Understanding the Impact of Partial Factorial Invariance on Selection Accuracy: An R Script

Mark H. C. Lai

University of Cincinnati

Oi-man Kwok, Myeongsun Yoon, and Yu-Yu Hsiao

Texas A&M University

Author Note

This is an Accepted Manuscript of an article published by Taylor & Francis in Structural Equation Modeling: A Multidisciplinary Journal on 22/05/2017, available online: https://www.tandfonline.com/10.1080/10705511.2017.1318703.

Mark H. C. Lai, School of Education, University of Cincinnati; Oi-man Kwok, Department of Educational Psychology, Texas A&M University; Myeongsun Yoon, Department of Educational Psychology, Texas A&M University; Yu-Yu Hsiao, Department of Educational Psychology, Texas A&M University.

Correspondence concerning this article should be addressed to Mark Lai, School of Education, University of Cincinnati, Cincinnati, OH 45221.

Email: mark.lai@uc.edu

Abstract

Much of the previous literature on partial measurement invariance has focused on (a) statistically detecting non-invariance and (b) modeling partial invariance to obtain correct inferences for latent mean comparisons across groups in a single research study. However, very little guidance is provided on the practical implications of partial invariance on the instrument itself in the context of selection. In a frequently cited paper, Millsap and Kwok (2004) provided a framework for evaluating the impact of partial invariance by quantifying the magnitude of non-invariance on the efficacy of the test for selection purposes, yet our literature review found that only a few of the citations have fully captured the essence of Millsap and Kwok's method. In this paper, we briefly review the selection accuracy analysis for partial invariance and provide a user-friendly R script (also available as a web application) that takes parameter estimates as input, automatically produces summary statistics for evaluating selection accuracy, and generates a graph for visualizing the results. Hypothetical and real data examples are provided to illustrate the use of the R script. The goal of this paper is to help readers understand Millsap and Kwok's framework of evaluating the impact of partial invariance through an accessible computer program and step-by-step demonstrations of the selection accuracy analysis.

Understanding the Impact of Partial Factorial Invariance on Selection Accuracy: An R Script

Measures in behavioral sciences, such as aptitude or personality tests, usually require evidence of validity across subpopulations before being established as a formal tool for research or selection purposes. One essential step in this process is to check for *measurement invariance*; that is, to make sure that the measures maintain similar measurement structures between the constructs of interest and the observed items across subpopulations. For psychological and behavioral measures, it is common to observe measurement invariance for some but not all items, a condition referred to as *partial invariance*.

Previous literature mainly focused on either the detection of non-invariant items (e.g., Byrne, Shavelson, & Muthén, 1989; Cheung & Rensvold, 1999; Kaplan, 1989; Yoon & Millsap, 2007) or the impact of partial invariance on parameter estimation for a given study (e.g., Guenole & Brown, 2014; Oberski, 2014); however, there have been relatively few discussions on the practical implications of partially invariant measures when using observed composite scores to select or classify individuals. An exception was the study by Millsap and Kwok (2004), who proposed an approach to evaluating how partial invariance affects the performance of a test (e.g., sensitivity and specificity) in selecting or classifying individuals based on an observed cutoff score, as compared to the performance when measurement invariance holds.

Although Millsap & Kwok's (2004) paper is frequently cited in measurement invariance research, only a small number of studies have directly and adequately applied their procedure. A probable reason for this void in the literature is the absence of user-friendly computer programs to implement the procedure. Therefore, the main goal of the present article is to provide a review of their procedure, along with an R script and a web application that researchers can easily use to perform that procedure to evaluate the effect of partial invariance.

Factorial Invariance

We first briefly review the definitions of measurement invariance and factorial invariance (i.e., measurement invariance under the common factor model), with similar notations in Millsap and Kwok (2004). For more in-depth discussion, please consult Meredith (1993) and Millsap (2011). Measurement invariance (Mellenbergh, 1989) is satisfied when the conditional probability distribution of the observed item score variable, X, given the latent variable to be measured, ξ, does not depend on the group membership variable, K. That is,

$$P(X \mid \xi, K = k) = P(X \mid \xi) \text{ for all } k.$$

In other words, for two individuals with the same score on the latent construct, the probability distributions of their respective observed item scores are the same regardless of their group membership.

Under the assumption that the observed variables conform to a common factor model (e.g., Thurstone, 1947), measurement invariance is equivalent to factorial invariance, namely, the invariance of measurement parameters in factor models. In subsequent sections we use measurement invariance and factorial invariance interchangeably.

For a psychological instrument with q observed variables measuring one latent construct in K groups, the common factor model has the form

$$\mathbf{X}_k = \boldsymbol{\tau}_k + \boldsymbol{\lambda}_k\xi_k + \boldsymbol{\delta}_k, \tag{1}$$

where $\mathbf{X}_k$ is a q × 1 column vector of item score variables for the kth subpopulation, ξ_k denotes the latent score random variable, $\boldsymbol{\tau}_k$ is a q × 1 column vector of measurement intercepts, $\boldsymbol{\lambda}_k$ is a q × 1 column vector of factor loadings that quantify the linear relationships between items and the latent variable, and $\boldsymbol{\delta}_k$ is a $q \times 1$ column vector of the unique factor random variables. Consistent with the notations in previous literature (Millsap, 2007; Widaman & Thompson, 2003), $\mathbf{X}_k$, ξ_k, and $\boldsymbol{\delta}_k$ are random variables and vectors. Let $\text{E}(\xi_k) = \kappa_k$ and $\text{Var}(\xi_k) = \varphi_k$ be the mean and the variance of the latent variable, respectively. Further, let the variance-covariance matrix among the unique factor variables be $\text{Cov}(\boldsymbol{\delta}_k) = \boldsymbol{\Theta}_k$, and assume that each unique factor variable has a zero mean, $\text{E}(\boldsymbol{\delta}_k) = \mathbf{0}$ for all k. In practice, researchers usually impose the local independence assumption such that $\boldsymbol{\Theta}_k$ is a diagonal matrix, meaning that the inter-item correlations are attributed solely to the variance of the underlying latent factor. It is also assumed that ξ_k and $\boldsymbol{\delta}_k$ are independent with $\text{Cov}(\xi_k, \boldsymbol{\delta}_k) = \mathbf{0}$, and together the model implies that $\text{E}(\mathbf{X}_k) = \boldsymbol{\tau}_k + \boldsymbol{\lambda}_k\kappa_k$ and $\text{Var}(\mathbf{X}_k) = \boldsymbol{\lambda}_k\varphi_k\boldsymbol{\lambda}_k' + \boldsymbol{\Theta}_k$. With the additional assumption that $\boldsymbol{\delta}_k$ is multivariate normal, factorial invariance implies that all measurement parameters (i.e., intercepts, loadings, and unique factor covariances) are identical, which can be expressed in mathematical notations:

$$\boldsymbol{\tau}_k = \boldsymbol{\tau},\ \boldsymbol{\lambda}_k = \boldsymbol{\lambda},\ \boldsymbol{\Theta}_k = \boldsymbol{\Theta} \text{ for all } k, \tag{2}$$

a condition commonly known as *strict* factorial invariance (Meredith, 1993).

