Accepted manuscript published in *International Journal of Research & Method in Education*, https://doi.org/10.1080/1743727X.2021.1973992

On the Merits of Longitudinal Multiple Group Modeling: A Valid and Often Preferred Approach to Intervention Evaluations

Todd D. Little, Daniel Bontempo, Charlie Rioux,
Texas Tech University

Allison Tracy
2M Research

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was funded by the Centers for Disease Controls and Prevention (CDC) [Grant Number: 20IPA20-09434]. Charlie Rioux was supported by fellowships from the Canadian Institutes of Health Research and the Fonds de Recherche du Québec - Santé. The authors would like to thank Keith F. Widaman for instilling many of the ideas that evolved into this work.

On the Merits of Longitudinal Multiple Group Modeling: An Alternative to Multilevel Modeling for Intervention Evaluations

Abstract

Multilevel modeling (MLM) is the most frequently used approach for evaluating interventions with clustered data. MLM, however, has some limitations that are associated with numerous obstacles to model estimation and valid inferences. Longitudinal multiple-group (LMG) modeling is a longstanding approach for testing intervention effects using cluster-sampled data that has been superseded by the rise of MLM approaches, but the LMG approach can have advantages when research questions do not pertain to predicting variability at the higher levels. In this paper, we first review the advantages and limitations of MLM and LMG approaches. Second, steps in the estimation of a LMG model are presented, with some recent upgrades and changes in the modeling strategy that have particular utility for evaluating interventions. We discuss the advantages of the LMG approach as a guided confirmatory model-testing framework and how the approach places a premium on avoiding Type II errors, particularly when complex interactions are potentially at play.

On the Merits of Longitudinal Multiple-Group Modeling: An Alternative to Multilevel Modeling for Intervention Evaluations

Organizations providing oversight for randomized control trials emphasize a conservative evaluation paradigm: avoid overstating program effects (Type I error). In this regard, researchers are encouraged to choose very few hypotheses and are required to register these hypotheses a priori (Moher et al., 2010). This overly conservative paradigm is often at odds with many evaluation efforts, doing little to avoid overlooking program effects (Type II error; see Little et al., 2017, Niolon et al., 2019). A great deal of expense and effort goes into conducting a rigorous evaluation study in naturalistic settings, and stakeholders are keen to answer many research questions using the resulting data. There are often multiple outcomes, multiple time points, multiple subgroups, and even multiple interventions of interest. The traditional evaluation of many program effects can unduly inflate Type II errors, particularly when nested data structures are involved (e.g., time within person, person within clinic, student within school, etc.).

Common wisdom dictates the use of multilevel modeling (MLM, also referred to as hierarchical linear modeling; Peugh, 2010, Snijders & Bosker, 2011) when nested data structures exist. In fact, the pre-registration process of many school-based clinical trials would likely be questioned if an MLM was not indicated as the primary analytic approach. This paper describes an alternative to the MLM modeling approach that provides balanced management of Type I and Type II errors in the context of evaluating nuanced program effects. In particular, we highlight the advantages of the traditional structural equation modeling (SEM; Hoyle, 2012; Lei & Wu, 2007) approach to fitting a longitudinal multiple-group (LMG) model to evaluate an intervention or similar evaluation. As discussed below, the LMG approach is particularly useful when moderating factors such as gender, multiple cohorts, and the like are examined as part of the overall treatment evaluation (Niolon et al., 2019).

Below, we first discuss the rationale and limitations of multilevel approaches and how LMG can be advantageous over these methods. Second, we highlight a number of important advances and options in the general SEM framework that optimize an evaluation and then describe the steps in analyzing data using LMG.

Statistical approaches to evaluating interventions with nested data structures

Rationale and Limitations of Multilevel Modeling and Multilevel Structural Equation Modeling

Current statistical approaches that are considered state-of-the art for nested data structures include MLM and multilevel structural equation modeling (ML-SEM). MLM takes into account shared variance in nested data, providing more accurate estimates of standard errors compared to traditional single-level regression approaches that ignore dependency in the data, and allows relationships within and between clusters to be examined (Gelman, 2006; Greenland, 2000; Osborne, 2000). This feature makes it possible to examine research questions related to both lower levels (e.g., predictors of student-level differences in the magnitude of a treatment affect) and higher levels (e.g., predictors of school-level differences in the magnitude of a treatment effect). In most intervention evaluations, however, an MLM approach is used only to account for the dependency in the data structure that is caused by nesting. As we will describe below, when research questions do not pertain to predicting variability at the higher levels, the need for MLM is only minimally justified because the issue of biased standard errors is readily addressed with other statistical approaches. In such contexts, where only standard error corrections is the goal, a number of compelling arguments can be levied against relying on MLM as the best or only way to generate valid inferences from nested-data structures because SEM methods rely on fewer assumptions and have more flexibility (see Table 1).

[Table 1 about here]

ML-SEM (Hox, 2013; Rabe-Hesketh, Skrondal & Zheng, 2012), where time is represented in a wide format and nesting factors are represented as higher-level units, also takes

into account shared variance in nested data and would give the same results as MLM for an identical model (Curran, 2003). By taking advantage of the strengths of the SEM framework, however, ML-SEM can rectify several limitations of MLM (Curran, 2003), thus improving multilevel analyses (see Table 1). An important advantage is that a measurement model (with latent factors) can be specified, which reduces measurement error and allows testing the assumption of measurement invariance (Tomarken & Waller, 2005; Van De Schoot et al., 2015). It remains limited, however, in specifying very complex models and in significance testing. Indeed, because all standard errors in the ML-SEM model, including those associated with treatment effects, are dependent on the parameter constraints selected to set the scale of latent variables, significance tests based on standard errors are inadequate for inferences (Gonzalez & Griffin, 2001). Furthermore, while a major strength of ML-SEM over MLM is that it allows testing for measurement invariance, when latent factors are specified at Level 1 (e.g., measures of student outcomes), factorial invariance between Level 2 units (e.g., schools) is still assumed and not testable. Furthermore, model fit evaluation is complicated by the Within-Level and Between-Level model components, which is compounded with additional levels of nestedness (Ryu, 2014). In addition, complex treatment effects (multi-way interactions) are difficult to properly specify, estimate, and evaluate. The LMG framework described below is an SEM model and thus shares the strengths of ML-SEM, but also remedies some of its limitations (see Table 1).

