Journal of Family Psychology

The Role of Multiple-Group Measurement Invariance in Family Psychology Research

Justin L. Kern, Brent A. McBride, Daniel J. Laxman, W. Justin Dyer, Rosa M. Santos, and Laurie M. Jeans

Online First Publication, January 21, 2016. http://dx.doi.org/10.1037/fam0000184

CITATION

Kern, J. L., McBride, B. A., Laxman, D. J., Dyer, W. J., Santos, R. M., & Jeans, L. M. (2016, January 21). The Role of Multiple-Group Measurement Invariance in Family Psychology Research. *Journal of Family Psychology*. Advance online publication. http://dx.doi.org/10.1037/fam0000184

Journal of Family Psychology

© 2016 American Psychological Association
0893-3200/16/$12.00 http://dx.doi.org/10.1037/fam0000184

The Role of Multiple-Group Measurement Invariance in Family Psychology Research

Justin L. Kern and Brent A. McBride
University of Illinois at Urbana–Champaign

Daniel J. Laxman
University of Wisconsin–Madison

W. Justin Dyer
Brigham Young University

Rosa M. Santos
University of Illinois at Urbana–Champaign

Laurie M. Jeans
St. Ambrose University

Measurement invariance (MI) is a property of measurement that is often implicitly assumed, but in many cases, not tested. When the assumption of MI is tested, it generally involves determining if the measurement holds longitudinally or cross-culturally. A growing literature shows that other groupings can, and should, be considered as well. Additionally, it is noted that the standard techniques for investigating MI have been focused almost exclusively on the case of 2 groups, with very little work on the case of more than 2 groups, even though the need for such techniques is apparent in many fields of research. This paper introduces and illustrates a model building technique to investigating MI for more than 2 groups. This technique is an extension of the already-existing hierarchy for testing MI introduced by Meredith (1993). An example using data on father involvement in 5 different groups of families of children with and without developmental disabilities from the Early Childhood Longitudinal Study–Birth Cohort dataset will be given. We show that without considering the possible differential functioning of the measurements on multiple developmental groups, the differences present between the groups in terms of the measurements may be obscured. This could lead to incorrect conclusions.

Keywords: measurement invariance, structural equation modeling, multiple-groups analysis, father involvement

The use of latent variables in early intervention, early childhood special education, and family science research is becoming increasingly common (Connell et al., 2008). Latent variables, in some broad sense, are variables that are unobservable, or not directly measureable. The concept is that the latent variable is something a person (for instance, the father) possesses in some quantity; the "amount" of that variable that a person possesses probabilistically determines how the person responds behaviorally to a question or a task related to that variable. In a more practical sense, by combining the observed indicators one can identify a single latent variable, summarizing the indicators' information about a latent variable into a single score. Researchers can then test theoretically meaningful hypotheses more simply than would be possible by looking at the items separately. For instance, one question that could be asked with the data in our example is, "Do fathers of children with disabilities differ significantly from fathers of typically developing children on literacy involvement?" Furthermore, using latent variables allows for measurement error, which is abundant in behavioral science data, to be partially accounted for. If measurement error is not accounted for, then effects of dependent variables can be underestimated, or worse, in the wrong direction (Millsap, 1998). As the number of variables in a model increases, the consequences of measurement error become exacerbated (Bollen, 1989). Truly, as the number of scales and complexity of research questions being addressed by family science scholars increases, the machinery of latent variables becomes invaluable from the perspective of the researchers interested in family processes.

One of the downsides to using latent variables is that the scale of the construct is somewhat arbitrary, meaning that once the analysis is completed, the units of the latent variable are not necessarily meaningful. The units are determined primarily from

Justin L. Kern, Department of Psychology, University of Illinois at Urbana–Champaign; Brent A. McBride, Department of Human and Community Development, University of Illinois at Urbana–Champaign; Daniel J. Laxman, Waisman Center, University of Wisconsin–Madison; W. Justin Dyer, School of Family Life, Brigham Young University; Rosa M. Santos, Department of Special Education, University of Illinois at Urbana–Champaign; Laurie M. Jeans, Department of Early Childhood Education, St. Ambrose University.

This research was supported by a grant from the Institute for Education Sciences of the U.S. Department of Education to Brent A. McBride, Rosa M. Santos, Sungjin Hong, and W. Justin Dyer (R324A120174). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Institute for Education Sciences.

Correspondence concerning this article should be addressed to Justin L. Kern, Department of Psychology, University of Illinois at Urbana–Champaign, Champaign, IL 61820. E-mail: kern4@illinois.edu

the items to which the latent variable is related, so it is often necessary to try to make the items as similarly scaled as possible. As will be discussed shortly, a special consideration comes from multiple-group analyses in that the latent variable might not be on the same scale across groups, and may not even measure the same construct across groups. Since it is often the case that questions involving different groups are important in research examining families, this problem has to be sufficiently addressed, otherwise one should be suspicious of the conclusions that one comes to when using of latent variables across groups, like with differences between mothers and fathers or, like our example, different groups of fathers. An intuitive example illustrates the problem. Suppose the lengths of two pencils are both found to be 5, and the question, "Are the two pencils the same length?" is posed. We want to say yes, but in truth, we cannot, because we do not know if the lengths are measured in centimeters, inches, or some other unit of length. Without comparable units, direct numerical comparisons are meaningless. This problem underlying latent variable research creates challenges when trying to make comparisons across groups; that is, the scale used by mothers and fathers could be different. These challenges can by partially allayed through the use of measurement invariance techniques.

Measurement Invariance Defined

Measurement invariance (MI) is a key concern for analyses using latent variables (Vandenberg & Lance, 2000), particularly when studying group differences. MI is defined as the property that given a subject's factor score, his or her observed score is independent of their group membership. Thus, MI implies that the latent variables used truly measure the same constructs for all groups. It is often assumed that the latent variables are invariant across groups. If a latent variable is not the same across groups, then analyses using that latent variable must be interpreted with extreme caution. For example, most common statistical tests, such as the *t* test, assume MI. However, if this assumption is not met, then two broad categories of problems become apparent. First, because the comparison of the means of dissimilar constructs may not be quantitatively tractable, the usual statistical conclusions of interest (e.g., statistical significance, Type I error rate, power) are suspect. Second, and possibly of greater concern, the comparison of dissimilar constructs is meaningless, and therefore interpretation of mean differences is highly questionable. For instance, suppose that a measurement (say, how often a father plays with toys with his child) depends not just on his level of play involvement, but also on his group membership (i.e., being a father of a normally developing child vs. a father of a child with disabilities). In this sense, then the differences in how often a father plays with toys with his child are functions both of differences in level of "play involvement" (i.e., the construct we are interested in investigating) and the properties of the measurement for the group itself, and thus, differences in play involvement are confounded by the measurement. This is a big concern.

