American Journal on Intellectual and Developmental Disabilities

Measurement Invariance in IDD Research

--Manuscript Draft--

Manuscript Number:	AJIDD-D-23-00063
Article Type:	Perspectives
Keywords:	measurement invariance; differential item functioning; psychometrics; item response theory; structural equation modeling
Corresponding Author:	Cristan A. Farmer National Institute of Mental Health Bethesda, MD UNITED STATES
First Author:	Cristan A. Farmer
Order of Authors:	Cristan A. Farmer
	Aaron J Kaat
	Michael C Edwards
	Luc Lecavalier
Manuscript Region of Origin:	UNITED STATES
Abstract:	Measurement invariance (MI) is a psychometric property of an instrument indicating the degree to which scores from an instrument are comparable across groups. In recent years, there has been a marked uptick on publications using MI in intellectual and developmental disability (IDD) samples. Our goal here is to provide an overview of why MI is important to IDD researchers and to describe some challenges to evaluating it, with an eye towards nudging our subfield into a more thoughtful and measured interpretation of studies using MI.

Abstract

Measurement invariance (MI) is a psychometric property of an instrument indicating the degree to which scores from an instrument are comparable across groups. In recent years, there has been a marked uptick on publications using MI in intellectual and developmental disability (IDD) samples. Our goal here is to provide an overview of why MI is important to IDD researchers and to describe some challenges to evaluating it, with an eye towards nudging our subfield into a more thoughtful and measured interpretation of studies using MI.

Keywords

Measurement invariance, differential item functioning, validity theory, psychometrics, item response theory, factor analysis, structural equation modeling

Much ink has been spilled about the challenges of assessing behavioral constructs in our small and heterogeneous intellectual and developmental disabilities (IDD) samples (Kelleher & Wheeler, 2020; Lecavalier et al., 2014). Psychometric properties are not inherent to an instrument but are instead emergent properties of the interaction between items and a particular (sub-)population of respondents. A central question when mounting a validity argument (i.e., accumulating evidence of the validity of a particular interpretation of a score in a particular context) is whether scores from an instrument are comparable across different (sub-)populations. The type of evidence that supports such comparisons is measurement invariance (MI).

<< Insert Figure 1 about here >>

We have observed a marked uptick in MI analyses in the subfield of IDD research (**Figure 1**). Because of the relevance of MI analyses to our subfield, we believe that this rate will continue to increase. Our goals here are to provide a nontechnical explanation of MI, describe why and when MI should matter to IDD researchers, and to highlight important considerations, with an eye towards nudging our subfield into a more thoughtful and measured use and interpretation of MI.

What is Measurement Invariance (MI)?

MI refers generally to situations where the parameters of a model are the same as a function of some other variable. This “some other variable” is often a categorical grouping variable, like diagnosis, but it could also be a continuous variable like age (Bauer, 2017). In a simple regression context, MI can be described as the absence of an interaction: For MI to be a reasonable assumption, the slope for some predictor is the same (or close enough) regardless of group membership. In this manuscript, we focus on MI in the psychometric context. In this case, MI is assumed for measurement models like those expressed using classical test theory (CTT), structural equation modeling (SEM), or item response theory (IRT). Each of these

approaches conceptualizes psychological domains as unobserved—or latent—constructs. What are observable—or manifest—are the behaviors that indicate the individual's magnitude or severity of the construct. Generally speaking, if the same magnitude or severity of the construct yields a similar expected probability of endorsing a behavior item regardless of group membership, then the assumption of MI is tenable.

By way of example, let us imagine items on a depression scale. The assumption of MI for an item on this scale is tenable if someone with IDD and someone without IDD who have the same amount of depression have the same expected response to the item. MI would fail if, despite having the same level of depression, a person with IDD had a different expected response to the item than a person without IDD. At the item-level, one useful way to interpret a lack of invariance (i.e., noninvariance) is that it indicates the item works differently for someone with IDD than for someone without IDD. Consider an item that requires language (e.g., "Talks about feelings of guilt"). For someone with fluent language, this question asks how often they convey feelings of guilt to others. For someone with IDD, who is more likely to have limited language, this question is asking both about their ability to speak and about how often they use that ability to convey feelings of guilt. This failure of MI means that on average, people with IDD would have a different response to that item than non-IDD folks with the same level of depression.