Another issue that obscures the utility of both MLM and ML-SEM approaches is the behavior of these models under multiple imputation. Users of MLM and ML-SEM generally handle missing data using full information maximum likelihood estimation (FIML; Enders, 2010; Enders, 2001), which has been shown to be an efficient missing data technique for missing nested data (Larsen, 2011) and is the most commonly used missing data treatment in the analysis of psychosocial clinical trial (Rioux & Little, 2019). In the context of complex moderation and extensive missing data, however, FIML can fail. In these situations, the

presence of many unique missing data patterns can prevent FIML estimation from converging. Moreover, the FIML estimator utilized by MLM approaches often cannot address the missing data demands because many auxiliary variables are needed to address potential missing at random (MAR; Seaman et al., 2013) mechanisms (Enders, 2010; Graham, Cumsille, & Shevock, 2012).

Under these circumstances, multiple imputation (Enders, 2017; Enders, 2010; van Buuren, 2018) is often required. Indeed, when planned missing designs are used (e.g., multi-form designs, skip and fill questions, cohort sampling; Rhemtulla & Hancock, 2016; Rioux et al., 2020) and/or when unplanned missing data patterns are extensive, multiple imputation is needed due to its ability to handle very large numbers of missing data auxiliary variables. Further, nearly all imputation software is designed to impute data with a single-level structure and multilevel imputation as implemented in a limited number of available software packages is less user-friendly than single-level imputation (Drechsler, 2015; Grund, Lüdtke & Robitzsch, 2016). To accommodate nestedness in a single-level imputation model, many moderating effects are required, one per each nesting unit. This largely increases predictors in the imputation model, especially as the number of clusters increases, which is an issue since MI can fail when there is a high number of variables relative to the number of observations (Enders, 2010; Graham, Cumsille, & Shevock, 2012; van Buuren, 2018). To the best of our knowledge, the principal component auxiliary variables approach (Howard et al., 2015), as implemented using the PcAux package (Lang, et al., 2018), is the only method that can successfully address imputation in this context. PcAux can recover both MAR and missing completely at random (MCAR) missing data patterns with extensive use of auxiliary variables that account for complex interactions, such as cross-level dependencies (see Lang et al., 2018, for details). Importantly, when properly specified with clear delineation of moderators and other design elements, the missing data model and analysis model will match when the evaluation is conducted in an LMG framework but not in an MLM or ML-SEM framework. When any of the above considerations

are at play in a given evaluation context, the LMG approach may be a preferred alternative to use.

Overview and merits of the Longitudinal Multiple-Group Framework

The LMG approach builds on established methods in the field of program evaluation (McArdle & Hamagami, 1996; Thompson & Green, 2006). We describe, however, a number of extensions of the LGM approach that allow researchers to test sophisticated hypotheses about treatment effects. When data are collected in a nested sampling framework and when higher levels of analysis (e.g., schools) can be considered as simply a potential confound or "nuisance" factor, LMG is arguably more flexible than its MLM counterpart. As we explain in more detail below, higher-level nesting can be statistically adjusted prior to entering the substantive modeling phase (see also Niolon et al., 2019). By controlling for dependencies in the data due to nesting as a data preparation step, the evaluation can be focused solely on the lower level unit (e.g., students), allowing explicit modeling of measurement error, evaluation of measurement invariance over time and across groups, and straightforward examination of any model parameters across treatment groups and time. As we highlight in more detail below, such testing is done using a principled confirmatory model-testing framework, which minimizes the number of hypothesis tests and places a premium on avoiding Type II error while holding Type I error at bay (Niolon et al., 2019).

In the LMG framework, time is represented in a wide format; constructs are represented for each measurement occasion in a straightforward panel model. Modeling time in this way, a simple 2-level repeated measures model is rendered as a single-level SEM panel model. That is, occasions of measurement, which are nested within participant, are explicitly represented as multiple waves or panels of the constructs. In SEM, the measurement model of the latent constructs can be explicitly estimated for each occasion and factorial invariance (or measurement equivalence) can be tested and enforced (Little, 2013).

In the LMG framework, intervention condition and other key moderating factors (e.g.,

gender, cohort) in the evaluation are treated as grouping variables, where the parameters are estimated in each group as fixed effects rather than as random effects. When other nesting factors, such as classrooms and schools, are not the focus of the evaluation, they are treated as nuisance variables thereby addressing the issue of nestedness (see Table 1 and below).

As indicated in Table 1, the LMG approach can address nested data issues effectively. In addition, the strong factorial invariance offered by LMG allows mean-level information to be estimated and compared over time and across groups. Strong factorial invariance also adds stability and power to the analysis model because the measurement model parameters are defined by the full sample of participants at each wave of data collection. These power gains can begin to outstrip the power demands needed in MLM to detect intervention effects (see power gains below).

