of the most popular methods, meta-analysis (see Glass, 1976), allows for researchers to analyze results, such as summary statistics or point estimates and standard errors that have been presented in different but comparable studies. As described by Cooper and Patall (2009), this traditional approach to meta-analysis involves using aggregated data, in which a researcher: (1) systematically searches and collects studies that have been conducted on the topic of interest, (2) extracts effect sizes based on reported summary statistics, and (3) combines these estimates using statistically sound techniques to obtain a single average effect size and confidence interval (Cooper, 2009). Notably, however, Cooper and Patall (2009) differentiate between meta-analysis based on aggregated data, as described above, *vs.* meta-analysis based on individual participant-level data. The latter, often called data synthesis, is referred to in this paper as integrative data analysis, which involves "the central collection, checking, and re-analysis of the raw data from each study to obtain combined results" (Cooper & Patall, 2009, p. 166). By accessing the raw data from each study, IDA

CONTACT Joseph Kush jkush1@jhu.edu Johns Hopkins Bloomberg School of Public Health, 624 N Broadway Room 841, Baltimore, MD 21205, USA.

allows researchers to replicate analyses performed in the original studies, to improve precision in effect estimates by increasing the sample size and statistical power, as well as to estimate both within-study and between-study effects, among other benefits.

When pooling data across multiple studies for IDA, a great deal of effort is required in preparing the data for statistical analyses. The first basic step of any pre-statistical data harmonization process involves collecting the datasets and obtaining codebooks from the included studies. Next, it is often useful to construct a concordance table, documenting which variables are collected across studies. This table acts as a central codebook, allowing researchers to confirm response scale types, identify common data domains, and examine variables over time if data were collected longitudinally. Finally, variables are harmonized across studies, or made more homogenous to allow for a more straightforward analytic plan. This step involves the identification of relevant domains and instruments, developing uniform variable names and labels, creating mergeable data, etc. For example, consider two studies that measured aggressive behavior in first-grade students. Both studies shared the same item stem: "This student exhibits aggressive behaviors, such as breaking rules, harming property, and teasing others." However, the two studies differed in the response options: Study A measured this item on a five-point Likert scale (0 = "Never" to 4 = "Almost always"), while Study B measured this item as a binary variable (0 = "False," 1 = "True"). In this example, a researcher conducting IDA may consider collapsing a sparsely distributed ordinal variable in Study A into a binary variable to ensure the two studies have equivalent response categories. This is often referred to as "logical harmonization" throughout the literature, in which the original items are transformed to have logically equivalent response scales (see Hussong et al., 2013). Only after data have been appropriately harmonized may researchers move on to establishing a theoretical construct on a common scale across studies. MNLFA provides researchers with an advanced modeling approach that can accommodate such data in an effort to estimate latent factors.

2. Factor Analysis and the 2-PL Model

The methodological development of MNLFA is rooted in psychometric theory, with contributions from both the linear and generalized factor analysis framework (Bollen, 1989; Muthén, 1984; Rabe-Hesketh et al., 2004; Skrondal & Rabe-Hesketh, 2004), as well as from the two-parameter logistic (2-PL) model from item response theory (IRT; Birnbaum, 1968; Bock & Aitkin, 1981; Lord & Novich, 1968). As a starting point, consider the following simple linear factor model (Jöreskog, 1967):

$$\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

in which i indexes individual observations, $\boldsymbol{y}_i$ is a p-dimensional vector of observed continuous variables, $\boldsymbol{\nu}$ is a p-dimensional vector of measurement intercepts, $\boldsymbol{\Lambda}$ is a $p \times m$ matrix of factor loadings, $\boldsymbol{\eta}_i$ is an $m \times 1$ vector of latent variables, and $\boldsymbol{\varepsilon}_i$ is a p-dimensional vector of measurement errors. It is assumed that $E(\boldsymbol{\varepsilon}_i) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\eta}_i, \ \boldsymbol{\varepsilon}_i) = \mathbf{0}$, with a model implied variance-covariance matrix given by

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{2}$$

where $V(\boldsymbol{\eta}_i) = \boldsymbol{\Phi}$, and $V(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Theta}$, a diagonal matrix. This model is identified so long as a unique solution can be obtained for each parameter. The scale of the latent factor may be set in several ways, dependent upon the central research question. For example, fixing the intercept and factor loading of the first item to 0 and 1, respectively, allows the researcher to freely estimate the factor mean and variance. Conversely, the factor mean and variance may be fixed to 0 and 1, respectively, placing the factor on a standard normal distribution.

An important concept in factor analysis is measurement invariance (also known as factorial invariance or measurement equivalence). Measurement invariance represents the degree to which observed item distributions are dependent only upon an individual's latent variable and no other characteristics and may be used for investigating correspondence in factor models across studies through traditional techniques (Bauer et al., 2020; Jöreskog, 1971; Meredith, 1993; Sörbom, 1974). This represents one of the most fundamental concerns of IDA, namely, ensuring that the hypothesized factor being measured is in fact common across studies, as well as across other covariates (e.g., sex, race, or time). Levels of measurement invariance include dimensional invariance (same number of factors across groups), metric invariance (same factor loadings across groups to allow for comparisons of factor variances/covariances), and strong factorial invariance (same item intercepts to allow for unbiased group differences in factor means), among others (Little et al., 2006). The process of testing for measurement invariance in factor analysis involves the following general procedure. First, a baseline two-group model (with constraints for dimensional and configural invariance) is established, in which a log-likelihood value is estimated using maximum likelihood for example, and stored. Then, one progressively specifies more stringent constraints, again storing the log-likelihood value. Finally, a likelihood ratio difference test is calculated for the current model in comparison to the prior model (in which the current model is nested) to determine the effect of a given level of measurement invariance. If the likelihood ratio difference test is non-significant (e.g., $p > .05$), measurement invariance is established for a given constraint. By demonstrating invariance across more stringent constraints, one can provide more evidence that the same factor structure is being measured in both groups.

We next consider the 2-PL model from the IRT literature. Although both factor analysis and the 2-PL model assume continuous latent variables, these models differ in item scale type; linear factor models require continuous items, while 2-PL models require binary items. Consider the following 2-PL model, in which we assume a single continuous latent trait (i.e., factor) underlies a set of observed binary item responses. Assuming each item follows a conditional Bernoulli distribution, the probability of observing a correct response (i.e., score of 1) for a single item y can be given as

$$P(y_i = 1|\theta_i) = \frac{e^{[\alpha(\theta_i - \delta)]}}{1 + e^{[\alpha(\theta_i - \delta)]}}, \tag{3}$$

in which i indexes individual observations, θ_i is the latent trait score, α is the discrimination parameter, and δ is the difficulty parameter. It is assumed that the item response is independent across observations and conditionally independent across the item, conditional on the latent trait. To set the scale of the latent factor, it is typical to assume $\theta_i \sim N(0,1)$, allowing the discrimination and difficulty parameters to be freely estimated. In the 2-PL model, the discrimination parameter represents the rate at which the probability of a correct response increases as the latent trait increases. In general, a highly discriminating item can better distinguish between different values of the latent trait (particularly when θ_i is near δ). The difficulty parameter may be defined as $P(y_i = 1|\theta_i = \delta) = .5$, and represents the value of the latent trait at which the probability of a correct response equals .5. Although we focus our attention on the 2-PL model for simplicity, there are numerous alternative IRT models that may be of interest to the researcher. For example, the three-parameter logistic (3-PL) model includes an additional lower asymptote parameter χ, often referred to as the guessing parameter. Likewise, the rating scale model or partial credit model may be appropriate for polytomous items, such as Likert-scale responses.

As with factor analysis, it is important to establish measurement invariance for the 2-PL model. Within the IRT framework, measurement invariance can be investigated through differential item functioning (DIF; Holland & Wainer, 1993). After controlling for the latent trait, an item without bias should perform the same for two individuals, regardless of a group membership. When the probability of a correct response for an item differs over groups with equal values of the latent trait, the item is said to exhibit DIF. Although there are a variety of methods that can be used to assess DIF in the 2-PL model, DIF in the discrimination and difficulty parameters is typically jointly tested using a likelihood ratio test (Belzak, 2020; Thissen et al., 1993). As a general strategy similar to the process described earlier for factor analysis, the first step involves fitting a baseline 2-PL model with all parameters varying between the reference and focal groups. Next, a (nested) restricted model is estimated, in which the parameters for a single item are constrained to be equal between groups. Finally, the likelihood ratio test statistic is computed, with a significant (e.g., $p < .05$) test statistic indicating DIF.

3. Factor Estimation Considerations in IDA

There are four major issues researchers face when using traditional measurement models and invariance testing techniques, all of which can be appropriately dealt with using MNLFA: (1) some items not being shared across studies, (2) continuous covariates, (3) sequential testing of covariates, and (4) items with different scale types. First, researchers conducting IDA often encounter some items not being shared across studies (i.e., an item exists in one study but does not in another), which may be considered as missing data and handled through maximum likelihood estimation (Graham, 2003; Schafer & Graham, 2002). A similar difficulty includes scenarios in which there are no items that are common to all studies being used in IDA. However, this may be addressed by linking or chaining the studies together. For example, perhaps Study A includes items x1–x5, Study B includes items x5–x10, and Study C includes items x10–x15. While there are no items common to all studies, item x5 is common to Study A and Study B, while item x10 is common to Study B and Study C. Thus, by including parameter invariance constraints on these parameters, estimated factor scores across the three studies based on a common metric may be obtained.