How is MI Assessed?

Thus far, we have referred to MI being "tenable" or "reasonable" because MI is an assumption, or something in which we must believe for our inferences based on the model to be valid. Returning to the simple regression context, the assumption of MI is somewhat analogous to the assumption of residual independence, normality, and homoscedasticity. Like MI, these assumptions are not things which are or are not true. Instead, they are ideals with which the observed data are more or less consistent. Just as in the case of regression assumptions,

where it is incumbent upon the modeler to evaluate their data and determine whether those assumptions are reasonable, it is incumbent upon the user of an instrument to look to the available data to determine whether MI is reasonable.

There are many ways in which this can be done. Most approaches are technically complex, and so their details are outside the scope of this manuscript. For interested readers, we have included a reading list (**Table 1**). For the remainder of the manuscript, we refer generally to SEM and IRT-based methods. These methods typically involve a series of model comparisons where some model parameter is first allowed to be different between the groups (freely estimated) and then forced to be the same between the groups (constrained). The specific parameters depend on the model, but the parameters in question describe the relationship between the construct and observed item responses. In theory, parameters are either the same (invariant) or different (noninvariant) between populations. When working with finite samples and imperfect measurement, it is never this simple. What we end up asking is something along the lines of "Does forcing these two parameters to be the same not make the model fit too much worse?" The task of defining "too much worse" is also complex and has its own literature (Gunn, Grimm, & Edwards, 2020; Meade, 2010; Millsap & Kwok, 2004; Nye & Drasgow, 2011). From a methodological perspective, it is important to note that the notion of scale scores being invariant depends on items being invariant which in turn depends on model parameters about those items being invariant. Generally speaking, we test parameter invariance to assess item invariance to in turn assess scale score invariance.

<<Insert Figure 2 About Here>>

The Assumption of MI Is Not Tenable – What Now?

Consider a hypothetical situation in which MI is not supported between children with and without autism spectrum disorder (ASD) for some items of the ABC Irritability subscale (Aman & Singh, 2017). There are (roughly) four options available to a researcher in this position (Figure

2). Option 1, not using the subscale at all, is defensible from the MI perspective, but likely to be problematic in many situations when evaluated from other dimensions. Option 2, dropping an item (or items) that exhibit a lack of MI, can be an effective way to address the violation of MI, but losing items inevitably lowers reliability and reduces the content covered by the scale. As the validity arguments for score interpretation all depend upon the item content of a scale quite heavily, it is easy to imagine that dropping items can very quickly undermine any existing validity evidence/arguments.

Under what circumstances might option 3, ignoring evidence of noninvariance, be reasonable? One situation might be if no comparison is to be made between the groups exhibiting noninvariance (Borsboom, 2006). For example, noninvariance between children with and without ASD has no bearing on the interpretation of correlations between the ABC Irritability score and other variables within an ASD sample. However, our hypothetical researcher may need to avoid comparing those results to similar analyses in children without ASD, because they would understand that if MI fails, ABC Irritability scores might mean something different in the two groups.

Noninvariance might also be ignored if its functional impact is very small. As in other areas of statistics, a *statistically detectable* degree of noninvariance is not necessarily a *meaningful* degree of noninvariance. Although noninvariance exists in degrees, it is common in the IDD literature (and generally) to dichotomize based on some statistical threshold, for which a number of guidelines exist (Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). To illustrate the idea that statistical noninvariance – a change in fit that exceeds whatever threshold the user has selected – is not always practically important, let us return to our researcher who has found statistical evidence of noninvariance (i.e., the change in fit indices exceeded some threshold) in items of the ABC Irritability score.

The ABC manual states that Irritability sum scores can range from 0 to 45, and the average score in children with IDD is around 9 with a SD of around 8 (Aman & Singh, 2017).