When researchers have investigated the assumption of MI, they often focused on cross-cultural groupings (Davidov, Schmidt, & Billiet, 2011). There is a consensus that to compare groups cross-culturally the measurements used for these comparisons should be invariant; this is considered an essential step in this field (Tran, 2009). This is also true within the longitudinal data analysis community, because modeling change across time assumes MI. Examining MI across other groupings is still not the norm within the parenting and family psychology literature, but an acknowledgment of its importance is becoming more commonplace elsewhere. Some examples of unique groupings include gender (Beyers & Goossens, 2008; Bagheri, Jafari, Tashakor, Kouhpayeh, & Riazi, 2014), chronic conditions (Schuler et al., 2014), disability status (Randall & Engelhard, 2010; MacLean, McKenzie, Kidd, Murray, & Schwannauer, 2011; Reynolds, Ingram, Seeley, & Newby, 2013), and test accommodations (Randall & Engelhard, 2010). It can be argued that regardless of the outcome in the study, the simple acknowledgment that MI should be examined is a step in the right direction.

While invariance is a desirable property, noninvariant latent variables can also be informative and provide evidence of differential processes across groups, an important issue when constructing measurement tools of all purposes. Noninvariant structures show that the construct under investigation has response biases in some way for the different groups. If invariance is not checked, this response bias could be completely overlooked. For example, it may be that fathers think about the interactions with their children differently dependent on the child's disability status. As a result, survey questions about those interactions may themselves be interpreted differently. Rather than necessarily showing a meaningful quantitative difference in the father–child interaction, this gives evidence that responses from fathers of children with different disability statuses rely on different underlying response processes. As such, a lack of invariance in a particular factor focuses the researcher to a specific place in which they can carry out further studies.

Testing Measurement Invariance

For testing MI, a common framework is the multiple-group factor analysis model. The constraints for testing are convenient to integrate into the model, and a large literature on model fitting is already well established (Bollen, 1989). Furthermore, many statistical software programs can easily accommodate this framework (e.g., Mplus, LISREL, SPSS Amos). For this paper, a multiple-group structural equation model (SEM) framework will be used. For an introduction to this, see Millsap (2011).

One standard MI procedure within the multiple-group SEM framework involves a four-stage nested hierarchy of parameter constraints (Meredith, 1993), which will be elaborated on later. Unfortunately, this hierarchy makes the assumption that the overall structure of the model is the same for all groups. While this is reasonable for two groups, in many of the research areas found in family psychology this seems to be a very strict assumption when there are more than two groups. For example, consider the situation in which there are three groups. Examples could include biological fathers, adoptive fathers, and father figures; mothers, fathers, and step-parents; White families, Black families, and Latino families; or normally developing children, children with cerebral palsy, and children with autism spectrum disorder. While a measure may have strong invariance between two groups, there is no reason to believe that that same structure must be present with respect to the third group. It does not seem far-fetched to believe that some items have different meaning for the third group than they do on the first two groups. By simply following the four-stage hierarchy, this possibility is ignored. On the other hand, when different stages of the hierarchy are met, nice properties and interpretations are present. Thus, there needs to be reconciliation between the ability to apply Meredith's (1993) hierarchy and the recognition that many intermediate forms of invariance are missing from that hierarchy.

It is possible to consider this problem as a model-building problem. As a simple approach in our research on fathers, we propose using pairwise modeling of groups rather than simply using an omnibus approach in which all groups are considered together. In this way, each pairing is considered under the invariance hierarchy, and a decision is made about how the two groups compare to each other. After decisions are made comparing the pairwise groups, these unitary decisions may then be recombined into a single model.

In the following, we describe in detail the method of multiple-group MI, and provide an example of this method using five different groups of fathers of children with and without disabilities. The example is derived from the Early Childhood Longitudinal Study–Birth Cohort (ECLS-B) dataset. This is followed by a discussion and recommendations on how to proceed with a multiple-group MI study.

Pairwise Method for Multiple-Group Measurement Invariance Analysis

The process for conducting a multiple-group MI analysis involves two steps: (1) a traditional MI analysis is done for each pair of groups and (2) a model is built that contains (as closely as possible) all the levels of invariance contained in the pairwise MI analyses.

A traditional MI analysis (that is, an MI analysis for two groups) under the SEM framework is given in the following, involving a four-stage nested hierarchy of parameter constraints (Meredith, 1993). First, we examine whether the factor structure under the two groups is similar, that is, that the same observed measures load onto the same factors, and only on those factors. If that is the case, then the factor structure demonstrates *configural invariance* with respect to the two groups. If this model holds, then we try to determine in what ways the factor structure is the same (or different) for the two groups. To do this, a series of constraints are placed. This series of constraints defines a set of nested models that is designed to meet increasingly higher levels of invariance. Traditionally, consecutive models in the hierarchy are each tested against each other using a chi-square difference test to determine if the model improved (or rather, if the model's fit did not significantly drop).

The first set of constraints is used to define *weak invariance* (Widaman & Reise, 1997), or *metric invariance* (Thurstone, 1947). Here, the factor loadings for the same items on the same factors are set equal to each other across both groups. For a statistical test of whether this constraint should be set, a chi-square difference test between the configural invariance model and the weak invariance model can be conducted. If this model holds, then group differences in latent variable variances and covariances become invariant under rescaling, and thus, any substantive interpretation of group differences in variances or covariances among latent variables will remain invariant over rescalings of the latent variables. However, no true MI conditions, as defined by Meredith (1993), are met.

In the second set of constraints, we set the intercepts to be equal across the two groups. These constraints, along with the factor loading constraints in the weak invariance model, define *strong invariance*, also called *scalar invariance* (Steenkamp & Baumgartner, 1998). As before, a chi-square difference test between the strong invariance model and the weak invariance model can be used to test if these extra constraints can be placed. When the model holds, the group differences in means for the observed variables are attributable solely to population differences in common factor means, and so there is no effect of measurement bias on the observed means (Millsap, 1998). Furthermore, though different latent variable scalings lead to different latent variable means, the relative group differences in the means will remain, and, thus, model interpretations of the means will be invariant to rescaling. As such, Widaman and Reise (1997) argued that strong invariance is the minimum stage necessary for mean comparisons of latent variables across groups.

In the last set of constraints considered here, the error variances for corresponding observed variables are set equal to each other across the groups. These constraints in conjunction with the previous two sets of constraints (loadings and intercepts) define the *strict invariance* model. A final chi-square difference between the strict invariance model and the strong invariance model tests whether invariance of the error variances across groups is supported by the data. If this model holds, then differences in group covariances in the manifest variables are due solely to group differences on the latent variables. This is also equivalent to failing to detect measurement bias.

There are a few points to recognize while making the decision of which level of invariance the groups meet. First, one must note while advancing through the models that *more* constraints are placed, and so we expect that the model fit in terms of the chi-square statistic will decrease. Thus, to transition from one model to the next in the hierarchy in terms of the chi-square difference tests, the test should be found to be nonsignificant at some prespecified level. Second, it is important to keep in mind the discussion about when to use statistical tests and when to use model-fit indices when deciding the "significance" of a model. In summary, the argument is that large samples are often necessary for usage of many statistical tests, but that by virtue of the large samples, any divergence from the null hypothesized model will result in a significant test result. The goal with chi-square tests in structural equation modeling, however, is often to *not* reject a null hypothesized model. Thus, even though the researcher may see little difference between model-implied moments (i.e., means and covariances) and sample moments, a model may be deemed unsuitable, statistically. To counteract this, many researchers have proposed model-fit indices, such as comparative fit index (CFI), root-mean-square error of approximation (RMSEA), and Akaike information criterion (AIC), that fit the goals of the researchers (that is, to find a model that suitably represents the relationships present in the data) better than the chi-square tests. As such, we also find it important to take into account the values on many model-fit indices when making decisions pertaining to MI. In our example, we will use CFI, RMSEA, and AIC, along with differences in CFI and McDonald's noncentrality index (MNCI; Cheung & Rensvold, 2002) for making decisions. This framework is general enough to be used in many contexts.

