Measurement Invariance Testing Using Confirmatory Factor Analysis and Alignment Optimization: A Tutorial for Transparent Analysis Planning and Reporting

Raymond Luong and Jessica Kay Flake

Department of Psychology, McGill University

*** This paper was accepted for publication in *Psychological Methods* on 09/07/2021. This is the pre-copy edited, peer-reviewed version. ***

Author Note

Raymond Luong https://orcid.org/0000-0001-6587-6159

Jessica Kay Flake https://orcid.org/0000-0002-3498-615X

We have no conflicts of interest to disclose. Materials and data are openly available on the Open Science Framework here. An early version of this tutorial was presented virtually at the 2020 Canadian Psychological Association Conference, Montreal, Quebec, Canada. There was no other prior dissemination of this manuscript.

Correspondence concerning this article should be addressed to Jessica Kay Flake and Raymond Luong, 2001 Avenue McGill College, Montréal, Quebec, Canada, H3A 1G1. Emails: jessica.flake@mcgill.ca, raymond.luong@mail.mcgill.ca

Abstract

Measurement invariance—the notion that the measurement properties of a scale are equal across groups, contexts, or time—is an important assumption underlying much of psychology research. The traditional approach for evaluating measurement invariance is to fit a series of nested measurement models using multiple-group confirmatory factor analyses. However, traditional approaches are strict, vary across the field in implementation, and present multiplicity challenges, even in the simplest case of two groups under study. The alignment method was recently proposed as an alternative approach. This method is more automated, requires fewer decisions from researchers, and accommodates two or more groups. However, it has different assumptions, estimation techniques, and limitations from traditional approaches. To address the lack of accessible resources that explain the methodological differences and complexities between the two approaches, we introduce and illustrate both, comparing them side by side. First, we overview the concepts, assumptions, advantages, and limitations of each approach. Based on this overview, we propose a list of four key considerations to help researchers decide which approach to choose and how to document their analytical decisions in a preregistration or analysis plan. We then demonstrate our key considerations on an illustrative research question using an open dataset and provide an example of a completed preregistration. Our illustrative example is accompanied by an annotated analysis report that shows readers, step-by-step, how to conduct measurement invariance tests using *R* and *Mplus*. Finally, we provide recommendations for how to decide between and use each approach and next steps for methodological research.

Keywords: measurement invariance, measurement equivalence, multiple-group confirmatory factor analysis, alignment, differential item functioning

Measurement Invariance Testing Using Confirmatory Factor Analysis and Alignment Optimization: A Tutorial for Transparent Analysis Planning and Reporting

Measurement invariance (also known as *measurement equivalence*) refers to the notion that the psychometric properties of a scale are equal (i.e., invariant or equivalent) across groups and/or measurement occasions like contexts or time. Without it, interpreting group differences raises questions: Is an observed difference across groups due to a group difference on the construct or due to differences in how the scale is measuring the construct? Ignoring measurement non-invariance can lead to incorrect conclusions about comparisons between groups, such as erroneously concluding one group is higher on a construct than the other (Chen, 2008; Steinmetz, 2013). Thus, measurement invariance is important to consider in a variety of contexts, including longitudinal research, research on diverse groups, cross-cultural psychology including translated instruments, and in experimental designs to evaluate assumptions to ensure comparability across treatment and control groups. As such, it is broadly applicable to many areas of psychology.

There are a variety of psychometric methodologies for assessing measurement invariance across two or more groups, with most using model comparisons in confirmatory factor analyses (CFA) or item response theory (IRT)[1] to test the equality of measurement properties across groups or time (for an overview, see Millsap, 2011). We will refer to this model comparison approach as the traditional approach. To address challenges in applying the traditional approach, Asparouhov and Muthén (2014) developed an alternative, more automated approach known as the alignment method.

[1] IRT is used specifically for binary or polytomous indicators and emphasizes identifying non-invariant items (known as *differential item functioning*). In this tutorial, we focus on CFA due to the propensity of Likert-type scales in psychology that are commonly treated as continuous rather than polytomous. Item scores are also usually combined into composites (e.g., sum scores or averages) for analysis.

The alignment method makes no assumptions about the number of groups and can accommodate two or more groups easily. Simulation studies showed good performance in two-group cases for recovering factor model parameters (i.e., unbiased point estimates, and near or above 95% coverage; Asparouhov & Muthén, 2014). The alignment method is also ideal for smaller numbers of groups for which the data would not satisfy assumptions for a random effects approach (e.g., multilevel measurement models which require many groups; see Muthén & Asparouhov, 2018). Thus, the alignment method can be implemented as an alternative or accompanying method to traditional approaches when there are only two groups. Despite the potential for the alignment method's use with two groups, it has generally not been considered as a two-group alternative by applied researchers and use thus far has focused on many-groups cases (Lomazzi, 2018; Muthén & Asparouhov, 2018). As of writing, there is no guidance or side-by-side comparison of the two approaches for the two-group case. The alignment method has also only very recently received a comprehensive methodological comparison to the traditional approach with moderate numbers of groups (see Magraw-Mickelson et al., 2021). Few accessible resources exist that aim to assist substantive researchers in considering when the alignment might be better suited for certain research contexts.

The purpose of this tutorial is to provide a non-technical introduction to two different approaches to measurement invariance testing, with a focus on testing two groups. Researchers can use this as a resource to assist in planning, choosing between, implementing, and interpreting either approach. We aim to facilitate the ease of appropriately using these methods as well as support transparent practices for the planning and reporting of measurement invariance testing consistent with Transparency and Open Practices Guidelines adopted by American Psychological Association journals in 2021 (Center for Open Science, 2020). We will first explain and compare

the conceptual basis of each method and highlight their key similarities and differences in assumptions and implementation. We will then provide an illustrative preregistered data analysis example of measurement invariance testing using both methods on an open cross-national dataset. Through this example, we will offer recommendations on how researchers can appropriately decide between and then use either approach. We will close with recommendations for the methods and suggest next steps for methodological research.

Approaches to Measurement Invariance Testing in Psychology

Confirmatory Factor Analysis: A Primer

CFA is fundamental to both the traditional factor analytic approaches and the alignment method. First, consider the confirmatory factor analysis model for continuous items in one group, expressed in notation used by Asparouhov and Muthén (2014) for ease of reference:

$$y_{ip} = \nu_{pk} + \sum_{k=1}^{K} \lambda_{pk}\eta_{ik} + \varepsilon_{ip} \quad (1)$$

In Equation 1, the factor model is represented as a linear regression of the items on the factors (or latent variables). Here, $i = 1,\ldots,I$ where I is the total number of people (or observations), $p = 1,\ldots,P$ where P is the total number of items (or indicators), and $k = 1,\ldots,K$ where K is the total number of factors. y_{ip} is the observed score for person i on item p, ν_{pk} is the intercept for item p of factor k, λ_{pk} is the factor loading for item p on factor k, η_{ik} is a factor score for person i on factor k, and ε_{ip} is the residual for person i of their observed score of item p (which is y_{ip}).

The multiple-group CFA (MGCFA) extends the one-group CFA to accommodate multiple groups:

$$y_{ipg} = \nu_{pg} + \sum_{k=1}^{K} \lambda_{pg}\eta_{ig} + \varepsilon_{ipg} \quad (2)$$

Equation 2 shows that MGCFA is represented in the same way as a one-group CFA with the addition of a group subscript g to indicate group membership, where $g = 1,\ldots,G$ and G is the total number of groups. Furthermore, we assume that the residuals ε_{ipg} are normally distributed with a mean of 0 and some variance θ_{pg} and that the factors η_{ig} are normally distributed with some group-specific factor mean α_g and variance Ψ_g.

Traditional Factor Analytic Approaches

The traditional factor analytic approaches involve conducting a series of MGCFAs and using them to test the equality of measurement properties (i.e., factor structure, loadings, intercepts, and uniquenesses/residual variances) across groups in increasingly strict stages. The equality tests for model parameters are conducted on like items, meaning the same items across groups (e.g., Item 1 in group 1 vs. Item 1 in group 2). Hence, under these approaches, measurement invariance is a hierarchical property, and the level of measurement invariance for a measure is determined by the best comparatively fitting model. This hierarchy is depicted in Figure 1: The fit of the MGCFA corresponding to each level of measurement invariance is compared to the next sequentially, starting from the bottom of the hierarchy and compared to the level exactly above it (i.e., configural vs. metric, metric vs. scalar, scalar vs. strict). Below, we provide a conceptual overview of these levels as per van de Schoot et al. (2012), Muthén and Asparouhov (2018), and Bialosiewicz et al. (2013). Then in our illustrative data analysis example, we present testing each level, for which accompanying data analysis code is reported in the Supplementary Materials.

Figure 1

Hierarchy of the Four Levels of Measurement Invariance

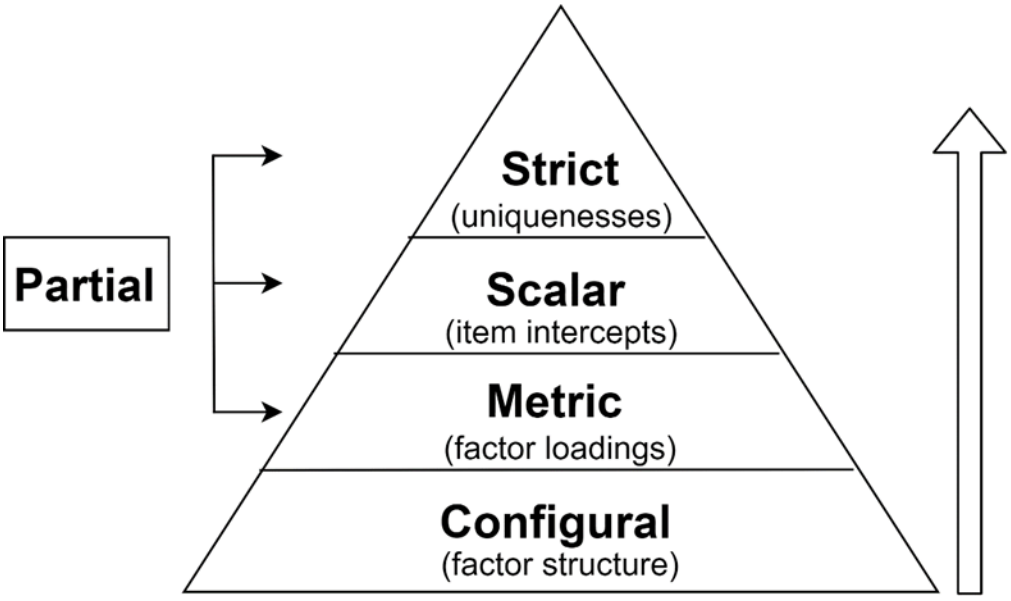


Figure 1 shows the four hierarchal levels of measurement invariance: configural, metric, scalar and strict (Horn & McArdle, 1992; Meredith, 1993). The first and lowest level of the hierarchy is *configural* invariance (Horn & McArdle, 1992), which means that the configuration of the indicators to their factors is the same across groups—that is to say, the number of latent constructs and the specific items loaded onto them are the same across groups. Configural non-invariance precludes comparisons of a scale's scores (latent or observed) across groups: Having different numbers or configurations of items to factors plainly suggests that different constructs are being measured in different groups and scores from different constructs are not comparable. Configural non-invariance may reflect a theoretical inconsistency such that further research is required to understand the nature of the construct, including the content of the construct and the construct's meaning to different groups. This type of inquiry is well suited for qualitative or mixed methods research with the populations of interest.