Dealing with Nestedness. Treating school or classroom as a set of dichotomous covariates addresses dependency issues related to nesting but does not diminish potential treatment effect evaluation. In many studies, a set of nesting covariates will be too large to easily incorporate into a substantive model. To best accomplish the adjustment, a 2-stage approach can be implemented as a modeling simplification (see Niolon et al., 2019, for an empirical example). In the first stage, indicators of a latent variable are regressed on the set of nesting unit covariates, and the unstandardized individual-level residuals are calculated and saved. Indicators can be items or parcels and do not need to be tau-equivalent (Little, Bovaird, & Widaman, 2006). In the second stage, the unstandardized residuals are used as the indicators in hypothesis testing or substantive exploration models (Little, Bovaird, & Widaman, 2006). When there a substantial number of additional covariate controls required (e.g., variability in assessment timing, student demographics), a broad number of additional covariates can be readily included in the pre-analysis adjustment phase. In LMG, it is advisable to adjust indicators for covariate effects separately for each group and time point to match the substantive model. This approach is equivalent to freely estimating covariate effects in a traditional analysis

model but does not burden the model with the additional parameter estimates and the potential for non-convergence when many covariates are involved.

If the substantive analysis involves the means structure (e.g., comparison of group means), it is important to strategically consider selecting the zero point of each covariate. Since residuals are centered at zero, the only parameter remaining to inform the means is the intercept. Therefore, mean-centering or effects-coding the covariates produce an intercept that can be interpreted as the sample average. Adding the intercept to the residuals preserves the scale of the construct. In LMG, centering covariates can be done either at the grand mean or the group mean. However, if schools are the unit of random assignment to treatment, school effects should be coded within treatment groups to preserve treatment effects shared across students within schools (see Niolon et al., 2019, for an empirical example).

Robust significance testing when standard errors are untrustworthy. In the LMG approach, model constraints are placed on any parameters of a model across time and/or groups to test for differences across time and/or groups using the change in chi-squared value, as a nested model comparison, to evaluate significance. Standard errors in SEM models are untrustworthy because they are influenced by the scale-setting constraint; however, the nested model change in chi-squared value is accurate regardless of scaling constraint and when multiple imputation of missing data is used (Gonzalez & Griffin, 2001).

As mentioned above, standard errors in SEM models generally are scale dependent and are not a robust method of evaluating significance (Gonzalez & Griffin, 2001). Even estimation procedures (e.g., MLR, WL-MSV) to correct standard errors for potential bias due to factors such as distributional violations are still inaccurate because of the scale-setting dependency but difference testing between nested models can sometimes be done. Fortunately, although standard errors are not trustworthy, the nested-model chi-squared difference test under maximum likelihood or FIML estimation is an unbiased estimate of significance regardless of scaling method and is quite robust to violations of distributional assumptions. Although the

power of the chi-squared itself is tied to sample size, the difference test using the chi-squared value is robust to other potential effects such as model misspecification and provides a statistical rationale for tests of differences (as opposed to a modeling rationale such as change in CFI). With any highly powered test, however, practical significance should be emphasized by focusing on effect sizes.

Measurement equivalence and power in the Longitudinal Multiple-Group approach. In the LMG approach, factorial invariance (aka, measurement equivalence) across treatment and time (as well as other potential grouping variables such as cohort and gender) is specified, which ensures that comparisons are psychometrically equivalent (regardless of whether the constructs' indicators are tau equivalent or congeneric) and, quite importantly, measurement error is corrected. In many longitudinal and multiple-group evaluation contexts, the assumption of factorial invariance is a critical assumption that warrants explicit testing (Little, 2013). Two benefits of testing and enforcing factorial invariance are a) it ensures that the constructs are psychometrically equivalent in the definition and meaning over time and across groups and b) further tests of differences in the latent variable parameters that are considered as part of the evaluation are done with "considerably" more power than either MLM or ML-SEM. The precise amount of power gained is somewhat open-ended (but can be easily examined using a Monte Carlo simulation on a case by case bases); however, the fixed effect parameter tests conducted in the LMG framework are done so in a space where invariance of loading and intercepts across time and groups provides a common grounding across time and groups that renders the parameter space anchored in the common loadings and intercepts. When invariance holds, the principled modeling of the latent parameters can then be conducted.

Fixed effects modeling. Because the analysis model of the LMG approach treats treatment as a grouping variable and each level of the potential moderators (e.g., cohort and gender) as additional grouping variables, each sub-group (and time of measurement) is estimated uniquely. Here, the nestedness of these factors are explicitly included as discrete

(fixed) estimates in the baseline (freely estimated) analysis model. That is, each element is treated as a fixed effect parameter that can be tested for differences across each group and each time of measure in subsequent nested model specifications. By utilizing the power and robustness of the chi-squared difference test, all manner of parameter comparisons (via nested-model constraints) can be evaluated with adequate statistical precision because the measurement model with factorial invariance in place grounds the latent parameter space in a measurement model derived from the total sample (i.e., loadings and intercepts are estimated from the whole sample giving a common factor score metric for all latent parameters).

Having addressed the advantages and disadvantages of MLM, ML-SEM, and LGM, we now turn to more specifics of the model constraining process, which is the hallmark advantage of the LMG framework.

Estimating the Longitudinal Multiple-Group Model

The key method for testing hypotheses in an LMG model is to compare the fit of a freely estimated model with the fit of a nested model in which parameters are constrained across groups, using the chi-square difference test. Such tests can be conducted in many different ways from multiple pair-wise comparisons (which has problematic error rates) to structured tests of uniform trends, to omnibus tests of no differences etc. On the other hand, we present an alternative testing paradigm that builds on the logic of confirmatory model testing and adds the systematic inclusion of a set of circumscribed modeling steps that are developed and followed during the evaluation testing phase (see Niolon et al., 2019, for a recent empirical example).

