Quantifying the Impact of Partial Measurement Invariance in Diagnostic Research: An Application to Addiction Research

Mark H. C. Lai[a], George B. Richardson[b], Hio Wa Mak[c]

[a]Department of Psychology, University of Southern California, Los Angeles, CA, USA

[b]School of Human Services, University of Cincinnati, Cincinnati, OH, USA

[c]Department of Human Development and Family Studies, The Pennsylvania State University, University Park, PA, USA

Author Note

This is an Accepted Manuscript of an article by Elsevier in Addictive Behavior on 19/11/2018, available online: https://www.sciencedirect.com/science/article/abs/pii/S0306460318313455

Correspondence concerning this article should be addressed to Mark Lai, Department of Psychology, University of Southern California, Los Angeles, CA 90089-1061. Email: hokchiol@usc.edu

Abstract

Establishing measurement invariance, or that an instrument measures the same construct(s) in the same way across subgroups of respondents, is crucial in efforts to validate social and behavioral instruments. Although substantial previous research has focused on detecting the presence of noninvariance, less attention has been devoted to its practical significance and even less has been paid to its possible impact on diagnostic accuracy. In this article, we draw additional attention to the importance of measurement invariance and advance diagnostic research by introducing a novel approach for quantifying the impact of noninvariance with binary items (e.g., the presence or absence of symptoms). We illustrate this approach by testing measurement invariance and evaluating diagnostic accuracy across age groups using DSM alcohol use disorder items from a public national data set. By providing researchers with an easy-to-implement R program for examining diagnostic accuracy with binary items, this article sets the stage for future evaluations of the practical significance of partial invariance. Future work can extend our framework to include ordinal and categorical indicators, other measurement models in item response theory, settings with three or more groups, and via comparison to an external, "gold-standard" validator.

Keywords: measurement invariance; diagnostic accuracy; alcohol use disorder; practical significance; addiction research

Quantifying the Impact of Partial Measurement Invariance in Diagnostic Research: An Application to Addiction Research

1. Overview

1.1. Measurement Invariance

Measurement is the foundation of science. Establishing measurement invariance, or that an instrument or a test measures the same construct(s) in the same way across subgroups of respondents, is crucial in efforts to validate social and behavioral instruments (Meredith, 1993; van de Vijver & Poortinga, 1997; Vandenberg, 2002). When the properties of an instrument vary across subgroups of participants with different characteristics (e.g., language spoken, age, gender, and other demographic characteristics), the observed scores in different groups will be on different metrics and cannot be directly compared—any observed group differences in means, regression coefficients, or prevalence rates of disorders will be confounded with the inconsistent properties of the instrument (e.g., Millsap, 2011). Thus, at least some degree of measurement invariance is needed for meaningful interpretations of research findings (Schmitt & Kuljanin, 2008).

With the advance of methodological research in evaluating measurement invariance (e.g., Millsap, 2011) and increased awareness from substantive researchers, measurement invariance evaluation has been applied to a growing array of instruments designed to measure social and behavior constructs (e.g., personality traits, Nye et al., 2008; prosocial behaviors, Carlo, Knight, McGinley, Zamboanga, & Jarvis, 2010). Measurement invariance has also been evaluated in research of psychopathologies such as depression (Borsboom, 2008; Crockett, Randall, Shen, Russell, & Driscoll, 2005) and substance use disorder (e.g., Aiken, Stein, & Bentler, 1994). Although evaluations of measurement invariance is becoming more common, researchers in

many disciplines still frequently compare groups without establishing equivalence of measurement (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Vandenberg & Lance, 2000).

1.2. Partial Invariance

Full measurement invariance holds when all the measurement parameters of all items are the same across all groups. This is rather difficult to achieve empirically. Often researchers find that measurement invariance holds for only a subset of items, a condition known as *partial measurement invariance* (also known as differential item functioning in item response theory; see Penfield & Lam, 2000). Whereas it is possible to make valid group comparisons (e.g., on means, path coefficients) using latent variable models when only some of the items are invariant (Byrne, Shavelson, & Muthén, 1989), group comparisons based on observed scores (e.g., scale or composite scores; Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009) may yield misleading conclusions, as noninvariant items are not precluded from biasing comparisons of means (Schmitt & Kuljanin, 2008), path coefficients (Hsiao & Lai, 2018), and prevalence rates across groups.