Regarding the second (continuous covariates) and third (sequential testing of covariates) issues highlighted above, there are two apparent limitations to traditional approaches to measurement invariance and DIF testing. Typically, when using multi-sample approaches, (a) discrete groups are (b) tested in succession. First, while discrete group testing may be a natural way to compare group membership (e.g., treatment *vs.* control, old *vs.* young, or females *vs.* males), it may be desirable to consider invariance testing across continuous covariates, such as age or IQ. Additionally, even if researchers are interested in establishing invariance across multiple different group comparisons, this process is typically conducted sequentially (e.g., invariance between treatment *vs.* control is tested, then invariance between old *vs.* young is tested, etc.). With a large number of groups, the number of tests to conduct may become unwieldy, while the cell sizes of the various strata may become small.

A final consideration of factor estimation in IDA deals with item scale type. For the two models presented, only continuous items are appropriate for linear factor analysis, while only binary items are appropriate for the 2-PL model. The limitations of these requirements become apparent for IDA of studies with potentially different instruments, different items, and different item scale types. Consider a model in which the latent factor is measured by a combination of item scale types (e.g., both continuous and binary items). Neither single model presented thus far is capable of handling such data, as is often encountered when pooling datasets. In direct response to the concerns raised here, MNLFA has been developed from the generalized factor analysis and non-linear item-level factor analysis literature, and represents a more flexible approach ideal for use in IDA.

4. Moderated Non-linear Factor Analysis

Originally proposed by Bauer and Hussong (2009), MNLFA builds upon generalized factor analysis (Muthén, 1984; Rabe-Hesketh et al., 2004; Skrondal & Rabe-Hesketh, 2004), with models that can accommodate items of different scale types (e.g., binary, ordinal, or continuous), as well as items of mixed scale types (e.g., both binary and continuous). Consider the following generalized factor model:

$$g_i(\mu_{ij}) = \upsilon_i + \lambda_i \eta_j, \tag{4}$$

in which μ_{ij} is the expected value of item i for observation j, $g_i(\cdot)$ specifies the desired link function, υ_i is the measurement intercept, λ_i is the factor loading, and η_j is the latent factor, assumed to be normally distributed as $\eta_j \sim N(\alpha, \psi)$. While the generalized factor model does not include a specific error term, measurement error is implicitly taken into account by modeling the conditional-response distribution for a given item. For example, for a set of continuous indicators, a normal conditional response distribution, $y_{ij}|\eta_j \sim N\left(\mu_{ij}, \sigma_i^2\right)$, with the identity link function, $g_i(\mu_{ij}) = \mu_{ij}$, gives

$$\mu_{ij} = \upsilon_i + \lambda_i \eta_{j.} \tag{5}$$

Notice Equation (5) represents a reparameterization of the linear factor model presented in Equation (1). Likewise, consider a set of binary indicators, in which a conditional Bernoulli response distribution, $y_{ij}|\eta_j \sim Ber(\mu_{ij})$, with the logit link function, $g_i(\mu_{ij}) = \ln\left[\mu_{ij}/(1-\mu_{ij})\right]$, gives

$$\ln\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \upsilon_i + \lambda_i \eta_{j.} \tag{6}$$

We note that generalized factor models may also be expressed in terms of the inverse link function, $g_i^{-1}(\cdot)$:

$$\mu_{ij} = g_i^{-1}\left(\upsilon_i + \lambda_i \eta_j\right). \tag{7}$$

Now, substituting the inverse logit link function (i.e., logistic function) into Equation (6), the expected value can be more naturally expressed as

$$\mu_{ij} = \frac{1}{1 + e^{\left[-\left(\upsilon_i\ +\ \lambda_i \eta_j\right)\right]}} \tag{8a}$$

$$= \frac{e^{\left(\upsilon_i\ +\ \lambda_i \eta_j\right)}}{1 + e^{\left(\upsilon_i\ +\ \lambda_i \eta_j\right)}} \tag{8b}$$

$$= \frac{e^{\left\{\lambda_i\left[\eta_j - \left(-\upsilon_i/\lambda_i\right)\right]\right\}}}{1 + e^{\left\{\lambda_i\left[\eta_j - \left(-\upsilon_i/\lambda_i\right)\right]\right\}}}. \tag{8c}$$

Comparing Equations (8a)–(8c) with Equation (3), it can be seen that this model represents a reparameterization of the 2-PL model, in which $\upsilon_i = -\alpha\delta$, $\lambda_i = \alpha$, and $\eta_j = \theta_i$.

One of the greatest benefits of the generalized factor analysis framework is the ability to model different response distributions and link functions for different items simultaneously. The flexibility of this model requires the assumption of conditional independence for the items, in which individual univariate distributions are modeled for each item, rather than assuming a multivariate distribution for the set of items. For example, one could choose a normal distribution with an identity link function for a continuous item, a Bernoulli distribution with a logit link function for a binary item, a Poisson distribution with a log link function for a count item, and a multinomial distribution with a logit link function for a nominal (e.g., unordered polytomous) item, with all parameters, estimated simultaneously.

4.1. MNLFA

One limitation of the generalized factor model is the assumption of parameter invariance across individuals. Examining Equation (4), the four parameters that define the model (α = latent factor mean, ψ = latent factor variance, υ_i = intercept for item i, and λ_i = factor loading for item i) are assumed equal between groups (e.g., treatment and control units, males and females, or between individuals from different studies). MNLFA extends the generalized factor analysis framework by allowing the four model parameters to vary as a function of covariates.

We first focus on allowing observed covariate moderation of the latent factor mean and variance. Here, the latent factor is assumed to be normally distributed as $\eta_j \sim N\left(\alpha_j, \psi_j\right)$, with parameters defined as (Bauer et al., 2020; Bauer & Hussong, 2009; Curran et al, 2014):

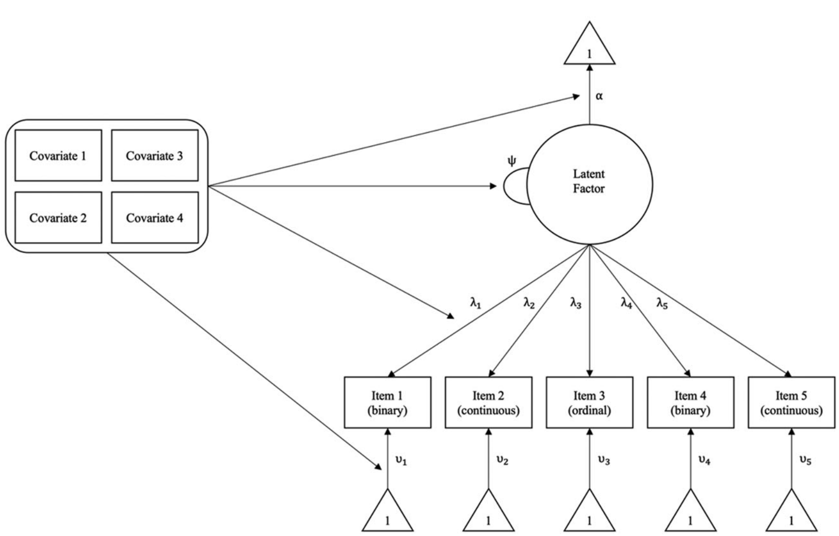

Figure 1. Example MNLFA path diagram.

$$\alpha_j = \alpha_0 + \sum_{w=1}^{W} \alpha_w x_{wj}, \tag{9}$$

$$\psi_j = \psi_0 + e^{\left(\sum_{w=1}^{W} \beta_w x_{wj}\right)}, \tag{10}$$

in which x_w denotes observed moderator x, with W total moderators, α_0 and ψ_0 are the factor mean and variance, respectively, when all moderators are equal to 0, and α_w and β_w represent the effect of the moderator on the factor mean and variance, respectively. To ensure a non-negative variance estimate, Equation (10) is modeled as a log-linear function of the moderators. Again, the scale of the latent factor is typically set by constraining the factor mean and variance to 0 and 1, respectively.

Next, we consider the observed covariate moderation of items. With the addition of j subscripts to the item intercept (υ_{ij}) and factor loading (λ_{ij}), the generalized factor model from Equation (4) is extended to

$$g_i(\mu_{ij}) = \upsilon_{ij} + \lambda_{ij}\eta_j, \tag{11}$$

allowing the intercept and loading for observation j to uniquely differ as a (linear) function of the moderators, expressed as:

$$\upsilon_{ij} = \upsilon_{0i} + \sum_{w=1}^{W} \upsilon_{wi} x_{wj}, \tag{12}$$

$$\lambda_{ij} = \lambda_{0i} + \sum_{w=1}^{W} \lambda_{wi} x_{wj}. \tag{13}$$

Now, υ_{0i} and λ_{ij} are the item intercept and factor loading for individual j when all moderators are equal to 0, respectively, and υ_{wi} and λ_{wi} represent the effect of the moderator on the item intercept and factor loading, respectively. See Figure 1 for an example MNLFA path diagram.

There are two specific considerations of the MNLFA model worth noting. First, it is possible to allow for different moderators in Equations (9)–(13). For example, one could model sex as a moderator of the factor mean but not the factor variance. Likewise, one could model study membership as a moderator of the intercept of the first item, but not of the factor loading of the first item. Moreover, different items may also have different moderators (e.g., race moderates the intercept and loading of item 1, but only the loading of item 2). Second, it is not required that the conditional-response distribution for an item belonging to the exponential family, although this is assumed for the generalized linear factor model. For example, for a non-normally distributed continuous item with heavy tails, the Student's t distribution may be used.