The innovation that principled model constraints offers is to leverage model constraints to impose theory-guided parsimony on many model parameters simultaneously. The goal here is to arrive at a model that (a) is associated with a single or a few hypothesis testing comparisons, thereby minimizing the Type I error rate and (b) derives interpretable comparisons yielding nuanced evaluation of any program effects across multiple time points, subgroups, and/or outcomes, thereby minimizing the Type II error rate. This principled constraining approach is

very generalizable and can encompass a wide variety of models and hypothesis tests. There is a need, however, for structured guidelines to avoid overfitting the models to the data at hand.

The model constraining process is a theory-guided procedure that places a premium on avoiding Type II error (Little et al., 2017). In our view, this confirmatory-model-testing approach provides an optimal approach to modeling the nature of any treatment effects. For example, base-line equivalence can be tested and evaluated by constraining the mean-levels of the outcome variables at Time 1 to be equivalent across groups (e.g., Niolon et al., 2019).

Parameter Constraining in the Longitudinal Multiple-Group Framework

As mentioned, the model constraining process is designed to impose a parsimonious theory-guided structure on any of the parameters across the groups and across time. The order and hierarchy of importance in the possible constraints should be established a priori. For intervention evaluations, the assumption of baseline equivalence is probably the paramount expectation under random assignment to treatment versus control. Of penultimate importance is the guiding principle of when and how a treatment effect should emerge. The third most important principle is allowing for moderation by other design (e.g., cohort) or sample (e.g., gender) characteristics. Time of measurement influences such as within school-year exposure should also be considered. Additional influences such as change in school (e.g., transition from middle school to high school) can then be added to the hierarchy of principles to follow in the constraining process.

In the constraining process, the parameter space of interest needs to be clearly delineated (e.g., mean-levels, associations, and/or variances of the latent constructs). In other words, depending on the nature of the research question, the parameters can be mean-levels or they can be covariance/regression coefficients (or variance estimates for that matter). The latent parameters in these models need to be estimated in meaningful and comparable metrics. A number of modeling steps are particularly advantageous to facilitate model parameter comparisons across groups and time in the LMG framework: Using parcels as indicators,

rescaling the manifest scales into percent of maximum scoring, using effects-coded scaling constraints, and estimating rescaling constructs are four important aspects of the general SEM framework that facilitate a valid and meaningful evaluation. Although these techniques are not unique to LMG (e.g., they are useful in an ML-SEM framework as well), they are briefly described here because they are important steps to conduct an effect analysis using LMG.

Using parcels as indicators. An in-depth discussion of the use of parcels is beyond the scope of the recommendations we are describing here (see Little, Cunningham, Shahar, & Widaman, 2002; Matsunaga, 2008 for details of their pros and cons). Parceling is a pre-analysis step that involves averaging two or more items per parcel to create a reduced set of optimal indicators of the latent constructs. Ideally, three indicators are derived through parceling to create a just identified latent construct that can be included at many time points and in many groups, whereby the estimation process is rendered tractable and robust with regard to various potential assumption violations (multivariate non-normality, local independence, item-level non-invariance, etc.; see Little, 2013). When a theory-guided parceling scheme is followed (e.g., using facet-representative parcels), the resulting indicators provide meaningful information that is more reliable, better distributed, and are more robust estimation-wise than are equivalent models fit to item-level data. Briefly, the use of parcels has several psychometric and model-level benefits.

Psychometrically, parcels tend to have better distributional properties. Indeed, when items are averaged, non-normal distributions become more normally distributed and scales become more continuous since intervals increase in number while also becoming smaller and more equal. Compared to item-level data, parcels also have higher reliability and communality, as well as a higher ratio of common-to-unique factor variance (Little et al., 2002; Matsunaga, 2008). This increase in reliability is because when parcels are computed, the proportion of the "true" score variance relative to the specific variance and random error is higher than in the individual items. Indeed, the variance related to the construct measured, that is, the "true"

score, which is shared between the items, is preserved. At the same time, the specific variance and random error variance, which are not shared between the items, are reduced (Little et al., 2002).

At the model level, advantages of parcels over items include lower likelihoods of correlated residuals and dual factor loadings, and reduced sources of both sampling and parsimony errors. Parceling will reduce type II error where researchers would conclude that cross-loadings and residual correlations exist in the population when they do not (Little et al., 2002). Although parcel-level models can have lower power than item models to detect within-factor misspecifications and single cross-loadings, they have higher power to detect multiple cross-loadings and structural model misspecifications (Rhemtulla, 2016). As long as parcels are thoughtfully constructed, and the item-level data is examined to identify potential sources of misspecification and used to inform the parceling scheme, misspecification due to using parcels can be avoided easily (Bandalos & Finney, 2001; Rhemtulla, 2016) and their use is then advantageous. Moreover, considering that using three indicators to model a just-identified latent variable is considered optimal (Little, 2013), parceling can be used to reduce a larger number of items to three indicators.

For evaluating the intervention effects, the parceling scheme needs to be the same across groups and measurement occasions. This common parceling scheme is necessary for factorial invariance, an aspect of the model for which parcels can also be advantageous. Indeed, parcel-level data may show strong factorial invariance even when differential item functioning or lack of measurement invariance is found at the item level (Meade & Kroustalis, 2006). Lastly, we generally recommend doing the covariate adjustment step discussed above to parcel level indicators that will be used in the modeling process. As mentioned above, when covariate adjustments to the constructs' indicators are made as a preliminary step, adding the intercept of this residualizing step maintains the mean-level information (Little, Bovaird et al., 2006).