1.3. Practical Significance of Partial Invariance

Although substantial previous research has focused on detecting the presence of partial invariance (e.g., Byrne, Shavelson, & Muthén, 1989; Stark, Chernyshenko, & Drasgow, 2006), discussion of the *practical significance* (e.g., Ferguson, 2009; Kirk, 1996) of partial invariance, that is, whether the degree of noninvariance is large enough in a practically meaningful way, has been scarce. Specifically, just as a statistically significant *t* test may indicate a negligible mean difference, or in other words a trivial effect size (e.g., Kirk, 1996), statistical indication of measurement invariance violations may or may not have substantial practical impact on research

findings and the efficacy of the instrument. As Millsap and Kwok (2004) discussed, the practical significance of partial invariance should be interpreted "in relation to the purpose of the measure" (pp. 94–95).

1.4. Diagnostic Accuracy Analysis

Even less attention has been devoted to the practical impact of partial invariance on diagnostic accuracy based on observed scores (Millsap, 2011). Millsap and Kwok (2004; see also Lai, Kwok, Yoon, & Hsiao, 2017) first proposed to examine the practical significance of partial invariance by comparing the accuracy of selection (of individuals for, e.g., job promotion, or classification of drinking behaviors) based on a given partially invariant instrument to the accuracy of selection when full measurement invariance is assumed, and the framework can be extended to cover diagnostic accuracy of partially invariant instruments. If diagnostic accuracy in one or more groups changes a lot due to noninvariant items, then the partial invariance has a relatively large practical impact. Note that in this article we used the term diagnostic accuracy to mean the estimated performance of a test in the absence of an external validator (e.g., actual clinical diagnosis and "gold standard" measure; see Faraone & Tsuang, 1994).

1.5. Current Study

Above we have drawn additional attention to the importance of measurement invariance as well as identified a significant gap in the literature with respect to the impact of partial invariance on diagnostic accuracy. Next we advance diagnostic research by introducing a novel approach for quantifying the impact of noninvariance with binary items (e.g., the presence or absence of symptoms). We illustrate this approach by testing measurement invariance and then evaluating diagnostic accuracy across age groups using Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association [APA], 1994) alcohol dependence items

from a public national data set. The current application to addiction research is significant because (a) addiction researchers and clinicians often classify people into substance use disorder (SUD) categories by summing the SUD criteria and comparing the scores to cut-points (e.g., endorsing > 3 items), which requires measurement invariance to hold (Midanik, Greenfield, & Bond, 2007), and (b) although research suggests the DSM-IV criteria for alcohol, marijuana, and cocaine use disorders are measurement invariant between community and selected (i.e., clinical) samples (Derringer et al., 2013), evidence also suggests that SUD criteria (e.g., DSM-IV alcohol dependence tolerance and time spent) function differentially across age groups, gender, and race/ethnicity (e.g., Harford, Yi, Faden, & Chen, 2009; Martin, Chung, Kirisci, & Langenbucher, 2006; Saha, Chou, & Grant, 2006). The current study is also significant given recent calls in addiction research for analyses that examine item and measure performance across critical populations (e.g., Baker, Breslau, Covey, & Shiffman, 2012; Burlew, Feaster, Brecht, & Hubbard, 2009). Importantly, we are aware of no studies that have examined the impact of partial invariance on SUD diagnosis. If SUD is only partially invariant, diagnostic/classification estimates for subpopulations may not be comparable across groups. There is, therefore, a critical need for studies addressing this gap in the literature.

2. Application to Addiction Research

2.1. Definition and Model Notations

Formally, measurement invariance (Mellenbergh, 1989) holds 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. In other words, for participants with the same score on the latent construct, their group membership plays no role in determining how they respond to an item.

In addiction research, items in a scale or an instrument, such as the diagnostic criteria for alcohol use disorder (AUD), are sometimes formulated with a binary response format with 0/1 = absence/presence of a symptom. Then a common choice of measurement model under the structural equation model (SEM) framework is to assume a continuous unobserved response variate, X_{ij}^*, underlying the observed response for person i on the jth binary item X_{ij}, so that

$$\begin{cases} X_{ij} = 1, \text{if } X_{ij}^* > \tau_j \\ X_{ij} = 0, \text{if } X_{ij}^* \leq \tau_j \end{cases}'$$

where τ_j is a threshold parameter of the unobserved variate over which the observed score will be 1; in other words, when the degree of symptom (X_{ij}^*) is above the threshold, a participant will respond with a "1." With the continuous X_{ij}^*, one can then apply the common factor model with the form

$$X_{ij}^* = \lambda_j \xi_i + \varepsilon_{ij},$$

where ξ_i denotes the latent score (e.g., true degree of AUD) for person i; λ_j is the factor loading for item j, which is the regression slope of X_{ij}^* on ξ_i; and ε_{ij} is the unique factor score that captures the influence of construct-irrelevant factors on the item response (e.g., machine error).[1] Here we assume that ξ_i and ε_{ij} for all items are jointly normal and independent to each other.[2] Measurement invariance thus would imply that the thresholds, factor loadings, and unique factor variances are all the same across groups.