Overall, the MNLFA model improves upon traditional measurement invariance testing by allowing multiple covariates to moderate different item and factor parameters simultaneously. Additionally, covariates may include both categorical and continuous variables, a limitation of more traditional measurement invariance testing techniques. It is also possible to reduce certain MNLFA model specifications to more familiar models, such as linear factor models or IRT models. However, by allowing for specifications of different response distributions and link functions for different items, MNLFA offers an extremely flexible alternative modeling approach to more traditional factor models. Importantly, the ability to estimate factors based on potentially different items pooled across studies, as well as allow for the moderation of multiple covariates simultaneously on the factor and item parameters makes MNLFA extremely suitable for use within IDA.

5. Applied Example: Aggressive-Disruptive Behavior among Elementary Students

We now present an applied example using empirical data from five independent prevention trials to estimate the effect of a latent aggressive-disruptive behavior factor on

Table 1. Sample demographic characteristics.

	Sample sizes across study				
	Race		Sex		
	Black	White	Female	Male	Total
Study 1	432	385	254	563	817
Study 2	7	444	234	217	451
Study 3	1,322	562	1,016	868	1,884
Study 4	556	83	302	337	639
Study 5	144	13	86	71	157

	Item endorsement rates across study					
	Study					
	Study 1	Study 2	Study 3	Study 4	Study 5	Total
Breaks rules	.895	.525	.616	.521	.703	.661
Harms property	–	.175	.362	.202	.351	.307
Breaks things	.540	.175	.316	.152	.359	.331
Takes property	.668	.220	.417	.227	.487	.431
Fights	.816	.375	.360	.291	.487	.458
Lies	.738	.255	.438	.236	.583	.466
Yells at others	.816	.370	.490	.335	.506	.530
Stubborn	.876	.745	.620	.336	.551	.631
Teases others	.821	.575	.542	.382	.532	.578

Note. Item "harms property" was not measured in Study 1.

high school graduation. Previous research has demonstrated the Authority Acceptance subscale of the Teacher Observation of Classroom Adaptation-Revised (TOCA-R; Werthamer-Larsson et al., 1991), an instrument commonly used in schools to assess student behavior, to be related to several negative outcomes including later delinquent behavior and criminal justice involvement (Petras et al., 2004; 2005), as well as high school drop and unemployment (Bradshaw et al., 2010). Bradshaw & Kush (2020) found the TOCA-R to be a highly valid and reliable measure of aggressive and disruptive behavior among students as early as kindergarten. For the current applied example, we hypothesized baseline measures of aggressive-disruptive behaviors would be associated with later high school dropout.

To conduct integrative data analyses, data were drawn from the following five independent school-based cluster-randomized prevention trials: (1) JHU Center for Prevention and Early Intervention first-generation trial (Kellam et al., 1998), (2) JHU Center for Prevention and Early Intervention second-generation trial (Ialongo et al., 1999), (3) Schools and Families Educating Children Study (Tolan et al., 2004), (4) Fast Track Project (Conduct Problems Prevention Research Group, 2019), and (5) Linking the Interests of Families and Teachers Study (Eddy et al., 2003). All studies administered similar versions of the TOCA-R. For this example, we focus on teacher ratings of first-grade students exclusively. In addition to increased statistical power through a larger combined sample size, conducting IDA on the combined data was useful for increasing two aspects of heterogeneity. First, racial subgroup sample sizes varied dramatically across studies; for example, Black students comprised ∼13% of the sample in Study 5, but about 70% of the sample in Study 3 (see Table 1). Second, item endorsement rates also differed substantially across studies. For example, the item “Takes property” was endorsed by ∼22% of students in Study 2, yet about 67% of students in Study 1. By combining data from multiple studies, we were able to collect more robust and nuanced findings than any single study could have provided. Thus, the overall goal was to establish a theoretical aggressive-disruptive behavior construct that has been placed on a common scale across studies. Statistical programming was conducted in R version 4.1.1 (R Core Team, 2020), relying on the MplusAutomation package (Hallquist & Wiley, 2018) to facilitate conducting analyses in Mplus version 8.4 (Muthén & Muthén, 2017). All syntax and data used for analyses are freely available at: https://github.com/jmk7cj/SEM-mnlfa.

(a)

```
! note: comments are designated by exclamation points
title: CFA model for Study 4 ! title of analysis

data:
file = data_cfa.dat; ! name of datafile

variable:
names = id study_id study_1-study_5 sex race x1-x9 hs; ! names of columns in datafile
usevariables = x1-x9; ! only need to use variables x1-x9
categorical = x1-x9; ! variables x1-x9 are categorical outcome variables
useobservations = study_id == 4; ! constrain to individuals from study 4
missing = all (-999); ! missing data identifier

analysis:
estimator = wlsmv; ! weighted least squares with mean and variance adjusted fit statistics
processors = 1; ! number of cores / processors for parallel processing

model:
Factor BY x1-x9; ! latent variable 'Factor' is measured by items 1 through 9

output:
standardized; ! can view standardized output (in addition to IRT parameterization)
stdyx;
```

(b)

```
SUMMARY OF ANALYSIS

Number of groups                                 1
Number of observations                         639

Number of dependent variables                    9
Number of independent variables                  0
Number of continuous latent variables            1

Observed dependent variables

  Binary and ordered categorical (ordinal)
   X1        X2        X3        X4        X5        X6
   X7        X8        X9

Continuous latent variables
   FACTOR

Estimator                                          WLSMV
Maximum number of iterations                       1000
Convergence criterion                              0.500D-04
Maximum number of steepest descent iterations      20
Maximum number of iterations for H1                2000
Convergence criterion for H1                       0.100D-03
Parameterization                                   DELTA
Link                                               PROBIT


THE MODEL ESTIMATION TERMINATED NORMALLY

MODEL FIT INFORMATION

Number of Free Parameters                   18

Chi-Square Test of Model Fit

        Value                          92.346*
        Degrees of Freedom                 27
        P-Value                        0.0000

RMSEA (Root Mean Square Error Of Approximation)

        Estimate                             0.062
        90 Percent C.I.                      0.048  0.076
```