Following configural invariance, *metric invariance* (Horn & McArdle, 1992; also known as *weak (factorial) invariance* as per Meredith, 1993) is the next level of measurement invariance. In addition to equality of the factor model configuration across groups by configural invariance, achieving metric invariance means that the specific statistical relationships between the scale's items and their associated latent constructs also stay the same across groups—that is to say, factor loadings are equal across groups. Metric non-invariance can bias observed factor variances, factor covariances, and factor means (French & Finch, 2016; Shi et al., 2019; Yoon & Millsap, 2007), which can lead to erroneous conclusions on downstream statistical tests.

Typically, after metric invariance is tested, *scalar invariance* (Steenkamp & Baumgartner, 1998; also known as *strong* (*factorial) invariance* as per Meredith, 1993) is the next level of measurement invariance. In addition to equality of the factor model across groups by configural invariance and equality of factor loadings across groups by metric invariance, scalar invariance is achieved when the meaning of the levels of item responses are also equal across groups—that is to say, both the factor loadings and intercepts are equal across groups. If scalar invariance is achieved, then groups can be compared by their observed or latent scores for the construct; the former is the most frequent application in psychological research. Scalar non-invariance precludes any observed mean comparisons; even one non-invariant intercept can bias the results of a mean comparison (Steinmetz, 2013).

Finally, following scalar invariance is *strict invariance* (Meredith, 1993; also known as *error variance invariance* as per Steenkamp & Baumgartner, 1998, or *full uniqueness measurement invariance* as per van de Schoot et al., 2012), the strictest level of measurement invariance. Strict invariance is achieved when the unexplained variance for each item is equal across groups. This would imply identical measurement at the item level of the construct across

groups. Because strict invariance is complete equivalence of the measurement model, it guarantees comparability of a scale across groups, but it has been considered too strict achieve in practice. There is some disagreement on whether scalar invariance is sufficient for mean comparisons in general (Deshon, 2004; Lubke et al., 2003), but scalar invariance remains the commonly accepted standard of measurement invariance in psychology to permit the use of observed scores.

The evaluations of fit and model selection from these levels are like other applications of confirmatory factor analysis, such as chi-square tests, the comparative fit index (CFI), and root mean squared error of approximation (RMSEA) (e.g., Chen, 2007; van de Schoot et al., 2012). For instance, if a chi-square model fit test comparing two invariance models is not statistically significant, then the stricter higher-level invariance model is supported because it has more equality constraints on measurement properties (fewer parameter estimated freely) and is therefore more parsimonious than the lower-level invariance model. Although confirmatory factor analysis forms the foundation of the approaches that will be discussed in this tutorial, the various applications of this approach fall under a family because there is significant variability in how CFAs have been used to assess measurement invariance.

Partial Invariance

Although scalar invariance is the commonly accepted level of invariance for comparing observed means, it is also in itself still a strict criterion that is rarely achieved in practice (van de Schoot et al., 2015), in part because traditional factor analytic approaches test exact equality of all model parameters. A poorly fitting scalar invariance model, for example, does not necessarily imply that all the items are non-invariant; only one non-invariant item in the scale could be enough to result in poor fit of the model. This reasoning similarly applies to the metric and strict

invariance models. Accommodating the possibility that parts of a scale may achieve measurement invariance is the core idea behind *partial invariance*.

Under partial invariance, the model in which measurement invariance fails is examined more closely and statistically adjusted to systematically identify and specify a model in which the specific parameter estimate(s) that are non-invariant are estimated freely (Byrne et al., 1989; Steenkamp & Baumgartner, 1998). Researchers may wish to identify the non-invariant parameter estimate(s) for specific item(s) to remove them from the measure in a scale development study, or they may wish to retain the item(s) on the measure but also estimate a model in which they are estimated freely. A correctly specified partial invariance model can statistically adjust for non-invariance and compare groups on latent (but not observed) means or variances: Once non-invariant item parameters are identified, the invariant items are used as anchors (known as *anchor items* or *referent items*), which correctly sets the scale across groups and allows for unbiased estimates of latent means and variances (Bryne et al., 1989).

There are different methods for identifying which items are non-invariant, which can include backward selection via factor-ratio tests, modification indices, and forward selection (Jung & Yoon, 2016). In all approaches, the measurement invariance model is adjusted by removing the equality constraints for the identified non-invariant items. The factor-ratio test by Rensvold and Cheung (1998) involves testing models representing each possible combination of anchor item and potentially non-invariant item(s) against the configural invariance model, where significant differences in model fit (e.g., chi-square ratio tests) indicate that the new model may contain noninvariant items. Backward selection, as shown by Yoon and Millsap (2007), involves using the largest modification index on a fully constrained metric or scalar invariance model and relaxing the constraints until the largest modification index is no longer statistically significant.

Forward selection, an approach proposed by Jung and Yoon (2016), is analogous to backward selection but tests in order of additions of constraints rather than removals and simplifies the use of multiple tests in data analysis with confidence intervals.

Researchers should consider several points when using partial invariance models. First, we recommend that partial invariance models only be used to make latent comparisons and not justify comparisons with observed scores. Simulation studies indicate that non-invariant items bias observed score comparisons even when a partial invariance model can be specified to adjust latent comparisons (e.g., Chen, 2008; Hsiao & Lai, 2018; Guenole & Brown, 2014; Steinmetz, 2013). Second, there is considerable contention and uncertainty regarding how many non-invariant items are acceptable in a partial invariance model to make valid group comparisons at the latent level, and this problem requires future investigation. On one hand, it is generally agreed that latent comparisons are *statistically* justified with just one invariant item in addition to the anchor item that is assumed to be invariant because they set a comparable scale across groups (Bryne et al., 1989; Steenkamp & Baumgartner, 1998). On the other, it is unclear how many non-invariant items are acceptable for group comparisons to be *conceptually* justified in that the originally operationalized construct has the same meaning as what is being compared with the partial invariance model. Is a construct measured by an entire scale across groups the same as the construct measured with two invariant items? Is a construct measured by five highly non-invariant items across groups the same as the construct measured by the same five items with only slight non-invariance? From this standpoint, researchers have suggested that at least a majority of items should be noninvariant, confidence decreases as the number and degree of noninvariant items increases, and analyses should be supplemented by qualitative theory-based evaluation of the non-invariant items whenever possible (e.g., Chen, 2008; Shi et al., 2019;

Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

The Alignment Approach

Asparouhov and Muthén (2014) developed the alignment method as an alternative to traditional factor analytic approaches for data structures with many groups. We outline the conceptual basis of the alignment method as described by Asparouhov and Muthén (2014), Muthén and Asparouhov (2018), and Lomazzi (2018).

Under the traditional factor analytic approaches, mean comparisons in observed scores across groups are justified if the factor model configuration, factor loadings, and item intercepts are equivalent across groups (i.e., scalar invariance is achieved). Researchers can have different goals when evaluating measurement invariance, but often the goal is to make unbiased factor mean comparisons. The alignment approach works to address this by producing a factor model that is sufficient to make factor mean comparisons—that is, a model with factor loadings and item intercepts that are as close to equivalent as possible. Framed another way, the alignment approach assumes that measurement non-invariance can be minimized, so minor measurement differences (*approximate measurement invariance*) present at the item levels across groups are assumed and adjusted for, i.e., “aligned”.

Alignment Optimization Procedure

Here, we describe the alignment optimization procedure in a non-technical fashion (for mathematical details, see Appendix A; for complete details, see Asparouhov & Muthén, 2014). The alignment optimization procedure involves two models—the original model and the optimized model—which we will denote as M0 and M1 respectively. M0 is produced by transforming a baseline configural model which assumes the same configuration of items to factors across groups, and M1 is produced by optimizing M0. The alignment optimization

procedure produces M1 by minimizing the differences between factor loadings and item intercepts across groups. The factor means and variances that correspond to M1 are then used to make group comparisons. Recall that scalar invariance in a traditional MGCFA requires invariant factor configuration, factor loadings, and item intercepts. The logic of the alignment is that an adequate configural model that has minimal differences in factor loadings and intercepts across groups (i.e., has a majority of factor loadings and intercepts that are approximately equal) should be good enough to make factor mean comparisons. There are no loading, intercept, or residual equality constraints placed on the configural model, so model fit of the original M0 is unaffected by alignment optimization and equal to the model fit of M1.

The optimization procedure works in a similar manner to rotation algorithms used in exploratory factor analyses. Rotation algorithms are designed to extract factors from items that load highly on those factors, but not on others (i.e., to achieve a solution with simple structure and no cross loading). To achieve a simple structure, rotation algorithms maximize big loadings and minimize small loadings such that items load highly on one factor, but not others. The alignment optimization works similarly to achieve a different kind of simple structure: one that minimizes the differences between loadings and intercepts across groups. Just as rotation attempts to select a loading matrix with large loadings on one factor and small loadings on the others, the alignment attempts to find a solution in which most item parameters are approximately equal and there are only a few larger intercept/loading differences across groups.

Overall, the alignment approach is not necessarily a measurement invariance testing procedure, but is rather a treatment of measurement invariance as an optimization problem: It produces a factor model that is good enough to make unbiased latent mean comparisons by selecting factor means and variances that minimize measurement non-invariance of the item-

level parameters. This is done such that most factor loadings and item intercepts are approximately invariant, with a minority of item parameters that have substantial differences across groups. As a result, there are enough invariant items to use this factor model to produce aligned latent scores that are comparable across groups without achieving exact scalar invariance or needing to identify a partial invariance model.

After the alignment procedure produces optimized model M1, there is a separate ad-hoc item-level testing algorithm. This algorithm produces item-level significance tests and non-invariance effect size estimates for all possible pairs of factor loadings and intercepts across groups. Given possibly large numbers of comparisons, these significance tests are interpreted at the .001 level of significance. The non-invariance effect size estimates, denoted as R^2 values by Asparouhov and Muthén (2014), range from 1.00, indicating complete invariance, to 0, indicating non-invariance. This testing algorithm is largely automated and does not require researcher input, contrasting with the traditional approach which involves manual model specification for partial invariance.

There are four key points for applying the alignment method due to how the optimization procedure works. First, the alignment method does not optimize uniquenesses[2] because the primary goal is to estimate unbiased latent factor means for valid group comparisons. Second, configural invariance is an assumption of alignment optimization because only factor loadings and intercepts are optimized in the procedure and an adequate configural model M0 is required for this process. Third, because the optimization procedure works analogously to rotation methods in exploratory factor analyses, the presence of a few large noninvariant parameters and many approximately invariant parameters is another assumption of alignment optimization.

[2] There is an extension of the alignment method which applies to uniquenesses ("alignment-within-CFA") but will not be discussed here. For interested readers, see Marsh et al. (2018).

Fourth, the alignment optimization model can be identified in two ways, which requires researcher input (discussed later in the illustrative example): The factor mean and variance of the reference group can either be fixed to 0 and 1 respectively (FIXED alignment optimization option) or the factor mean can be estimated freely (FREE alignment optimization option).