Through further modeling, the researcher determines that the noninvariance observed in a couple of the items leads to an Irritability score that is 6 points lower for a child without ASD than for a child with ASD *with the same amount of the underlying trait*. Considering what we know about the distribution of the ABC Irritability score in the population, 6 points is a large difference that would substantively affect the interpretation of any resulting analysis comparing the ABC Irritability score between ASD and non-ASD groups. This degree of noninvariance would be both statistically and practically meaningful and would indicate that the noninvariance should not be ignored. But, that same degree of model misfit (statistical noninvariance) could also have been associated with a much smaller practical impact; what if the researcher had instead identified that the bias in scores for children with ASD was only one-half of a point on the sum scale? This value is tiny relative to the between-subject variability in scores, and despite being statistically significant, might well be ignorable as just another source of measurement error. And so, *we recommend that researchers strive not to stop with a categorical result of an MI analysis.* Instead, should the analysis indicate a statistically appreciable degree of noninvariance, we encourage researchers to determine the impact of observed noninvariance on the actual response scale (e.g., Edelen, Stucky, & Chandra, 2015; Nye & Drasgow, 2011; Teresi, Ramirez, Jones, Choi, & Crane, 2012). We note here that specifying a measurement model that reflects sum-score usage when evaluating MI may facilitate this endeavor (e.g., Christensen, Kreiner, & Mesbah, 2012; McNeish & Wolf, 2020).

Whether or not a researcher can make use of option 4, proactively modeling MI, depends on whether they are using an adequate psychometric model and whether they have identified the exact model parameter(s) that stand in the way of MI. Here, it is useful to contrast IRT and SEM approaches to MI. In general, IRT approaches proceed item by item, then parameter by parameter (or parameter type by parameter type). In the SEM framework, the testing proceeds parameter type by parameter type – and then often stops. In the example above, this would amount to the researcher concluding from a couple of noninvariant items that

MI across ASD does not hold for ABC Irritability (or, if the researcher is more MI savvy, it may lead to them concluding a certain *level* of invariance is unsupported– see (Meredith, 1993) for more discussion on the typical levels of invariance considered in the SEM-based MI tests). This is common across psychology; one review indicated that only about one-third of publications about MI explored the item/specific parameter-level source of invariance (Putnick & Bornstein, 2016).

What does it mean to proactively model the lack of MI? This process is called "partial invariance," and it consists of allowing some parameters to vary across group while holding others invariant. Partial invariance has a long history in the MI literature. As long as enough parameters are invariant between groups, other parameters can be allowed to vary while still maintaining all the benefits of MI (Edwards & Wirth, 2009). In our experience, discussion of partial invariance is far more prevalent in the IRT literature than the SEM literature, which is likely due to the ease with which most IRT-based MI methods lead to identifying and modeling partial invariance. To be clear, it is completely possible to pursue a partial-invariance strategy in a SEM framework – it just is not currently common. Whether or not partial invariance is considered, *we recommend that researchers using an SEM approach to MI take care to discuss the results at the item/parameter level, rather than at the aggregate (scale) level.*

What Makes the Evaluation of MI So Hard?

We earlier justified our 30,000-foot point of view in this manuscript with the statement that the details of MI evaluation are technically complex. Here, we will expand upon one of the main hurdles IDD researchers face when implementing and interpreting MI analyses: the small size of the subpopulations we study.

To fully evaluate MI within the multiple group framework, one must be able to fit the SEM or IRT model in each group separately, prior to instituting any constraints. Commonly cited guidelines for SEMs vary from a minimum of 100 to 200 or more (Norris & Lecavalier, 2010),

and recommendations for IRT models are usually slightly larger, with a minimum of 250 or more (Svetina & Dai, 2022). Thus, most MI simulation studies suggest that sample sizes should range anywhere from 150 to 500 participants *per group* (Meade & Bauer, 2007; Meade et al., 2008; Woods, Cai, & Wang, 2013).

One sample size issue is unique to researchers using SEM methods, because these approaches were developed in the context of factorial model with *continuous* indicators (IRT approaches inherently focus on categorical item-level responses). Unfortunately for IDD researchers, the MI methods for continuous indicators are too simplistic for the ordinal data produced by our measures (e.g., ABC item level scores) (Millsap & Yun-Tein, 2004; Wu & Estabrook, 2016). While we recommend IDD researchers used methods specific to ordered-categorical indicators (Svetina, Rutkowski, & Rutkowski, 2020), these models can often require larger samples sizes to offset increases in model complexity (e.g., nonlinear models, more parameters).