Figure 2. (a) Mplus input file of CFA model for Study 4. (b) Select Mplus output file of CFA model for Study 4.

```
          Probability RMSEA <= .05          0.078

     CFI/TLI

          CFI                        0.994
          TLI                        0.992

     Chi-Square Test of Model Fit for the Baseline Model

          Value                      11431.494
          Degrees of Freedom                36
          P-Value                       0.0000

     SRMR (Standardized Root Mean Square Residual)

          Value                      0.040

     Optimum Function Value for Weighted Least-Squares Estimator

          Value                  0.42026358D-01


     MODEL RESULTS

                                         Two-Tailed
                    Estimate      S.E.  Est./S.E.    P-Value

      FACTOR   BY
         X1            1.000      0.000    999.000    999.000
         X2            1.153      0.037     31.230      0.000
         X3            1.151      0.040     28.971      0.000
         X4            1.114      0.037     30.462      0.000
         X5            1.108      0.037     29.864      0.000
         X6            1.075      0.038     28.056      0.000
         X7            1.017      0.039     26.281      0.000
         X8            0.901      0.042     21.240      0.000
         X9            1.012      0.042     24.025      0.000

      Thresholds
         X1$1         -0.053      0.050     -1.068      0.285
         X2$1          0.835      0.056     14.803      0.000
         X3$1          1.029      0.060     17.032      0.000
         X4$1          0.749      0.055     13.623      0.000
         X5$1          0.550      0.052     10.499      0.000
         X6$1          0.718      0.055     13.174      0.000
         X7$1          0.426      0.051      8.320      0.000
         X8$1          0.422      0.051      8.242      0.000
         X9$1          0.301      0.050      5.965      0.000

      Variances
         FACTOR        0.688      0.041     16.717      0.000


     IRT PARAMETERIZATION

                                         Two-Tailed
                    Estimate      S.E.  Est./S.E.    P-Value

      Item Discriminations

      FACTOR   BY
         X1            1.486      0.143     10.425      0.000
         X2            3.283      0.598      5.488      0.000
         X3            3.199      0.647      4.944      0.000
         X4            2.414      0.293      8.231      0.000
         X5            2.334      0.270      8.657      0.000
         X6            1.969      0.219      8.991      0.000
         X7            1.574      0.159      9.906      0.000
         X8            1.125      0.111     10.166      0.000
         X9            1.544      0.159      9.704      0.000

      Item Difficulties
         X1$1         -0.064      0.060     -1.065      0.287
         X2$1          0.873      0.061     14.400      0.000
         X3$1          1.078      0.067     16.056      0.000
         X4$1          0.811      0.062     13.081      0.000
         X5$1          0.599      0.058     10.272      0.000
         X6$1          0.806      0.065     12.389      0.000
         X7$1          0.505      0.063      8.029      0.000
         X8$1          0.565      0.074      7.665      0.000
         X9$1          0.358      0.061      5.843      0.000

      Variances
         FACTOR        1.000      0.000      0.000      1.000


     STANDARDIZED MODEL RESULTS

     STDYX Standardization

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

Figure 2. (Continued).

6. Sample, Item Selection, and Pre-Statistical Harmonization

Data from $N = 3{,}948$ students ($n = 1{,}892$, 48% female, $n = 2{,}461$, 62% Black), with an average sample size per the study of 790 students (min = 157, max = 1,884, $SD = 659$) were analyzed. Based on prior research (Petras et al., 2004), we considered a total of 9 items measured along a 6-point Likert scale (1 = "almost never," 6 = "almost always") as comprising the aggressive-disruptive behavior subscale, including items, such as "*takes others' property*" and "*teases classmates*." In an effort to reduce sparseness in extreme responses, all items were collapsed into binary variables (see DiStefano et al., 2021). Endorsement rates of items across all studies ranged from 0.32 ("*harms property*") to 0.66 ("*breaks rules*"). See Table 1 for additional demographic characteristics.

7. Confirmatory Factor Analysis

As an initial step, confirmatory factor analyses (CFA) based on the nine binary items were conducted independently for each study. Parameters were estimated using a probit link function with weighted least squares with mean and variance adjusted chi-square fit statistics (Muthén, 1984). As an example, Figure 2a provides an annotated Mplus input file of the CFA model for Study 4. Here, the *useobservations* = -*study_id* == *4;* option of the *variable:* command is used to restrict the analyses to those from Study 4, while the

FACTOR BY				
X1	0.830	0.025	33.434	0.000
X2	0.957	0.015	64.619	0.000
X3	0.954	0.017	55.551	0.000
X4	0.924	0.016	56.190	0.000
X5	0.919	0.016	55.836	0.000
X6	0.892	0.020	43.848	0.000
X7	0.844	0.025	34.448	0.000
X8	0.747	0.032	23.036	0.000
X9	0.839	0.026	32.827	0.000
Thresholds				
X1$1	-0.053	0.050	-1.068	0.285
X2$1	0.835	0.056	14.803	0.000
X3$1	1.029	0.060	17.032	0.000
X4$1	0.749	0.055	13.623	0.000
X5$1	0.550	0.052	10.499	0.000
X6$1	0.718	0.055	13.174	0.000
X7$1	0.426	0.051	8.320	0.000
X8$1	0.422	0.051	8.242	0.000
X9$1	0.301	0.050	5.965	0.000
Variances				
FACTOR	1.000	0.000	999.000	999.000

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value	Residual Variance
X1	0.688	0.041	16.717	0.000	0.312
X2	0.915	0.028	32.310	0.000	0.085
X3	0.911	0.033	27.776	0.000	0.089
X4	0.854	0.030	28.095	0.000	0.146
X5	0.845	0.030	27.918	0.000	0.155
X6	0.795	0.036	21.924	0.000	0.205
X7	0.712	0.041	17.224	0.000	0.288
X8	0.559	0.049	11.518	0.000	0.441
X9	0.704	0.043	16.413	0.000	0.296

Figure 2. (Continued).

measurement model is defined with the line *Factor BY x1-x9;* in the *model:* command. Figure 2b provides the selected Mplus output of the same model. Note that in addition to probit parameterization, results are parameterized as IRT (i.e., item difficulty and discrimination) and standardized. Results of the Study 4 CFA model demonstrate that the model fit the data relatively well, with factor loadings ranging from .75 to .96 (RMSEA = .062, CFI = .994, TLI = .992, SRMR = .040). A final CFA was fit to all (pooled) observations simultaneously (by removing the *useobservations* option), with factor loadings ranging from .82 to .91 (RMSEA = .049, CFI = .996, TLI = .995, SRMR = .027). Overall, the CFA results demonstrate that a single factor model is an adequate model when fit to each student independently, as well as when fit to the pooled data. Having established our model, we move on to building a model for MNLFA.

8. MNLFA Model Building

After establishing a measurement model, it is important to follow a methodical approach to model building for MNLFA. Guided by theory and prior findings, we focus on three moderators of aggressive-disruptive behaviors: sex, race, and study membership. Here, sex was a binary variable coded (0 = female, 1 = male), while the race was a binary variable coded (0 = Black, 1 = White). Multiple dummy coded indicator variables were created for Study 2 through Study 5 such that estimates were in reference to Study 1. This resulted in a total of six moderators. In what follows, we present an item-by-item testing approach for developing a final MNLFA model, as recommended by Curran et al. (2014) and Gottfredson et al. (2019) (see also Finch, 2005; Thissen, 2001). However, we note the possibility of different approaches and procedures that may be used to establish invariance resulting in equivalent measurement models (e.g., Vandenberg & Lance, 2000).

8.1. Baseline MNLFA Model

As a first step, we estimate the covariates' effects on the latent mean and latent variance, recording the log-likelihood value, as well as the estimated coefficients and associated *p*-values for each covariate effect. This represents a baseline model, to which future models are compared. Figure 3a provides an annotated Mplus input file of the baseline MNLFA model. The statement *Factor ON study_2 – study_5 sex race;* in the *model:* command is used to allow the covariates to moderate the factor mean. This command corresponds to Equation (9). Allowing the covariates to moderate the factor variance requires additional steps. To avoid negative variance estimates, a log-linear constraint will be used with the covariates. This is first implemented using the *constraint = study_2 – study_5 sex race;* option in the *variable:* command. Then, in the *model:* command, a label is referred to in parentheses following the variance estimate of the factor: *Factor (factor_variance)*. Next, in the *model constraint:* command, new labels are given referencing the parameters of the moderators: *new (f_study_2 f_study_3 f_study_4 f_study_5 f_sex f_race)*. Finally, the factor variance moderation is implemented using the following command: *factor_variance = EXP(f_study_2*study_2 + f_study_3*study_3 + f_study_4*study_4 + f_study_5*study_5 + f_sex*sex + f_race*race)*. This command corresponds to Equation (10). Similarly, Figure 3b provides select Mplus output of the baseline MNLFA model. Examining the model results, the covariate moderation of the factor mean is given in the *FACTOR ON* subsection, while the covariate moderation of the factor variance is given in the *New/Additional Parameters* subsection. Here it can be seen that all six covariates were significant moderators of the factor mean, while all covariates except for *study_2* were significant moderators of the factor variance.

8.2. Item Specific MNLFA Models

Next, leaving each of the covariate effects on the latent mean and variance (regardless of significance), we explore item moderation by allowing the covariates to additionally moderate the item intercept and factor loading of the *first*

(a)

```
title: Baseline MNLFA model

data:
file = data_mnlfa.dat;

variable:
names = id study_id study_1-study_5 sex race x1-x9 hs;
usevariables =study_2 - study_5 sex race x1-x9;
categorical = x1-x9;
missing = all (-999);
constraint = study_2 - study_5 sex race; ! to be used as moderators of factor variance

analysis:
! estimator = wlsmv; cannot be used with certain model constraints
estimator = mlr; ! use maximum likelihood with robust standard errors instead
link = logit;
processors = 1;

model:
Factor BY x1-x9; ! measurement model

! allow covariates to moderate factor mean (linear function)
Factor ON study_2 - study_5 sex race;
[Factor@0]; ! constraint factor mean to zero to identify model

! factor variance implicitly set to one to identify model
Factor (factor_variance); ! estimate factor variance and define new label

model constraint:
new (f_study_2 f_study_3 f_study_4 f_study_5 f_sex f_race); ! label parameters of moderators

! allow covariates to moderate factor variance
! use log-linear function to avoid negative variance
factor_variance = EXP(f_study_2*study_2 + f_study_3*study_3 +
f_study_4*study_4 + f_study_5*study_5 + f_sex*sex + f_race*race);

output:
sampstat;
svalues;
tech1;
```

(b)

```
MODEL RESULTS

                                                Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 FACTOR   BY
    X1                 1.000      0.000    999.000    999.000
    X2                 1.182      0.080     14.836      0.000
    X3                 1.922      0.154     12.513      0.000
    X4                 1.816      0.125     14.496      0.000
    X5                 1.583      0.101     15.702      0.000
    X6                 1.673      0.108     15.480      0.000
    X7                 1.462      0.095     15.396      0.000
    X8                 1.171      0.075     15.570      0.000
    X9                 1.404      0.087     16.176      0.000

 FACTOR   ON
    STUDY_2           -1.069      0.165     -6.489      0.000
    STUDY_3           -1.268      0.103    -12.292      0.000
    STUDY_4           -2.605      0.164    -15.846      0.000
    STUDY_5           -1.210      0.224     -5.400      0.000
    SEX                0.658      0.078      8.443      0.000
    RACE              -0.842      0.094     -8.979      0.000

 Intercepts
    FACTOR             0.000      0.000    999.000    999.000

 Thresholds
    X1$1              -2.166      0.120    -18.006      0.000
    X2$1              -0.075      0.112     -0.664      0.507
    X3$1              -0.120      0.189     -0.632      0.527
    X4$1              -1.016      0.173     -5.883      0.000
    X5$1              -1.088      0.153     -7.104      0.000
    X6$1              -1.280      0.162     -7.884      0.000
    X7$1              -1.709      0.154    -11.101      0.000
    X8$1              -2.218      0.139    -15.907      0.000
    X9$1              -2.108      0.152    -13.882      0.000