In practice, however, strict invariance does not commonly hold, and for certain testing purposes only a subset of parameters (e.g., factor loadings, measurement intercepts) need to be equal across groups for meaningful group comparisons (Meredith, 1993; Steenkamp & Baumgartner, 1998). Therefore, four stages of factorial invariance are usually defined for different applications (Millsap, 2007). The first stage is *configural* invariance, which requires that the factor structures be the same across groups, including the same number of factors and the same composition of items for each factor. The second stage is *metric* invariance (Horn & McArdle, 1992; also called weak measurement invariance, Meredith, 1993; or pattern invariance, Millsap, 1995). This requires, in addition to configural invariance, that the factor loadings have the same magnitudes in all groups (i.e., $\boldsymbol{\lambda}_k = \boldsymbol{\lambda}$ for all k). As such, metric invariance ensures that a unit difference in the latent construct is comparable across subpopulations. The third stage is *scalar* invariance (also called strong factorial invariance; Meredith, 1993). This requires, in addition to metric invariance, that the measurement intercepts are equal across groups (i.e., $\boldsymbol{\tau}_k = \boldsymbol{\tau}$ for all k). Scalar invariance ensures that a given measure has the same origin or zero point. The final stage is strict invariance as discussed in the previous paragraph, where the unique factor variances (and covariances, if applicable) are also identical (i.e., $\boldsymbol{\Theta}_k = \boldsymbol{\Theta}$ for all k).

Partial Measurement Invariance

For a particular stage of factorial invariance, when invariance holds only for a subset of items (e.g., eight items have scalar invariance but 2 items do not), one only obtains partial measurement invariance (e.g. Byrne et al., 1989).[1] Similar issues were also studied in the *differential item functioning* (DIF; cf. Hambleton, 2006; Millsap & Everson, 1993; Penfield & Lam, 2000) literature in item response theory (IRT), a framework for formulating measurement models for dichotomous and polytomous items; we refer to the DIF literature in our discussion when appropriate.

Although many authors have provided guidance on how to identify partial measurement invariance using SEM (e.g., Asparouhov & Muthén, 2014; Byrne et al., 1989; Cheung & Rensvold, 1999; Jak, Oort, & Dolan, 2014; Kaplan, 1989; Merkle, Fan, & Zeileis, 2014; Merkle & Zeileis, 2013; Stark, Chernyshenko, & Drasgow, 2006; Van De Schoot, Schmidt, & De Beuckelaer, 2015; Yoon & Millsap, 2007), there is relatively less guidance on understanding the impacts of partial measurement invariance. This is analogous to the difference between *statistical significance* and *practical significance* (and effect size), as the detected partial invariance may practically make no difference when interpreting the test scores, or vice versa.

As noted by Millsap and Kwok (2004), the evaluation of partial measurement invariance should be made “in relation to the purpose of the measure” (pp. 94–95). In the behavioral sciences, common purposes of a measure include (a) to quantify constructs in scientific research and (b) to select or identify individuals based on their relative standings or absolute scores on the test (Crocker & Algina, 2008). However, most of the existing literature on partial measurement invariance addressed (a), with little attention paid to (b); (b) is exactly the problem studied in Millsap and Kwok (2004).

Practical Significance of Partial Measurement Invariance in a Single Study

As has been well documented in previous research (Steenkamp & Baumgartner, 1998; Vandenberg, 2002), certain applications of test scores are valid only when a certain stage of measurement invariance holds. For example, observed difference between groups could simply be an artifact of scalar non-invariance and may vanish or even be reversed if the researchers use a different measure that is scalar-invariant across groups (e.g., Steinmetz, 2013; Wicherts, Dolan, & Hessen, 2005). Invariances of loadings and unique factor variances, on the other hand, are needed for comparing associations between test scores and other external variables across groups, such as in multiple regression and path analysis (Guenole & Brown, 2014). Recent research efforts have started to evaluate the degree to which partial invariance affects parameter estimations for a single study. For example, Oberski (2014) proposed the expected parameter change-interest index for evaluating the sensitivity of means or path coefficients of interest when one relaxes an invariance constraint on a non-invariant item.

Although the effects of ignoring partial invariance can be detrimental, for a given study it is still possible to obtain correct inferences at the latent-variable level if one uses a correct partial invariance model (Byrne et al., 1989) by placing equality constraints across groups only on the invariant items in multiple-group SEM. Alternatively, one can utilize the recently developed technique of approximate invariance to align a measure with many non-invariant items across a large number of groups and estimate the latent means (Asparouhov & Muthén, 2014; Van de Schoot et al., 2013).

Measurement non-invariance can also be related to theoretically justified factors, and it is important to understand what sources are responsible for the differences in measurement parameters across groups.[2]

Practical Significance of Partial Measurement Invariance in the Context of Selection

Another perspective to quantify the practical significance of partial invariance is to assess its impact on the validity of using test scores for selection, placement, or classification purposes. Psychological and behavioral measures are commonly used for, for example, identifying people with depressive symptoms, selecting or promoting employees, or providing support for college admissions decisions. Although selection is an important purpose for psychological and behavioral testing, the majority of the literature on measurement invariance has focused more on obtaining valid inferences on mean comparisons and path coefficients for research studies (Schmitt & Kuljanin, 2008), and there has been relatively little guidance on what to do with partially invariant tests in the context of selection (Millsap & Kwok, 2004).

There are some issues specific to using a test for selection as opposed to using it for a single research study, including (a) in making a decision, one usually uses the whole observed composite score distribution, rather than simply the observed or latent variable means; (b) a dichotomous decision (e.g., select or not; need to treat or not) is often made at the individual level; (c) the observed scores are compared to a prespecified cutoff. Each of these issues has implications for the impact of partial invariance. For (a), whereas the majority of the measurement invariance literature has focused on the detection and impact of non-invariant *items* (or DIF in IRT), for selection the focus is on the aggregate bias of the test (or differential *test* functioning in IRT; Raju et al., 1995; Stark, Chernyshenko, & Drasgow, 2004). It is possible that the biases introduced by multiple items are somewhat canceled out, resulting in little overall impact on selection using the observed test scores. For (b), the focus is no longer on the mean of each subpopulation, but rather on the classification accuracy at the individual level, as we discuss in the remaining of the article. For (c), rather than using a summary index to quantify the impact of partial invariance on the latent mean difference or the path coefficient, in selection one evaluates the impact at a specific cutoff, as the impact of partial invariance may be different for different cutoffs chosen (Stark et al., 2004).