Another way to talk about sample size is to talk about the power of the statistical test to detect an effect when it occurs. There is no guarantee that the statistical thresholds used to generate sample size rules-of-thumb pertain to clinically meaningful effects for the investigator. The goal for any sample size / power calculation for any statistical test is to identify the smallest possible sample that would allow the researcher to statistically detect a clinically meaningful effect with a pre-determined level of confidence. It is therefore reasonable to think that researchers intending to evaluate the assumption of MI for a particular instrument should have a working idea of the degree of noninvariance that would be unacceptable for their proposed use of the scale (the “clinically meaningful effect”), and that they should be able to estimate just how many participants that might require. Unfortunately, sample size / power calculations for MI analyses are non-trivial (Wang & Rhemtulla, 2021). A major reason for this is that the sources of noninvariance are at the parameter level, and so as the complexity of the measurement model

increases, the number of possible ways to observe clinically meaningful noninvariance increases exponentially.

One might wonder why, if an evaluation of MI is competently performed, power matters? The standard rules around power in the context of null hypothesis significance testing should apply here; when an analysis was underpowered, we are less able to detect a true effect but, paradoxically, a statistically significant result is likely to be an overestimate of the true effect size ("winners curse") (Button et al., 2013). Here, we encounter a wrinkle to the application of these standard rules around power. Unlike traditional null hypothesis significance testing, in MI analyses we hope to *fail* to reject the null hypothesis of no difference between groups. This means that an underpowered study is biased *towards* the desired outcome (i.e., concluding that MI is supported). The nature of underpowered tests would suggest that numerically small sample sizes, or even numerically large sample sizes with significantly imbalanced group sizes, or a low participant-to-model parameter ratio, are more likely to yield evidence in support of MI (Jobst, Bader, & Moshagen, 2023). This puts consumers of IDD MI studies in a bind – without some indication of the statistical power of an individual study, it is very difficult to weigh a study supporting MI against the possibility that it failed to reject the null hypothesis due to low power. Given the small samples sizes in IDD research, this could be true even if a series of studies supported MI. Or, conversely, if noninvariance is observed, consumers must grapple with the possibility that the result may only have been statistically significant because it is an overestimate of the problem. Given all of this, investigators must carefully consider whether a potentially underpowered evaluation of MI is better than no evaluation of MI. We propose that at best, underpowered MI analyses might be considered uninterpretable; at worst, underpowered MI analyses might be used to justify incorrect psychometric arguments. *The fact that we often cannot (or cannot practically) know how well-powered our MI analyses are is a limitation with which we must seriously engage when planning new analyses or interpreting their results*.

In this manuscript, we have focused on the multiple group approach to detecting MI. In IDD, these groups are likely to be diagnostic (e.g., Phelan McDermid Syndrome versus not) or classifications of some continuous construct (e.g., language level). But, technically speaking, group membership is likely serving as a proxy for some set of variables. That is to say, when MI is evaluated across those with and without Phelan McDermid Syndrome, the underlying theory is not likely that alterations in *SHANK3* lead to differences in psychometric performance, but rather that people with similar alterations in *SHANK3* tend to share characteristics that affect psychometric performance. Recognizing this nuance in the underlying theory allows a researcher to take a broader view of how the MI analysis contributes to the validity argument for a particular measure (see Houts, Bush, Edwards, & Wirth, 2022 for further explanation). For example, a researcher might wish to compare scores from the ABC between groups of individuals with Phelan McDermid Syndrome and CLN3 disease. Intellectual disability is part of the behavioral phenotype of both conditions, but blindness is common only in CLN3, and so the researcher might formulate a hypothesis that the vision loss might be the main source of potential noninvariance between Phelan McDermid and CLN3. Rather than conduct an arduous study to obtain ABC ratings from inevitably small and heterogeneous samples of these rare disease populations, the investigator might do desk research to identify studies supporting MI across samples with or without vision impairment (regardless of etiology).

Conclusion

The genesis of this commentary was the increasingly common evaluation of MI in the IDD literature. While MI is an important assumption for many types of comparisons, the interpretation of the results of MI analysis is challenged by our small and heterogenous samples. We hope that by raising these “yellow flags,” we will encourage IDD researchers to approach the use and interpretation of MI analyses in a thoughtful manner, considering whether and when they are useful or necessary.