Traditional Approach versus Alignment

There are several decisions that affect the choice of how to investigate and consider measurement invariance, and as a result, there are many ways that researchers could decide to conduct their analyses that could produce different results (i.e., many researcher degrees of freedom). This makes planning an analysis and navigating those decisions difficult, particularly if the researcher wants to develop an analysis plan before opening the data. Though it can be difficult to develop a priori analysis plans for complex models, having some plan is better than having no plan (Nosek et al., 2019). To address this, we provide an explicit list of considerations and decisions researchers can use to plan their analysis and increase their transparency when choosing between the traditional factor analytic approach, the alignment method, or a combination of both. Then, using an illustrative dataset, we walk through a detailed example of making these decisions and implementing them in a preregistered analysis plan. The list of considerations and decisions is briefly summarized in Table 1 and the preregistration is in Appendix B.

Table 1

Summary of Considerations for Measurement Invariance Testing in Analysis Planning

Decision 0: Prerequisites		
Considerations	Traditional	Alignment
Factor structure	Cite previous studies and conduct CFA on current sample	Same as traditional
Sample size	Requires large sample sizes based on literature review and/or simulation studies	Same as traditional
Assumptions	Check number of scale points and multivariate normality	Same as traditional
Configural invariance	Test configural invariance	Same as traditional
Decision 1: Research Goal		
Observed or factor scores	Compare observed scores and/or compare factor scores	Compare factor means and variances
Model complexity	Use with longitudinal designs, covariates, or cross-loadings	Cannot use with longitudinal designs[a], covariates, or cross-loadings
Decision 2: Model Identification		
Identification: CFA	Choose marker item or variance standardization	Same as traditional
Identification: MGCFA	Consider based on research goal	Use FIXED option if 2 groups, FREE otherwise
Anchor item	Consider theory-based, iterative, or significance-based selection strategies	No anchor items
Decision 3: Model Evaluation		
Configural model	Check model chi-squared and fit indices (e.g., point estimates, permutation tests, dynamic, equivalence tests)	Same as traditional
Metric/scalar/strict models	Check model fit differences (e.g., chi-squared difference test and model fit index differences)	No subsequent models; check number of non-invariant items (e.g., 25% rule, R^2) and impact of non-invariance
Partial invariance models	Check model fit differences (e.g., modification indices)	No partial invariance models

[a] See Lai (in press) for a very recent extension of the alignment method for longitudinal models.

Decision 0: Prerequisites for Both Methods

Before considering a measurement invariance analysis, researchers must consider the basic psychometric requirements that are shared by both the traditional factor analytic approach and the alignment method. Specifically, a tenable configural invariance model is fundamental to both methods, and so a configural invariance test is the starting point for either approach. Because an MGCFA underlies the configural model, the requirements for MGCFAs carry over to both methods. Thus, before researchers can consider any measurement invariance analysis, they should check and account for these three requirements in study planning and data.

Evidence of Factor Structure. Researchers should only consider measurement invariance testing for scales that have a known factor structure in at least one group or sample, ideally with existing confirmatory evidence (i.e., confirmatory factor analyses). Issues with factor structure can be avoided by selecting developed scales with strong validity evidence, but this is not always possible. However, regardless of whether previous evidence is available, we recommend that researchers confirm the factor structure of the scale in their own sample by conducting a confirmatory factor analysis on the entire sample. This is because a known factor structure for the scale is a necessary requirement for testing configural invariance. There is little point overall in testing measurement invariance across multiple groups if the scale's factor structure cannot be supported in even one group. There is also no way to test measurement invariance if the factor structure is not known because it would be impossible to specify the factor models in either method. Moreover, this preliminary check helps catch mistakes that can cause subtle but disastrous downstream analytical errors—mistakes such as mislabeled items, mistakenly mis-specified factor models, and scoring errors—so that they can be corrected before conducting and interpreting the more complex measurement invariance analyses.

Sample Size. Researchers should have a large sample size for each group when using either approach because latent variable models rely on large sample sizes to achieve adequate statistical power and precision. Existing simulation studies based on the traditional approach appear to suggest a minimum of 400 participants per group (e.g., French & Finch, 2006; Meade & Bauer, 2007; Meade et al., 2008; Koziol & Bovaird, 2018), but we emphasize that this should be used as a starting point, and there is a need for further research and consideration of other aspects that impact sample size requirements. For the traditional approach, sample size requirements can increase depending on the complexity of the analysis because statistical error rates are inflated by additional hypotheses. This can include when there are many items in the scale, when there are more than two groups of interest, and when there are partial invariance analyses. For the alignment method, such multiple comparisons are avoided as it was designed with many-groups analyses in mind, but there is a trade-off as a result: Type I error is adjusted in the item-level analyses, so as the amount of items and groups increases, statistical power decreases, thus increasing the required sample size. The nature of this trade-off is not yet well understood and requires further research (e.g., Flake & McCoach, 2018). Overall, both methods are generally large-sample techniques, and this should be accounted for in study design and before considering any measurement invariance analyses.

Assumption Checks. Researchers should check the assumptions of MGCFAs before using either approach. The two most pertinent assumptions pertain to maximum likelihood estimation: The items should be measured on a continuous scale (or can safely be treated as continuous) and follow a multivariate normal distribution. Multivariate normality can be tested in various ways, including but not limited to examination of item-level distributions and normality hypothesis tests. Likert-type items are, by definition, measured on an ordinal scale

(i.e., discrete or categorical), but methodological research suggests that they can be acceptably treated as continuous for confirmatory factor analyses if they are measured on at least five scale points (e.g., Rhemtulla et al., 2012). Violations of these assumptions can affect model fit tests and fit indices, which consequently affect measurement invariance results (Lubke & Muthén, 2004). Researchers can account for this under both methods by selecting an alternative estimation strategy for the MGCFA such as weighted least squares (Flora & Curran, 2004) or robust maximum likelihood estimation.

Planning Measurement Invariance Analyses

Once the prerequisites are met, researchers can then consider which approach they should use and how to conduct the analysis. We present three decisions, in temporal order, that researchers should consider when planning a measurement invariance analysis, whether it is the traditional approach, the alignment method, or both.

Decision 1: Choosing the Best Approach(es) for the Research Goal

Perhaps the most important consideration when deciding between the two approaches is the goal and purpose of the measurement invariance investigation. We suggest researchers consider two main types of goals: (1) developing and evaluating a scale to modify or improve it by ensuring there is invariance and/or (2) obtaining a model that allows for group mean comparisons either via observed scores or latent scores. The researcher may have both goals or may focus on one over the other. We discuss how these goals can guide choosing between when and how to use each approach.

Traditional. The traditional approach can be used to meet both goals and accommodate the use of observed or latent scores to make comparisons of means and variances. The traditional approach is more amenable to the first goal of scale development and modification because

targeted item-level analyses can be conducted to identify which items are non-invariant. Through partial invariance testing, researchers can compare models with different specifications and levels of non-invariance. However, the traditional approach requires the researcher to specify which models to execute, in what order, and what item-level follow-up tests will be conducted. Through this process, the researcher could determine a set of invariant items to continue in the scale development process. We recommend replication analyses of any such model, given the exploratory nature of the analyses and the number of model comparisons needed.

If the goal of the researcher is to evaluate whether a scale's observed scores can be used to compare groups, that can be achieved with the traditional approach by focusing on evaluating scalar or strict invariance. If scalar or strict invariance is not met to justify the use of observed scores, researchers can compare and test a series of models to identify a partially-invariant model. A correctly specified partial invariance model accommodates comparisons of latent means and variances.

Alignment. The alignment method can be used to meet both goals in most cases but is more amenable to meeting the goal of using latent scores to make group comparisons of factor variances and means. The alignment method does not allow for the testing of specific models with differing levels of measurement invariance, but instead fully automates the procedure of identifying non-invariant items. The alignment method is appropriate for practical use to answer substantive research questions using optimized latent means and variances, particularly when metric or scalar invariance fails under the traditional approach (Marsh et al., 2018).

Though the results indicate which items are non-invariant, the alignment optimization was not designed to evaluate whether instruments can produce unbiased observed group means. The optimization assumes that most items are approximately invariant to estimate unbiased

latent means. Thus, it is unclear whether items are invariant enough to produce unbiased observed means if the item testing procedure results indicate all items are approximately invariant. Further, no research points to what pattern of results would indicate that the instrument will produce unbiased observed scores (e.g., number of tolerable non-invariant items). This is an important area for future investigation, but we currently cannot recommend that the alignment results inform the usage of observed scores. The alignment method could be used as an exploratory analysis to identify non-invariant items, but we suggest that if researchers want to evaluate the use of observed scores, they should plan to conduct a sensitivity analysis comparing any latent estimates to observed estimates. If results differ, that may suggest the observed scores are biased. Further, the alignment method cannot accommodate longitudinal models[3] (Marsh et al., 2018) or models with cross-loadings or covariates.

Decision 2: Identifying the Model

Structural equation modeling always requires model identification decisions. The traditional and alignment approach differ in their assumptions regarding identification. Researchers can consider this ahead of time to plan their analysis.

Traditional. The same challenges of model identification from confirmatory factor analysis and structural equation modeling more broadly are present in the traditional approach (Bollen, 2014), as is the requirement of setting a scale to provide a metric for the latent construct (Johnson et al., 2009). Additionally, to compare the measurement of items across the groups, at least one item in the scale must be fixed as an anchor item and assumed to be equal across groups (Johnson et al., 2009). However, anchor items carry with them the assumption of invariance that cannot be understated but is also rarely substantiable: How can a researcher be

[3] The alignment method was very recently extended to apply to longitudinal models (an extension of "alignment-within-CFA") but will not be discussed here. For interested readers, see Lai (in press).

sure that their selection of an anchor item is correct? Approaches to selecting an anchor item or items can vary (e.g., theory-based, iterative, significance-based), and the results and performance of measurement invariance tests can vary based on the choice of anchor item (e.g., Wang & Yeh, 2003; Meade & Lautenschlager, 2004; Stark et al., 2006; Meade & Wright, 2012). Specific details on methods for choosing anchor items and their implications are beyond the scope of this tutorial, so for demonstration purposes, we will opt for an informal content review of the items.

Alignment. The alignment method assumes minimal non-invariance: Most of the items should be approximately invariant, but researchers do not indicate any specific non-invariant items ahead of time. However, researchers must choose how to identify the model with respect to the scaling of the latent factor means and variances. There are two options: The factor mean and variance of the first group can either be fixed to 0 and 1 respectively (FIXED alignment optimization) or can be estimated freely (FREE alignment optimization). As per Asparouhov and Muthén (2014), the decision is generally straightforward and can be made by the number of groups being compared: FIXED must be used if there are only two groups, and FREE can be used if there are three or more groups.

Decision 3: Evaluating the Model

Traditional. CFA underlies all aspects of the traditional approach, making model fit criteria crucial. However, researchers are faced with a variety of recommendations: Many cite guidelines such as from Hu and Bentler (1999) to compare a set of model fit indices (e.g., the configural model might be considered to fit well if its CFI > .95, RMSEA < .06, and standardized root mean square residual (SRMR) < .08). This is because chi-square model fit tests are sensitive and almost always rejected with large sample sizes, and CFA is a large-sample technique, meaning the test will likely be rejected in most cases. Thus, for the configural model, we recommend that

the chi-square test be reported but the evaluation of model fit be based primarily on model fit indices.