Introduction to Selection Accuracy Analysis

Millsap and Kwok (2004) introduced a novel approach to understanding the practical significance of partial invariance by evaluating the change in selection accuracy using observed composite scores. Their approach is based on the assumption that the selection would be made based upon a cutoff on the composite of the observed item scores applied to all subpopulations. Although their discussion focused only on the unweighted sum of the item scores, the procedure can easily be applied to scale scores that are weighted sums of the item scores. The approach also assumes that all items measure one single latent construct, which ideally could be used to make the selection decision; in other words, the measure is unidimensional. If the items measured multiple dimensions of a construct, the researchers may treat the dimensions as separate and perform the selection accuracy analysis for each of the dimensions that exhibit lack of measurement invariance.

We will first review Millsap and Kwok's (2004) procedure following the notations in their paper. The selection accuracy analysis proceeds by first deriving the joint distribution of the latent construct, ξ, and the observed composite, Z, which is bivariate normal under the common factor model as defined in equation (1). From equation (1), the mean of the composite for the kth subpopulation is $\mu_{zk} = \mathbf{1}'\boldsymbol{\tau}_k + \mathbf{1}'\boldsymbol{\lambda}_k\kappa_k$, where $\mathbf{1}$ is a q × 1 unit vector, and the variance of the composite is $\sigma^2_{zk} = (\mathbf{1}'\boldsymbol{\lambda}_k)^2\varphi_k + \mathbf{1}'\boldsymbol{\Theta}_k\mathbf{1}$, with the first term due to the latent factor and the second term representing the sum of unique factor variances of all items forming the composite. The correlation between ξ and Z in the kth subpopulation is $\rho_{z\xi k} = (\mathbf{1}'\boldsymbol{\lambda}_k)\varphi_k^{1/2} / \sigma_{zk}$, which is the ratio of (a) the standard deviation (*SD*) attributed to the latent factor and (b) the *SD* of Z.

Because a single cutoff is applied to all subpopulations for selection, if the joint distribution of ξ and Z is constant for all k, selection accuracy is the same across subpopulations. With bivariate normality, the parameters (μ_{zk}, κ_k, σ^2_{zk}, φ_k, $\rho_{z\xi k}$) completely determine the joint distribution of ξ and Z in the kth subpopulation. This means that even when measurement invariance holds (i.e., $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$, $\boldsymbol{\Theta}$ are constant across groups), differences in the distributions of the latent construct (i.e., differences in κ_k and φ_k) still lead to differences in the joint distribution of ξ and Z, a phenomenon studied by Borsboom, Romeijn, and Wicherts (2008) and Millsap (1995).

Consider a simple example of selection wherein school counselors try to identify 25% of students most in need of counseling for depression in two subpopulations (e.g., native English speakers vs. English learners). Using the terminology of differential item functioning (Holland & Thayer, 1988), we refer to the majority subpopulation as the *reference group* and the minority subpopulation as the *focal group*; the latter is assumed to be at a disadvantage when non-invariance is present. Ideally the counselors should select the 25% of participants with highest true (latent) depression scores in the combined population, but has to make the decision based on the cutoff corresponding to the 75th percentile (i.e., top 25%) on the observed test scores. Therefore, one can imagine dividing the joint distribution into four quadrants using the cutoff scores on ξ and Z, as shown in Figure 1. The challenge lies in determining the respective cutoff scores on ξ and Z, as the joint distributions of ξ and of Z in a combined population of two or more groups would be a mixture of normal distributions and are not standard. The R script discussed in this paper (Appendix A) automates those calculations.

In Figure 1, the top right quadrant labelled as A is the area of *true positive*, where individuals have scores above the cutoffs on both ξ and Z. The top left quadrant labelled as B is the area of *false positive*, where individuals have scores above the cutoff on Z but not on ξ. The bottom left quadrant labelled as C is the area of *true negative*, where individuals have scores below the cutoffs on both ξ and Z. Finally, the bottom right quadrant labelled as D is the area of *false negative*, where individuals have scores above the cutoff on ξ but not on Z.

Using terminologies in signal-detection theory (Swets et al., 1979) and diagnostic testing (Altman & Bland, 1994a, 1994b), one can summarize the selection accuracy for subpopulation k ($k = r$ or f for the reference/focal groups) using four criteria: proportion selected (*PS*), success ratio or positive predictive value (*SR*), sensitivity (*SE*), and specificity (*SP*), where

$$PS_k = p(A_k) + p(B_k); \tag{3}$$

$$SR_k = p(A_k) / [p(A_k) + p(B_k)]; \tag{4}$$

$$SE_k = p(A_k) / [p(A_k) + p(D_k)]; \tag{5}$$

$$SP_k = p(C_k) / [p(C_k) + p(B_k)]. \tag{6}$$

Continuing with our example, *PS* refers to the proportion of students identified as in need of counseling by the inventory. *SR* is the proportion of students who are truly in need of the service among all the identified students; thus, a low success ratio means some wasted effort in providing service to students who do not need it. *SE* is the proportion of students who are identified among all the students in need of the service; thus, a low sensitivity means a failure to provide service to many of the students in need. Finally, *SP* is the proportion of students who are not identified among all the students who do not need the service; so a low specificity means that many students who are classified as not being at risk by the inventory are actually in need of a service. In this hypothetical example one can argue that sensitivity may be more important among the indices, but in practice different combinations of the indices may matter most, depending on the purpose of the test.

Thus, researchers can better understand the impact of partial invariance by examining the changes in these four indices for the two subpopulations from the strict invariance model to the partial invariance model. Even when strict invariance holds, the four indices would generally be different across groups due to differences in the distributions of the latent construct, as previously discussed. However, when partial invariance is present and the two groups differ in some loadings and/or intercepts, the differences in the four indices can become larger.

Although one can get a rough idea of how the four selection indices may change when one isolates an intercept or a loading, with non-invariance on multiple intercepts, loadings, and uniqueness of varying magnitudes and potentially different signs, it is preferable to resort to computer programs to evaluate the impact of partial invariance on selection accuracy, which has not been available before this article. This can potentially be a reason that we found only one paper (Alkemade, Bowden, & Salzman, 2015) performed the actual selection accuracy analysis on real data, out of the 79 published articles located from Web of Science Core Collection that studied measurement invariance for empirical data and that cited Millsap and Kwok (2004).