2.2. Measurement Invariance With Binary Items

In many areas of diagnostic research, binary items that assess the presence or absence of a symptom are used. Importantly, there are substantial differences in the stages of measurement invariance under a factor model for binary items and for continuous items. With continuous items, researchers commonly distinguished between four stages of invariance (e.g., Vandenberg & Lance, 2000): configural, metric, scalar, and strict invariance. However, with binary and

categorical items, these four stages may not be fully applicable because of (a) the addition of thresholds and (b) the observed scores having limited categories. As discussed in Wu and Estabrook (2016), with binary items only three stages of measurement invariance can be tested, and one way to organize the three stages would be to first establish configural invariance (having the same factor structure but no constraints on the measurement parameters), followed by scalar invariance (equal factor loadings and thresholds), and then finally strict invariance (scalar invariance + equal unique factor variances of the unobserved response variates). Unlike in the case of continuous items where scalar invariance is sufficient for valid mean comparisons, with binary items, only when strict invariance holds would respondents from two populations with the same level of the latent trait have equal probabilities of responding "1." As far as we are aware, no studies have developed approaches for evaluating the impact of partial invariance on diagnostic accuracy in the context of binary items.

2.3. Diagnostic Accuracy Indices With Two Populations

Consider an example of a diagnostic test used to screen White and Asian participants in the United States for AUD. Following previous literature, we call the majority population (i.e., White) the *reference* population and the minority population (i.e., Asian) the *focal* population. For each population, one can conceptualize the relation between the observed score on the diagnostic test and the true latent AUD score on a graph, as shown in Figure 1. For each population, the relation between the latent score and the observed test score is represented with an ellipse. The horizontal and vertical lines, which denote the cutoffs for the observed diagnostic test and for the true latent AUD score, divide the total area into four quadrants. The area above the horizontal line (i.e., Quadrants A and B) represents those receiving a positive diagnosis whereas area on the right of the vertical line (i.e., Quadrants A and D) represents individuals who

truly have AUD. Thus, Quadrant A represents *true positives*, and Quadrant C denotes the *true negatives*. Quadrant B represents the *false positives*—individuals who do not have AUD but are incorrectly diagnosed by the instrument as having AUD. Finally, Quadrant D denotes the *false negatives*—individuals who truly have AUD but are not diagnosed by the instrument. Conventional terminologies used in diagnostic testing (Altman & Bland, 1994a, 1994b) provide four indices (see Table 1) that summarize diagnostic accuracy: proportion selected (*PS*; or proportion diagnosed), success ratio (*SR*) or positive predictive value, sensitivity (*SE*), and specificity (*SP*).

Proportion selected (A + B) in our example refers to the proportion of individuals being diagnosed as having AUD by the diagnostic criteria, which is the prevalence rate of AUD for each population. Success ratio (A / [A + B]) is the proportion of individuals who truly have AUD among all individuals being positively diagnosed by the diagnostic criteria; a low success ratio means that many individuals who do not have AUD are given the diagnosis, which may lead to potential stigmatization as well as wasted resources. Sensitivity (A / [A + D]) is the proportion of individuals who are correctly positively diagnosed among all individuals who truly have AUD; a low sensitivity indicates many individuals with AUD are undiagnosed, which means failure to provide help and resources to those who are truly in need. Finally, specificity (C / [B + C]) is the proportion of individuals who are correctly ruled out among all individuals who do not have AUD; a low specificity indicates that many individuals without AUD are being labelled as having such a problem, potentially leading to stigmatization and wasted resources.

2.4. Heuristic Example

Now consider an example of five AUD items tested across two groups (populations). Assume that the latent factor means are 0 and -0.5 for the reference and the focal groups.