After decisions are made regarding the pairwise levels of invariance, an attempt to build a single model, called the combined model, containing all groups and their relationships with each other simultaneously is undertaken. The goal is to preserve the relationships between the groups (in terms of invariance) as closely as possible, which amounts to finding the set of constraints that allows for these relationships. First, we find what each of the individual invariance decisions means in terms of constraints. This is a simple process of recording the sets of parameters that are implied to be equal to each other. For instance, suppose we have three groups and that the measurement is strictly invariant to Groups 1 and 2, strongly invariant to Groups 1 and 3, and weakly invariant to Groups 2 and 3. Then, we know the following: $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2$, $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2$, $\mathbf{\Theta}_1 = \mathbf{\Theta}_2$, $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_3$, $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_3$, and $\mathbf{\Lambda}_2 = \mathbf{\Lambda}_3$, where $\mathbf{\Lambda}_\mathrm{g}$ is the matrix of factor loadings for group g, $\boldsymbol{\tau}_\mathrm{g}$ is the vector of intercepts for group g, and $\mathbf{\Theta}_\mathrm{g}$ is the matrix of error variances for group g. Note that implicitly we have also determined a set of nonequivalences, by virtue of the invariance decisions made previously. Here, they are: $\boldsymbol{\tau}_2 \neq \boldsymbol{\tau}_3$, $\mathbf{\Theta}_1 \neq \mathbf{\Theta}_3$, and $\mathbf{\Theta}_2 \neq \mathbf{\Theta}_3$. From these relationships, we can then determine for certain that $\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2 = \mathbf{\Lambda}_3$ and $\mathbf{\Theta}_1 = \mathbf{\Theta}_2$.

Notice that for the intercepts there is an inconsistency in our usual notions of equality; a cursory glance shows that $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2$ and $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_3$, but that $\boldsymbol{\tau}_2 \neq \boldsymbol{\tau}_3$. However, these are vectors and so one way to reconcile this is to consider the possibility of partial invariance. Partial invariance implies that some of the parameters in a set are invariant to the group, but not all of them. Thus, only some of the intercepts are equal, and the task is to find the ones that are. From our decisions ($\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2$ and $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_3$), it is readily shown that this implies that $\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2 = 0$ and $\boldsymbol{\tau}_1 - \boldsymbol{\tau}_3 = 0$. Our approach, then, is to have all intercepts freely estimated, test the differences in intercepts, and sequentially add equality constraints by constraining the intercepts with the smallest difference to be equal. The model is refit with the new constraint and the fit statistics are recorded. This process is continued until the model's fit no longer improves. In general, this will be done for any set of constraints that are inconsistent with the properties of equality. For this portion of the process, we used AIC as the guiding fit statistic (i.e., as AIC decreased, we continued the process); other statistics are possible to use as well, but in our experience, AIC does the best job at model selection in this process. After this process is completed, the full model can be interpreted for MI. The full pairwise method is shown in Figure 1.

The pairwise method described here uses only factor loadings, intercepts, and error variances to assess MI. While not discussed here, other parameters can also be used in this process, such as latent variable covariances, and thresholds in the case of item factor analysis. The underlying method for assessing MI will be similar, but with additional steps in the decision-making in the first step.

Simulation Study

Design

To determine the viability of this method, a simulation study was designed to assess performance. For each of 1,000 replications in the simulation, three groups were sampled according to sample size and factorial invariance complexity. Sample size had four levels (50, 100, 200, and 300 samples per group). Invariance complexity was at four levels: and all groups configurally invariant to each other (IC1), all groups strictly invariant (IC2), two groups strictly invariant to each other with one group configurally invariant to the others (IC3), and two groups strictly invariant to each other with one group strongly invariant to the others (IC4). The simulation was limited to having only five indicators on a single latent variable.

For each replication, we proceeded to build the invariance model using both the standard method (i.e., placing constraints at each of the four levels of the invariance hierarchy for all groups simultaneously) and with the pairwise method. The performance of the methods at capturing true invariance was measured via the proportion of constraints correctly placed on the model; that is, we simply counted the number of constraints/nonconstraints correctly placed and divided by the total number of constraints/nonconstraints. Altogether, there were (3 × 3 × 5 =) 45 possible constraints/nonconstraints (three pairs of groups with three parameters for each of five items). MNCI, RMSEA, and CFI were collected for each replication.

Results

For each combination of conditions (invariance complexity, sample size, and MI method), CFI, RMSEA, MNCI, and match percentage were all averaged over all replications. Overall, we find that the pairwise method outperforms, or does approximately the same as, the standard methods for all conditions, except for IC1, on all metrics; this is shown in Table 1.

The main metric to consider when evaluating the methods' performances is the match percentage, as this directly measures the performance of the method. Investigation shows that as sample size increases, match percentage increases. We also find that the conditions with all groups at the same level of invariance (IC1 and IC2) have a higher match percentage than the other more complex invariance structures (IC3 and IC4). Furthermore, within the less complex invariance conditions, the standard method has a higher match percentage the pairwise method, though the difference is slight for IC2. On the other hand, for the more complex invariance conditions, the pairwise method has a much higher match percentage than the standard method.

While the match percentage is the best metric for evaluating these methods, it is not available when the true invariance structure is unknown, as is the case when encountering real data. Therefore, it is also important to know the performance of the methods in terms of standard metrics, such as the standard fit measures CFI, RMSEA, and MNCI. We find that these perform very similarly to match percentage in all cases, except for the IC2 and for a sample size of 50 in the IC1 condition; match percentage slightly prefers the standard method, whereas CFI, RMSEA, and MCNI all slightly prefer the pairwise method. For the IC2 condition, however, the preference is very slight, with only a difference of about 1% in the match percentage. The other conditions have much higher discrepancies in match percentage, and thus, the best course of action seems to be to always prefer the model with the best set of fit measures, as this will lead to a higher match percentage, on average.