To make the selection accuracy analysis more accessible, in this paper we demonstrate the use of a user-friendly R script to perform the procedure with both hypothetical and real data examples. We also describe in Appendix B an example of selection accuracy analysis from a fitted `lavaan` (Rosseel, 2012) object in R using the `PartInv.lavaan` function, which avoids the need for manual input. For readers who are not familiar with R, we also provide a web application of the program on https://sites.google.com/site/partialinvarianceselection that does not require installations of R and relevant R packages.

Hypothetical and Real Data Examples

Hypothetical Example 1: Strict Invariance

Consider the strict invariance example in Millsap and Kwok (2004) with a one-factor four-indicator model for two groups, where $\kappa_r = 0.5$ for the reference group and $\kappa_f = 0$ for the focal group, and φ_r and φ_f are both 1.0. Strict invariance implies that the factor loadings are equal, $\boldsymbol{\lambda}_r = \boldsymbol{\lambda}_f = [0.3\ 0.5\ 0.9\ 0.7]'$, the intercepts are equal, $\boldsymbol{\tau}_r = \boldsymbol{\tau}_f = [0.225\ 0.025\ 0.010\ 0.240]'$, and the unique factor covariance matrices are also equal and follow a diagonal matrix, $\boldsymbol{\Theta}_r = \boldsymbol{\Theta}_f = \text{diag}[0.96\ 0.96\ 0.96\ 0.96]$. In this example, we assume that the measure is used to select the top 25% of the combined population, and that the two subpopulations are of similar size.

We will detail the steps needed to execute this analysis with the direct application of the R script in Appendix A, but the information and instructions also apply if one performs the same analysis using the web application. A prerequisite to using the script is to have the R package `mnormt` (Azzalini & Genz, 2016) installed and loaded, which computes the densities and quantiles for multivariate normal distributions.

To perform the procedure in R, one first sources the R script file given in Appendix A by entering the following command:

```
source("PartInv.R")
```

Then one calls the function `PartInv` to perform the analysis:

```
PartInv(propsel = .25, kappa_r = 0.5, kappa_f = 0, phi_r = 1, lambda_r = c(.3, .5, .9, .7),
        tau_r = c(.225, .025, .010, .240), Theta_r = diag(.96, 4), pmix_ref = 0.5)
```

where the first argument, `propsel`, is the proportion to be selected in the combined population (i.e., .25 in this example), `kappa_r` and `kappa_f` take values for κ_r and κ_f, respectively, and `phi_r`, `lambda_r`, `tau_r`, and `Theta_r` expect inputs of φ_r, $\boldsymbol{\lambda}_r$, $\boldsymbol{\tau}_r$, and $\boldsymbol{\Theta}_r$ for the reference group. The function also has optional arguments for the focal group: `phi_f`, `lambda_f`, `tau_f`, and `Theta_f`; however, if no inputs are provided, by default they are assumed equal to their counterparts in the reference group. Finally, the argument `pmix_ref` allows one to specify the proportion of the reference group at the population level, which has a default value of .5 (can be dropped in the above syntax as the input is the same as the default), meaning that the sizes of the two subpopulations are equal; in a later example readers can see that the value can be changed to also reflect unequal subpopulation sizes. Please see the documentation in Appendix A (i.e., the part in `PartInv` preceded by "#"s) for more details about the arguments.

The above R call results in the following output, as also illustrated by the solid and dotted ellipses in Figure 2 (a) and (b):

```
$cutoffs

  propsel  cutpt_xi   cutpt_z

0.2500000 0.9457938 3.2292682


$summary
                     Reference Focal
A (true positive)        0.222 0.110
B (false positive)       0.089 0.079
C (true negative)        0.583 0.748
D (false negative)       0.106 0.062
Proportion selected      0.311 0.189
Success ratio            0.714 0.580
Sensitivity              0.677 0.638
Specificity              0.868 0.904
```

The first value in the output, `propsel`, is simply a reprint of the selection proportion. The second and third values, `cutpt_xi` and `cutpt_z`, give the cutoff values on ξ and Z, respectively. What follows is a summary table providing the proportions of the four quadrants, $p(A_k)$, $p(B_k)$, $p(C_k)$, and $p(D_k)$, as well as the four summary statistics for selection accuracy, PS_k, SR_k, SE_k, and SP_k, for both the reference group and the focal group, which are the major results with Millsap and Kwok's (2004) procedure. Finally, the R call also generates the graph in Figure 2 (a) and (b) with the latent score ξ on the x-axis and the observed composite score Z on the y-axis, with the corresponding cutoff values on ξ and on Z; the two cutoff lines divide the two-dimensional space into the four quadrants A, B, C, and D, as previously discussed. The regions enclosed by the solid ellipse and the dotted ellipse are 95% confidence regions for the joint distributions of ξ and Z for the reference group and the focal group, respectively.

Two observations of the graph are worth mentioning. First, compared to the reference subpopulation, there are less true positives ($p(A)$) but more true negatives ($p(C)$) in the focal subpopulation, because the reference subpopulation has a higher latent mean. This again shows that strict measurement invariance does not imply invariance in selection, unless the latent score distributions are identical for the subpopulations. Therefore, it is important for researchers to obtain the selection statistics for both the partial invariance model and the strict invariance to correctly evaluate the impact of partial invariance on selection. Second, we observe that $SR_f <$ SR_r (.58 vs. .71), $SE_f < SE_r$ (.64 vs. .68), and $SP_f >$ SP_r (.90 vs. .87) from the text output, suggesting similar sensitivity and specificity of the measure across the two subpopulations (see also Table 1). The summary statistics for this strict invariance example will be used as a basis for evaluating the impact of partial invariance in the next two examples.

Hypothetical Example 2: Partial Scalar Invariance

Now modify the previous example so that three of the four items are scalar non-invariant, with $\boldsymbol{\tau}_f = [0.225\ {-0.050}\ 0.240\ {-0.025}]'$ and $\boldsymbol{\tau}_r = [0.225\ 0.025\ 0.010\ 0.240]'$. Also note that the differences in measurement intercepts have different magnitudes and directions. To investigate the influence of such partial invariance on selection accuracy, one can again call in R:

```
PartInv(propsel = .25, kappa_r = 0.5, kappa_f = 0, phi_r = 1, lambda_r = c(.3, .5, .9, .7),
        tau_r = c(.225, .025, .010, .240), tau_f = c(.225, -.05, .240, -.025),
        Theta_r = diag(.96, 4))
```

Notice that this time we drop `pmix_ref`, invoking the default value of 0.5. Also, one needs to input `tau_f` as it is no longer invariant across subpopulations. The results were shown in the middle column of Table 1 and the solid and dashed ellipses in Figure 2 (a).