Assume that the factor variances are both 1.0, and the unique factor variances are all 1.0 for both groups. Also assume that the loadings (λ_1 to λ_5) are invariant and equal 1, 0.8, 1.2, 0.9, and 1.3, but only the first three indicators are scalar invariant across the two groups with thresholds (τ_1 to τ_3) of 1, 2, and 0.5. For the remaining two items, assume that the thresholds (τ_4 and τ_5) are 2.1 and 2.4 for the reference group and are 1.4 and 1.5 for the focal group. Thus, Asians would be more likely to endorse the forth and the fifth items than White people at the same latent AUD level (as illustrated in Figure 2). However, knowing that the test is partially invariant in that the thresholds are lower for the focal group by differences of 0.7 and 0.9 for items 4 and 5 does not help researchers understand the practical impact, and diagnostic accuracy analysis can provide such information in the context of screening or diagnostic tests with binary items.

Assume that the test would screen an individual endorsing three or more items. To perform diagnostic accuracy analysis, one needs to obtain the diagnostic accuracy indices for the partially invariant parameters as well as if the parameters were invariant.[3] The `PartInv_cat()` R function in Supplemental material 1 automates this process by taking as input the parameter estimates from standard SEM output based on the partial invariance model. A tutorial on using the R script can be found in Supplemental material 2, with the resulting diagnostic indices shown in Table 2.

For this example, if the test were fully invariant, 10.7% of the reference group and 4.7% of the focal group would have been diagnosed using the cutoff score of 3, compared to 8.1% and 6.4% for the actual partial invariant test. Therefore, the partial invariance has a substantial impact as it causes the test to positively classify fewer people in the reference group but more people in the focal group, mainly because the two noninvariant items are more likely to be endorsed by the focal group at the same level of the latent variable. Also, while the

noninvariance did not have a strong influence on the specificity of the test (from .960 to .974 for the reference group and from .979 to .964 for the focal group), it has a relatively big impact on its sensitivity: it drops from .648 to .551 for the reference group while increases from .614 to .740 for the focal group.

Therefore, with the diagnostic accuracy analysis, one has more information regarding the impact of the partial invariance, which in this example seems matter most if one is interested in obtaining accurate prevalence rates for both groups and/or if one is concerned about identifying people who should be diagnosed based on their true latent scores (potentially for treatment and interventions) in the reference group (in which high sensitivity is needed).

3. Real Data Illustration

To further illustrate the diagnostic accuracy analysis, we used alcohol use data from the 2014 National Survey on Drug Use and Health (NSDUH), which is an annual nationally representative survey of substance use and health in the United States conducted by the Substance Abuse and Mental Health Services Administration (SAMHSA). We examined diagnostic accuracy across two age groups (12-25 vs. 26 and older) given evidence of differential item functioning between older and younger participants on some of the items (e.g., slope was lower among older compared to younger participants for tolerance and greater in the older group for activities given up; Saha, et al., 2006).

3.1. Sample

Participants were noninstitutionalized civilians aged 12 or older sampled across all 50 states and the District of Columbia. More details regarding NSDUH study design and survey procedures are available elsewhere (Center for Behavioral Health Statistics and Quality, 2015). There were 55,271 participants in the original data set; however, after excluding participants who

had never used alcohol or had not used alcohol in the past 12 months, or had missing data on all AUD items (n = 293),[4] the final sample of the real data illustration consisted of 28,629 participants (50.1% females), with 36.1% (n = 10,340) aged between 12 to 25 years and 63.9% (n = 18,289) aged 26 years or above. In the final sample, participants reported their race as White (66.3%), Hispanic (14.7%), Black/African American (10.7%), Asian (3.3%), Native American (1.5%), Native HI/Pacific Islanders (0.4%), mixed (3.1%).

3.2. Measure

Participants were asked about their alcohol use via the DSM-IV alcohol dependence criteria (APA, 1994). Following the NSDUH coding manual for substance dependence, we combined and recoded participants' responses to 10 alcohol related items to form the seven binary DSM-IV alcohol dependence criteria, and participants were considered to have alcohol dependence if they met three of the seven dependence criteria.[5]

3.3. Tests of Measurement Invariance

As mentioned, we evaluated measurement invariance of the seven binary DSM-IV items for diagnosing alcohol dependence across two age groups. A configural invariance one-factor model was first fitted to the data using Mplus 7.4 with the robust weighted least square estimator (ESTIMATOR=WLSMV) and the THETA parameterization.[6] The model fit was good, but an examination of the modification indices (MIs) showed that a large unmodeled unique factor covariance between Item 2 and Item 4, especially for the group with older age (MI = 383.38). Freeing this unique factor covariance in both groups resulted in substantially improved model fit, $\chi^2(df = 26) = 170.18$, $\Delta\chi^2(df = 2) = 225.129$, $p < .001$, RMSEA = .020, 90% CI [.017, .023], CFI = .995.