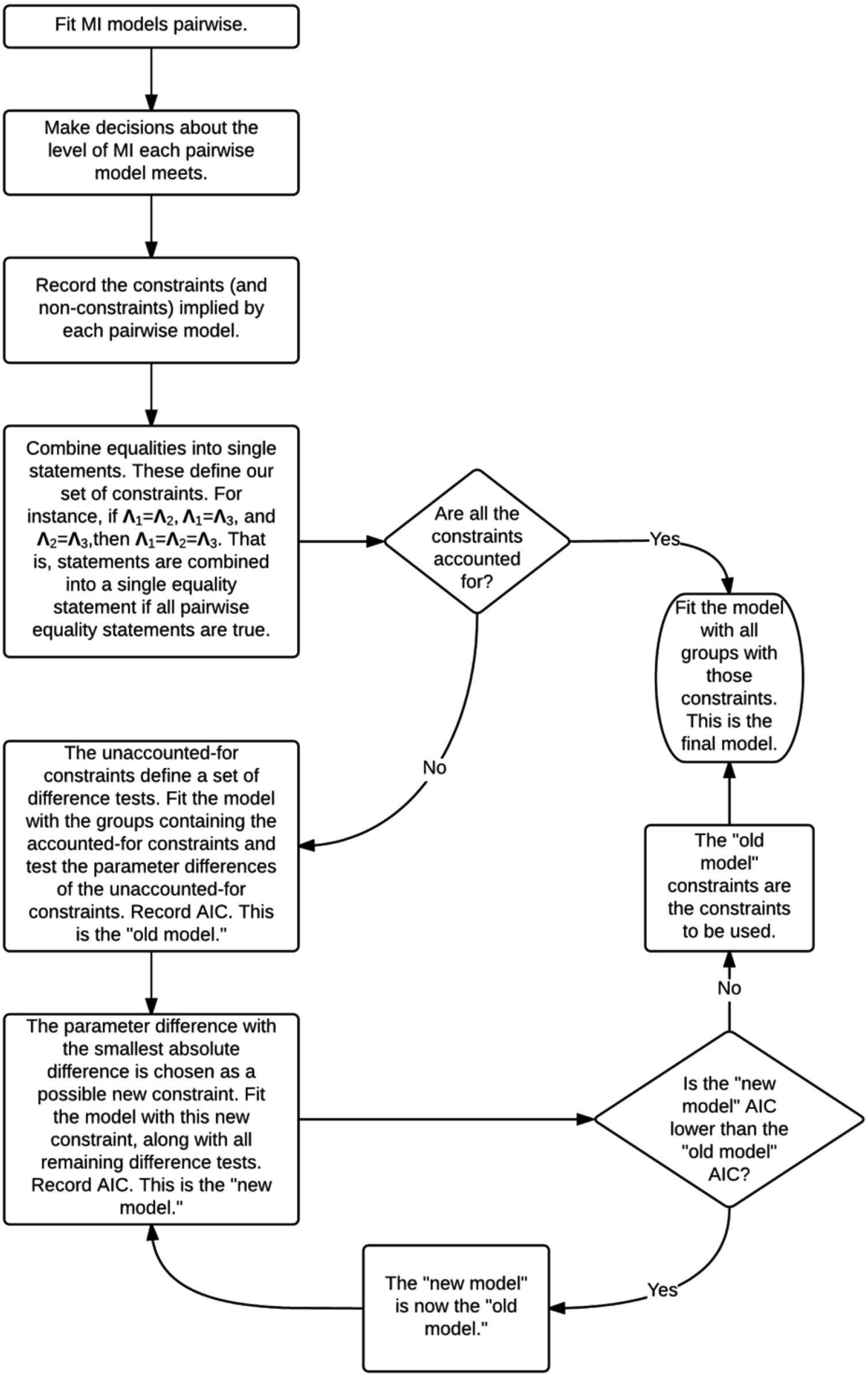
Fit MI models pairwise.
Make decisions about the level of MI each pairwise model meets.
Record the constraints (and non-constraints) implied by each pairwise model.
Combine equalities into single statements. These define our set of constraints. For instance, if $\Lambda_1=\Lambda_2$, $\Lambda_1=\Lambda_3$, and $\Lambda_2=\Lambda_3$,then $\Lambda_1=\Lambda_2=\Lambda_3$. That is, statements are combined into a single equality statement if all pairwise equality statements are true.
Are all the constraints accounted for?
Yes
Fit the model with all groups with those constraints. This is the final model.
No
The unaccounted-for constraints define a set of difference tests. Fit the model with the groups containing the accounted-for constraints and test the parameter differences of the unaccounted-for constraints. Record AIC. This is the "old model."
The parameter difference with the smallest absolute difference is chosen as a possible new constraint. Fit the model with this new constraint, along with all remaining difference tests. Record AIC. This is the "new model."
Is the "new model" AIC lower than the "old model" AIC?
No
The "old model" constraints are the constraints to be used.
Yes
The "new model" is now the "old model."

Figure 1. Flowchart of the pairwise method for multiple-group measurement invariance (MI). AIC = Akaike information criterion.

Table 1
Simulation Results for All Conditions

Condition	Sample size	Method	CFI	RMSEA	MNCI	Match %
IC1	50	Standard	.968	.039	.994	88.43
IC1	50	Pairwise	.972	.034	.998	82.01
IC1	100	Standard	.986	.027	.998	96.77
IC1	100	Pairwise	.980	.032	.996	87.54
IC1	200	Standard	.995	.016	1.000	100.00
IC1	200	Pairwise	.989	.023	.998	91.50
IC1	300	Standard	.996	.014	1.000	100.00
IC1	300	Pairwise	.994	.018	.999	92.54
IC2	50	Standard	.977	.024	1.001	84.37
IC2	50	Pairwise	.988	.013	1.015	84.02
IC2	100	Standard	.986	.018	1.000	94.33
IC2	100	Pairwise	.991	.011	1.005	92.84
IC2	200	Standard	.992	.013	1.000	98.77
IC2	200	Pairwise	.995	.010	1.001	97.04
IC2	300	Standard	.994	.011	1.000	99.67
IC2	300	Pairwise	.995	.010	1.000	98.96
IC3	50	Standard	.966	.040	.991	52.27
IC3	50	Pairwise	.980	.022	1.004	71.86
IC3	100	Standard	.981	.033	.994	53.74
IC3	100	Pairwise	.988	.018	1.001	82.59
IC3	200	Standard	.992	.023	.998	55.94
IC3	200	Pairwise	.994	.013	1.000	92.33
IC3	300	Standard	.995	.017	.999	56.47
IC3	300	Pairwise	.996	.012	1.000	96.35
IC4	50	Standard	.975	.027	.999	78.43
IC4	50	Pairwise	.985	.016	1.010	82.67
IC4	100	Standard	.986	.020	1.000	85.30
IC4	100	Pairwise	.991	.013	1.004	92.63
IC4	200	Standard	.992	.015	1.000	87.62
IC4	200	Pairwise	.994	.011	1.001	96.76
IC4	300	Standard	.995	.012	1.000	88.46
IC4	300	Pairwise	.995	.010	1.000	98.66

Note. CFI = comparative fit index; RMSEA = root-mean-square error of approximation; MNCI = McDonald's noncentrality index. All groups configurally invariant to each other (IC1), all groups strictly invariant (IC2); the conditions are as follows: two groups strictly invariant to each other with one group configurally invariant to the others (IC3), two groups strictly invariant to each other with one group strongly invariant to the others (IC4).

Illustration of Multiple-Group Measurement Invariance

Here, we will describe in the data used in our example in detail, and then illustrate how to apply our multiple-group MI technique to these data. We used Mplus for our analyses. A series of pairwise invariance steps are made using the MI hierarchy introduced by Meredith (1993) and discussed above (some mathematical detail appears in the Appendix). After decisions regarding the level of invariance between the groups are made, an attempt to reunite the groups into a single model is made. Because this is essentially a model selection procedure rather than a hypothesis testing procedure, and due to recommendations in the literature (Cheung & Rensvold, 2002), all intermediary models and the final model are assessed via model fit indices (RMSEA, CFI, MNCI).