Because the parameters for the reference group stay the same, the solid ellipse stay the same in Figure 2 (a). As three of the four measurement intercepts are non-invariant for the focal group, readers may at first expect a big change on the joint distribution of ξ and Z . However, as illustrated, the dashed ellipse barely changes from Example 1, and from Table 1, $p(A)$, $p(B)$, $p(C)$, and $p(D)$ are comparable to those in Example 1 for both the reference and the focal group. As a result, in this example, partial invariance has little effect in the focal population on the proportion selected (from .189 to .184), success ratio (from .580 to .587), sensitivity (from .638 to .627), and specificity (from .904 to .926), which can also be observed in Table 1.

The small impact of partial invariance in this example may be explained by the fact that the non-invariances for items 3 and 4, which are relatively large in magnitude but have different directions (i.e., item 3: 0.240 for focal and 0.010 for reference; item 4: −0.025 for focal and 0.240 for reference) and roughly cancel out, and the non-invariance on item 2 is relatively small (i.e., −0.050 for focal and 0.025 for reference). Note that the impact of partial scalar invariance can be much more dramatic with a different pattern of non-invariances on the intercepts, and selection accuracy analyses help summarize the consequence of partial invariance.

Hypothetical Example 3: Partial Metric Invariance

Consider another example where items 2, 3, and 4 are scalar non-invariant with mixed directions and magnitude, with $\boldsymbol{\tau}_f = [0.225\ {-0.225}\ 0.240\ {-0.025}]'$ and $\boldsymbol{\tau}_r = [0.225\ 0.025\ 0.010\ 0.240]'$, and items 3 and 4 are also metric non-invariant with smaller factor loadings on for the focal group such that $\boldsymbol{\lambda}_f = [0.3\ 0.5\ 0.7\ 0.5]'$ and $\boldsymbol{\lambda}_r = [0.3\ 0.5\ 0.9\ 0.7]'$. To investigate the influence of such non-invariances on selection accuracy, one can again call R:

```
PartInv(propsel = .25, kappa_r = 0.5, kappa_f = 0, phi_r = 1,
        lambda_r = c(.3, .5, .9, .7), lambda_f = c(.3, .5, .7, .5),
        tau_r = c(.225, .025, .010, .240), tau_f = c(.225, -.225, .240, -.025),
        Theta_r = diag(.96, 4))
```

Aside from a different input for `tau_f`, one also needs to input `lambda_f` as it is no longer invariant across subpopulations. The above R call computes the summary statistics in the last column of Table 1 and generates the solid and dashed ellipses in Figure 2 (b).

Again, the solid ellipse stays the same in Figure 2 (b). As two of the factor loadings for the focal group get smaller, the correlation between Z and ξ is reduced, so the dashed ellipse rotates slightly clockwise around its center and the ratio of its major axis to its minor axis gets smaller. The reduced correlation mainly reduces $p(A)$ and increases $p(C)$, while the impacts on $p(B)$ and $p(D)$ are negligible. In addition, the net effect of the non-invariances on the intercepts and the loadings shifts the dashed ellipse downward. As a result, fewer individuals (from .189 to .161) are selected from the focal group, and there is a large reduction in SE_f (from .638 to .529). A decreased sensitivity is perhaps most problematic for screening tools for identifying individuals in need of interventions. For example, for a measure identifying individuals with high suicidal ideation, less sensitivity would mean that among the focal subpopulation (e.g., females) with high risk of committing suicide, a lower proportion can be detected from the test. For such a test purpose and given similar results as this example, a researcher may decide not to use this measure for screening, especially for the focal subpopulation.

The partial metric invariance also slightly reduces SR_f; that is, among individuals in the reference group who are identified by the test, fewer are actually in need of an intervention. There are also changes in the selection accuracy indices in the reference subpopulation due to the change of cutoff on Z from 3.23 to 2.99 (which results from the changes in the measurement parameters), but the differences are less dramatic than those for the focal subpopulation.

Real Data Example

We now demonstrate the use of selection accuracy analyses using results on partial measurement invariance reported in an empirical study. Zhang et al. (2011) studied the measurement equivalence of the 4-factor, 20-item Center for Epidemiological Studies Depression (CES-D) Scale (Radloff, 1977) across a Chinese sample (N = 4, 903) and a Dutch sample (N = 1,903) of elderly groups. Comparing the metric invariance model to the baseline (configural) model, the authors found that changes in fit indices were small, with ΔCFI = 0.004 and ΔRMSEA = 0.003, thus concluding that metric invariance held. However, when comparing the scalar invariance model and the metric invariance model, they got ΔCFI = 0.014 and ΔRMSEA = 0.0148, and concluded that that scalar invariance did not hold. They then searched for scalar non-invariant items, which resulted in a partial scalar invariance model with one item on the Depressive Affect factor ("failure") and one item on the Positive Affect factor ("good") being scalar non-invariant. We use the Positive Affect items to illustrate our R script in the present paper. Note that the demonstration below is for illustration purpose only, and should not be taken as an accurate account of the selection bias of the CES-D across Chinese and Dutch populations.[3]

Table 2 shows the estimated factor loadings, intercepts, and uniqueness of the four items. Note that the items were all reversely coded in the original analyses in Zhang et al. (2011), so the factor may be better understood as *Lack of* Positive Affect. Although the correctly specified partial scalar invariance model still allows for valid comparisons of the latent factor means between the two samples in Zhang et al.'s (2011) study, with the presence of non-invariant items it is not clear whether the CES-D should still be used as a screening tool for the Chinese and the Dutch elderly populations.

To understand the impact of the non-invariance, one can perform selection accuracy analyses using the `PartInv` function with the parameter estimates in Table 2. Although the CES-D includes four factors and the selection accuracy analysis assumes a one-factor model, as suggested by Millsap and Kwok (2004) one can conduct the analyses from a selection approach separately for each factor. In this paper, we illustrate the impact of the non-invariant items measuring the (Lack of) Positive Affect factor. We also assume that the ratio between the two target populations is similar to the sample size ratio of 5 to 2 for Chinese and Dutch elderly for illustrative purposes; however in empirical studies and it is important to incorporate knowledge about the target populations and the intended usage of the test in deciding the mixing proportion.