We then tested for scalar invariance by constraining the factor loadings and the thresholds to be equal across groups. The χ^2 difference test was statistically significant, with $\Delta\chi^2(df = 5) = 56.50$, $p < .001$, and MIs showed evidence for threshold noninvariance of item 3 (MI = 97.17). A partial scalar invariance model with unequal item 3 threshold across groups demonstrated good fit, with $\chi^2(df = 30) = 165.56$, RMSEA = .018, 90% CI [.015, .020], CFI = .996. For the remaining scalar invariant items, we found evidence of noninvariance for the unique factor variances with $\Delta\chi^2(df = 6) = 138.33$. Modification indices suggested noninvariance of the unique factor variance for item 4 (MI = 143.55), and further sequential investigation (Yoon & Millsap, 2007) suggested noninvariance for item 1, 2, and 5 (MIs = 18.19 to 33.19). In other words, only items 6 and 7 were considered invariant in the final partial strict invariance model.

The latent factor mean and variance were fixed to be 0 and 1.0 for the younger group and were estimated to be -1.29 and 3.16 for the older group, so the older group (aged 26+) had lower alcohol dependence tendency and was more heterogeneous. The model parameters were shown in Table 3, with the thresholds for item 3 estimated to be 0.76 for the younger group and 1.18 for the older group, and the unique factor variances for the noninvariant items were between 1.41 and 1.83. Thus, with the same degree of alcohol dependence, an older participant would be less likely to endorse item 3 than a younger participant.

3.4. Diagnostic Accuracy Analysis

Despite the statistical evidence of noninvariance, it is not clear what practical impact is associated with the noninvariant threshold and unique factor variances, so we conducted a diagnostic accuracy analysis (see Supplemental material 3). As shown in Table 4, if the items were measurement invariant, 8.8% in the younger population and 5.4% in the older population would be identified as potentially having alcohol dependence. The statistically significant

noninvariant items changed the rates to 8.6% and 5.5%, and most people would agree that a change of 0.2 percentage points is small unless very high precision is needed. If measurement invariance held, the sensitivity values were .645 and .716 for the younger and older populations, respectively, and the specificity values were .965 and .986. With the partial invariance found on the items, the sensitivity (.663 and .698) and specificity values (.968 and .983) were still quite comparable. Therefore, with the diagnostic accuracy analysis, one can see that the statistically significant noninvariance had a small impact on the efficacy of the items for diagnostic purpose.

4. Discussion

4.1. Statistical and Practical Significance of Noninvariance

Measurement invariance is particularly crucial in ensuring inferences about cross-group differences are correct (e.g., gender, age, and culture; Milfont & Fischer, 2010; Widaman & Reise, 1997). Despite increasing awareness from methodologists and addiction researchers on the need for measurement invariance testing, there has been little methodological discussion regarding metrics that quantify the practical significance of violations of measurement invariance(Putnick & Bornstein, 2016; Vandenberg, 2002). In this article, we reviewed one approach to answer the practical significance question—selection/diagnostic accuracy analysis. This approach directly links noninvariance to the efficacy of the instrument for diagnostic, selection, or classification purposes. Whereas the original framework was developed for continuous and normally distributed indicators, we extended the framework to accommodate binary items and developed an R program to perform the analysis. Furthermore, through heuristic and real data examples, we illustrated how diagnostic accuracy analysis can provide more concrete information about how measurement noninvariance impacts the diagnostic results

given by tests, and showed that statistically significant violations of measurement invariance may or may not be practically significant.

It should be emphasized that the interpretation of practical significance depends on the context and the constructs being measured. For example, a difference of one percentage points in proportion selected may be considered small for relatively low stake traits (e.g., attitudes), but this difference may be substantial in more high-stake contexts where instruments play a major role in diagnosis (e.g., SUD and other mental disorders), despite Millsap and Kwok's (2004) suggestion that changes under five percentage points "are unlikely to be meaningful in most applications" (p. 111).[7] The question of practical significance should also be answered differently depending on intended use of the test (Millsap & Kwok, 2004). Whereas diagnostic accuracy analysis would be highly suitable for instruments used for diagnosis and classifications, it may be less meaningful for instruments intended to score individuals on a continuum that does not involve a cutoff score, and researchers should consult alternative metrics for quantifying practical impact of invariance violations (e.g., Nye & Drasgow, 2011; Oberski, 2014; Stark, Chernyshenko, & Drasgow, 2004).