Data

Data come from the first three waves (9 months, 2 years, and 4 years) of the ECLS-B, a nationally representative, longitudinal dataset of children born in 2001 in the United States. Children in the ECLS-B sample come from a variety of socioeconomic and racial/ethnic backgrounds, and the study included oversampling of low- and very-low-birth weight children, Chinese children, other Asian and Pacific Islander children, American Indian and Alaska Native children, and twins. Data was extracted from the ECLS-B for 3,750[1] children that had the same resident father/father figure at all three time-points.

Disability status of the child. For the purposes of the current study, children were categorized into five groups based on their disability status: typically developing (TYP; N = 2900), autism spectrum disorder (ASD; N = 50), cerebral palsy (CP; N = 50), other developmental delay (OthDD; N = 100), and other (OTH; N = 450; Hvidtjørn et al., 2009).

Father involvement. Fathers were asked at the 9-month, 2-year, and 4-year time-points to report how frequently they engaged in several different activities. Response categories were based on 4-point or 6-point scales ranging from never to every day or more than once a day. In the original metric, a 1-point difference may represent the difference between once a week and two or three times a week, or between seven times a week and 14 or more times per week. Thus, the original metric is not equal-interval. To address this limitation, father involvement items were rescaled to represent the approximate number of times per week the father engaged in the activity. Although this approach is not perfect given that number of times per week has to be approximated (e.g., it is assumed that "few" means twice a month and "more than once a day" means twice a day), rescaling yields a more meaningful scale that is closer to reality. A similar approach to rescaling can be found in previous research using ECLS-B data with items that use these same response categories (e.g., Dyer, McBride, Santos, & Jeans, 2009).

For our example, we will focus on three constructs which were chosen to illustrate different levels of model-building complexity within our method. This can be a complicated process when there are "in-between" or "transition" groupings, particularly if we insist on being able to have subsets of the groups have some form of traditional invariance (weak, strong, or strict) to one another. The different involvement constructs in order of complexity are literacy involvement (engagement in reading or other language activities with their children) at 9 months, literacy involvement at 4 years, and routine caregiving involvement (engagement in providing care for their child's functional needs) at 2 years. Table 2 shows the items used in constructing each measure in our example.

Analysis Plan

For purposes of comparison, we will do an overall "omnibus" MI procedure—the standard method including all five of the groups simultaneously—in addition to the multiple-group MI procedure. We will call the models built in this way the *five-group models*. This will be conducted separately for each of the father involvement factors. All decisions are made using a variety of model fit indices (viz., CFI, RMSEA, MNCI), because this is a model-building approach rather than a hypothesis-testing approach; the model-building approach has traditionally relied on a

[1] Per National Center for Educational Statistics requirements when using ECLS-B data, all Ns in this paper were rounded to the nearest 50.

Table 2
Items Used in the Construction of the Father Involvement Factors

9-Month literacy	4-Year literacy	2-Year routine caregiving
How often do you . . .	How often do you . . .	How often do you . . .
1. read books to your child?	1. read books to your child?	1. help your child get dressed?
2. tell stories to your child?	2. tell stories to your child?	2. feed your child?
3. sing songs to your child?	3. sing songs to your child?	3. put your child to sleep?
		4. give your child a bath?

combination of model fit indices for decision-making purposes. Standard cut-off values for CFI, MNCI, and RMSEA (CFI and MNCI >0.95, and RMSEA <0.06) were used (Hu & Bentler, 1999; Hooper, Coughlan, & Mullen, 2008).

Furthermore, Cheung and Rensvold (2002) have suggested using differences in CFI and MNCI between models to guide decision-making. They suggest that values less than or equal to −0.01 and −0.02 for ΔCFI and ΔMNCI, respectively, indicate that the more constrained model is appropriate, but that these values are flexible. However, it must be noted that while these cut-offs, the standard cut-offs for CFI and RMSEA, and the value for MNCI (higher is better) were considered, decisions were made in the overall context of the other indices used for making decisions so that our judgment would be less cloudy (Crowley & Fan, 1997). Generally speaking, a model was chosen when a majority of indices indicated that that model was best in comparison to the others.

Finally, it should be noted that parts of this example may have small sample problems, particularly in the ASD and CP groups. It could be argued, at least, that the use of RMSEA and CFI are fairly resistant to the effects of small samples even with moderate amounts of model misspecification (Hu & Bentler, 1998; Fan & Sivo, 2005). However, these data mainly serve as an example of usage of the multiple-group MI method and as an example of why disability groupings should be examined for MI; future research with larger samples within each group is needed to make more definitive conclusions.

Results

For each of the three factors we will focus on, our method was applied. The 9-month literacy involvement factor is an example of the simplest case for our method, where constraints are applied uniformly for all groups. For the other examples, the reconstruction process is more complex. The main goal is to have all of the individual decisions of invariance level between groups represented simultaneously in a single model. This amounts to placing constraints simultaneously to allow for this; that is, the constraints must be placed in such a way that when other groups are dropped, the pairwise invariance level decisions are still present.

For purposes of succinctness, statistical and model fit information for the models with all five groups for each time-point are in Table 3. Auxiliary information will be provided as necessary in subsequent tables.

Nine-Month Literacy Involvement

The five-group model showed strict invariance on the literacy involvement factor at 9 months (see Table 3). Further investigation showed that pairwise comparisons pointed to strict invariance for all groups.

In cases where all pairs have the same level of invariance with each other, such as for 9-month literacy involvement, the combined model would simply be the standard "omnibus" model, where the constraints are placed uniformly across groups. We find that strict invariance is met for all pairs. This would be enough evidence to suggest that when combined, the loadings, the intercepts, and the unique factor variances can all be constrained to be the same across all five groups, as in the five-group model.

Four-Year Literacy Involvement

As shown in Table 3, the five-group model showed strict invariance on the literacy involvement factor at 4 years. Pairwise comparisons, on the other hand, pointed to strict invariance for all groups, except CP. Table 4 shows the results of the pairwise models with the CP group at 4 years. In it, CP is shown to meet configural invariance with the other four groups, with some evidence that CP shows a higher level of invariance with TYP. The TYP and CP groups may have a higher level of invariance as indicated by RMSEA, and MCNI, but AIC and the drop in CFI shows that configural invariance is appropriate.

A quick examination shows that the pairwise decisions have no contradictions with respect to constraint placement, and so we can readily fit the combined model implied by the decisions by setting

Table 3
The Five-Group Model for All Time-Points on All Three Father Involvement Factors

Model	MNCI	CFI	RMSEA	AIC
9-Month literacy				
Configural	1.000	.999	.019	45,912.455
Weak	1.000	1.000	.000	45,911.843
Strong	.999	.996	.013	45,923.404
Strict	.998	.995	.011	45,921.333
4-Year literacy				
Configural	1.000	1.000	.000	45,217.378
Weak	.995	.988	.036	45,240.955
Strong	.987	.972	.040	45,260.316
Strict	.987	.969	.031	45,257.230
2-Year routine caregiving				
Configural	.956	.960	.095	73,324.529
Weak	.972	.975	.051	73,324.351
Strong	.974	.977	.039	73,321.924
Strict	.968	.972	.036	73,369.108

Note. MNCI = McDonald's noncentrality index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; AIC = Akaike information criterion.