For the latent factor (Lack of) Positive Affect, the mean and the *SD* were 0 and 0.354, respectively, for the Chinese sample and −0.125 and 0.329, respectively, for the Dutch sample. To perform the selection accuracy analysis, assuming that this time one is interested in identifying 14% of the combined population with highest depressive affect (which corresponds to an observed subscale score of 8 or above), one can first use the following R call:

```
PartInv(propsel = .14, kappa_r = 0, kappa_f = -0.125, phi_r = 0.354^2, phi_f = 0.329^2,
        lambda_r = c(1.00, 1.66, 2.30, 2.29),
        tau_r = c(1.54, 1.36, 1.16, 1.08),
        tau_f = c(0.68, 1.36, 1.16, 1.08),
        Theta_r = diag(c(1.20, 0.81, 0.32, 0.32)),
        Theta_f = diag(c(0.72, 0.81, 0.32, 0.32)),
        pmix_ref = 5 / 7)
```

Note that one can replace the `propsel = .14` argument with `cut_z = 8`, which is the way to provide a prespecified cutoff on the observed composite score (8 in this case) for selection/diagnosis. The argument `pmix_ref = 5 / 7` specifies that the Chinese elderly population is assumed to represent 5/7 of the combined population. As shown in Table 3, the summary statistics under the partial scalar invariance model are, for the Chinese and the Dutch elderly population respectively, .173 and .049 for proportions selected, .646 and .695 for success ratios, .688 and .456 for sensitivities, and .927 and .984 for specificities.

Although the numbers in the previous analysis are interpretable on their own, to appreciate the impact of the partial scalar invariance one needs to compare these numbers with those under a strict invariance model. As suggested by Millsap and Kwok (2004), one can use the weighted averages of the non-invariant parameters as the common parameter values when invariance holds. In our example, we assume that the measurement intercept for “good” would be 1.54 × 5/7 + 0.68 × 2/7 ≈ 1.29 if scalar invariance held, and the unique factor variances would similarly be the weighted averages of the estimates from the Chinese and the Dutch samples.

Under the strict invariance model with the latent means and variances unchanged, the summary statistics are, for the Chinese and the Dutch elderly population, respectively, .152 and .081 for proportions selected, .673 and .535 for success ratios, .656 and .609 for sensitivities, and .941 and .960 for specificities (see also Table 3). Therefore, for the Dutch (focal) group, the sensitivity drops by 15.3 percentage points (i.e., from .609 to .456), so the partial invariance has a relatively big effect on the selection accuracy of the CES-D Positive Affect factor. In other words, if one intends to identify people with truly low positive affect (e.g., for the purpose of receiving a particular type of intervention) in a combined Chinese and Dutch elderly population, the four CES-D Positive Affect items will likely do a poor job, especially for the Dutch group, as only 45.6% of the Dutch individuals who are truly with low positive affect are identified to receive the intervention, compared to 60.9% if the scale is strict invariant.

Summary of the Steps for Analyzing Partial Invariance With the Selection Approach

In summary, we suggest the following steps for performing the selection accuracy analysis when partial measurement invariance is identified:

1. Obtain parameter estimates for each group (i.e., latent factor means and variances, factor loadings, measurement intercepts, and uniqueness) under the partial measurement invariance model using regular SEM programs.
2. Determine the mixing proportion in the population.
3. Determine the intended percentage of selection from the combined population (i.e., `propsel` in `PartInv`) or a specific cutoff on the observed composite score (i.e., `cut_z`).

4. Call the R function `PartInv` (or use the web application) with the input of the parameter estimates under the partial invariance model, mixing proportion, and the intended percentage of selection to obtain the results from the selection accuracy analysis.
5. Obtain plausible parameter estimates under the strict invariance model by replacing the non-invariant parameters with the weighted averages of the estimates under the partial invariance model (or with other sensible values).
6. Repeat step 4 but use the plausible parameter estimates obtained in step 5.
7. Compare the results from steps 4 (partial invariance) and 6 (strict invariance) to evaluate the impact of partial invariance (e.g., examine the changes in the proportion selected, success ratio, sensitivity and specificity).

Note that if one uses `lavaan` for parameter estimations in steps 1, 4, 5 and 6 may be automated using the function `PartInv.lavaan` described in Appendix B.

Conclusions

Despite abundant research efforts to develop guidance for detecting partial measurement invariance and documenting the consequences of incorrectly modeling the partial invariance for research purposes, few attempts have been made to answer the question, "What actual impact would the detected partial invariance have on the selection accuracy when using the same test across groups?"

We agree with Millsap and Kwok (2004) that the impact of non-invariance should be evaluated in relation to the purpose of a given test, and should be assessed not only by its presence but also by its practical significance. Millsap and Kwok made one of the early attempts to answer the question and proposed a framework for evaluating the impact of measurement non-invariance by quantifying the magnitude of non-invariance with respect to selection accuracy.

Although the approach has been recognized as important by a number of authors (Borsboom, 2006; Bowden, 2013; Meade & Bauer, 2007; Schmitt & Kuljanin, 2008), our literature review showed that only one study (Alkemade, Bowden, & Salzman, 2015) actually conducted the corresponding analysis with real data. One likely reason for this is the lack of computer programs to automate the many steps involved in the analysis.

The present paper briefly reviewed the theoretical impact of partial invariance on selection in a combined population of two subpopulations. More important, we provide an R script and a web application to automatically compute the cutoff scores on the observed composite and the latent factor, the proportions selected, success ratios, sensitivities, and specificities of the test for each subpopulation, and a diagram visualizing the selection, taking as inputs parameter estimates of the partial invariance model, which can be obtained from standard SEM software. In addition to simply carrying out hypothesis tests for measurement invariance models, we encourage researchers to perform the selection accuracy analyses to understand the practical impact of partial invariance on selection, just like researchers should report effect sizes in addition to *p*-values to understand the practical significance of the results of data analyses.

References

Alkemade, N., Bowden, S. C., & Salzman, L. (2015). Scoring correction for MMPI-2 Hs scale with patients experiencing a traumatic brain injury: A test of measurement invariance. *Archives of Clinical Neuropsychology*, *30*(1), 39–48. http://doi.org/10.1093/arclin/acu058

Altman, D. G., & Bland, J. M. (1994a). Diagnostic tests 1: Sensitivity and specificity. *British Medical Journal*, *308*, 1552. http://doi.org/10.1136/bmj.308.6943.1552

Altman, D. G., & Bland, J. M. (1994b). Diagnostic tests 2: Predictive values. *British Medical Journal*, *309*, 102. http://doi.org/10.1136/bmj.309.6947.102

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

Azzalini, A., & Genz, A. (2016). The R package mnormt: The multivariate normal and ’t distributions (R package version 1.5-4).