Using POMS scoring. When scales are measured with a non-meaningful zero point,

we also recommend a monotonic data transformation such as converting Likert-like scales into proportion of maximum scaling (POMS). For many applications, POMS is preferred over original metric because POMS provides a meaningful zero and max that can be useful such as when developing relative risk ratios between treatment and control groups. POMS involves establishing zero as the minimum and 1.0 as the maximum through simple subtraction followed by division (see Little, 2013 for steps in doing POMS metric that maintains group and time differences in mean-levels). After the monotonic transformation, using POMS metric in the LMG framework places a meaningful zero to the scale and mean levels are readily interpreted as the proportion of the maximum for the construct in question. For example, a seven-point Likert scale (coded 1-7) would be transformed by subtracting 1.0 from each score and then dividing by 6 (the new maximum after subtracting 1.0). A score of .75 would indicate a .75 proportion of the 1.0 maximum.

Using Effects-Coded Scaling Constraints. When the mean-level coefficients are addressed as key elements of the LMG evaluation, using the effects-coded scaling constraint is preferred because the inherent scale meaning of the mean-levels is maintained (Little, Slegers et al., 2006). Effects-coded constraints keep the inherent meaning of the indictors' metric by restricting the loading to average 1.0 and the intercepts to average 0.0. When all indicators share a meaningful metric, the effects-coded constraints estimate the latent mean and variance in the observed metric (e.g., POMS metric) of the indicators.

Using Rescaling Constructs. If the parameters of interest for evaluating the intervention are covariances or regression coefficients, an additional necessary step is to specify the constructs in the models using standardization constructs (aka, phantom constructs, Rindskof, 1984; see associated figure) . Standardization constructs convert all covariances or regression coefficients into common metrics that become actual parameter estimates in the model (i.e., they are not post-estimation transformations). These standardization constructs provide model estimates of the variance/covariance information that are on the same metric of

comparison and that can be tested using model constraints and evaluated using the nested-model chi-squared difference test. Differences in variances across groups or times are estimated as standard deviations that rescale the standardization constructs to have a variance of 1.0 and a mean of zero (the mean-level information is still retained on the lower-order constructs that are already in a comparable mean-level metric). This model estimation "trick" puts the regression linkages among standardization constructs on a comparable metric across groups and time; similarly, any covariances are now standardized estimates of the model (i.e., estimated as correlations). Importantly, the parameters among the standardization constructs are direct estimates in the analysis model and, as such, can be constrained in a nested-model manner to test differences in the magnitudes of the parameter estimates being constrained. Here, when the chi-squared value is significant, the constrained parameters are different in magnitude and, when the chi-squared value is non-significant, the constrained parameters in question are statistically equivalent in magnitude.

The choice of parameters should be circumscribed by the research question and theoretical relevance. A research question about mean-level differences between groups, for example, would not include parameters related to the variance/covariance part of the model and a research question about cross-lagged and auto-regressive pathways of prediction, similarly, would not include the mean-level parameters in the constraining process. Even more principled and circumscribed subsets of parameters could be modeled sequentially or hierarchically. The parameters to be constrained are matters of choice and justification on the part of the investigation team and should be articulated in the guiding principles of the constraining process.

[Figure 1 about here]

Establish a rigorous measurement model. In the LMG framework, one way to examine whether an intervention is effective is to systematically test parameters of interest for equality and to constrain to equality those parameters that do not differ meaningfully from each

other. The hypothetical model presented in Figure 1 would be represented and estimated in each group of the intervention evaluation. In this model, strong factorial invariance is specified as denoted by the common superscripted letters associated with the measurement model parameters. The ability to test and enforce strong factorial invariance has a number of advantages in the LMG approach. First, (as mentioned above) invariance ensures that the parameters of the model are estimated in a psychometrically equivalent manner across time and groups. Second, these measurement model parameters are derived from the conjoint estimation across all groups and therefore anchor the latent factor information within each group to the power offered across the full sample. That is, this across-group estimation of the loadings and intercepts provides stability and power for all latent model parameters even when subgroup sizes are relatively small (e.g., ~50). Third, as a multiply indicated measurement model, the latent parameter estimates are now estimated as disattenuated and error free parameters. Fourth, the measurement model can allow for congeneric indicators across time and groups.

The logic of parameter constraining

Once a measurement model is established using (a) an optimal parceling scheme, (b) standardization constructs, (c) effects-coded scaling constraints and (d) strong factorial invariance constraints, the estimated latent construct parameters can be examined. The idea in the constraining process is to provide sets or "bands" of jointly constrained estimates of the functionally equivalent parameters (i.e., clearly similar at $p >= .2$). Again, the bands are defined by the set of guiding principles and their hierarchy of implementation.

Here, tests of program effects are based on the differences between the constrained parameters in a given band against the constrained parameters in a neighboring band (i.e., the next group of similar/like parameters), rather than on differences between individual point estimates. This simplification of the model comparisons results in fewer formal hypothesis tests, which minimize the risks of both Type I errors and Type II errors. Although not a common approach, it has been used in past research (Little & Lopez, 1997) and in recent evaluations

(Niolon et al., 2019). In other words, constrained bands are confirmatory tests of program effects that essentially eliminate the Type II errors that can occur under nil hypothesis testing frameworks.

Once the constrained bands are identified, they are then evaluated for differences between bands. Depending on the nature of the modeling framework, either a chi-squared difference test can be employed or one can use a post hoc Wald test ($p < .01$). When testing sets of constrained parameters, the scaling constraint to define the metric of the constructs no longer unduly impacts the estimates of the constrained parameters' standard errors, particularly when strong factorial invariance is imposed in the measurement model. Creating heat maps of the discrepancies between the constrained and unconstrained point estimates facilitates overall evaluation of the adequacy of the constraints in representing the unconstrained parameters.

When neighboring bands of parameters do not differ, either the individual means are evaluated for alternative placement or the total number of bands is decreased. This process continues until the neighboring bands are the least number needed and where each band obtains a significant separation from each other. That is, bands that remain different reflect unequivocal differences between the grouped values. When interpreting the meaning of any treatment effects, it is the preponderance of evidence across time and between groups that indicates the nature of a treatment effect or the lack of one.