Table 4
The Cerebral Palsy Pairwise Models for the 4-Year Time-Point on Literacy Involvement

Model	MNCI	CFI	RMSEA	AIC
CP vs. TYP				
Configural	1.000	1.000	.000	37317.11
Weak	.997	.985	.054	37334.85
Strong	.993	.963	.060	37353.65
Strict	.994	.971	.040	37351.69
CP vs. ASD				
Configural	1.000	1.000	.000	834.57
Weak	.851	.292	.328	844.89
Strong	.899	.536	.206	842.84
Strict	.899	.531	.163	846.76
CP vs. OthDD				
Configural	1.000	1.000	.000	2002.71
Weak	.943	.811	.243	2022.46
Strong	.899	.660	.231	2030.27
Strict	.906	.685	.168	2029.91
CP vs. OTH				
Configural	1.000	1.000	.000	6495.68
Weak	.987	.925	.095	6513.44
Strong	.970	.828	.111	6535.15
Strict	.965	.801	.101	6532.89

Note. MNCI = McDonald's noncentrality index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; AIC = Akaike information criterion; CP = cerebral palsy; TYP = typically developing; ASD = autism spectrum disorder; OthDD = other developmental delay; OTH = other.

the loadings, intercepts, and unique factor variances equal to each other for the Groups TYP, OTH, OthDD, and ASD, while allowing all these parameters to be freely estimated (i.e., not equal to the other four groups) for CP. This model was fit and was shown to be superior to the final five-group model for 4-year literacy, as shown in Table 5. Table 5 also includes the parameter estimates for this model.

Two-Year Routine Caregiving Involvement

For the routine caregiving involvement factor at 2 years, the five-group models indicated that strict invariance was met as shown in Table 3. However, an investigation of the pairwise group comparisons told a slightly different story; there is strict invariance for all groups except CP and ASD. The results of the pairwise models for the CP and ASD groups are shown in Table 6. For the CP, ASD, and OthDD groups, the invariance structure seems to be complex; while CP is configurally invariant with respect to the other groups, ASD actually meets strict invariance with respect to the OTH and TYP groups, but only configural with the other two groups.

From these pairwise decisions, we can determine that the loadings, intercepts, and error covariances for TYP and OTH are equal to each other, and also that CP has no constraints with any other group without any contradictions. However, because we also find that ASD and OthDD are both strictly invariant with respect to TYP and OTH, but only configurally invariant with each other, there is a contradiction in the implied model constraints. Thus, we find ASD and OthDD have some form of partial invariance to TYP and OTH. That means that some parameters are constrained to be equal while other parameters are allowed to be freely estimated. We used the pairwise method, as shown in Figure 1, to determine which parameters to constrain in the combined model. For this example, this process translated to (1) fitting a model with the loadings, intercepts, and error covariances for the TYP and OTH groups set equal to each other— the CP group has no constraints with any other group, and stays that way throughout; (2) testing the differences of parameter estimates between TYP/OTH and

Table 5
The Estimated Parameters for the Combined Model of 4-Year Literacy Involvement

Model	AIC	MNCI	CFI	RMSEA
Five-group model	45,257.23	.987	.969	.031
Combined model	45,221.81	.998	.997	.012

	Items	Loadings	Intercepts	Residual variances	Latent mean	Latent variance
TYP	Read books	1.23 (.07)[a]	2.73 (.06)[a]	2.95 (.16)[a]	.00 (NA)	1.00 (NA)
	Tell stories	1.73 (.11)[a]	2.60 (.06)[a]	1.51 (.30)[a]		
	Sing songs	1.08 (.06)[a]	2.98 (.06)[a]	4.24 (.17)[a]		
ASD	Read books	1.23 (.07)[a]	2.73 (.06)[a]	2.95 (.16)[a]	−.06 (.31)	.98 (.31)
	Tell stories	1.73 (.11)[a]	2.60 (.06)[a]	1.51 (.30)[a]		
	Sing songs	1.08 (.06)[a]	2.98 (.06)[a]	4.24 (.17)[a]		
CP	Read books	.29 (.26)	1.84 (.20)	1.18 (.49)	.00 (NA)	1.00 (NA)
	Tell stories	2.02 (.25)	2.83 (.47)	.00 (NA)		
	Sing songs	1.67 (.42)	3.59 (.66)	3.26 (.94)		
OthDD	Read books	1.23 (.07)[a]	2.73 (.06)[a]	2.95 (.16)[a]	.08 (.17)	1.19 (.26)
	Tell stories	1.73 (.11)[a]	2.60 (.06)[a]	1.51 (.30)[a]		
	Sing songs	1.08 (.06)[a]	2.98 (.06)[a]	4.24 (.17)[a]		
OTH	Read books	1.23 (.07)[a]	2.73 (.06)[a]	2.95 (.16)[a]	−.05 (.09)	1.00 (.15)
	Tell stories	1.73 (.11)[a]	2.60 (.06)[a]	1.51 (.30)[a]		
	Sing songs	1.08 (.06)[a]	2.98 (.06)[a]	4.24 (.17)[a]		

Note. AIC = Akaike information criterion; MNCI = McDonald's noncentrality index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; TYP = typically developing; ASD = autism spectrum disorder; CP = cerebral palsy; OthDD = other developmental delay; OTH = other; NA = parameter estimate was fixed for identification purposes. Parentheses contain the *SE*s for the estimates.
[a] Estimate was constrained to be equal across groups.

Table 6
The Autism Spectrum Disorder Pairwise Models for the 2-Year Time-Point on Routine Caregiving Involvement

Model	Model	MNCI	CFI	RMSEA	AIC
ASD vs.					
TYP	Configural	.994	.989	.053	60,333.48
	Weak	.994	.988	.041	60,328.73
	Strong	.993	.985	.038	60,332.43
	Strict	.992	.984	.033	60,333.39
CP	Configural	.841	.785	.294	1,506.05
	Weak	.854	.804	.212	1,508.62
	Strong	.844	.790	.184	1,512.87
	Strict	.650	.466	.248	1,527.05
OthDD	Configural	.806	.666	.328	3,176.12
	Weak	.845	.741	.219	3,173.53
	Strong	.838	.725	.188	3,175.75
	Strict	.832	.716	.162	3,173.85
OTH	Configural	.981	.958	.099	10,103.90
	Weak	.985	.968	.066	10,099.56
	Strong	.974	.945	.072	10,105.00
	Strict	.977	.951	.058	10,103.03
CP vs.					
TYP	Configural	.991	.979	.066	60,642.85
	Weak	.992	.981	.048	60,653.81
	Strong	.994	.984	.036	60,649.43
	Strict	.987	.968	.043	60,696.96
ASD	Configural	.841	.785	.294	1,506.05
	Weak	.854	.804	.212	1,508.62
	Strong	.844	.790	.184	1,512.87
	Strict	.650	.466	.248	1,527.05
OthDD	Configural	.632	.250	.479	3,485.49
	Weak	.798	.630	.254	3,492.38
	Strong	.858	.749	.175	3,489.41
	Strict	.851	.734	.152	3,501.40
OTH	Configural	.955	.889	.152	10,413.27
	Weak	.966	.916	.100	10,424.38
	Strong	.978	.947	.066	10,419.29
	Strict	.942	.859	.092	10,464.66

Note. RMSEA = root-mean-square error of approximation; MNCI = McDonald's noncentrality index; CFI = comparative fit index; AIC = Akaike information criterion; ASD = autism spectrum disorder; TYP = typically developing; CP = cerebral palsy; OthDD = other developmental delay; OTH = other.