4.2. Conclusions

In the past few decades there has been a strong emphasis on effect size and practical significance for quantitative research in behavioral sciences (e.g., Ferguson, 2009; Henson, 2006; Lai & Kwok, 2016). Similarly, for measurement invariance, we urge addiction researchers to not only conduct invariance testing for psychological instruments, which is extremely important to ensure valid research findings, but to also evaluate and report the practical significance of any detected invariance violations as a function of the purposes of tests and the associated research contexts, in the same manner as interpreting effect size statistics. We hope that by (a)

successfully raising researchers' awareness on recent developments in the measurement invariance literature, (b) highlighting the importance of understanding and reporting the practical significance of the diagnostic accuracy indices to violations of invariance for diagnostic tools, and (c) providing researchers with an easy-to-implement R program to perform such analyses, this article helps to increase the extent to which future research evaluates and reports the practical significance of partial invariance. Finally, we expect future methodological work to extend the framework to include ordinal and categorical indicators, other measurement models in item response theory, settings with three or more groups, and via comparison to an external, "gold-standard" validator.

Acknowledgement

Hio Wa Mak was supported by the National Institute on Drug Abuse (T32 DA017629 and P50 DA039838). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse or the National Institutes of Health.

References

Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology*, *62*, 488–499. https://doi.org/10.1037/0022-006X.62.3.488

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

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

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.). Washington, DC: Author.

Baker, T. B., Breslau, N., Covey, L., & Shiffman, S. (2012). DSM criteria for tobacco use disorder and tobacco withdrawal: A critique and proposed revisions for DSM-5. *Addiction*, *107*, 263–275. https://doi.org/10.1111/j.1360-0443.2011.03657.x

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology, 64*, 1089–1108. https://doi.org/10.1002/jclp.20503

Burlew, A. K., Feaster, D., Brecht, M. L., & Hubbard, R. (2009). Measurement and data analysis in research addressing health disparities in substance abuse. *Journal of Substance Abuse Treatment*, *36*, 25–43. https://doi.org/10.1016/j.jsat.2008.04.003

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*, 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Carlo, G., Knight, G. P., McGinley, M., Zamboanga, B. L., & Jarvis, L. H. (2010). The multidimensionality of prosocial behaviors and evidence of measurement equivalence in Mexican American and European American early adolescents. *Journal of Research on Adolescence, 20*, 334–358. https://doi.org/10.1111/j.1532-7795.2010.00637.x

Center for Behavioral Health Statistics and Quality. (2015). 2014 National Survey on Drug Use and Health: Methodological summary and definitions. Retrieved from https://www.samhsa.gov/data/

Crockett, L. J., Randall, B. A., Shen, Y. L., Russell, S. T., & Driscoll, A. K. (2005). Measurement equivalence of the center for epidemiological studies depression scale for Latino and Anglo adolescents: a national study. *Journal of Consulting and Clinical Psychology, 73*(1), 47–58. https://dx.doi.org/10.1037/0022-006X.73.1.47

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Sociology*, *40*, 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

Derringer, J., Krueger, R. F., Dick, D. M., Agrawal, A., Bucholz, K. K., Foroud, T., . . . & Nurnberger, J. I. (2013). Measurement invariance of DSM-IV alcohol, marijuana and cocaine dependence between community-sampled and clinically overselected studies. *Addiction*, *108*, 1767–1776. https://doi.org/10.1111/add.12187

Faraone, S. V., & Tsuang, M. T. (1994). Measuring diagnostic accuracy in the absence of a "gold standard." *American Journal of Psychiatry, 151,* 650–657. https://doi.org/10.1176/ajp.151.5.650

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*, 532–538. https://doi.org/10.1037/a0015808

Harford, T. C., Yi, H., Faden, V. B., & Chen, C. M. (2009). The dimensionality of DSM-IV alcohol use disorders among adolescent and adult drinkers and symptom patterns by age, gender, and race/ethnicity. *Alcoholism: Clinical and Experimental Research, 33*, 868–878. https://doi.org/10.1111/j.1530-0277.2009.00910.x

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist*, *34*, 601–629. https://doi.org/10.1177/0011000005283558

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

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746–759. https://doi.org/10.1177/0013164496056005002

Lai, M. H. C., & Kwok, O. (2016). Estimating standardized effect sizes for two- and three-level partially nested data. *Multivariate Behavioral Research, 51*, 740–756. https://doi.org/10.1080/00273171.2016.1231606