When using the chi-squared difference test to evaluate the tenability of the constraints, the conclusion that a parsimonious set of parameters adequately describes the freely estimated parameters is supported when the chi-squared difference test is nonsignificant at $p >= .2$. Given the power of this large-sample approach, a *p*-value of >= .2 would indicate essential equivalence between the freely estimate parameters and the constrained parameters, thereby strengthening any conclusions regarding where parameters are equal and where they are different (see e.g., Niolon et al., 2019). As mentioned, the tolerance of the constrained parameters is further evaluated by examining the residual difference between the constrained and unconstrained

parameters (which should average zero and be normally distributed) as well as the difference in fit of the model with constrained bands compared to the model with no constrained parameters. Again, the constrained model should differ from the freely estimated and unconstrained parameters at $p >= .2$, although the level of this criteria is a matter of choice on the part of the investigator. This approach to hypothesis testing reduces the sampling variability of pair-wise comparisons across discretely estimated means that may be of little practical value when statistical power is high. The choice of p-value is a matter of choice, but we recommend >= .2 to make a strong case that the constrained estimates faithfully reproduce the unconstrained estimates.

In most evaluations of a treatment, baseline equivalence across groups would be tested first. If the chi-squared difference test revealed significant decrement in fit ($p < .2$), then the set of constraints is not supported and baseline equivalence across all groups would be re-evaluated. When indicated, baseline equivalence would then be examined as moderated by known grouping factors such as cohort and/or gender. In other words, constraints would be imposed within gender and/or within cohorts for baseline equivalence between treatment and control. After moderated baseline equivalence is established then the tests of treatment effects would be specified and the same pattern of moderation of baseline equivalence would be allowed to carry through the latter time points of the treatment evaluation.

The constraining process is best conducted in a “wesearch” framework (Little, 2015). Namely, the team of investigators mutually examine the model constraints and parameter deviations in order to avoid local misspecification obscured by global model fit – the hallmark of model evaluation when confirmatory modeling is employed. The rigor of a confirmatory approach is in the forethought that the team of investigators brings to the modeling process. As with any modeling endeavor, the goal of this approach is overall error management. The traditional balance of Type I and II errors is no longer relevant in the broader confirmatory modeling context (Little et al., 2017) where the premium is placed on avoiding an omnibus Type

II error – a failure to ascertain when and how an intervention worked when it really did.

Limitations and Caveats

With the approach outlined here, a number of potential pitfalls and abuses could occur. First, two or more approximately equivalently fitting models are possible and can be constructed for a given evaluation. Second, unscrupulous application of the approach is possible. There are also several aspects of LMG that require more methodological and simulation research to properly guide researchers. While the LMG approach allows balanced management of Type I and Type II errors, the extent to which enforcing parameter parsimony contributes to balancing these errors for specific program effectiveness tests is unknown. There are also no established guidelines for directly comparing competing constraint selections, which can for now only be done through reference to the unconstrained model. Research into the use of relative fit indices for comparing constraints could allow the establishment of better guidelines. There are also no established guidelines for guarding against local misfit in the model when isolated parameters are inappropriately constrained. Finally, the universal application of the LMG approach to other modeling procedures such at Latent Transition Analysis and Mixture Modeling has not been scrutinized.

As with the general confirmatory modeling framework, theory, principles, and "wesearch" teams (Little, 2015) generate optimal models to be tested against data and wise judgement after careful scrutiny of local misfit is obligatory. Also, transparency via abundant online support material and full data sharing is warranted. With the LMG approach the sufficient statistics of each group modeled can be shared with impunity. LMG can be done on sufficient statistics only; raw data are not required. Sharing would allow another "wesearch" team the opportunity to generate a principled modeling approach that may lead to alternative conclusions but more likely would generate a validation of the carefully considered principles employed in the model building process. In our view, alternate team replication would be a more useful check and balance to unscrupulous application than even today's mandates for pre-

registration. The current zeitgeist of pre-registration does not presage the unanticipated changes to a protocol that too frequently are necessitated by the wants and whims of non-laboratory trials such as school- or community-based interventions. Pre-registering the approach outlined here would allow flexibility to principally accommodate the unexpected and still provide a rigorous and effective evaluation.

Conclusion

Multilevel approaches (MLM and ML-SEM) are quite useful to analyze clustered data, particularly when level 2 predictors are of interest. Their usefulness, however, does have limits when the primary goal is standard error correction. The LMG approach outlined here does not examine level 2 associations but allows controlling for the clustered nature of the data. LMG puts a premium on avoiding an overall Type II error of failing to identify when and where any program effects are evident. In turn, this focus has several related implications that warrant considering the approach. First, minimizing Type II error can maximize return on costly evaluation research. It also reduces the likeliness of obscuring isolated effects (or non-effects) for key subgroups, thus contributing to establishing equity. The LMG approach can also be framed as exploratory, to be followed by replication for interesting groups or time points. Accordingly, while the LMG approach is rarely utilized for evaluation research in many fields, it is a viable alternative modeling approach, which for many applications in the educational sciences may even be a preferred approach over MLM or ML-SEM. Moreover, we encourage researchers to consider this approach during the pre-registration process, which should detail the hierarchy of principles that would be employed during the formal intervention evaluation.

References

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269-296). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research, 38*(4), 529-569. doi:10.1207/s15327906mbr3804_5

Drechsler, J. (2015). Multiple Imputation of Multilevel Missing Data—Rigor Versus Simplicity. *Journal of Educational and Behavioral Statistics, 40*(1), 69-95. doi:10.3102/1076998614563393

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(1), 128-141. doi:10.1207/S15328007SEM0801_7

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York, NY: The Guildford Press.