To determine whether metric, scalar, and strict measurement invariance are supported, researchers would conduct chi-square model fit difference tests between successive models at $\alpha = .05$ and examine changes in fit indices between the models. Here, failing to reject the null implies that the two models fit equally well and thus provides support to the higher measurement invariance model (fewer estimated parameters or higher degrees of freedom makes the higher model preferable due to parsimony). Researchers should consider how much model misfit is needed to reject the next model. Chen (2007) suggests increases in RMSEA by more than .015 or decreases CFI by more than .01, can be interpreted as failure to support the higher-level measurement invariance model. Conventionally, we recommend that researchers report all three methods and clearly specify decision rules for how they will interpret them ahead of time. For example, researchers could specify that they will report both chi-squared model fit difference tests and model fit index difference guidelines but provide rationale for their interpretation (e.g., acceptable model fit index differences will be interpreted as adequate fit regardless of the chi-squared test results due to large sample sizes). Though these decisions rules are difficult to develop a-priori, they provide guidance in the face of conflicting findings and can limit the inclination to cherry pick results.

Note, however, that there is considerable contention regarding these conventional recommendations. Hu and Bentler's (1999) guidelines, for example, are popular but are one of several guidelines of only a subset of fit indices (e.g., Hooper et al., 2008; Kline, 2015) and only apply to the specific conditions that the original authors investigated (Hu & Bentler, 1998, p. 446). These points are also true for the model fit comparison criteria suggested by Chen (2007).

Indeed, recent research suggests the use of dynamic fit index cut-offs that are computed based on the characteristics of the examined factor model and not universally fixed (McNeish & Wolf, 2021). Moreover, equivalence testing approaches with multi-group structural equation modeling have demonstrated some evidence of superior performance to both the chi-square test and fixed fit index approaches with respect to error control, but may require greater sample sizes to achieve adequate statistical power (Yuan & Chan, 2016; Counsell et al., 2020). Permutation methods, which generate empirical distributions for model fit measures, also present Type I error control advantages over conventional approaches (Jorgensen et al., 2018). Overall, we additionally recommend that the choice of model fit criteria be clearly specified a priori and, if feasible, in consideration of model and design characteristics.

Partial Invariance. Model fit criteria are also necessary for researchers to determine whether partial invariance analyses will be conducted. Here, we recommend that researchers specify the following: (1) whether partial invariance analyses will be conducted or not upon failure of achieving metric or scalar invariance based on the specified criteria, (2) how non-invariant items will be identified and accounted for, and (3) how the final partial invariance model will be used to address the research goal, e.g., to remove non-invariant items or to retain them but estimate them freely in a structural equation model. We encourage researchers to consider under what circumstances they will conduct a partial invariance analysis ahead of time because downstream results (latent versus observed means) could differ across models. For example, a preregistration could specify that a partial invariance analysis will only be conducted if one of the model evaluation criteria indicates a lack of invariance, or only if all model evaluation criteria converge to a conclusion of failing to meet invariance.

Alignment. Model fit criteria are relevant only to finding a well-fitting baseline model.

The fit does not change from the baseline configural model because alignment does not apply constraints or formally test any additional models. Like the traditional approach, researchers should focus on deciding their criteria for a well-fitting measurement and configural model ahead of time. The other aspect of model evaluation for the alignment is ensuring minimal non-invariance: The performance of the alignment solution is evaluated via assumption checks and item-level analyses, primarily the number of significantly non-invariant items, their degree of non-invariance, and the contribution of each item to total non-invariance. Based on Monte Carlo simulations, Muthén and Asparouhov (2014) suggested a rule of thumb that no more than 25% of items should be non-invariant based on the item-level significance tests for good performance (interpreted at $\alpha = .001$). This was supported in simulations from Flake and McCoach (2018) with good performance when less than 29% of items are non-invariant.

Furthermore, researchers can assess the R^2 invariance effect size measure, which quantifies how much variability in the item parameter estimates can be explained by the groups' factor means and variances. An R^2 near 1 indicates complete invariance because the variability in item parameters is completely explained by group mean differences, whereas an R^2 near 0 indicates that group mean differences explain none of the variability in the item parameter. However, exact guidelines for assessing this degree of invariance or performance are not yet clearly established and require further investigation. Because of this, we also recommend examining the magnitude of the item differences via raw and/or standardized effect sizes (e.g., Gunn et al., 2020) for each item-level test to gauge whether potential deviations due to non-invariance are meaningful.

Illustrative Example: Consideration for Future Consequences Scale Across Sexes

Next, we demonstrate the conceptual and empirical implications of the traditional model

comparison and alignment method approach by illustrating how to plan a measurement invariance analysis using the Consideration for Future Consequences Scale (CFC). The CFC measures how people consider the future consequences of their current behavior and how much their behaviors are influenced by those future consequences (Strathman et al., 1994). Participants indicate their agreement to 12 items on a 5-point scale (1 = *Extremely uncharacteristic*, 5 = *Extremely characteristic*). Construct validation evidence from Petrocelli (2003) and Joireman et al. (2008) suggests that the CFC scale, as originally developed, measures two future consequence constructs: a future concern sub-factor, which is measured with four items (e.g., "I am willing to sacrifice my immediate happiness or well-being in order to achieve future outcomes."); and an immediate concern sub-factor, which is measured with eight items (e.g., "I only act to satisfy immediate concerns, figuring the future will take care of itself."). For simplicity of illustration, we limit our example to a test of one of the subscales across two groups. We evaluate the measurement invariance of the 8-item immediate concern sub-scale ("CFC-Immediate") across sex (male and female) with the goal of comparing mean scores (latent or observed) on consideration for future consequences across males and females.

The data for the CFC was acquired from the Open Source Psychometrics Project (openly available at https://openpsychometrics.org/_rawdata/). For illustration purposes, we removed missing data on any of the eight items of interest or on sex on a listwise basis, resulting in an effective sample size of 14,598 participants (54% female; original n = 15,035). We performed the analyses for the traditional factor analytic approach using *R* version 4.0.3 with the *lavaan* package version 0.6-7 (as of writing, the alignment method can only be correctly implemented in *Mplus*). We duplicated the analyses for the traditional factor analytic approach and performed the alignment method in *Mplus* version 8.4. All materials can be accessed in the Supplementary

Materials here: https://osf.io/3p7n9/ .

Illustrative Analysis Plan Example

Below, we walk through the decisions in example form of how researchers could structure, develop, and rationalize an analysis plan with each approach. Though we provide examples of decisions researchers can make, we want to emphasize that other decisions can be made with adequate justification. Our goal is to demonstrate how to make and justify decisions ahead of time to develop an a priori analysis plan, not to dictate the only way one can proceed with a measurement invariance analysis. This can be used as an example template for a preregistration of a measurement invariance analysis (see Appendix B). First, we will examine the prerequisites to determine whether measurement invariance testing is feasible with either approach. We will then walk through the decisions for both the traditional factor analytic approach and the alignment method.

Decision 0: Prerequisites for Both Methods

Evidence of Factor Structure. The CFC scale is a relatively well-known scale with a known factor structure substantiated by some confirmatory evidence. Construct validation evidence from Petrocelli (2003) and Joireman et al. (2008) suggests that the CFC scale, as originally developed, measures two future consequence constructs: a future concern sub-factor, which is measured with four items; and an immediate concern sub-factor, which is measured with eight items. We subsequently conducted a CFA on the overall sample using this factor structure specification (estimated with MLR due to multivariate non-normality; see Assumption Checks). As per Hu and Bentler (1999), we deemed the CFA to fit well if its CFI > .95, RMSEA < .06, and standardized root mean square residual (SRMR) < .08. We found that the factor structure was indeed supported in our sample with good model fit, $\chi^2_{Y-B}(20) = 919.74$, $p < .001$,

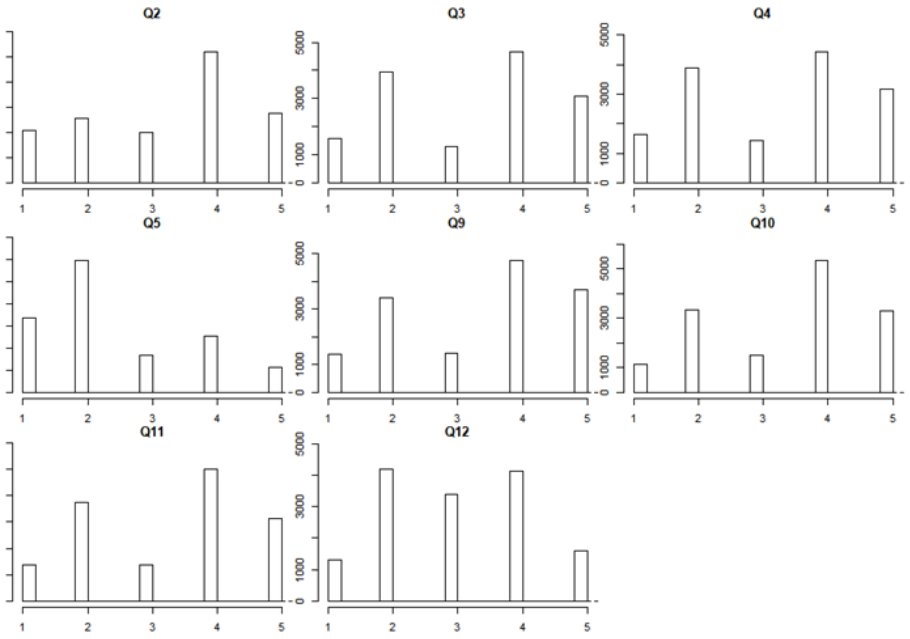
Robust CFI = 0.972, Robust RMSEA = 0.060, 90% CI [0.057, 0.064], SRMR = 0.023. Overall, we can conclude that there is adequate knowledge and evidence of factor structure of the CFC scale to consider conducting measurement invariance tests.

Sample Size. We had over 7,000 female participants and over 6,000 male participants, which far exceeds the suggested sample size of 400 participants per group as determined from our review of simulation studies in the measurement invariance literature (e.g., French & Finch, 2006; Meade & Bauer, 2007; Meade et al., 2008; Koziol & Bovaird, 2018). We were also only investigating two groups with a single 8-item subscale, which greatly minimizes the possible complexity of the analyses, even when considering possible partial invariance analyses. Overall, we could justify that we had an adequate sample size to consider conducting measurement invariance tests.

Assumption Checks. The CFC-Immediate subscale is a 5-point Likert-type scale, which meets the minimum amount of scale points required to be safely treated as continuous. However, we found that our data violated the assumption of multivariate normality (e.g., clearly non-normal item-level distributions, which necessarily imply multivariate non-normality; see Figure 2). To account for this, we used robust maximum likelihood estimation with the Yuan-Bentler scaled chi-squared statistic (MLR; Yuan & Bentler, 2000) and robust standard errors for all CFAs and measurement invariances tests. Overall, we could conclude that we have met the assumptions required to consider measurement invariance tests.

Figure 2

Item Score Distributions for the CFC-Immediate Scale.