 Residual Variances
    FACTOR           999.000      0.000    999.000    999.000

New/Additional Parameters
    F_STUDY_2          0.185      0.180      1.028      0.304
    F_STUDY_3          1.002      0.101      9.889      0.000
    F_STUDY_4          1.194      0.118     10.104      0.000
    F_STUDY_5          1.444      0.210      6.872      0.000
    F_SEX              0.475      0.080      5.925      0.000
    F_RACE             0.535      0.100      5.347      0.000
```

Figure 3. (a) Mplus input file of baseline MNLFA model. (b) Select Mplus output file of baseline MNLFA model.

item only (i.e., "*breaks rules*"). Again, model estimates and *p*-values are recorded. Figure 4a provides an annotated Mplus input file of the first item MNLFA model. In addition to the moderation of factor parameters, moderation of the item intercept is implemented using the statement *x1 ON study_2 – study_5 sex race;* in the *model:* command. This command corresponds to Equation (12). Allowing the covariates to moderate the item loading requires model constraints. First, a label is referred to in parentheses following the loading estimate of the first item: *Factor BY x1 (x1_loading);* in the *model:* command. Next, in the *model constraint:* command, new labels are given referencing the parameters of the item moderators: *new (x1_int x1_study_2 x1_study_3 x1_study_4 x1_study_5 x1_sex x1_race)*. Finally, moderation of the first item loading is implemented using the following command: *x1_loading = x1_int + x1_study_2*study_2 + x1_study_3*study_3 + x1_study_4*study_4 + x1_study_5*study_5 + x1_sex*sex + x1_race*race*. This command corresponds to Equation (13). Similarly, Figure 4b provides select Mplus output of the first item MNLFA model. Examining the model results, the covariate moderation of the item intercept is given in the *X1 ON* subsection, while the covariate moderation of the item loading is given in the *New/Additional Parameters* subsection. Here it can be seen that no covariates were significant moderators of the item intercept, while *study_4* and *sex* were significant moderators of the item loading.

Now, a likelihood ratio test (LRT) is conducted, comparing the change in model fit between the current model (*breaks rules* item and factor moderation), and the baseline model (factor moderation only). For example, the baseline model has 29 parameters, a log-likelihood of −13,603.1, and a scaling factor of 1.050; the *breaks rules* model has 42 parameters, a log-likelihood of −13,467.8, and a scaling factor of 1.022. Thus, the chi-square difference test based on log-likelihood values with a scaling factor obtained from maximum likelihood estimates with robust standard errors

(a)

```
title: Item 1 MNLFA model

data:
file = data_mnlfa.dat;

variable:
names = id study_id study_1-study_5 sex race x1-x9 hs;
usevariables =study_2 - study_5 sex race x1-x9;
categorical = x1-x9;
missing = all (-999);
constraint = study_2 – study_5 sex race;

analysis:
estimator = mlr;
link = logit;
processors = 1;

model:
Factor BY x1-x9;

! allow covariates to moderate factor mean (linear function)
Factor ON study_2 - study_5 sex race;
[Factor@0];

Factor (factor_variance);

! allow covariates to moderate item 1 intercept
x1 ON study_2 - study_5 sex race;
Factor BY x1 (x1_loading); ! label used for moderation of item 1 factor loading

model constraint:
new (f_study_2 f_study_3 f_study_4 f_study_5 f_sex f_race);
new (x1_int x1_study_2 x1_study_3 x1_study_4 x1_study_5 x1_sex x1_race);

! allow covariates to moderate factor variance
! use log-linear function to avoid negative variance
factor_variance = EXP(f_study_2*study_2 + f_study_3*study_3 +
f_study_4*study_4 + f_study_5*study_5 + f_sex*sex + f_race*race);

! allow covariates to moderate factor loading of item 1
x1_loading = x1_int + x1_study_2*study_2 + x1_study_3*study_3 +
x1_study_4*study_4 + x1_study_5*study_5 + x1_sex*sex + x1_race*race;
```

(b)

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

 FACTOR   BY
    X1               999.000     0.000    999.000    999.000
    X2                 1.917     0.131     14.630      0.000
    X3                 3.282     0.265     12.372      0.000
    X4                 3.000     0.218     13.755      0.000
    X5                 2.596     0.174     14.940      0.000
    X6                 2.772     0.187     14.860      0.000
    X7                 2.426     0.165     14.711      0.000
    X8                 1.997     0.135     14.793      0.000
    X9                 2.388     0.165     14.497      0.000

 FACTOR    ON
    STUDY_2           -0.654     0.092     -7.130      0.000
    STUDY_3           -0.848     0.065    -13.092      0.000
    STUDY_4           -1.701     0.113    -15.019      0.000
    STUDY_5           -0.872     0.138     -6.314      0.000
    SEX                0.403     0.047      8.556      0.000
    RACE              -0.576     0.058     -9.931      0.000

 X1       ON
    STUDY_2           -0.899     1.030     -0.873      0.383
    STUDY_3           -0.703     0.507     -1.387      0.166
    STUDY_4           -0.794     0.579     -1.370      0.171
    STUDY_5           -0.714     0.908     -0.787      0.431
    SEX               -0.282     0.312     -0.905      0.365
    RACE               0.668     0.428      1.559      0.119

 Intercepts
    FACTOR             0.000     0.000    999.000    999.000

 Thresholds
    X1$1              -4.098     0.565     -7.251      0.000
    X2$1              -0.278     0.122     -2.275      0.023
    X3$1              -0.567     0.218     -2.597      0.009
    X4$1              -1.380     0.198     -6.962      0.000
    X5$1              -1.389     0.171     -8.143      0.000
    X6$1              -1.617     0.184     -8.804      0.000
    X7$1              -1.998     0.174    -11.449      0.000
    X8$1              -2.491     0.160    -15.559      0.000
    X9$1              -2.432     0.178    -13.635      0.000

 Residual Variances
    FACTOR           999.000     0.000    999.000    999.000

 New/Additional Parameters
    F_STUDY_2         -0.470     0.188     -2.503      0.012
    F_STUDY_3          0.073     0.109      0.666      0.505
    F_STUDY_4          0.429     0.136      3.147      0.002
    F_STUDY_5          0.581     0.223      2.603      0.009
    F_SEX              0.180     0.082      2.187      0.029
    F_RACE             0.124     0.091      1.363      0.173
    X_INT              3.505     0.484      7.248      0.000
    X_STUDY_2         -0.106     0.892     -0.119      0.905
    X_STUDY_3         -0.437     0.462     -0.947      0.344
    X_STUDY_4         -1.241     0.486     -2.555      0.011
    X_STUDY_5         -1.132     0.679     -1.667      0.095
    X_SEX             -0.613     0.239     -2.569      0.010
    X_RACE             0.215     0.330      0.651      0.515
```

Figure 4. (a) Mplus input file of item 1 "breaks rules" MNLFA model. (b) Select Mplus output file of item 1 "breaks rules" MNLFA model.

can be computed as:

$$LRT = \frac{-2 \times (-13,603.1 - -13,467.8)}{\left[(29 \times 1.050) - (42 \times 1.022)\right] \div (29 - 42)} = 282.01 \tag{14}$$

As a chi-square distribution of 282.01 with $42 - 29 = 13$ degrees of freedom results in a p-value <.05, we conclude the *breaks rules* model fits the data significantly better than the baseline model and thus will include any item-specific covariate effects with significant p-values in future models. Next, the covariate effects for the first item are removed but are now added for the *second item only* (i.e., "*harms property*"). Again, with model estimates and p-values recorded, an LRT is conducted comparing the current model (*harms property* item and factor moderation) to the baseline model (factor moderation only). This process is continued for each item model individually. All models used maximum

Table 2. Examining DIF in factor and item parameters using a sequential model building approach.