Enders, C. K. (2017). Multiple imputation as a flexible tool for missing data handling in clinical research. *Behaviour Research and Therapy, 98*, 4-18. doi:10.1016/j.brat.2016.11.008

Gelman, A. (2006). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics, 48*(3), 432-435. doi:10.1198/004017005000000661

Gonzalez, R. & Griffin, D. (2001). Testing parameters in structural equation modeling: Every ”one" matters. *Psychological Methods*, 6, 258-265.

Graham, J. W., Cumsille, P. E., & Shevock, A. E. (2012). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of Psychology, Vol. 2: Research Methods in Psychology* (2nd ed.). New York, NY: Wiley.

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology, 29*(1), 158-167. doi:10.1093/ije/29.1.158

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple Imputation of Multilevel Missing Data: An Introduction to the R Package pan. *SAGE Open, 6*(4). doi:10.1177/2158244016668220

Howard, W. J., Rhemtulla, M., & Little, T. D. (2015). Using principal component analysis (PCA) to obtain auxiliary variables for missing data estimation in large data sets. *Multivariate Behavioral Research, 50,* 285-299. doi: 10.1080/00273171.2014.999267

Hox, J. J. (2013). Multilevel regression and multilevel structural equation modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods, Vol. 2: Statistical Analysis* (pp. 281-294). New York, NY: Oxford University Press.

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. New York, NY: The Guilford Press.

Lang, K. M., Little, T. D., & PcAux Development Team (2018). PcAux: Automatically extract auxiliary features for simple, principled missing data analysis [R Package]. Retrievable from https://github.com/PcAux-Package/PcAux/

Larsen, R. (2011). Missing Data Imputation versus Full Information Maximum Likelihood with Second-Level Dependencies. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(4), 649-662. doi:10.1080/10705511.2011.607721

Lei, P.-W., & Wu, Q. (2007). Introduction to structural equation modeling: Issues and practical considerations. *Educational Measurement: Issues and Practice, 26*(3), 33-43. doi:10.1111/j.1745-3992.2007.00099.x

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: Guilford Press.

Little, T. D. (2015). Methodological practice as matters of justice, justification, and the pursuit of verisimilitude. *Research in Human Development, 12,* 268-273. doi:10.1080/15427609.2015.1068044

Little, T. D., Bovaird, J. A., & Widaman, K. F. (2006). On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables.

Structural Equation Modeling, 13, 497-519. doi:10.1207/s15328007sem1304_1

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*(2), 151-173. doi:10.1207/s15328007sem0902_1

Little, T. D., & Lopez, D. F. (1997). Regularities in the development of children's causality beliefs about school performance across six sociocultural contexts. *Developmental Psychology, 33*, 165-175. doi:10.1037/0012-1649.33.1.165

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

Little, T. D., Widaman, K. F., Levy, R., Rodgers, J. L., & Handcock, G. R. (2017). Error, error, in my model, who's the fairest of them all. *Research on Human Development, 14,* 271-286. doi:10.1080/15427609.2017.1370965

Matsunaga, M. (2008). Item parceling in structural equation modeling: A primer. *Communication Methods and Measures, 2*(4), 260-293. doi:10.1080/19312450802458935

McArdle, J. J., & Hamagami, F. (1996). Multilevel models from a multiple group structural equation perspective. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Structural Equation Modeling: Issues and Techniques* (pp. 89-124). Mahwah, NJ: Lawrence Erlbaum Associates.

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods, 9*(3), 369-403. doi:10.1177/1094428105283384

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., . . . Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, Article no. c869. doi:10.1136/bmj.c869

Niolon, P.H., Vivolo-Kantor, A.M., Tracy, A., Latzman, N.E., Little, T.D., DeGue, S.,

Lang, K.M., Estefan, L.F., Ghazarian, S. R., McIntosh, W. L., Taylor, B., Johnson, L., Kuoh, H. Burton, T., Fortson, B., Mumford, E. A., Nelson, S., Joseph, H. Valle, L. A. & Tharp, A.T. (2019). An RCT of dating matters: Effects on teen dating violence and relationship behaviors. *American Journal of Preventive Medicine*, *57*, 13-23

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research, and Evaluation, 7*(1), Article 1. doi:10.7275/pmgn-zx89

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85-112. doi:10.1016/j.jsp.2009.09.002

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2012). Multilevel structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 512-531). New York, NY: The Guildford Press.

Rhemtulla, M. (2016). Population performance of SEM parceling strategies under measurement and structural model misspecification. *Psychological Methods, 21*, 348-368. doi:10.1037/met0000072

Rhemtulla, M., & Hancock, G. R. (2016). Planned Missing Data Designs in Educational Psychology Research. *Educational Psychologist, 51*(3-4), 305-316. doi:10.1080/00461520.2016.1208094

Rioux, C., Lewin, A., Odejimi, O. A. & Little, T. D. (2020). Reflection on modern methods: Planned missing data designs for epidemiological research. *International Journal of Epidemiology*. Advance online publication. doi:10.1093/ije/dyaa042.

Rioux, C., & Little, T. D. (2019). Missing data treatments in intervention studies: What was, what is, and what should be. *International Journal of Behavioral Development*. Advance online publication. doi:10.1177/0165025419880609

Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology, 5*, Article no. 81. doi:10.3389/fpsyg.2014.00081

Seaman, S., Galati, J., Jackson, D., & Carlin, J. (2013). What is meant by "Missing at

Random"? *Statistical Science, 28*(2), 257-268. doi:10.1214/13-sts415

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE Publications Inc.