Decision 1: Choosing the Best Approach(es) for the Research Goal

Now we can decide between the traditional factor analytic approach and/or the alignment method. As mentioned previously, the illustrative goal is to evaluate the measurement invariance of the 8-item immediate concern subscale across sex to ultimately compare mean scores (latent or observed) on consideration for future consequences across males and females.

Traditional. The traditional approach can accommodate this research goal regardless of whether the comparison is made on latent or observed means. If we can conclude at least complete scalar invariance of the model we can use the observed means or if we can identify a partially invariant model, we can use the latent means.

Alignment. There are no expected cross loadings, covariates, or other sources of model complexity that the alignment method cannot accommodate. Therefore, the alignment method can accommodate this research goal by comparing latent means.

Decision 2: Identifying the Model

Traditional. In practice, it is helpful to track the number of parameters and degrees of freedom based on the data and varying model identification strategies available to researchers under the traditional approach (see Supplementary Materials). To identify each model, we fixed the loading of the anchor item to 1 and factor means to 0 respectively to both groups. As mentioned previously, we reviewed the content of the items and selected the item Q2 that was deemed least likely to be non-invariant as the anchor item.

Alignment. We fixed the factor mean and variance to 0 and 1 respectively because we were only comparing two groups (i.e., the FIXED alignment configuration).

Decision 3: Evaluating the Model

Traditional. We followed the most popular conventional recommendations for model fit indices, chi-squared model fit tests, and model fit differences. For all models, we reported both the chi-square model fit test and multiple additional fit indices. To evaluate the overall factor model across both groups as well as the baseline configural model, we reported the total model chi-square and the CFI, RMSEA, and SRMR. If the chi-square test was significant, which was likely given the large sample size, we deemed the overall factor model and configural model to have acceptable fit to move forward with invariance testing if CFI > .95, RMSEA < .06, and standardized root mean square residual (SRMR) < .08 (Hu & Bentler, 1999). Then, to determine whether metric, scalar, and strict measurement invariance were supported, we reported the chi-squared model fit difference tests and model fit index differences between successive models. We concluded that the next level of invariance was not supported if the chi-square test was significant at $\alpha = .05$ and/or the higher-level model increased RMSEA by more than .015 or decreased CFI by more than .01 (Chen, 2007). Thus, if the two criteria disagreed, we returned to

the level of measurement invariance that failed and conduct a partial measurement invariance analysis.

Partial Invariance. Given that the research goal was to compare means across sex, regardless of whether they are latent or observed, we planned to proceed with partial invariance analyses if metric or scalar invariance was not supported by either the chi-square difference test or differences in model fit indices. We employed a backward-selection approach using modification indices to identify non-invariant items.

Specifically, we returned to the model in which that level of measurement invariance failed, identified the first item that is most non-invariant (i.e., the item parameter with the greatest modification index), constrained the loadings and/or intercepts of all items except the non-invariant item to be equal across groups, and compared the fit of the new model to the old model in which measurement invariance was achieved. If there was no evidence that the models differed in fit, as determined by chi-squared model fit difference tests and differences in model fit indices, then partial invariance was established. However, if there was still a comparative difference in fit between the new and old model, we proceeded to the next most non-invariant item, allowed its loading and/or intercept to freely vary alongside the first item, and re-tested the new model's fit again against the model in which measurement invariance was achieved. We repeated this process until partial invariance is established or modification indices no longer indicated significant improvements in model fit (MIs < 3.84, which is the critical value for chi-squared tests for $df = 1$ at $\alpha = .05$). Once the final partial invariance model was established, we used it to estimate latent factor scores to use for statistical analysis instead of the observed scores.

Alignment. For the baseline configural model, we followed the most popular

conventional recommendation for model fit indices: As per Hu and Bentler (1999), we deemed the configural model to fit well if its CFI > .95, RMSEA < .06, and standardized root mean square residual (SRMR) < .08. For evaluating the performance of the alignment optimization, we followed Muthén and Asparouhov's (2014) rule of thumb in which no more than 25% of parameters are non-invariant to conclude good performance.

Presentation and Interpretation of Results

Traditional Factor Analytic Approach

We show how overall model fit comparison results can be summarized in a manuscript in Table 2.

Table 2

CFC-Immediate Fit Indices for Configural, Metric, and Scalar Invariance Models

Model	χ^2_{Y-B}	df	*p*	CFI	RMSEA (90% CI)	SRMR
1. Configural	943.05	40	< .001	0.97	0.061 [0.057, 0.064]	0.023
2. Metric	982.67	47	< .001	0.97	0.056 [0.053, 0.059]	0.024
1 vs. 2	7.80	7	.350	< .001	-.0047	
3. Scalar	1049.17	54	< .001	0.97	0.053 [0.051, 0.056]	0.025
2 vs. 3	56.50	7	< .001	.001	-.0026	
4. Strict	1080.09	62	< .001	0.97	0.050 [0.047, 0.053]	0.026
3 vs. 4	15.75	8	.0462	< .001	.0034	

Note. Fit indices are robust forms.

Configural Invariance. To test CFC-Immediate for configural invariance across sex, we conducted a multi-group confirmatory factor analysis where all loadings, intercepts, and error variances are freely estimated (only the Q2 loading and factor means are constrained to equality across sex for identification). The configural invariance model met our criteria for good fit based on fit indices, $\chi^2_{Y-B}(40) = 943.05$, $p < .001$, Robust CFI = .972, Robust RMSEA = .061, 90% CI [.057, .064], SRMR = .023. As discussed previously, the chi-square test is likely to be rejected even in the presence of adequate fit indices. Based on our model evaluation criteria, we

determined that configural invariance is supported.

Metric Invariance. Because the scale satisfied configural invariance, we proceeded to test metric invariance. To test the scale for metric invariance across sex, we built upon the previous multi-group confirmatory factor analysis by constraining the seven loadings to be equal across sex. After specifying this new model, we determined whether metric invariance was supported by comparing the configural invariance and metric invariance models using the Yuan-Bentler scaled chi-squared model fit difference test and the differences in CFI and RMSEA. Results indicated no significant difference in model fit between models, $\Delta\chi^2_{Y-B}(7) = 7.80$, $p = .350$, ΔCFI = < .001, ΔRMSEA = .0047. Therefore, metric invariance is supported.

Scalar Invariance. Next, because CFC-immediate also satisfied metric invariance, we proceeded to test scalar invariance. To test scalar invariance across sex, we again built upon the previous multi-group confirmatory factor analysis by additionally constraining the seven item intercepts to be equal across sex. Like testing metric invariance, we determined whether scalar invariance is supported by comparing the metric invariance and scalar invariance models by again using the Yuan-Bentler scaled chi-squared model fit difference test and the differences in CFI and RMSEA. Results from the chi-square test but not the fit indices indicated that the metric invariance model fit significantly better than the scalar invariance model, $\Delta\chi^2_{Y-B}(7) = 56.50$, $p < .001$, ΔCFI = .001, ΔRMSEA = .0026.

There are two possible interpretations: Scalar invariance was not supported due to the rejection of the chi-square test, or scalar invariance was supported due to no deterioration of model fit indices when comparing the metric to the scalar model. If we conclude the former for illustration, then we can make observed mean comparisons of the observed scale scores across sex: There was no evidence that males ($M = 3.19$) differed from females ($M = 3.17$) on

consideration for immediate future consequences, $t(13{,}898) = 1.78$, $p = .0748$, $d = -0.03$ (95% CI [-0.06, 0.00]). However, because we have a case of conflicting evidence between model comparison tests and fit indices, we proceeded to conduct partial invariance analyses and compare factor means from the final partial invariance model as per our analysis plan.

Partial (Scalar) Invariance. To establish a partial scalar invariance model, we revisited the original scalar invariance model and computed modification indices to identify the most non-invariant item intercept. Modification indices indicated that freeing the Q9 intercept would result in the greatest significant model fit improvement (MI = 16.29), so we freed that parameter, compared the new partial scalar invariance model to the metric invariance model, and repeated the process until an acceptable model was achieved. Through this iterative process, we established a partial scalar invariance model by freely estimating item intercepts for Q9, Q12, Q2, Q5, and Q10 (see Table 3).

Table 3

CFC-Immediate Partial Scalar Invariance Model Comparisons

Model (Freed Intercept)	Modification Index	$\chi^2_{Y-B}(1)$	p
Model 1 (Q9)	16.29	38.13	< .001
Model 2 (Q12)	12.44	24.08	< .001
Model 3 (Q2)	8.34	14.65	.0055
Model 4 (Q5)	5.94	7.92	.0477
Model 5 (Q10)	6.63	0.42	.810

Note: Partial scalar invariance models were compared to the metric invariance model.

Based on this partial scalar invariance model, males reported greater latent immediate consideration for future consequences than females, $\Delta M_{F-M} = -0.030$, $p = .021$. Though this mean difference comparison is statistically significant whereas the observed score analysis was not, the results do not conflict substantively. The latent mean difference of .03 and the observed

standardized difference of .03 are nearly identical and small. Given the large sample size, an applied researcher could interpret these results as consistent: There is no meaningful difference between males and females on this construct. Comparing the latent mean difference from the partial invariance model to the observed score difference provides a sensitivity analysis and demonstrates that the mean difference (latent or observed) is not sensitive to the effect of the non-invariant intercepts, even when the majority of items were non-invariant. This is logical given that there were conflicting model fit results and the differences in the intercepts was small (0.048 to 0.080; see Supplementary Materials for full output).

Strict Invariance. Scalar invariance was partially supported as per our evaluation criteria, so we proceeded to report the strict invariance model for illustrative purposes. We built upon the full scalar invariance model by additionally constraining the uniquenesses of each of the eight items to be equal across sex. Like testing scalar invariance, we determined whether strict invariance is supported by comparing the scalar invariance and strict invariance models using the Yuan-Bentler scaled chi-squared model fit difference test and the differences in CFI and RMSEA. Results from the chi-square test but not the fit indices indicate that the scalar invariance model fits significantly better than the strict invariance model, $\Delta\chi^2_{Y-B}(8) = 15.75$, $p = .0462$, ΔCFI < .001, ΔRMSEA = .0034. Similarly, the strict invariance model indicated no evidence that males differed from females in consideration for immediate future consequences, $\Delta M_{F-M} = -0.024$, $p = .056$.

Alignment Method

Configural Invariance. The first assumption of the alignment method is configural invariance. Therefore, before beginning an alignment, we followed the same first steps for establishing configural invariance as the traditional approach (i.e., the MGCFA across groups

with no constrained parameters). As was illustrated for the traditional approach, the configural invariance model fit well, and this identification strategy does not affect model fit, so configural invariance is established. We were thus justified to proceed with the alignment method.

Alignment. Because configural invariance was established, we used the configural model for alignment with the FIXED specification (required when testing two groups). As discussed previously, alignment produces a solution that allows for factor mean comparisons and ad-hoc item invariance analysis, accounting for small amounts of measurement non-invariance. There are three results of interest: pairwise comparisons for factors means in each group, pairwise comparisons for invariance of factor loadings between each group, and pairwise comparisons for invariance of item intercepts between each group. Prior to examining the factor means, we first examined the pairwise comparisons for loadings and intercepts to identify any noninvariant items. As per Asparouhov and Muthén (2014), these pairwise comparisons are corrected for multiplicity in the alignment algorithm and interpreted at $\alpha = .001$.