					Factor mean		Factor variance		Item intercept		Item loading	
Model	Covariate	Parameters	LL	SF	Est	*p*	Est	*p*	Est	*p*	Est	*p*
Baseline		29	−13,603.1	1.050								
	Study 2				−1.07	<.01	0.19	.30				
	Study 3				−1.27	<.01	1.00	<.01				
	Study 4				−2.61	<.01	1.19	<.01				
	Study 5				−1.21	<.01	1.44	<.01				
	Sex				0.66	<.01	0.48	<.01				
	Race				−0.84	<.01	0.54	<.01				
Breaks rules		42	−13,467.8	1.022								
	Study 2				−0.65	<.01	−0.47	.01	−0.90	.38	−0.11	.91
	Study 3				−0.85	<.01	0.07	.51	−0.70	.17	−0.44	.34
	Study 4				−1.70	<.01	0.43	<.01	−0.79	.17	**−1.24**	.01
	Study 5				−0.87	<.01	0.58	.01	−0.71	.43	−1.13	.10
	Sex				0.40	<.01	0.18	.03	−0.28	.37	**−0.61**	.01
	Race				−0.58	<.01	0.12	.17	0.67	.12	0.22	.52
Harms property		41	−13,517.2	1.046								
	Study 2				−1.11	<.01	0.15	.40	**1.30**	.03	1.43	.36
	Study 3				−1.37	<.01	1.05	<.01	**1.13**	<.01	**−1.48**	<.01
	Study 4				−2.65	<.01	1.14	<.01	**1.72**	<.01	−0.37	.39
	Study 5				−1.26	<.01	1.40	<.01	**0.93**	.01	−0.20	.76
	Sex				0.65	<.01	0.48	<.01	0.04	.76	−0.20	.08
	Race				−0.84	<.01	0.54	<.01	0.08	.67	0.14	.29
Breaks things		41	−13,594.4	0.980								
	Study 2				−1.05	<.01	0.23	.21	0.55	.23	−0.03	.97
	Study 3				−1.28	<.01	1.00	<.01	**1.25**	<.01	0.56	.19
	Study 4				−2.60	<.01	1.22	<.01	**0.72**	.02	0.74	.13
	Study 5				−1.20	<.01	1.47	<.01	**0.96**	.01	**0.50**	<.01
	Sex				0.65	<.01	0.48	<.01	0.16	.35	−0.31	.21
	Race				−0.84	<.01	0.52	<.01	0.23	.44	0.81	.05
Takes property		41	−13,577.4	1.039								
	Study 2				−1.09	<.01	0.21	.26	−0.23	.58	−0.18	.68
	Study 3				−1.35	<.01	0.97	<.01	**0.79**	<.01	−0.16	.58
	Study 4				−2.66	<.01	1.18	<.01	0.42	.17	−0.27	.41
	Study 5				−1.29	<.01	1.39	<.01	**1.38**	.03	0.48	.55
	Sex				0.67	<.01	0.50	<.01	−0.30	.10	−0.23	.23
	Race				−0.83	<.01	0.56	<.01	−0.23	.29	−0.11	.64
Fights		41	−13,520.6	1.032								
	Study 2				−1.04	<.01	0.15	.43	1.24	.22	0.83	.30
	Study 3				−1.16	<.01	1.08	<.01	**−1.59**	<.01	**−0.62**	.01
	Study 4				−2.56	<.01	1.17	<.01	0.25	.54	0.07	.84
	Study 5				−1.15	<.01	1.46	<.01	−0.69	.13	−0.40	.33
	Sex				0.62	<.01	0.48	<.01	**0.41**	.01	−0.23	.10
	Race				−0.83	<.01	0.50	<.01	0.10	.63	0.11	.54
Lies		41	−13,571.0	1.025								
	Study 2				−1.03	<.01	0.25	.17	**−0.99**	.01	−0.55	.21
	Study 3				−1.30	<.01	1.05	<.01	−0.32	.15	**−0.96**	<.01
	Study 4				−2.63	<.01	1.24	<.01	−0.59	.05	**−0.89**	.01
	Study 5				−1.33	<.01	1.55	<.01	0.41	.40	**−1.31**	<.01
	Sex				0.69	<.01	0.48	<.01	**−0.40**	.03	−0.21	.16
	Race				−0.87	<.01	0.54	<.01	−0.11	.57	**−0.41**	.01
Yells at others		41	−13,567.4	1.033								
	Study 2				−0.99	<.01	0.08	.67	2.35	.17	2.59	.10
	Study 3				−1.23	<.01	1.02	<.01	**−0.97**	<.01	**−0.65**	<.01
	Study 4				−2.62	<.01	1.23	<.01	**−0.75**	.03	**−0.73**	<.01
	Study 5				−1.14	<.01	1.44	<.01	**−1.17**	.01	−0.44	.26
	Sex				0.70	<.01	0.47	<.01	**−0.77**	<.01	−0.17	.19
	Race				−0.84	<.01	0.54	<.01	−0.31	.14	**−0.29**	.04
Stubborn		41	−13,507.2	1.037								
	Study 2				−1.27	<.01	0.49	.01	1.06	.23	−0.44	.24
	Study 3				−1.26	<.01	1.05	<.01	**−0.90**	.01	**−0.64**	.01
	Study 4				−2.53	<.01	1.22	<.01	**−2.25**	<.01	**−0.97**	<.01
	Study 5				−1.13	<.01	1.51	<.01	**−2.22**	<.01	**−1.04**	<.01
	Sex				0.71	<.01	0.44	<.01	**−0.77**	<.01	−0.11	.28
	Race				−0.86	<.01	0.56	<.01	0.01	.97	−0.16	.18
Teases others		41	−13,576.5	1.036								
	Study 2				−1.20	<.01	0.32	.08	0.69	.27	−0.40	.23
	Study 3				−1.27	<.01	1.02	<.01	**−0.60**	.02	**−0.47**	.02
	Study 4				−2.65	<.01	1.24	<.01	−0.63	.09	**−0.61**	.01
	Study 5				−1.16	<.01	1.45	<.01	**−1.43**	.00	**−0.67**	.03
	Sex				0.66	<.01	0.51	<.01	−0.36	.09	**−0.35**	.01
	Race				−0.82	<.01	0.51	<.01	**−0.58**	.01	−0.18	.19

Note. LL: log-likelihood; SF: scaling factor used in likelihood ratio test; significant item moderation effects are bolded.

likelihood estimation with robust standard errors (i.e., sandwich estimator) and chi-square test statistic (Muthén & Muthén, 2017). See Table 2 for the number of parameters, log-likelihood, and scaling factor for each model, as well as parameter estimates and p-values for the factor mean, factor variance, item intercept, and factor loading for each model.

If any item-moderation model results in a non-significant LRT, no covariate effects are estimated for that item intercept or factor loading, even if the p-values for a given effect are significant. If an item-moderation model does result in a significant LRT, as is the case for all item models examined in Table 2, only item-specific covariate effects with significant p-values are kept. For example, examining Table 2, it can be seen that for the item *lies*, the covariate *sex* significantly moderates the item intercept but not the item loading, while the covariate *race* significantly moderates the item loading but not the item intercept. As a result, for the item *lies*, the moderated effect of item loading on *sex* is removed, while the moderated effect of item intercepts on *race* is removed.

8.3. Final MNLFA Models

After removing any covariate effects for either of the two reasons described above, a new, penultimate model is estimated, in which all covariates moderate the factor mean and variance, while only item-specific covariate effects with significant p-values are kept (as described above). An annotated Mplus input file for the penultimate model is presented in Figure 5a. Again, only item-specific moderation effects presented in Table 2 are kept for this next-to-last model. Likewise, Figure 5b provides the select Mplus output of the next-to-last MNLFA model. Examining the model results, the covariate moderation of certain item intercepts is presented in the *X2 ON*, *X3 ON*, etc. subsections. For example, for item 2 "Harms property," it can be seen that *study_3* and *study_4* were significant moderators of the item intercept. Likewise, covariate moderation of certain item loadings is presented in the *New/Additional Parameters* subsection. For example, for item 6 "Lies," it can be seen that only *study_5* and *race* were significant moderators of the loading.

After examining the output from the next-to-last MNLFA model, all non-significant item moderation effects are discarded, leaving only significant moderators of item intercepts and loadings (in addition to always keeping moderation of factor parameters regardless of significance). This last pruning effort results in the final MNLFA model. Again, all models used maximum likelihood estimation with robust standard errors, with factor scores estimated using the expected *a posteriori* (EAP) method (i.e., mean of the posterior distribution) and saved for each individual. See Tables 3 and 4 for parameter estimates from the final MNLFA model.

(a)

```
title: Penultimate MNLFA model

data:
file = data_mnlfa.dat;

variable:
names = id study_id study_1-study_5 sex race x1-x9 hs;
usevariables =study_2 - study_5 sex race x1-x9;
categorical = x1-x9;
missing = all (-999);
constraint = study_2 - study_5 sex race;

analysis:
estimator = mlr;
link = logit;
processors = 1;

model:
Factor BY x1*1 (x1_loading); ! label for moderation of item 1 factor loading
Factor BY x2*1 (x2_loading); ! label for moderation of item 2 factor loading
Factor BY x3*1 (x3_loading); ! label for moderation of item 3 factor loading
Factor BY x4*1; ! no label needed, as no sig. moderation of item 4 factor loading
Factor BY x5*1 (x5_loading); ! label for moderation of item 5 factor loading
Factor BY x6*1 (x6_loading); ! label for moderation of item 6 factor loading
Factor BY x7*1 (x7_loading); ! label for moderation of item 7 factor loading
Factor BY x8*1 (x8_loading); ! label for moderation of item 8 factor loading
Factor BY x9*1 (x9_loading); ! label for moderation of item 9 factor loading

! allow covariates to moderate factor mean (linear function)
Factor ON study_2 - study_5 sex race;
[Factor@0];

Factor (factor_variance);

! moderation of item intercepts (previously determined from Table 2)
! no moderation of item x1 intercept
x2 ON study_2-study_5;
x3 ON study_3-study_5;
x4 ON study_3 study_5;
x5 ON study_3 sex;
x6 ON study_2 sex;
x7 ON study_3-study_5 sex;
x8 ON study_3-study_5 sex;
x9 ON study_3 study_5 race;

model constraint:
new (f_study_2 f_study_3 f_study_4 f_study_5 f_sex f_race);

! intercepts for loading moderation equation
new (int1 int2 int3 int5 int6 int7 int8 int9); ! no sig. moderators of x4 loading

new (x1_study_4 x1_sex);
new (x2_study_3);
new (x3_study_5);
! no loading moderation of x4
new (x5_study_3);
new (x6_study_3 x6_study_4 x6_study_5 x6_race);
new (x7_study_3 x7_study_4 x7_race);
new (x8_study_3 x8_study_4 x8_study_5);
new (x9_study_3 x9_study_4 x9_study_5 x9_sex);

! allow covariates to moderate factor variance - use log-linear function to avoid negative variance
factor_variance = EXP(f_study_2*study_2 + f_study_3*study_3 +
f_study_4*study_4 + f_study_5*study_5 + f_sex*sex + f_race*race);

! allow covariates to moderate factor loadings
x1_loading = int1 + x1_study_4*study_4 + x1_sex*sex;
x2_loading = int2 + x2_study_3*study_3;
x3_loading = int3 + x3_study_5*study_5;
! no loading moderation of x4
x5_loading = int5 + x5_study_3*study_3;
x6_loading = int6 + x6_study_3*study_3 + x6_study_4*study_4 +
x6_study_5*study_5 + x6_race*race;
x7_loading = int7 + x7_study_3*study_3 + x7_study_4*study_4 +
x7_race*race;
x8_loading = int8 + x8_study_3*study_3 + x8_study_4*study_4 +
x8_study_5*study_5;
x9_loading = int9 + x9_study_3*study_3 +
x9_study_4*study_4 + x9_study_5*study_5 + x9_sex*sex;
```

Figure 5. (a) Mplus input file of penultimate MNLFA model. (b) Select Mplus output file of penultimate MNLFA model.