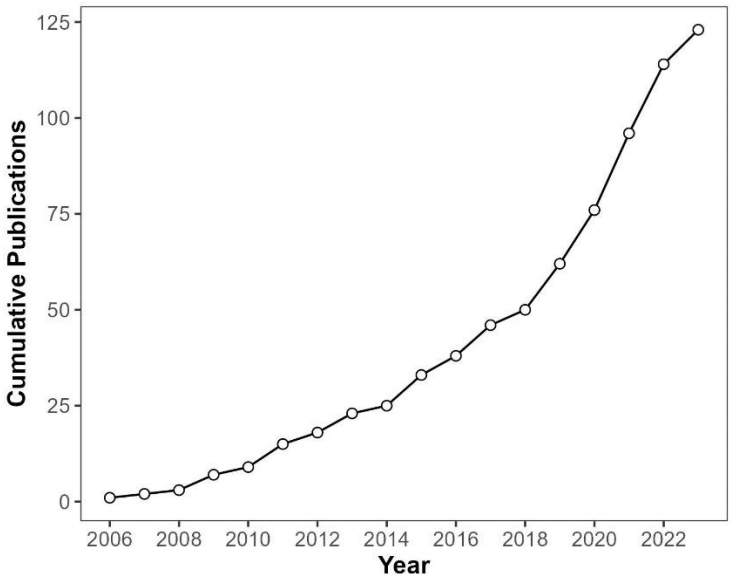

Figure 1. IDD Measurement Invariance Publications. A September 2023 Scopus all-time literature review searching key terms, titles, and abstracts for words related to measurement invariance and IDD (search specifications available upon request) yielded 156 results. After excluding duplicates and studies not related to IDD or measurement invariance, 123 papers from 48 journals remained. No publications were found prior to 2006.

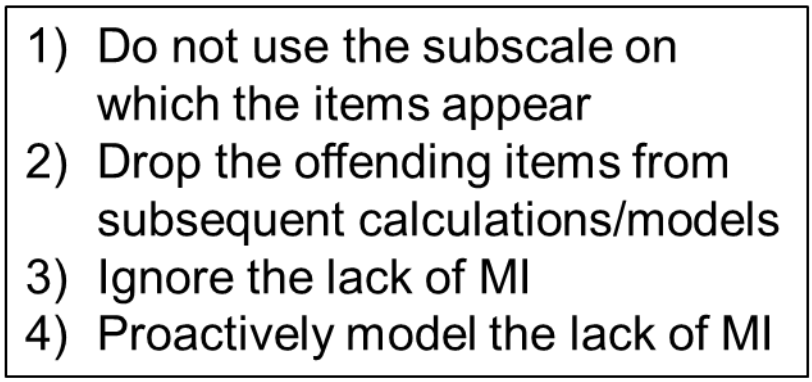

Figure 2. Potential courses of action when items on a measure exhibit noninvariance. The section, The Assumption of MI Is Not Tenable – What Now? describes the situations in which each action might be appropriate.

References

Aman, M. G., & Singh, N. (2020). *Manual for the Aberrant Behavior Checklist* (2nd edition). Slosson.

Aman, M., & Singh, N. (2017). *Aberrant Behavior Checklist Manual, Second Edition*. East Aurora, NY: Slosson Educational Publications, Inc.

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3), 507.

Borsboom, D. (2006). When does measurement invariance matter? *Medical care, 44*(11), S176-S181.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling, 9*(2), 233-255.

Christensen, K. B., Kreiner, S., & Mesbah, M. (2012). *Rasch models in health*: John Wiley & Sons.

Edelen, M. O., Stucky, B. D., & Chandra, A. (2015). Quantifying 'problematic' DIF within an IRT framework: application to a cancer stigma index. *Quality of Life Research, 24*(1), 95-103. doi:10.1007/s11136-013-0540-4

Edwards, M. C., & Wirth, R. (2009). Measurement and the study of change. *Research in Human Development, 6*(2-3), 74-96.

Gunn, H. J., Grimm, K. J., & Edwards, M. C. (2020). Evaluation of six effect size measures of measurement non-invariance for continuous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(4), 503-514.