Factor Loading Invariance. Table 4 shows the estimated factor loadings and pairwise comparisons between sexes. There was no evidence that factor loadings produced by the alignment solution differed across sex for any of the items, *ps* > .01. The R^2 statistic provides a measure for the degree of invariance for the parameter in that it quantifies how much variability in the parameter can be explained by the groups' factor means and variances. Higher values correspond to higher degrees of invariance, with values near 1 indicating complete invariance. The presence of items with high R^2 values is indicative of good performance of the alignment method, even if some items have low R^2 values (Muthén & Asparouhov, 2018). Indeed, most items here showed high R^2 values except Q9, which therefore indicates that the alignment method performed well.

Table 4

CFC-Immediate Pairwise Factor Loading Comparisons Across Sex

Item	λ_{Male}	λ_{Female}	$\lambda_{Female} - \lambda_{Male}$	$SE_{\lambda Female - \lambda Male}$	p	R^2
Q2	0.74	0.70	-0.044	0.021	.037	.73
Q3	1.06	1.07	0.006	0.015	.682	.99
Q4	1.01	1.02	0.015	0.016	.369	.89
Q5	0.55	0.55	-0.004	0.018	.803	.99
Q9	0.80	0.83	0.028	0.018	.130	< .01
Q10	0.87	0.87	-0.001	0.017	.972	1.00
Q11	1.08	1.09	0.003	0.014	.809	1.00
Q12	0.63	0.63	0.001	0.017	.933	1.00

Note. λ refers to factor loadings.

Item Intercept Invariance. Table 5 shows the estimated item intercepts and pairwise comparisons between sexes. There was evidence that three item intercepts produced by the alignment solution differed across sex for Q2, Q9, and Q12 (*ps* < .001), which exceeds our prespecified 25% rule of thumb (three non-invariant items out of eight). However, these intercept differences, although statistically significant, do not appear to be meaningful, especially with respect to the scale of the measure (e.g., less than 0.1 on a 5-point scale, or less than 2%), and they also differ in direction. This suggests that whatever bias may be present with these non-invariant items will not meaningfully affect interpretation of factor means. Indeed, the sum of the differences is about -0.012 points. We otherwise see several extremely low R^2 values despite all pairwise comparisons being nonsignificant, but the presence of high R^2 values such as Q3 indicate that the alignment procedure is performing well.

Table 5

CFC-Immediate Pairwise Item Intercept Comparisons Across Sex

Item	τMale	τFemale	τFemale - τMale	SEτFemale - τMale	p	R^2
Q2	3.32	3.26	-0.061	0.018	.001	
Q3	3.28	3.27	-0.006	0.012	.648	.98
Q4	3.27	3.25	-0.018	0.013	.185	.86
Q5	2.44	2.49	0.047	0.018	.009	< .01
Q9	3.38	3.46	0.071	0.018	< .001	
Q10	3.43	3.46	0.037	0.015	.011	< .01
Q11	3.36	3.34	-0.015	0.012	.204	.90
Q12	3.08	3.01	-0.067	0.017	< .001	

Note. τ refers to item intercepts.

Factor Mean Comparison. Though results indicated the alignment method did not produce a valid solution in line with our preregistered cut-off of 25% or less non-invariant items, our follow up investigation of the raw and standardized effect sizes of the item differences suggested the solution was valid because the item differences were extremely small. Thus, we compared the aligned factor means of the CFC-Immediate for each sex produced from the solution (Male as reference group). There was no evidence that males and females differed in immediate consideration for future consequences, $\Delta M_{F-M} = -0.029$, $p = .097$. When such instances arise, we recommend that researchers clearly state how their analyses, reporting, and interpretation deviated from the original plan. Then, in subsequent preregistrations, they can incorporate the added analyses or investigations into decision making criteria.

Discussion

In this tutorial, we described two approaches to measurement invariance testing: The traditional approach using MGCFAs and the alignment method. We then illustrated how to develop an analysis plan for both methods side-by-side by walking through considerations step-by-step. Here, we will describe key similarities, differences, and future areas of research that

would facilitate ease of use and interpretation for both methods.

Procedural Comparison

Similarities

As was illustrated in our step-by-step comparison, both methods begin with the same prerequisite checks. Overall, the traditional approach works with many of the same steps and considerations as the alignment method: The measure needs a confirmed factor structure and evidence of configural invariance before additional testing can be carried out. Should the configural model be untenable, both approaches would also not be feasible. Though not the focus of this tutorial, it is worth noting that if there are more than two groups, evaluating configural invariance is onerous, requiring an evaluation of the factor structure in each group and then in comparison across groups, but this is necessary for both methods. Therefore, both methods share the same planning requirements for prerequisites and configural invariance.

Differences

Perhaps the starkest difference is in labour and specialized knowledge that a researcher must possess to run and interpret the two methods. Whereas the traditional approach is largely directed by researcher decisions and model specifications at every step, the alignment method only requires specification of a configural model and otherwise handles the optimization procedure and item-level analyses automatically. The additional knowledge requirement and risk of error under the traditional approach is nontrivial: From a wide pool of options, researchers must decide on model identification strategies across multiple models, selection strategies for anchor items, and model fit criteria for interpreting many model comparisons—all of which are not decisions needed for the alignment method. While navigating all the decisions, researchers could also inadvertently engage in questionable measurement practices (Flake & Fried, 2020) as

they conduct many sets of slightly different analyses, potentially producing different downstream conclusions. With just one error of inference or misspecification, the researcher could continue along the wrong path and produce additional false positives (Asparouhov & Muthén, 2014; Simmons et al., 2011). For example, a researcher could select the wrong anchor item or flag the wrong non-invariant items when conducting the potentially dozens of statistical tests needed to identify a non-invariant item and then, uncertain of if they made the right decision, select different items and rerun the analysis. Overall, it is easier to get lost in a garden of forking paths with the traditional approach (Gelman & Loken, 2014), whereas there is less planning involved for the alignment method simply because there are fewer decisions that the researcher needs to make.

These risks were made clear by the partial scalar invariance analysis: Model evaluation criteria were conflicting, and so we could have reasonably decided to conduct the partial invariance analysis or not. Having taken a conservative approach by conducting the analysis if there were any conflict in criteria, we manually identified and tested five different partial scalar invariance models. Notably, the final partial invariance model produced a statistically significant group difference whereas our observed score and alignment analyses did not. These conflicts can put researchers in a difficult position: How do they decide if non-invariance is practically significant? Here we suggest researchers consider what differences at the item and factor levels would be substantively meaningful ahead of time. In our example, item-level intercept differences ranged from 0.006 to 0.080 across both methods (less than 2% of the scale). Mean differences were also consistently small across all methods: 0.026 points for the observed difference, 0.030 for the latent mean difference with partial invariance, and 0.029 for the latent mean difference with alignment (all less than 1% of the scale). Though these vary across

research questions, we encourage researchers to go through the same process while analysis planning by considering meaningful raw/standardized differences for their research question (e.g., Gunn et al., 2020) and to develop contingencies for interpreting results in the face of conflicts.

On one hand, the alignment method substantially decreases the burden and possible mistakes from the researcher by reducing input and number of manually-specified comparisons, as well as the number of errors of inference at the model comparison level—especially when there are large numbers of items or groups. But, on the other, we warn that this ease of use also renders the alignment procedure liable to misuse and misinterpretation. Indeed, the onus is largely on the researcher to properly interpret the performance and results of the item-level tests in context, and measures for performance of the procedure are still poorly understood and require understanding of the scale and context of its use for proper interpretation beyond rules of thumb. As we saw in our example, our results violated the 25% rule we specified ahead of time, but upon further consideration of the raw and standardized effect sizes of item differences, interpreting the latent mean difference seemed justified. Asparouhov and Muthén (2014) additionally suggested using simulation studies to evaluate performance, but this imposes a different requirement of specialized knowledge that is largely inaccessible to applied researchers. Therefore, although there is less planning involved for the alignment method, there are outcomes that make interpreting the results less straightforward, and researchers should be prepared to update and change analysis plans as the methodology evolves.

Using Both Methods: Possible Robustness Uses and Recommendations

Both methods essentially resulted in mean comparisons with similar conclusions: The latent and observed mean comparisons under the traditional approach found no meaningful

difference in CFC-Immediate between males and females, and the latent mean comparison under the alignment method found no difference as well. Despite ultimately arriving at these conclusions through holistic model evaluation as specified in our illustrative analysis plan, both methods also shared similar evidence suggesting the presence of small amounts of measurement non-invariance sourced from the same items. For the traditional approach, the chi-square model comparison test for scalar invariance was statistically significant, but the changes in model fit indices were trivial. For the alignment method, the optimization procedure appeared to perform poorly as per the 25% rule, flagging more than 25% of items, but the deviations in item parameter estimates were trivial, e.g., intercept differences of less than 0.1 on a 5-point scale that sum to a negligible effect on the overall score. Overall, the illustrative data analysis serves as an example of how the alignment method can be a viable alternative to the traditional approach in a two-group context. Because both approaches were viable analysis options, the similarity in results was not surprising.

Although either method alone would have led to the same conclusions, it is possible that we may have produced different results had we made different but defensible analytical decisions, such as different strategies/criteria for partial invariance analyses. Errors of inference are likely when specifying many models and following an analysis plan that is completely data driven, as is done with the traditional approach (MacCallum et al., 1992). Given this, we propose that the alignment method can be used as an exploratory tool to compliment the traditional approach, assuming that both methods are appropriate for the research problem. For example, the item-level tests from the alignment method can be used in an exploratory manner to empirically guide partial invariance analyses as a sole strategy or in tandem with the numerous existing strategies. If the non-invariant items identified by the alignment method match those that are

identified through the strategies decided by the researcher and the relative magnitudes of non-invariance also match, then there is additional evidence that the selected items are correct. In our illustrative example, we identified Q2, Q5, Q9, Q10, and Q12 as non-invariant items in the partial invariance analysis. Based on the alignment optimization results, these selections were defensible: Q2, Q9, and Q12 were flagged as non-invariant, and Q5 and Q10 had R^2 values close to zero.

Similarly, the alignment method can also compliment the traditional approach as an empirical robustness check or additional sensitivity analysis. For example, the alignment method can be used concurrently as a comparison to the traditional approach. If researchers expect to substantially inflate Type I error rates under the traditional approach—particularly because of numerous nested model comparisons—then the results can be compared against the alignment optimization. We recommend against this strategy if there is reason to believe that the sample size is too small due to the trade-off of Type I error control for increased Type II errors for the item-level analyses, i.e., the alignment method is more likely to fail to detect measurement non-invariance if it exists.

If both methods are used, it is important to match model evaluation criteria, including the fit criteria for the baseline model and interpretation of invariance with effect sizes. For both methods, we matched fit criteria for the configural model. Moreover, we considered not only whether measurement non-invariance was present, but also whether the amount of non-invariance is practically impactful on the downstream analyses with both methods. If using both methods or interpreting the results of the traditional approach and the alignment method, we recommend that researchers employ this holistic evaluation practice universally if results from both the traditional approach and alignment method are considered together, and we caution that

asymmetrical model evaluation strategies can produce conflicting results, as was possible even in simplified ideal cases such as the illustrative example (e.g., very large sample size, only two groups, simple 8-item scale).