(b)

MODEL RESULTS

	Estimate	S.E.	Two-Tailed Est./S.E.	P-Value
FACTOR BY				
X1	999.000	0.000	999.000	999.000
X2	999.000	0.000	999.000	999.000
X3	999.000	0.000	999.000	999.000
X4	3.284	0.246	13.374	0.000
X5	999.000	0.000	999.000	999.000
X6	999.000	0.000	999.000	999.000
X7	999.000	0.000	999.000	999.000
X8	999.000	0.000	999.000	999.000
X9	999.000	0.000	999.000	999.000
FACTOR ON				
STUDY_2	-0.653	0.096	-6.768	0.000
STUDY_3	-0.828	0.070	-11.793	0.000
STUDY_4	-1.485	0.110	-13.510	0.000
STUDY_5	-0.722	0.130	-5.549	0.000
SEX	0.436	0.049	8.940	0.000
RACE	-0.519	0.053	-9.708	0.000
X2 ON				
STUDY_2	0.513	0.336	1.527	0.127
STUDY_3	1.136	0.192	5.904	0.000
STUDY_4	1.471	0.259	5.687	0.000
STUDY_5	0.810	0.438	1.847	0.065
X3 ON				
STUDY_3	0.752	0.376	2.002	0.045
STUDY_4	0.054	0.402	0.136	0.892
STUDY_5	0.302	0.528	0.572	0.567
X4 ON				
STUDY_3	0.808	0.179	4.526	0.000
STUDY_5	0.678	0.335	2.025	0.043
X5 ON				
STUDY_3	-1.195	0.221	-5.399	0.000
SEX	0.451	0.131	3.434	0.001
X6 ON				
STUDY_2	-0.818	0.288	-2.844	0.004
SEX	-0.260	0.133	-1.960	0.050
X7 ON				
STUDY_3	-0.613	0.265	-2.312	0.021
STUDY_4	-0.649	0.346	-1.876	0.061
STUDY_5	-0.795	0.363	-2.189	0.029
SEX	-0.561	0.130	-4.310	0.000
X8 ON				
STUDY_3	-0.489	0.293	-1.668	0.095
STUDY_4	-2.055	0.306	-6.721	0.000
STUDY_5	-1.965	0.413	-4.762	0.000
SEX	-0.681	0.128	-5.302	0.000
X9 ON				
STUDY_3	0.062	0.258	0.240	0.811
STUDY_5	-0.984	0.468	-2.102	0.036
RACE	-0.283	0.132	-2.146	0.032
Intercepts				
FACTOR	0.000	0.000	999.000	999.000
Thresholds				
X1$1	-3.359	0.229	-14.672	0.000
X2$1	0.636	0.218	2.916	0.004
X3$1	0.119	0.402	0.296	0.767
X4$1	-0.841	0.207	-4.072	0.000
X5$1	-1.746	0.248	-7.043	0.000
X6$1	-1.704	0.214	-7.955	0.000
X7$1	-2.716	0.271	-10.018	0.000
X8$1	-3.558	0.265	-13.419	0.000
X9$1	-2.440	0.202	-12.058	0.000
Residual Variances				
FACTOR	999.000	0.000	999.000	999.000
New/Additional Parameters				
F_STUDY_2	-0.270	0.179	-1.512	0.131
F_STUDY_3	0.022	0.129	0.170	0.865
F_STUDY_4	0.267	0.169	1.584	0.113
F_STUDY_5	0.336	0.243	1.381	0.167
F_SEX	0.066	0.088	0.749	0.454
F_RACE	0.100	0.093	1.076	0.282
INT1	3.060	0.227	13.481	0.000
INT2	3.458	0.326	10.603	0.000
INT3	3.477	0.304	11.429	0.000
INT5	3.261	0.291	11.215	0.000
INT6	3.472	0.344	10.094	0.000
INT7	3.181	0.321	9.925	0.000
INT8	2.231	0.239	9.349	0.000
INT9	2.214	0.209	10.575	0.000
X1_STUDY_4	-0.519	0.179	-2.897	0.004
X1_SEX	-0.300	0.124	-2.408	0.016
X2_STUDY_3	-1.708	0.311	-5.493	0.000
X3_STUDY_5	-0.144	0.794	-0.182	0.856
X5_STUDY_3	-0.682	0.301	-2.268	0.023
X6_STUDY_3	-0.489	0.276	-1.773	0.076
X6_STUDY_4	0.087	0.372	0.234	0.815
X6_STUDY_5	-1.265	0.358	-3.532	0.000
X6_RACE	-0.408	0.177	-2.300	0.021
X7_STUDY_3	-0.667	0.322	-2.070	0.038
X7_STUDY_4	-0.777	0.410	-1.895	0.058
X7_RACE	0.053	0.153	0.345	0.730
X8_STUDY_3	0.137	0.269	0.511	0.610
X8_STUDY_4	-0.520	0.289	-1.798	0.072
X8_STUDY_5	-0.469	0.408	-1.148	0.251
X9_STUDY_3	0.501	0.262	1.914	0.056
X9_STUDY _4	0.376	0.218	1.726	0.084
X9_STUDY_5	0.061	0.533	0.115	0.909
X9_SEX	-0.036	0.131	-0.274	0.784

Figure 5. (Continued).

9. Incorporating Factor Scores into Subsequent Model Estimation

By allowing for covariate moderation of both item and factor parameters, the final MNLFA model provides estimates of a construct that has been scaled commensurately across studies. While it is possible to estimate the measurement model of the MNLFA within a more complex structural equation model, for example, this may not be feasible in practice due to the complexity of the model. As suggested by Bauer & Hussong (2009) and Curran et al. (2014), we used estimated factor scores for each individual as a predictor of high school graduation (0 = did not graduate, 1 = did graduate). We also included sex, race, and dummy indicators of study membership as predictors in the outcome model. This is informed by simulation findings by Curran et al. (2016, 2018), who found that including covariates from the MNLFA model as predictors in a subsequent outcome model resulted in little to no bias, but importantly, failure to include covariates that are correlated with the latent factor in either model resulted in substantial bias. We wish to emphasize that the outcome model presented here is provided as an illustration and that in practice, researchers should include covariates informed by existing literature and substantive knowledge. Results from the logistic regression model (see Table 5) indicated the main effect of baseline aggressive-disruptive behavior was significantly, negatively related to high school graduation (Odds Ratio = 0.77, $p < .001$). Sex was a significant covariate (OR = 0.76, $p = .002$), with males being less likely to complete high school than females. There were also significant differences across studies; students in Study 3 were more likely to graduate high school than those in Study 1 (OR = 3.40, $p < .001$), while students in Study 4 were less likely (OR = 0.05, $p < .001$).

Table 3. Final MNLFA model examining covariate effects on factor mean and variance.

Covariate effect	Estimate	*SE*	*p*
Factor mean			
Study 2	−0.62	0.09	<.01
Study 3	−0.87	0.07	<.01
Study 4	−1.51	0.10	<.01
Study 5	−0.67	0.13	<.01
Sex	0.44	0.05	<.01
Race	−0.54	0.06	<.01
Factor variance			
Study 2	−0.25	0.17	.15
Study 3	0.17	0.12	.13
Study 4	0.24	0.13	.07
Study 5	0.40	0.21	.06
Sex	0.07	0.08	.41
Race	0.12	0.09	.20

Note. Study 1 represents the reference group; Sex is coded (0 = female, 1 = male); Race is coded (0 = Black, 1 = White).

10. Discussion

This article provides an overview of MNLFA, demonstrating its flexibility for use within IDA where the goal is to develop a construct that has properly scaled across studies. We offer a tutorial on implementing this model in practice, demonstrating the steps involved in the admittedly complex model-building process of developing an appropriate MNLFA model. Further, to allow applied researchers to more easily implement and modify these models, all empirical data and code used for analyses are provided at: https://github.com/jmk7cj/SEM-mnlfa. This is in addition to the

Table 5. Logistic regression results of high school completion.

	Odds Ratio	95% C.I.	*p*
Study 2	1.52	[0.97, 2.39]	.065
Study 3	0.07	[0.86, 1.32]	.546
Study 4	3.40	[2.36, 4.89]	<.001
Study 5	0.05	[0.03, 0.08]	<.001
Male	0.76	[0.63, 0.91]	.002
White	1.19	[0.97, 1.46]	.091
Factor	0.77	[0.71, 0.84]	<.001

Note. Intercept = 1.14 logits (*SE* = 0.11).

Table 4. Final MNLFA model examining covariate effects on item intercepts and factor loadings.

	Intercept			Loading		
Covariate effect	Estimate	*SE*	*p*	Estimate	*SE*	*p*
1. Breaks rules	−3.33	0.21	<.01	2.91	0.21	<.01
Study 4	–	–	–	−0.43	0.12	<.01
Sex	–	–	–	−0.34	0.10	<.01
2. Harms property	0.50	0.21	.02	3.36	0.31	<.01
Study 3	0.98	0.18	<.01	−1.73	0.30	<.01
Study 4	1.43	0.22	<.01	–	–	–
3. Breaks things	0.06	0.24	.82	3.29	0.27	<.01
Study 3	0.66	0.20	<.01	–	–	–
4. Takes property	−0.84	0.20	<.01	3.13	0.22	<.01
Study 3	0.81	0.15	<.01	–	–	–
Study 5	0.43	0.28	.12	–	–	–
5. Fights	−1.71	0.25	<.01	3.19	0.27	<.01
Study 3	−1.21	0.21	<.01	−0.79	0.27	<.01
Sex	0.20	0.12	<.01	–	–	–
6. Lies	−1.54	0.19	<.01	2.98	0.21	<.01
Study 2	−1.05	0.25	<.01			
Study 5	–	–	–	−0.77	0.27	<.01
Race	–	–	–	−0.38	0.17	.02
7. Yells at others	−2.53	0.23	<.01	2.88	0.23	<.01
Study 3	−0.48	0.21	.02	−0.54	0.23	.02
Study 5	−0.93	0.30	<.01	–	–	–
Sex	−0.52	0.12	<.01	–	–	–
8. Stubborn	−3.04	0.18	<.01	2.09	0.15	<.01
Study 4	−1.15	0.14	<.01	–	–	–
Study 5	−1.39	0.27	<.01	–	–	–
Sex	−0.60	0.12	<.01	–	–	–
9. Teases others	−2.45	0.18	<.01	2.46	0.17	<.01
Study 5	−1.01	0.29	<.01	–	–	–
Race	−0.18	0.12	.13	–	–	–

annotated Mplus input and output files already discussed in detail. Results from the empirical demonstration support findings from prior research (Bradshaw et al., 2010), in which ratings of first-grade students on the aggressive-disruptive behavior subscale of the TOCA-R were found to have a negative association with high school graduation.