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11 Suppl 3), S176–S181. http://doi.org/10.1097/01.mlr.0000245143.08679.cc

Borsboom, D., Romeijn, J.-W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, *13*(2), 75–98. http://doi.org/10.1037/1082-989X.13.2.75

Bowden, S. C. (2013). Theoretical convergence in assessment of cognition. *Journal of Psychoeducational Assessment*, *31*(2), 148–156. http://doi.org/10.1177/0734282913478035

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

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*(1), 1–27. http://doi.org/10.1177/014920639902500101

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, Ohio: Cengage Learning.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, *2*, 199–215. http://doi.org/10.1080/15305058.2002.9669493

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, *38*, 164–187. http://doi.org/10.1111/j.1745-3984.2001.tb01121.x

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

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, *44*(Suppl 3), S182–S188. http://doi.org/10.1097/01.mlr.0000245443.86671.c4

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 129–145). Hillsdale, NJ: Erlbaum. http://doi.org/10.1017/CBO9781107415324.004

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. http://doi.org/10.1080/03610739208253916

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 31–39. http://doi.org/10.1080/10705511.2014.856694

Kaplan, D. (1989). Power of the likelihood ratio test in multiple group confirmatory factor analysis under partial measurement invariance. *Educational and Psychological Measurement*, *49*(3), 579–586. http://doi.org/10.1177/001316448904900308

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 611–635. http://doi.org/10.1080/10705510701575461

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. http://doi.org/10.1016/0883-0355(89)90002-5

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

Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika*, *79*, 569–584. http://doi.org/10.1007/s11336-013-9376-7

Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika, 78*, 59–82. http://doi.org/10.1007/s11336-012-9302-4

Millsap, R. E. (1995). Measurement Invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, *30*(4), 577–605. http://doi.org/10.1207/s15327906mbr3004_6

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*(4), 461–473. http://doi.org/10.1007/s11336-007-9039-7

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessiing measurement bias. *Applied Psychological Measurement*, *17,* 297–334.

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*(1), 93–115. http://doi.org/10.1037/1082-989X.9.1.93

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45–60. http://doi.org/10.1093/pan/mpt014

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: review and recommendations. *Educational Measurement: Issues and Practice*, *19*(3), 5–15. http://doi.org/10.1111/j.1745-3992.2000.tb00033.x

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385–401. http://doi.org/10.1177/014662167700100306

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353–368. http://doi.org/10.1177/014662169501900405

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48,* 1–36. http://doi.org/10.18637/jss.v048.i02

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222.

http://doi.org/10.1016/j.hrmr.2008.03.003

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, *89*(3), 497–508. http://doi.org/10.1037/0021-9010.89.3.497

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *The Journal of Applied Psychology*, *91*(6), 1292–306. http://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), 78–107. http://doi.org/10.1086/209528

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology*, *9*(1), 1–12. http://doi.org/10.1027/1614-2241/a000049

Swets, J. A., Pickett, R. M., Whitehead, S. F., Getty, D. J., Schnur, J. A., Swets, J. B., & Freeman, B. A. (1979). Assessment of diagnostic technologies. *Science, 205,* 753–759. http://doi.org/10.1126/science.462188

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing offwith Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*(OCT), 1–15. http://doi.org/10.3389/fpsyg.2013.00770

Van De Schoot, R., Schmidt, P., & De Beuckelaer, A. (Eds.). (2015). *Measurement Invariance*

[Special issue]. *Frontiers in Psychology*. Lausanne, Switzerland: Frontiers Media. http://doi.org/10.3389/978-2-88919-650-0

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*(2), 139–158. http://doi.org/10.1177/1094428102005002001

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, *89*(5), 696–716. http://doi.org/10.1037/0022-3514.89.5.696

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, *8*(1), 16–37. http://doi.org/10.1037/1082-989X.8.1.16

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. http://doi.org/10.1080/10705510701301677

Zhang, B., Fokkema, M., Cuijpers, P., Li, J., Smits, N., & Beekman, A. (2011). Measurement invariance of the Center for Epidemiological Studies Depression Scale (CES-D) among Chinese and Dutch elderly. *BMC Medical Research Methodology*, *11*, 74–83. http://doi.org/10.1186/1471-2288-11-74

Footnotes

[1] Note that an assumption of partial invariance is that the constructs being measured are conceptually comparable across groups. In a counterexample, Lommen, van de Schoot, and Engelhard (2014) discussed how theory may predict that a measure of posttraumatic stress disorder (PTSD) shows changes in its measurement properties across time, and argued for the possibility of treating pre-symptom scores as measuring a different construct than the post-symptom scores.

[2] For instance, in Wicherts et al. (2005) the measurement non-invariance across genders was attributed to stereotype threat, in which certain questions induce higher anxiety for test takers of one gender than of the other. In the DIF literature, Gierl and Khaliq (2001) found that measurement non-invariance in achievement tests across language groups can be predicted by various categories of translation factors (see also Ercikan, 2002). In these examples, the researchers studied measurement non-invariance per se rather than treating them as a nuisance that biased the research findings of interest. Such research efforts to uncover the sources of measurement non-invariance are important; however, it is beyond the scope of the present article as we focus more on the practical impact of measurement non-invariance for selection purposes.

[3] Although the CES-D items, measuring the frequency with which participants experienced depressive symptoms, are in a 4-point scale response format (1 = none, 2 = one or two days a week, 3 = three or four days per week, and 4 = five days or more per week), Zhang et al. (2011) treated the items as continuous as in regular CFA so the parameter estimates may be biased, but the impact of such misspecification is not known without access to the raw data. Also, the selection accuracy analysis so far discussed assumes a continuous and normally distributed observed composite, Z, which was clearly violated for the CES-D. However, as the composite

score is the sum of multiple ordered categorical items, the normality assumption should hold approximately with enough items.

Table 1

Proportions Selected, Success Ratios, Sensitivities, and Specificities for Hypothetical Examples 1, 2, and 3

	Strict Invariance	Partial Scalar Invariance	Partial Metric Invariance
PS – Reference	.311	.316	.339
PS – Focal	.189	.184	.161
SR – Reference	.714	.710	.691
SR – Focal	.580	.587	.566
SE – Reference	.677	.684	.715
SE – Focal	.638	.627	.529
SP – Reference	.868	.863	.844
SP – focal	.904	.908	.916

Note. *PS* = proportion selected; *SR* = success ratio; *SE* = sensitivity; *SP* = specificity.

Table 2

Factor Loadings, Intercepts, and Uniqueness of the CES-D Positive Affect Factor

	Factor Loadings	Intercepts		Uniqueness	
	All	Chinese	Dutch	Chinese	Dutch
Good	1.00	1.54	0.68	1.20	0.72
Hopeful	1.66	1.36		0.81	
Happy	2.30	1.16		0.32	
Enjoyed	2.29	1.08		0.32	

Note. *N* = 4,903 for the Chinese sample and *N* = 1, 903 for the Dutch sample. The four items were reversely coded.