However, the alignment method should not be treated as an accessory analysis that can be added onto any traditional approach analysis without proper consideration, nor should it be considered as a universal alternative. The alignment method imposes the restriction of unknown generalizability and analysis of only latent means, the former of which is an obstacle for generalizable research, and the latter of which is rarely practiced by psychologists using conventional parametric analyses (e.g., *t*-tests, ANOVAs, regression). Therefore, the alignment method should not be considered a universally superior option to the traditional approach, but it presents several procedural advantages should these considerations not be of concern.

Recommendations for Future Methodological Research

Though there is a rich literature on methodologies for measurement invariance, we identified gaps that are critical for researchers to plan, use, and interpret a measurement invariance analysis: sample size planning, model evaluation criteria, and the general necessity and role of the method in substantive research. First, sample size determination is currently difficult for both approaches with no complete and user-friendly calculation tool, resulting in overreliance on vague rules of thumb. More research is required to better understand how exactly to increase sample sizes in response to multiple comparisons from larger numbers of items and groups, and the measure’s psychometric properties. This is especially important for the alignment method, which has no studies to date on sample size determination given its relative novelty. Our starting sample size suggestion is only based on existing simulation studies pertaining to the baseline configural model, and there is no statistical power research available

yet for the item-level analyses. Future simulation studies should manipulate these aspects on varying levels of measurement non-invariance and group sizes to eventually incorporate them into a user-friendly sample size calculator for non-methods researchers.

Second, model evaluation criteria require further qualification across a larger pool of possible situations. There are multiple plausible model fit criteria for the traditional approach and determining what is an acceptable model using them is difficult. Here, for example, we employed the common criteria developed by Hu and Bentler (1999) for illustrative purposes. However, these criteria were developed on a limited set of models and may not generalize. Despite these well-known limitations, there are few alternatives with accessible implementations for applied researchers. As a result, different model fit criteria and/or the omission of certain strategies can produce conflicting or misleading results. Analysis planning can partially address this, and we provided our preregistration example to encourage researchers to consider which model fit criteria are pertinent to them and decide ahead of time how they will use and interpret them. We also noted that various new approaches to evaluating model fit are up and coming (McNeish & Wolf, 2021), and we encourage applied researchers to consider incorporating these approaches into their analysis plans.

With the alignment method, researchers not only need to evaluate the fit of the baseline configural model, but also the number of non-invariant items. Currently there is a rough 25% rule of thumb limit suggested by Muthén and Asparouhov (2014) based on limited empirical evidence. The alignment method also provides values such as the R^2 effect size measure of measurement invariance that are not yet well understood. As seen in our illustrative example, these important ambiguities include how to interpret this R^2 when the results seem to conflict with the significance test (e.g., non-significant invariance test but $R^2 = 0$). When these fringe

cases or conflicts occur, what specific criteria can be used to gauge “high” as opposed to “low” magnitudes of invariance? Further simulation research is needed to refine best practices for the alignment method.

Third, the practice of conducting and reporting measurement invariance testing in applied and substantive literature in psychology is limited despite the potential impacts of non-invariance on downstream analyses (e.g., Boer et al., 2018). This may be partially due to the lack of knowledge applied researchers have about measurement invariance testing, which is complex to navigate without advanced quantitative training. This tutorial was written to address that shortcoming by making these analyses accessible and incorporating modern open science practices into the process.

However, this is not the only reason these analyses are not often reported. Measurement non-invariance can vary in pattern and magnitude: In some cases, non-invariance will be trivial, whereas in others, not accounting for it will change the conclusion (Schmitt et al., 2011). More meta-scientific and methodological research is needed to understand the breadth of ramifications non-invariance can have in applied research and how researchers can and do use the methods to inform theory. From this, better guidelines for planning, use, and interpretation of such models can be developed. Overall, transparency and reporting of measurement details is lacking in the psychological literature (Flake & Fried, 2020; Flake et al., 2017), and while methodologists can encourage applied researchers to do more and do better, methodologists themselves can also do more to demonstrate the practical importance of such methods for applied researchers.

Recommendations for Improving Implementation

During the process of conducting the analyses for this illustration, we encountered two areas of improvement regarding the practical implementations of the traditional approach and

alignment in *lavaan* and *Mplus* that could be improved to facilitate measurement invariance testing in psychology. First, the alignment method is only available as originally specified by Asparouhov and Muthen (2014) in *Mplus*. To date, there is no existing package in R that replicates the alignment method functionality,[4] which makes accessibility to the alignment method difficult for researchers without the financial means to use *Mplus*. Second, the default software settings for the traditional approach vary drastically across software and within software (see Supplementary Materials for more details). Because of this, preregistrations and analysis plans must be clear and specific in their model specifications beyond broad statements—and ideally accompanied by the code to be used for analysis. Moreover, we recommend that models be specified manually in this code rather than by an automated or default function.

Conclusion

Measurement invariance analyses are applicable to many areas of psychology but are difficult to plan, conduct, and interpret. As psychologists move toward more transparent research practices, applying these practices to measurement invariance testing is an upcoming area for improvement. The alignment method shows exciting promise as an additional approach to assessing measurement invariance, but it also presents challenges with model selection, interpretation, and appropriate use. Here, we compared alignment to the traditional factor analytic approach to help researchers decide which to use, and we provided recommendations on how researchers can plan their measurement invariance analyses in a transparent manner. We hope that this tutorial helps applied researchers integrate measurement invariance assessment into their programs of research and facilitate transparent practices, consistent with the changing standards of contemporary research practices.

[4] The *sirt* package in R is closest but uses a procedure inspired by the alignment method in *Mplus*, requires manual configuration, and may produce different results.

References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Bialosiewicz, S., Murphy, K., & Berry, T. (2013). *An introduction to measurement invariance testing: Resource packet for participants*. Retrieved from http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=63758fed-a490-43f2-8862-2de0217a08b8

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713–734. https://doi.org/10.1177/0022022117749042

Bollen, K. A. (2014). *Structural Equations with Latent Variables*. John Wiley & Sons.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Center for Open Science. (2020, November 10). *APA Joins as New Signatory to TOP Guidelines*. https://www.cos.io/about/news/apa-joins-as-new-signatory-to-top-guidelines

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social*

Psychology, *95*(5), 1005–1018. https://doi.org/10.1037/a0013193

Counsell, A., Cribbie, R. A., & Flora, D. B. (2020). Evaluating equivalence testing methods for measurement invariance. *Multivariate Behavioral Research*, *55*(2), 312–328. https://doi.org/10.1080/00273171.2019.1633617

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46*(1), 137–149.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., & McCoach, D. B. (2018). An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 56–70. https://doi.org/10.1080/10705511.2017.1374187

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(3), 378–402. https://doi.org/10.1207/s15328007sem1303_3

French, B. F., & Finch, H. (2016). Factorial invariance testing under different levels of partial

loading invariance within a multiple group confirmatory factor analysis model. *Journal of Modern Applied Statistical Methods*, *15*(1), 511–538. https://doi.org/10.22237/jmasm/1462076700

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460. https://doi.org/10.1511/2014.111.460

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*, 980. https://doi.org/10.3389/fpsyg.2014.00980

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. https://doi.org/10.1080/10705511.2019.1689507

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53-60. https://doi.org/10.21427/D7CF7R

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3–4), 117–144. https://doi.org/10.1080/03610739208253916

Hsiao, Y.-Y., & Lai, M. H. C. (2018). The impact of partial measurement invariance on testing moderation for single and multi-level data. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00740

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

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika*, *71*(1), 173–191. https://doi.org/10.1007/s11336-003-1136-B

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 642–657. https://doi.org/10.1080/10705510903206014

Joireman, J., Balliet, D., Sprott, D., Spangenberg, E., & Schultz, J. (2008). Consideration of future consequences, ego-depletion, and self-control: Support for distinguishing between CFC-Immediate and CFC-Future sub-scales. *Personality and Individual Differences*, *45*(1), 15–21. https://doi.org/10.1016/j.paid.2008.02.011

Jorgensen, T. D., Kite, B. A., Chen, P.-Y., & Short, S. D. (2018). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. *Psychological Methods*, *23*(4), 708–728. https://doi.org/10.1037/met0000152

Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: Forward, backward, and factor-ratio tests. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 567–584. https://doi.org/10.1080/10705511.2015.1138092

Koziol, N. A., & Bovaird, J. A. (2018). The impact of model parameterization and estimation methods on tests of measurement invariance with ordered polytomous data. *Educational and Psychological Measurement*, *78*(2), 272–296. https://doi.org/10.1177/0013164416683754

Lai, M. H. C. (in press). Adjusting for measurement noninvariance with alignment in growth modeling. *Multivariate Behavioral Research*. https://quantscience.rbind.io/files/Lai_2021_mbr_awc_growth_am.pdf

Lomazzi, V. (2018). Using alignment optimization to test the measurement invariance of gender role attitudes in 59 countries. *Methods, Data, Analyses, 12*(1), 27. https://doi.org/10.12758/mda.2017.09

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology, 56*(2), 231–248. https://doi.org/10.1348/000711003770480020

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(4), 514–534. https://doi.org/10.1207/s15328007sem1104_2

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Magraw-Mickelson, Z., Hermida Carrillo, A., Weerabangsa, M. M., Owuamalam, C. K., & Gollwitzer, M. (2020, December 1). Comparing classic and novel approaches to measurement invariance. https://doi.org/10.31234/osf.io/pz8u9

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods, 23*(3), 524–545. https://doi.org/10.1037/met0000113

McNeish, D., & Wolf, M. G. (2021, February 15). Dynamic fit index cutoffs for confirmatory factor analysis models. https://doi.org/10.31234/osf.io/v8yru

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*, *14*(4), 611–635. https://doi.org/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. *The Journal of Applied Psychology*, *93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361–388. https://doi.org/10.1177/1094428104268027

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *The Journal of Applied Psychology*, *97*(5), 1016–1031. https://doi.org/10.1037/a0027934

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge/Taylor & Francis Group.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology*, *5*, 978. https://doi.org/10.3389/fpsyg.2014.00978

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research*, *47*(4), 637–664. https://doi.org/10.1177/0049124117701488

Noar, S. M. (2003). The role of structural equation modeling in scale development. *Structural*

Equation Modeling, *10*(4), 622–647. https://doi.org/10.1207/S15328007SEM1004_8

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Petrocelli, J. V. (2003). Factor validation of the Consideration of Future Consequences Scale: Evidence for a short version. *The Journal of Social Psychology*, *143*(4), 405–413. https://doi.org/10.1080/00224540309598453

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement*, *58*(6), 1017–1034. https://doi.org/10.1177/0013164498058006010

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Schmitt, N., Golubovich, J., & Leong, F. T. L. (2011). Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: An illustrative example using big five and RIASEC measures. *Assessment*, *18*(4), 412–427. https://doi.org/10.1177/1073191110373223

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

Shi, D., Song, H., & Lewis, M. D. (2019). The impact of partial factorial invariance on cross-group comparisons. *Assessment*, *26*(7), 1217–1233.

https://doi.org/10.1177/1073191117711020

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

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), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. https://doi.org/10.1086/209528

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(1), 1–12. https://doi.org/10.1027/1614-2241/a000049

Strathman, A., Gleicher, F., Boninger, D. S., & Edwards, C. S. (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology*, *66*(4), 742–752. https://doi.org/10.1037/0022-3514.66.4.742

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*(1), 4–70. https://doi.org/10.1177/109442810031002

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance.