Lai, M. H. C., Kwok, O., Yoon, M., & Hsiao, Y.-Y. (2017). Understanding the impact of partial factorial invariance on selection accuracy: An R script. *Structural Equation Modeling: A Multidisciplinary Journal, 24*, 783–799. https://doi.org/10.1080/10705511.2017.1318703

Martin, C. S., Chung, T., Kirisci, L., & Langenbucher, J. W. (2006). Item response theory analysis of diagnostic criteria for alcohol and cannabis use disorders in adolescents: implications for DSM-V. *Journal of Abnormal Psychology, 115*, 807–814. https://dx.doi.org/10.1037/0021-843X.115.4.807

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

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

Midanik, L. T., Greenfield, T. K., & Bond, J. (2007). Addiction sciences and its psychometrics: The measurement of alcohol-related problems. *Addiction*, *102*, 1701–1710. https://doi.org/10.1111/j.1360-0443.2007.01886.x

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*, 111–121. https://doi.org/10.21500/20112084.857

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

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

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology, 96,* 966–980. https://doi.org/10.1037/a0022955

Nye, C. D., Roberts, B. W., Saucier, G., & Zhou, X. (2008). Testing the measurement equivalence of personality adjective items across cultures. *Journal of Research in Personality, 42,* 1524–1536. https://doi.org/10.1016/j.jrp.2008.07.004

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis, 22*(1), 45–60. https://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, 5–15*. https://doi.org/10.1111/j.1745-3992.2000.tb00033.x

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41,* 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Saha, T. D., Chou, P. S., & Grant, B. F. (2006). Toward an alcohol use disorder continuum using item response theory: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Psychological Medicine, 36*, 931–941. https://doi.org/10.1017/S003329170600746X

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210–222. https://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. https://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*, 1292–306. https://doi.org/10.1037/0021-9010.91.6.1292

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: differences between educational groups in human values measurement. *Quality & Quantity, 43*, 599–616. https://doi.org/10.1007/s11135-007-9143-x

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*, 29–37. https://doi.org/10.1027/1015-5759.13.1.29

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

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70. https://doi.org/10.1177/109442810031002

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S.

G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika, 81*, 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

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

Footnotes

[1]With a continuous indicator there is a measurement intercept parameter that indicates the expected value of the item when $\xi_i = 0$. With categorical items, it is usually constrained to be zero for model identification purpose.

[2]By assuming a normally distributed unobserved response variate underlying each binary item, the model is described as having a probit link as the relation between each binary item and the latent variable resembles that of a probit regression. In Mplus, one can instead use a logit link by using the maximum likelihood estimator (ESTIMATOR=ML) with numerical integration. The logit link is also commonly used in the item response theory framework, for example with the two parameter logistic (2-PL) model.

[3]To obtain the diagnostic indices assuming measurement invariance, one needs to assume that the item loadings and thresholds were the same across the two groups. As suggested in Millsap and Kwok (2004) and Lai et al. (2017), one option is to replace the noninvariant parameters with the weighted averages by the relative proportions of the two populations.

[4]The missing data rates were between 1.05% and 1.67% (*n*s between 305 and 483) for the seven alcohol dependence items. From Mplus, the likelihood ratio test for the missing completely at random (MCAR) assumption for sex, age, and the seven items had $\chi^2(df = 1867, N = 28{,}922) = 255.61$, $p = 1.00$, suggesting that the MCAR assumption is tenable.

[5]The NSDUH data set also has items of the DSM-IV alcohol abuse, which together with alcohol dependence were integrated into a single alcohol use disorder (AUD) in DSM-V. Given the complexity in scoring the alcohol abuse items in the data set, we decided to demonstrate with only the alcohol dependence items. Refer to the pages 214–218 of the Codebook (https://www.datafiles.samhsa.gov/study-dataset/national-survey-drug-use-and-health-2014-

nsduh-2014-ds0001-nid16876) for the description of the alcohol dependence items, and page 269 for the procedure of recoding.

[6]In Mplus there are two ways for identifying models with ordered categorical items, which are called DELTA and THETA parameterizations. In this manuscript we assumed that the models are identified using the THETA parameterization as it resembles the model formulation with continuous indicators. With the THETA parameterization the unique factor variances of the unobserved response variates are fixed to 1.0. With the DELTA parameterization the total variances (as opposed to the unique factor variance) of the unobserved response variates are fixed to 1.0. In multiple-group analyses, such identification constraints are only needed for the reference group (or any one of the groups).