OthDD, and TYP/OTH and ASD; and (3) constraining the parameter with the smallest difference and evaluating the new model. If the model fit better, as determined by AIC, then Steps 2 and 3 were repeated.

This process was continued until the fit, in terms of AIC, stopped improving. Here, the outcome resulted in a combined model that fit better than the five-group model while allowing for the pairwise decisions to be (mostly) left intact. The fit of the combined model, as well as the final parameter estimates, are given in Table 7.

Discussion

In this paper, we extend the standard method of investigating MI of a latent father involvement variable with two groups to multiple groups. While this is not a case often studied in the methodological literature, multiple-groups analyses arise very often, especially in the fields of early intervention, early childhood special education, and family science research. Our particular example examines multiple disability groups, where it cannot be readily assumed that the particular measurements of father involvement are invariant to the groups. While it is easy to apply Meredith's (1993) MI hierarchy to two groups, as the comparisons are simple, it is a less straightforward task to apply it to three or more groups. The simple generalization of exploring invariance by including all groups into a single model often gives incomplete results. For instance, for many of our father involvement factors, the model including all five groups resulted in the conclusion that the factors were strictly invariant to the particular group. A more detailed description of invariance can be achieved by comparing groups in a pairwise fashion. This allows us to pinpoint particular groups for which only lower levels of invariance held in comparison to the other groups. While this can make MI more difficult to establish, we argue that this more detailed approach makes it easier to target subgroups for which the person-item interactions differ. In father involvement, for instance, when there are multiple disability groups or multiple subgroups of fathers, it may be difficult to determine if measured differences are due to actual differences (at the latent level) or simply due to differences in the measurement itself. Knowing those differences allows researchers to allocate research resources more efficiently. In family research, this can translate into more specific and targeted interventions. One thing to note in regards to usage is that having small sample sizes may make the method untenable, but in this case, the researcher should be very careful using any latent variable modeling method. According to Bollen (1989), there should be about five samples per free parameter when using a structural equation model.

The use of latent variables in research in family science is invaluable, particularly when observed variables alone will not answer the questions of concern. When the use of latent variables becomes necessary, the issues underlying their use must be of concern for the researcher. In this paper, we argue that MI is an important concern when using latent variables, particularly when the goal of the analysis is to make comparisons between groups, such as disability groups. Indeed, in any grouping when the measurement is intended to show differences between the groups, not just cross-cultural groupings, one should be very conscious to the possibility that shown differences may simply be a consequence of the interaction between the measurement instrument itself and the group rather than showing meaningful group differences. For instance, it is possible that the mean of two groups is significantly different when doing an unconstrained analysis, but when the latent variable is properly constrained to be invariant to the groups, and the model fits well, the mean difference disappears (Millsap, 1997, 1998). Without the invariance assumption taken into account, the difference is misleading to the investigator, and false conclusions can occur easily. This problem is exacerbated when a research question requires more than two groups, as different groups will have different scale relationships with one another, with scale values having very different meanings for every group.

As shown, not only is it true that when MI is not investigated results can be misleading, but it is also true that not considering all groups can also be misleading. This is particularly the case in early intervention, early childhood special education, and family studies as multiple-group comparisons are often the goal (e.g., cross-cultural groups, disability groups, etc.). Unfortunately, when more than two groups are the concern, such as in children with differing disabilities, the assumption of MI becomes increasingly difficult to

Table 7
The Estimated Parameters for the Combined Model of 2-Year Routine Caregiving Involvement

Model	AIC	MNCI	CFI	RMSEA		
Five-group model	73,369.108	.968	.972	.036		
Combined model	73,300.800	.983	.984	.033		
	Items	Loadings	Intercepts	Residual variances	Latent mean	Latent variance
TYP	Prep food	2.89 (.13)[a]	6.17 (.11)[a]	12.26 (.53)[a]	.00 (NA)	1.00 (NA)
	Dress child	3.12 (.10)[a]	6.25 (.09)[a]	4.82 (.42)[a]		
	Wash child	2.27 (.11)[a]	4.13 (.09)[a]	6.21 (.31)[a]		
	Put to sleep	2.54 (.10)[a]	6.96 (.09)[a]	7.69 (.48)[a]		
ASD	Prep food	1.99 (1.07)	5.02 (1.15)	12.26 (.53)[a]	.00 (NA)	1.00 (NA)
	Dress child	3.12 (.10)[a]	7.30 (.94)	10.25 (4.75)		
	Wash child	2.27 (.11)[a]	5.08 (.47)	2.89 (1.49)		
	Put to sleep	2.54 (.10)[a]	6.96 (.09)[a]	4.84 (2.38)		
CP	Prep food	5.20 (.65)	7.76 (1.62)	3.27 (4.58)	.00 (NA)	1.00 (NA)
	Dress child	4.26 (.90)	7.73 (1.42)	4.52 (3.41)		
	Wash child	3.50 (1.20)	5.35 (1.55)	9.79 (2.83)		
	Put to sleep	.14 (2.00)	9.25 (1.47)	23.97 (7.14)		
OthDD	Prep food	3.55 (.39)	6.17 (.11)[a]	9.45 (3.55)	−.16 (.16)	1.00 (NA)
	Dress child	3.12 (.10)[a]	6.25 (.09)[a]	7.75 (2.22)		
	Wash child	2.27 (.11)[a]	4.13 (.09)[a]	8.06 (3.09)		
	Put to sleep	2.54 (.10)[a]	6.96 (.09)[a]	7.69 (.48)[a]		
OTH	Prep food	2.89 (.13)[a]	6.17 (.11)[a]	12.26 (.53)[a]	−.11 (.09)	.87 (.14)
	Dress child	3.12 (.10)[a]	6.25 (.09)[a]	4.82 (.42)[a]		
	Wash child	2.27 (.11)[a]	4.13 (.09)[a]	6.21 (.31)[a]		
	Put to sleep	2.54 (.10)[a]	6.96 (.09)[a]	7.69 (.48)[a]		

Note. AIC = Akaike information criterion; MNCI = McDonald's noncentrality index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; TYP = typically developing; ASD = autism spectrum disorder; CP = cerebral palsy; OthDD = other developmental delay; OTH = other; NA = parameter estimate was fixed for identification purposes. Parentheses contain the *SE*s for the estimates.
[a] Estimate was constrained to be equal across groups.

establish, as invariance to the particular factor must be achieved for all groups simultaneously; however, using our method as laid out in this paper, it is possible. Our illustration of the method highlighting how strict MI in father involvement factors drawn from the ECLS-B was not met in all cases underscores the importance of this consideration. If the analysis simply took for granted the results from the five-group model, then we would not see that there is some evidence that the latent variables of concern are not valid for making comparisons across the disability groups. For instance, literacy involvement looks very different for fathers of children with cerebral palsy than other groups; either fathers engage in literacy activities with children with cerebral palsy at age 4 differently than fathers in other groups, or at least the items used have different properties for the CP group.