By combining raw data pooled across five separate prevention trials, we were likely able to produce findings that are more robust than any single study may have found. This was achieved through a larger overall sample size, increased frequencies of low base-rate behaviors (e.g., item "*breaks things*"), and the ability to account for measurement invariance or differential item functioning by allowing multiple covariates to simultaneously moderate item and factor parameters. As data repositories and open-source sharing of registered studies continue to grow in popularity (e.g., Registry of Efficacy and Effectiveness Studies), there is a rapid increase in the need for appropriate, advanced methodological tools. It is our hope this paper provides a tutorial on the model building process of MNLFA for IDA.

Acknowledgments

We thank Veronica Cole and the reviewers for their helpful comments on earlier versions of this manuscript.

Funding

This work was supported by the National Institute of Mental Health (1R01MH122214-01).

ORCID

Joseph M. Kush http://orcid.org/0000-0003-0183-494X
Katherine E. Masyn http://orcid.org/0000-0001-5337-1137
Masoumeh Amin-Esmaeili http://orcid.org/0000-0002-3888-3254
Ryoko Susukida http://orcid.org/0000-0003-0444-5368
Holly C. Wilcox http://orcid.org/0000-0003-2624-0654
Rashelle J. Musci http://orcid.org/0000-0001-7267-5822

References

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*, 507–526. https://doi.org/10.1037/met0000077

Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 43–55. https://doi.org/10.1080/10705511.2019.1642754

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*, 101–125. https://doi.org/10.1037/a0017642

Belzak, W. C. (2020). Testing differential item functioning in small samples. *Multivariate Behavioral Research*, *55*, 722–747. https://doi.org/10.1080/00273171.2019.1671162

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical theories of mental test scores*. Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459. https://doi.org/10.1007/BF02293801

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Bradshaw, C. P., & Kush, J. M. (2020). Teacher observation of classroom adaptation-checklist: Measuring children's social, emotional, and behavioral functioning. *Children & Schools*, *42*, 29–40. https://doi.org/10.1093/cs/cdz022

Bradshaw, C. P., Schaeffer, C. M., Petras, H., & Ialongo, N. (2010). Predicting negative life outcomes from early aggressive-disruptive behavior trajectories: Gender differences in maladaptation across life domains. *Journal of Youth and Adolescence*, *39*, 953–966. https://doi.org/10.1007/s10964-009-9442-8

Conduct Problems Prevention Research Group (2019). *The Fast Track program for children at risk: Preventing antisocial behavior*. Guilford Press.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, *14*, 165–176. https://doi.org/10.1037/a0015565

Cooper, H. M. (2009). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Sage.

Curran, P. J., Cole, V. T., Bauer, D. J., Rothenberg, W. A., & Hussong, A. M. (2018). Recovering predictor-criterion relations using covariate-informed factor score estimates. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 860–875. https://doi.org/10.1080/10705511.2018.1473773

Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 827–844. https://doi.org/10.1080/10705511.2016.1220839

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*, 81–100. https://doi.org/10.1037/a0015914

Curran, P. J., McGinley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., Sher, K., & Zucker, R. (2014). A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. *Multivariate Behavioral Research*, *49*, 214–231. https://doi.org/10.1080/00273171.2014.889594

DiStefano, C., Shi, D., & Morgan, G. B. (2021). Collapsing categories is often more advantageous than modeling sparse data: Investigations in the CFA framework. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 237–249. https://doi.org/10.1080/10705511.2020.1803073

Eddy, J. M., Reid, J. B., Stoolmiller, M., & Fetrow, R. A. (2003). Outcomes during middle school for an elementary school-based preventive intervention for conduct problems: Follow-up results from a randomized trial. *Behavior Therapy*, *34*, 535–552. https://doi.org/10.1016/S0005-7894(03)80034-5

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*, 278–295. https://doi.org/10.1177/0146621605275728

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3–8. https://doi.org/10.3102/0013189X005010003

Gottfredson, N. C., Cole, V. T., Giordano, M. L., Bauer, D. J., Hussong, A. M., & Ennett, S. T. (2019). Simplifying the implementation of modern scale scoring methods with an automated R package: Automated moderated nonlinear factor analysis (aMNLFA). *Addictive Behaviors*, *94*, 65–73. https://doi.org/10.1016/j.addbeh.2018.10.031

Graham, E. K., Rutsohn, J. P., Turiano, N. A., Bendayan, R., Batterham, P. J., Gerstorf, D., Katz, M. J., Reynolds, C. A., Sharp, E. S., Yoneda, T. B., Bastarache, E. D., Elleman, L. G., Zelinski, E. M., Johansson, B., Kuh, D., Barnes, L. L., Bennett, D. A., Deeg, D. J. H., Lipton, R. B., … Mroczek, D. K. (2017). Personality predicts mortality risk: An integrative data analysis of 15 international longitudinal studies. *Journal of Research in Personality*, *70*, 174–186. https://doi.org/10.1016/j.jrp.2017.07.005

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 80–100. https://doi.org/10.1207/S15328007SEM1001_4

Gross, A. L., Tommet, D., D'Aquila, M. D., Schmitt, E., Marcantonio, E. R., Helfand, B., … Jones, R. N. (2018). Harmonization of delirium severity instruments: A comparison of the DRS-R-98, MDAS, and MAC-S using item response theory. *BMC Medical Research Methodology*, *18*, 1–18. https://doi.org/10.1186/s12874-018-0552-4

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitation large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 621–638. https://doi.org/10.1080/10705511.2017.1402334

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Erlbaum.

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013). Integrative data analysis in clinical psychology research. *Annual Review of Clinical Psychology*, *9*, 61–89. https://doi.org/10.1146/annurev-clinpsy-050212-185522

Ialongo, N. S., Werthamer, L., Kellam, S. G., Brown, C. H., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression, and antisocial behavior. *American Journal of Community Psychology*, *27*, 599–641. https://doi.org/10.1023/A:1022137920532

Jöreskog, K. G. (1967). A general approach to confirmatory maximum likelihood factor analysis. *ETS Research Bulletin Series*, *1967*, 183–202. https://doi.org/10.1002/j.2333-8504.1967.tb00991.x

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426. https://doi.org/10.1007/BF02291366

Kellam, S. G., Mayer, L. S., Rebok, G. W., & Hawkins, W. E. (1998). Effects of improving achievement on aggressive behavior and of improving aggressive behavior on achievement through two preventive interventions: An investigation of causal paths. In B. P. Dohrenwend (Ed.), *Adversity, stress, and psychopathology* (pp. 486–505). Oxford University Press.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method for identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*, 59–72. https://doi.org/10.1207/s15328007sem1301_3

Lord, F. M., & Novich, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543. https://doi.org/10.1007/BF02294825

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variables indicators. *Psychometrika*, *49*, 115–132. https://doi.org/10.1007/BF02294210

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Petras, H., Chilcoat, H. W., Leaf, P. J., Ialongo, N. S., & Kellam, S. G. (2004). Utility of TOCA-R scores during the elementary school years in identifying later violence among adolescent males. *Journal of the American Academy of Child and Adolescent Psychiatry*, *43*, 88–96. https://doi.org/10.1097/00004583-200401000-00018

Petras, H., Ialongo, N., Lambert, S. F., Barrueco, S., Schaeffer, C. M., Chilcoat, H., & Kellam, S. (2005). The utility of elementary school TOCA-R scores in identifying later criminal court violence among adolescent females. *Journal of the American Academy of Child and Adolescent Psychiatry*, *44*, 790–797. https://doi.org/10.1097/01.chi.0000166378.22651.63

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190. https://doi.org/10.1007/BF02295939

Rose, J. S., Dierker, L. C., Selya, A. S., & Smith, P. H. (2018). Integrative data analysis of gender and ethnic measurement invariance in nicotine dependence symptoms. *Prevention Science*, *19*, 748–760. https://doi.org/10.1007/s11121-018-0867-8

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. https://doi.org/10.1037//1082-989X.7.2.147

Sibley, M. H., & Coxe, S. J. (2020). The ADHD teen integrative data analysis longitudinal (TIDAL) dataset: Background, methodology, and aims. *BMC Psychiatry*, *20*, 359. https://doi.org/10.1186/s12888-020-02734-6

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Chapman & Hall/CRC.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239. https://doi.org/10.1111/j.2044-8317.1974.tb00543.x

Thissen, D. (2001). *IRTLRDIF v.2.0.b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–114). Routledge.

Tolan, P., Gorman-Smith, D., & Henry, D. (2004). Supporting families in a high-risk setting: Proximal effects of the SAFEChildren Preventive Intervention. *Journal of Consulting and Clinical Psychology*, *72*, 855–869. https://doi.org/10.1037/0022-006X.72.5.855

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. https://doi.org/10.1177/109442810031002

Werthamer-Larsson, L., Kellam, S., & Wheeler, L. (1991). Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology*, *19*, 585–602. https://doi.org/10.1007/BF00937993