Houts, C. R., Bush, E. N., Edwards, M. C., & Wirth, R. (2022). Using validity theory and psychometrics to evaluate and support expanded uses of existing scales. *Quality of Life Research, 31*(10), 2969-2975.

Jobst, L. J., Bader, M., & Moshagen, M. (2023). A tutorial on assessing statistical power and determining sample size for structural equation models. *Psychological Methods, 28*(1), 207.

Kelleher, B. L., & Wheeler, A. C. (2020). Introduction to special issue on outcome measures for IDD: Where we have been, where we are now, and where we are heading. *American Journal on Intellectual and Developmental Disabilities, 125*(6), 413-417.

Lecavalier, L., Wood, J. J., Halladay, A. K., Jones, N. E., Aman, M. G., Cook, E. H., . . . Hallett, V. (2014). Measuring anxiety as a treatment endpoint in youth with autism spectrum disorder. *J Autism Dev Disord, 44*, 1128-1143.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior research methods, 52*, 2287-2305.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *95*, 728-743. doi:10.1037/a0018966

Meade, A. W., & Bauer, D. J. (2007). Power and Precision in Confirmatory Factor Analytic Tests of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 611-635. doi:10.1080/10705510701575461

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of applied psychology, 93*(3), 568.

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

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*(1), 93.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate behavioral research, 39*(3), 479-515.

Norris, M., & Lecavalier, L. (2010). Evaluating the use of exploratory factor analysis in developmental disability psychological research. *J Autism Dev Disord, 40*, 8-20.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: understanding the practical importance of differences between groups. *Journal of applied psychology, 96*(5), 966.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review, 41*, 71-90.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of applied psychology, 91*(6), 1292.

Svetina, D., & Dai, S. (2022). Number of Response Categories and Sample Size Requirements in Polytomous IRT Models. *The Journal of Experimental Education*, 1-32.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: an illustration using M plus and the lavaan/semtools packages. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 111-130.

Teresi, J. A., Ramirez, M., Jones, R. N., Choi, S., & Crane, P. K. (2012). Modifying measures based on differential item functioning (DIF) impact analyses. *Journal of aging and health, 24*(6), 1044-1076.

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. In (Vol. 6, pp. 1064): Frontiers Media SA.

Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science, 4*(1), 2515245920918253.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. Bryant, M. Windle, & S. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.

Willse, J. T., & Goodman, J. T. (2008). Comparison of Multiple-Indicators, Multiple-Causes– and Item Response Theory–Based Analyses of Subgroup Differences. *Educational and Psychological Measurement, 68*(4), 587-602. doi:10.1177/0013164407312601

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*(3), 532-547.

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika, 81*(4), 1014-1045.

Tables

Table 1. Measurement invariance suggested reading list

Reference	Description
Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. In (Vol. 6, pp. 1064): Frontiers Media SA.	Van De Schoot, Schmidt, De Beuckelaer, Lek, and Zondervan-Zwijnenburg (2015) contains a historical overview of MI. This is an editorial for a special issue on MI.
Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review, 41*, 71-90.	Putnick and Bornstein (2016) offer a broad description of how MI methods are used across psychology.
Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), The science of prevention: Methodological advances from alcohol and substance abuse research (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009	Widaman and Reise (1997) is a tutorial on SEM-based MI using mean-and-covariance matrices.
Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika, 81*(4), 1014-1045.	Wu and Estabrook (2016) contains best-practices recommendation for MI with categorical indicators.
Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: an illustration using M plus and the lavaan/semtools packages. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 111-130.	Svetina et al. (2020) is a tutorial for implementing the Wu & Estabrook procedures for categorial MI.
Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3), 507.	Bauer (2017) describes moderated nonlinear factor analysis (MNLFA) as a way to test MI and DIF in one method.
Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of applied psychology, 91*(6), 1292.	Stark, Chernyshenko, and Drasgow (2006) describes relationship between MI (using mean-and-covariance matrices) and DIF on the same dataset.
Willse, J. T., & Goodman, J. T. (2008). Comparison of Multiple-Indicators, Multiple-Causes– and Item Response Theory–Based Analyses of Subgroup Differences. *Educational and Psychological Measurement, 68*(4), 587-602. doi:10.1177/0013164407312601	Willse and Goodman (2008) is a describes multiple indicators multiple causes (MIMIC) models for evaluating MI.