Thompson, M. S., & Green, S. B. (2006). Evaluating Between-Group Differences in Latent Variable Means. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (pp. 119-169). Greenwich, CT: Information Age Publishing.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual review of clinical psychology, 1*, 31-65. doi:10.1146/annurev.clinpsy.1.102803.144239

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Boca Raton, FL: CRC Press.

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement Invariance. *Frontiers in Psychology, 6*(1064). doi:10.3389/fpsyg.2015.01064

Table 1. Strengths of the longitudinal multiple-group model (LMG), multilevel structural equation modeling (ML-SEM), and multilevel modeling (MLM)

Strengths	**LMG**	**ML-SEM**	**MLM**
Estimates relationships at lower and higher levels		X	X
Allows testing cross-level interactions		X	X
Addresses dependency issues related to nesting	X	X	X
Corrects for measurement error	X	X	
Allows testing the assumption of measurement invariance	X	X	
Provides comprehensive measures of model fit	X	X	
Flexibility in modeling the functional form of change over time	X	X	
Does not assume multivariate normal data	X	X	
Does not assume tau equivalent indicators of each construct	X	X	
Efficient for models with cross-classification over time	X	X	
Robust significance testing	X		
Allows testing the assumption of invariance across Level 2 units when latent factors are specified at Level 1	X		
Efficient for models with complex Level 1 interactions	X		
Efficient for models with more than 3 levels	X		
Nestedness treated as fixed effects	X		
High statistical power	X		

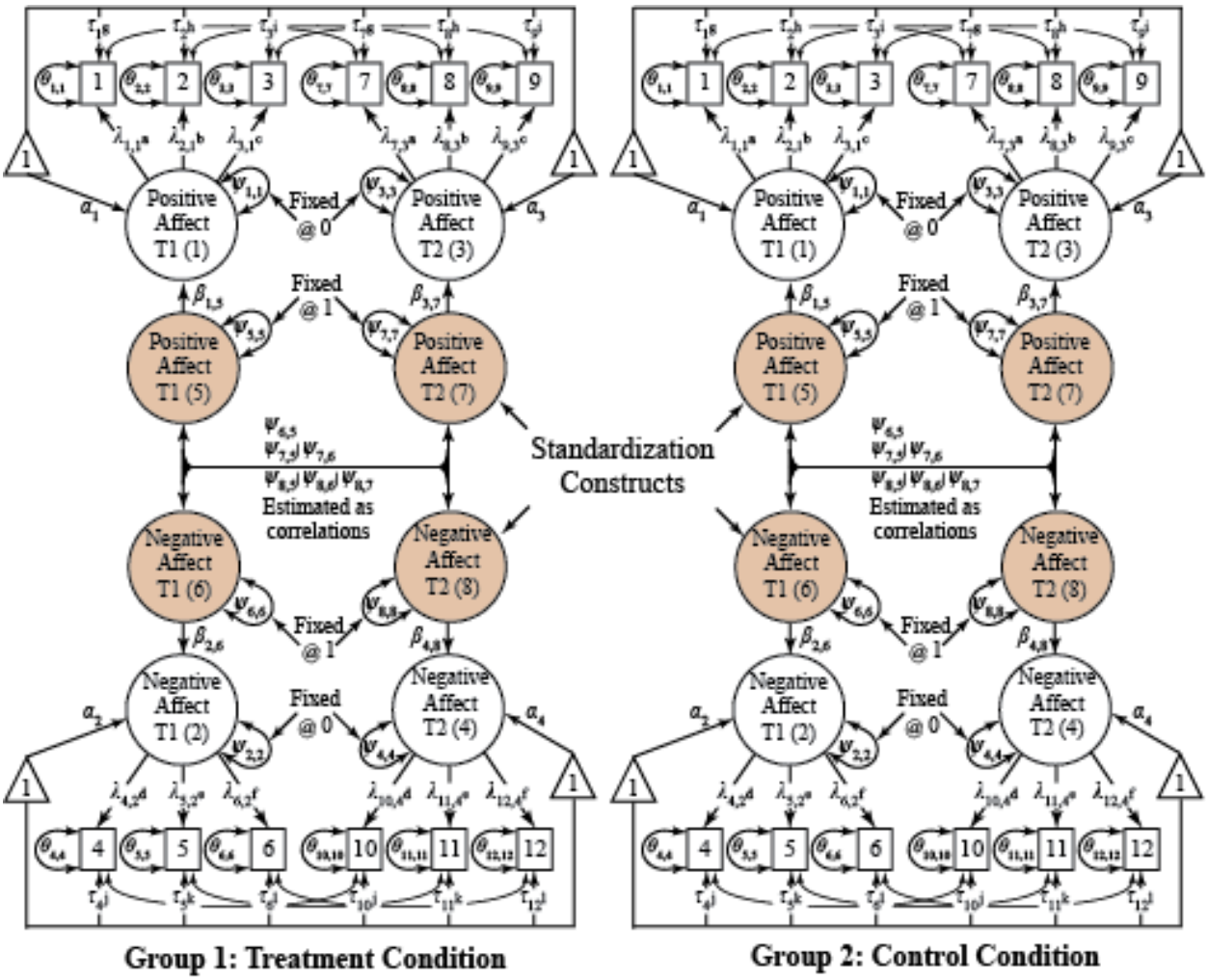

Figure 1. A Hypothetical Longitudinal Multiple Group Model for Evaluating an Intervention.

Note. Superscripted letters that are common indicate that the associated parameters are equated across time and across groups. Here, only a treatment and control group are represented but many more groups could be simultaneously estimated and tested. The scaling constraint for this model uses the effects-coded method of scaling (Little, Slegers et al., 2006). Indicators can be items or parcels. Invariance constraints derive loadings and intercepts from the whole sample providing a grounding and power for testing latent parameters across groups.