Table 3

Proportions Selected, Success Ratios, Sensitivities, and Specificities of the CES-D Positive Affect Factor

	Strict Invariance	Partial Scalar Invariance
PS – Chinese	.152	.173
PS – Dutch	.081	.049
SR – Chinese	.673	.646
SR – Dutch	.535	.695
SE – Chinese	.656	.688
SE – Dutch	.609	.456
SP – Chinese	.941	.927
SP – Dutch	.960	.984

Note. *N* = 4,903 for the Chinese sample and *N* = 1, 903 for the Dutch sample. PS = proportion selected; SR = success ratio; SE = sensitivity; SP = specificity.

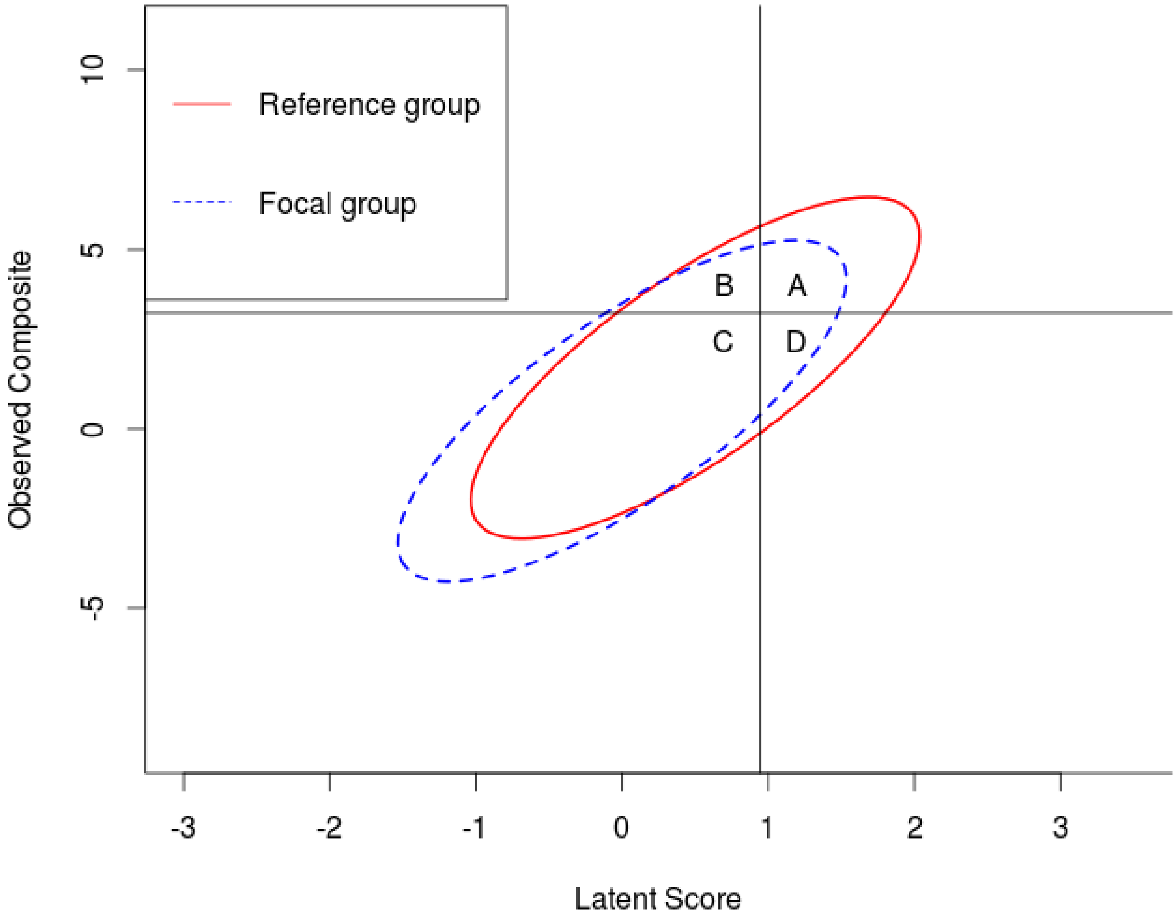
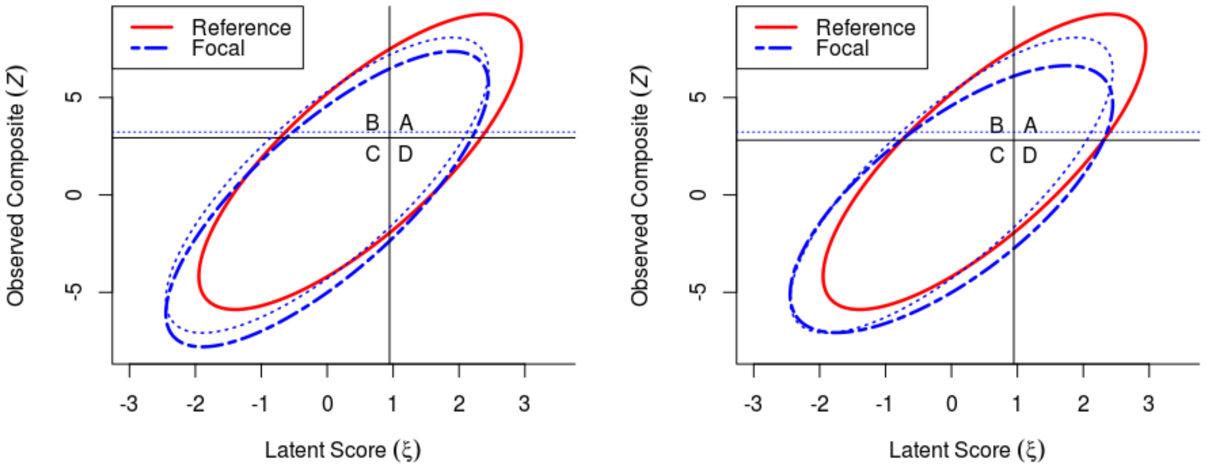


Figure 1. Recreation of Figure 2 from Millsap and Kwok (2004) showing the bivariate distribution of latent score and observed composite score with respect to two subpopulations.



(a) (b)

Figure 2. Bivariate distribution of latent score and observed composite score for (a) Example 2 with partial scalar invariance and (b) Example 3 with partial metric invariance. The thicker lines with long dashes show the distributions of the focal group under partial invariance, whereas the thinner dotted line show the distributions of the focal group under strict invariance.