European Journal of Developmental Psychology, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

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

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479–498. https://doi.org/10.1177/0146621603259902

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, *14*(3), 435–463. https://doi.org/10.1080/10705510701301677

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-lased methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*(1), 165–200. https://doi.org/10.1111/0081-1750.00078

Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods*, *21*(3), 405–426. https://doi.org/10.1037/met0000080

Appendix A

Mathematical Treatment of Alignment Optimization

Here we present a mathematical treatment of alignment optimization as per Asparouhov and Muthén (2014). First, recall Equation 2, which represented the multiple-group confirmatory factor model for the traditional factor analytic approach. M0 is estimated based on Equation 2 (MGCFA) where the factor in each group is transformed to have a factor mean of zero and variance of 1, $\alpha = 0$ and $\Psi_g = 1$ for every group g. Thus, in M0, factor loadings and intercepts are freely estimated and can be represented as follows:

$$\eta_{g0} = \frac{(\eta_g - \alpha_g)}{\sqrt{\Psi_g}} \tag{3}$$

Second, M0 is mathematically re-expressed to treat measurement invariance as an optimization problem. The end goal of the alignment optimization is to produce a new model with minimal measurement non-invariance, which we denoted as M1. The optimization process starts with a re-expression of the variance of items as

$$Var(y_{pg}) = \lambda_{pg}^2 \Psi_g = \lambda_{pg,0}^2 \tag{4}$$

and a re-expression of the mean (i.e., expectation or expected value) of items as

$$E(y_{pg}) = \nu_{pg} + \lambda_{pg} \alpha_g = \nu_{pg,0} \tag{5}$$

such that the loading estimates of the configural model M0, denoted as $\lambda_{pg,0}$, can be found by algebraically simplifying Equation 4,

$$\lambda_{pg,0} = \lambda_{pg} \sqrt{\Psi_g} \tag{6}$$

and the intercept estimates of the configural model M0, denoted as $\nu_{pg,0}$, can then be found by

substituting λ_{pg} from Equation 6 into Equation 5

$$\nu_{pg,0} = \nu_{pg} + \frac{\lambda_{pg,0}}{\sqrt{\Psi_g}}(\alpha_g) \tag{7}$$

For every set of group factor means α_g and variances Ψ_g, there are intercept parameters ν_{pg} and loading parameters λ_{pg} that yield the same likelihood as M0, the configural model. Therefore, we can obtain these loading parameters for M1, denoted $\lambda_{pg,1}$, by rearranging Equation 6

$$\lambda_{pg,1} = \frac{\lambda_{pg,0}}{\sqrt{\Psi_g}} \tag{8}$$

and these intercepts, denoted $\nu_{pg,1}$, by rearranging Equation 7

$$\nu_{pg,1} = \nu_{pg,0} - \frac{\lambda_{pg,0}}{\sqrt{\Psi_g}} \tag{9}$$

Third, Equations 8 and 9 can be used to create a total loss function F that represents total measurement non-invariance. Recall that scalar invariance requires invariant loadings and intercepts. F is thus the sum of the differences between factor loadings and intercepts across groups. Therefore, factor means α_g and variances Ψ_g for M1 can be selected that minimize the total loss function, and then they can be substituted into Equations 8 and 9 to find the optimal loadings and intercepts of M1. That is, the total loss function F is minimized with respect to α_g and Ψ_g in order to find the parameters for M1 that minimize total measurement non-invariance. For some pair of groups g_1 and g_2,

$$F = \sum_{p} \sum_{g_1 < g_2} w_{g_1,g_2} f(\lambda_{pg_1,1} - \lambda_{pg_2,1}) + \sum_{p} \sum_{g_1 < g_2} w_{g_1,g_2} f(\nu_{pg_1,1} - \nu_{pg_2,1}) \quad (10)$$

In Equation 10, the differences between factor loadings and intercepts are weighed by w, which is calculated by taking the square root of the product of the sample sizes of g_1 and g_2. This is done so that larger groups contribute more to F, the total loss function, than smaller groups, accommodating unequal group sizes. Additionally, f represents the component loss function (CLF), and these differences are scaled via the CLF. The CLF has been used in rotation methods in exploratory factor analysis to minimize differences in the loading matrix to find a solution with the simplest structure (e.g., Jennrich, 2006). The alignment method uses the following CLF

$$f(x) = \sqrt{\sqrt{x^2 + \varepsilon}} \quad (11)$$

with some small positive value for ε (e.g., .01). This specific type of value is chosen so that the CLF has a continuous first derivative, which mathematically simplifies the minimization of the total loss function F. Overall, F is minimized when there are only a few large noninvariant parameters and many approximately invariant parameters, so the presence of a few large noninvariant parameters and many approximately invariant parameters is an assumption of alignment.

Fourth, M1 is identified by estimating all group factor means and variances except for the first under the following constraint:

$$\Psi_1 \times \ldots \times \Psi_g = 1 \quad (12)$$

The alignment optimization procedure therefore takes two forms based on the decision to select the factor mean and variance of the first group or not. The factor mean and variance can either be fixed to 0 and 1 respectively (FIXED alignment optimization) or can be estimated freely (FREE

alignment optimization).

Appendix B

Measurement Invariance Analysis Plan Preregistration Example

This is an example of a preregistered measurement invariance analysis plan to supplement Luong and Flake (2021). This example focuses on a series of statistical models and does not include other details about a study that would go into a complete preregistration. General and detailed information about preregistration is available at https://cos.io/prereg. The example here corresponds to the "Variables" and "Analysis Plan, Statistical Models and Inference Criteria" sections of a full length preregistration, template available at: https://osf.io/preprints/metaarxiv/epgjd/.

Variables

Measured Variables

Grouping Variable

Sex, as self-reported by participants (2 groups: male, female). Non-binary gender identities will be excluded from analysis.

Outcome Variable

Consideration for future consequences, as measured using the Consideration for Future Consequences Scale (CFC; Strathman et al., 1994). The CFC measures two future consequence constructs: a future concern sub-factor, which is measured with 4 items (e.g., "I am willing to sacrifice my immediate happiness or well-being in order to achieve future outcomes."); and an immediate concern sub-factor, which is measured with 8 eight items (e.g., "I only act to satisfy immediate concerns, figuring the future will take care of itself."). Items are rated on 5-point scales (1 = *Extremely uncharacteristic*, 5 = *Extremely characteristic*). We will only analyze the immediate concern subscale.

Covariates

No covariates will be analyzed.

Indices

CFC-Immediate

We will combine the 8 immediate concern items from the CFC to create a single measure of concern for immediate consequences. We will use confirmatory factor analysis and alignment optimization to estimate concern for immediate consequences factors scores from the 8 items. If full scalar invariance is achieved, then we will also take the mean of the 8 immediate concern items from the CFC to create a single, observed score measure of concern for immediate consequences.

Analysis Plan

Summary

This analysis plan covers a two-group measurement invariance analysis with two methods: 1) a multiple group confirmatory factor model (MGCFA) and 2) an alignment optimization. It lists a series of decisions required for each method based on Luong and Flake (2021). Analysis code corresponding to this analysis plan is available at https://osf.io/3p7n9/ .

Prerequisites for Methods

Evidence of Factor Structure

We will test a one-factor model for the CFC-immediate factor, consistent with literature. Model fit will be considered acceptable at CFI > .95, RMSEA < .06, and SRMR < .08 (Hu & Bentler, 1999). The CFA will be identified using the marker method with Q2 (i.e., loading of Q2 fixed to 1.00).

Sample Size

Each group should have a sample size of at least 400 for the analysis to proceed (French & Finch, 2006; Meade & Bauer, 2007; Meade et al., 2008; Koziol & Bovaird, 2018).

Assumption Checks

Item level distributions will be used to assess normality visually. If items are non-normal, robust maximum likelihood estimation will be used with the Yuan-Bentler scaled chi-squared statistic (MLR; Yuan & Bentler, 2000) and robust standard errors for all CFAs and measurement invariance tests.

Research Goal

The research goal is to evaluate the measurement invariance of the 8-item immediate concern subscale across sex to ultimately compare mean scores (latent or observed) across males and females.

MGCFA

Using multiple group confirmatory factor analysis, we will compare latent means if partial but not full scalar invariance is achieved. We will compare observed means if full scalar invariance is achieved.

Alignment

We will use the alignment method to compare factor means.

Model Identification

MGCFA

We will fix the loading of the anchor item 1 and factor means to 0 respectively for both groups. As mentioned previously, we informally reviewed the content of the items and selected the item Q2 as the anchor item because we deemed it least likely to be non-invariant across groups.

Alignment

We will fix the factor mean and variance to 0 and 1 respectively because we are only comparing 2 groups (i.e., the FIXED alignment configuration).

Model Evaluation

MGCFA

For all models, we will report the chi-square model fit test and multiple additional fit indices. To evaluate the overall factor model across both groups as well as the baseline configural model, we will report the total model chi-square and the CFI, RMSEA, and SRMR. If the chi-square test is significant, which is likely given the large sample size, we will deem the overall factor model and configural model to have acceptable fit to move forward with invariance testing if CFI > .95, RMSEA < .06, and standardized root mean square residual (SRMR) < .08 (Hu & Bentler, 1999). Then, to determine whether metric, scalar, and strict measurement invariance are supported, we will report the chi-squared model fit difference tests and model fit index differences between successive model. We will conclude that the next level of invariance was not supported if the chi-square test is significant at $\alpha = .05$ and/or the higher-level model increases RMSEA by more than .015 or decreases CFI by more than .01 (Chen, 2007). Thus, if the two criteria disagree, we will return to the level of measurement invariance that failed and conduct a partial measurement invariance analysis.

Partial Invariance. Partial invariance analyses will use a backward-selection approach using modification indices to identify non-invariant items. Specifically, we will free the item parameter with the highest modification index first, then rerun the model with that freed. We will repeat this process until partial invariance is established (i.e., until the models no longer differ in fit) or modification indices no longer indicate significant improvements in model fit (MIs < 3.84,

the critical value for chi-squared tests for df = 1 at $\alpha = .05$). If we successfully establish a scalar partial invariance model, we will use it to compare the group factor means. We will also report a partial strict invariance model by constraining the uniquenesses of the invariant items to equality but will not continue further.

Alignment

To evaluate model fit for the baseline configural model, we will use the same criteria for the configural model using MGCFA. To evaluate the performance of the alignment optimization (i.e., determine that most items were approximately invariant), we will follow Muthén and Asparouhov's (2014) rule of thumb in which no more than 25% of parameters are non-invariant to conclude good performance. If we conclude good performance, we will use the aligned model to compare factor means.

If more than 25% of items are deemed non-invariant based on the item-level significance tests, we will examine the parameter differences to determine whether the amount of non-invariance is meaningful. For non-invariant intercepts, we will deem any differences meaningful if they exceed 0.25 points (5% of the 5-point scale). If the amount of non-invariance is not meaningful, we will use the aligned model to compare factor means.