This does not mean that not being able to compare certain groups on certain latent variables is the end of investigation; this means that the latent variable has response biases in some way for the different groups and further investigation into the nature of this difference is warranted. It is on us, as researchers, to then find why these biases exist, be it a difference in how the groups view latent variable, a problem in the definition of the construct under investigation itself, or something else. These are fundamental problems in measurement, and certainly not easy problems to solve, but not recognizing the issue should not be the standard. As seen in findings from the current study, scholars engaged in research on families and children with disabilities involving the use of latent variables would be well-advised to include an examination of MI as an important first step in their data analytic approach.

At this point, some discussion about how to approach a situation when strict invariance is not met is warranted. The level of invariance that is necessary to be met depends largely on the questions asked with the data at hand. If a researcher only wants to investigate group differences in latent variable (co)variances, then only weak invariance is necessary, but if he or she wants to investigate group differences in means and regressions, then strong invariance should be met. Finally, meeting strict invariance means that differences in group covariances (and means) in the manifest variables are entirely due to group differences on the latent variables, and so there is no differential effect of measurement. Thus, any possible comparison between the groups is possible when strict invariance is met. All of these conclusions are readily extended to the multiple group case. For instance, for routine caregiving at 2 years, the CP group is only configurally invariant to the other groups. In this case, group means and covariances for CP should not be directly compared with the other groups. For the TYP and OTH groups, factor means and covariances, as well as all observed means and covariances, can be compared. As for the levels of partial invariance, the questions that can be asked are highly dependent on how close to fully invariant the group is, though how close is close enough is a judgment call; Steenkamp and Baumgartner (1998) argue that at least one item (other than the fixed item) must be invariant to make the common inferences at that particular level. In our example, the OthDD group is very nearly strongly invariant with respect to TYP and OTH. It is not much of a leap to compare the means of the latent variables of these three groups with each other.

With our example, we hope to convey that incorrect conclusions can occur when simply following the standard steps in MI investigations for two groups when there are more than two groups. This method we have described is one way to carry out such an investigation, but it is still in its infancy. Explorations into the statistical properties of this method would be well-founded. In the future, a power study investigating the nature of MI in multiple groups would be a welcome addition to these arguments. Finally, a study of the use of fit statistics as measures of model misidentification with this method would be invaluable.

References

Bagheri, Z., Jafari, P., Tashakor, E., Kouhpayeh, A., & Riazi, H. (2014). Assessing whether measurement invariance of the KIDSCREEN-27 across child-parent dyad depends on the child gender: A multiple group confirmatory factor analysis. *Global Journal of Health Science, 6,* 142–153. http://dx.doi.org/10.5539/gjhs.v6n5p142

Beyers, W., & Goossens, L. (2008). Dynamics of perceived parenting and identity formation in late adolescence. *Journal of Adolescence, 31,* 165–184. http://dx.doi.org/10.1016/j.adolescence.2007.04.003

Bollen, K. (1989). *Structural equations with latent variables*. New York, NY: Wiley. http://dx.doi.org/10.1002/9781118619179

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5

Connell, A., Bullock, B. M., Dishion, T. J., Shaw, D., Wilson, M., & Gardner, F. (2008). Family intervention effects on co-occurring early childhood behavioral and emotional problems: A latent transition analysis approach. *Journal of Abnormal Child Psychology, 36,* 1211–1225. http://dx.doi.org/10.1007/s10802-008-9244-6

Crowley, S. L., & Fan, X. (1997). Structural equation modeling: Basic concepts and applications in personality assessment research. *Journal of Personality Assessment, 68,* 508–531. http://dx.doi.org/10.1207/s15327752jpa6803_4

Davidov, E., Schmidt, P., & Billiet, J. (Eds.). (2011). *Cross-cultural analysis: Methods and applications*. New York, NY: Taylor & Francis.

Dyer, W. J., McBride, B. A., Santos, R. M., & Jeans, L. M. (2009). A longitudinal examination of fathers' involvement with children with disabilities. *Journal of Early Intervention, 31,* 265–281. http://dx.doi.org/10.1177/0192513X09340386

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling, 12,* 343–367. http://dx.doi.org/10.1207/s15328007sem1203_1

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6,* 53–60.

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424–453. http://dx.doi.org/10.1037/1082-989X.3.4.424

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Hvidtjørn, D., Schieve, L., Schendel, D., Jacobsson, B., Svaerke, C., & Thorsen, P. (2009). Cerebral palsy, autism spectrum disorders, and developmental delay in children born after assisted conception: A systematic review and meta-analysis. *Archives of Pediatrics & Adolescent Medicine, 163,* 72–83. http://dx.doi.org/10.1001/archpediatrics.2008.507

MacLean, H., McKenzie, K., Kidd, G., Murray, A. L., & Schwannauer, M. (2011). Measurement invariance in the assessment of people with an intellectual disability. *Research in Developmental Disabilities, 32,* 1081–1085. http://dx.doi.org/10.1016/j.ridd.2011.01.022

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 30,* 419–440. http://dx.doi.org/10.1007/BF02289532

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2,* 248–260. http://dx.doi.org/10.1037/1082-989X.2.3.248

Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33,* 403–424. http://dx.doi.org/10.1207/s15327906mbr3303_5

Millsap, R. (2011). *Statistical approaches to measurement invariance*. New York, NY: Taylor & Francis.

Randall, J., & Engelhard, G., Jr. (2010). Using confirmatory factor analysis and the Rasch model to assess measurement invariance in a high stakes reading assessment. *Applied Measurement in Education, 23,* 286–306. http://dx.doi.org/10.1080/08957347.2010.486289

Reynolds, M. R., Ingram, P. B., Seeley, J. S., & Newby, K. D. (2013). Investigating the structure and invariance of the Wechsler Adult Intelligence Scales, Fourth Edition in a sample of adults with intellectual disabilities. *Research in Developmental Disabilities, 34,* 3235–3245.

Schuler, M., Musekamp, G., Bengel, J., Nolte, S., Osborne, R. H., & Faller, H. (2014). Measurement invariance across chronic conditions: A systematic review and an empirical investigation of the Health Education Impact Questionnaire (heiQ). *Health and Quality of Life Outcomes, 12,* 56. http://dx.doi.org/10.1186/1477-7525-12-56

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *The Journal of Consumer Research, 25,* 173–180. http://dx.doi.org/10.1086/209528

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.

Tran, T. V. (2009). *Developing cross cultural measurement*. Oxford, United Kingdom: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195325089.001.0001

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. http://dx.doi.org/10.1177/109442810031002

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance abuse domain. In K. J. Bryant (Ed.), *Alcohol and substance use research* (pp. 281–324). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10222-009

Received May 5, 2015
Revision received November 28, 2015
Accepted December 7, 2015 ■