[7]For exploratory work or contexts with limited prior information of what constitutes a practically significant change in diagnostic accuracy, one possible option is to compute Cohen's (1988) *h* effect size for the change in the diagnostic accuracy indices (e.g., sensitivity and specificity), $h = 2\arcsin\left(\sqrt{p_{\text{PI}}}\right) - 2\arcsin\left(\sqrt{p_{\text{MI}}}\right)$, where p_{PI} and p_{MI} are the values of the diagnostic accuracy index under the partial invariance and strict invariance models, respectively. The *h* effect size takes into account the fact that the same difference in an index depends on the base rate (e.g., a change from .05 to .10 is practically more significant than a change from .50 to .55, despite the same difference of .05). This can be obtained by passing the `show_effect_size = TRUE` argument to `PartInv_cat()`. When the base rate is close to .50, a difference of .05 in diagnostic accuracy index corresponds to $h \approx 0.10$, so we suggest, only when no other information is available, that the impact of partial invariance on a diagnostic accuracy is small when $h < 0.10$. For example, for the heuristic example in this article, the maximum *h*s are 0.20 and 0.27 for the reference and the focal groups on sensitivity, whereas for

the real data example all changes had $h < 0.08$. Note that the $h < 0.10$ rule is completely arbitrary and should be used with caution, and more systematic effort is needed to establish context-specific and domain-specific guidelines for determining what constitutes a practically significant change in diagnostic accuracy.

Table 1

Definition of the Diagnostic Accuracy Indices

	Mathematical Definition	Meaning
Proportion Selected (Diagnosed)	$PS_k = p(A_k) + p(B_k)$	Proportion of individuals being positively diagnosed; prevalence rates based on the criteria
Success Ratio	$SR_k = p(A_k) / [p(A_k) + p(B_k)]$	Proportion of individuals who are truly positive among all individuals being positively diagnosed by the diagnostic criteria
Sensitivity	$SE_k = p(A_k) / [p(A_k) + p(D_k)]$	Proportion of individuals who are correctly positively diagnosed by the diagnostic criteria among all individuals who truly are positive
Specificity	$SP_k = p(C_k) / [p(C_k) + p(B_k)]$	Proportion of individuals who are correctly ruled out by the diagnostic criteria among all individuals who are negative

Note: A = true positive, B = false positive, C = true negative, D = false negative, and $p(\cdot)$ represents the probability of a particular possibility. The index k denotes the population, where $k = r$ represents the reference population and $k = f$ represents the focal population.

Table 2

Diagnostic Accuracy Indices for the Heuristic Example

	Strict Invariance		Partial Scalar Invariance	
	Reference	Focal	Reference	Focal
Proportion diagnosed	.107	.047	.081	.064
Success Ratio	.672	.562	.715	.460
Sensitivity	.648	.614	.551	.740
Specificity	.960	.979	.974	.964

Table 3

Parameter Estimates of the Alcohol Use Disorder Items Based on the Partial Invariance Model

	Loadings	Thresholds		Unique Factor Variances[a]
		Younger	Older	Older
1. Great deal of time spent	1.00 (---)	1.01 (0.03)		1.70 (0.24)
2. Larger amounts than intended	0.84 (0.05)	2.19 (0.05)		1.41 (0.09)
3. Tolerance	0.74 (0.04)	0.76 (0.02)	1.18 (0.06)	1.76 (0.22)
4. Unsuccessful attempts to cut down	0.66 (0.04)	2.24 (0.05)		1.83 (0.10)
5. Continued use despite problems	1.21 (0.08)	2.57 (0.09)		1.63 (0.14)
6. Reduced or given up important activities	1.02 (0.06)	2.39 (0.06)		1.00 (---)
7. Withdrawal	0.82 (0.05)	2.47 (0.06)		1.00 (---)

Note: The younger group aged between 12-25 years, and the older group aged 26 years or above. Robust weighted least squares estimator (ESTIMATOR=WLSMV) and THETA parameterization were used. Numbers shown in parentheses were the standard errors.

[a]The unique factor variances were fixed to be 1.0 for the younger group.

Table 4

Diagnostic Accuracy Indices for the Real Data Example

	Strict Invariance		Partial Strict Invariance	
Age	12-25	26 or above	12-25	26 or above
Proportion diagnosed	.088	.054	.086	.055
Success Ratio	.636	.746	.660	.709
Sensitivity	.645	.716	.663	.698
Specificity	.965	.986	.968	.983

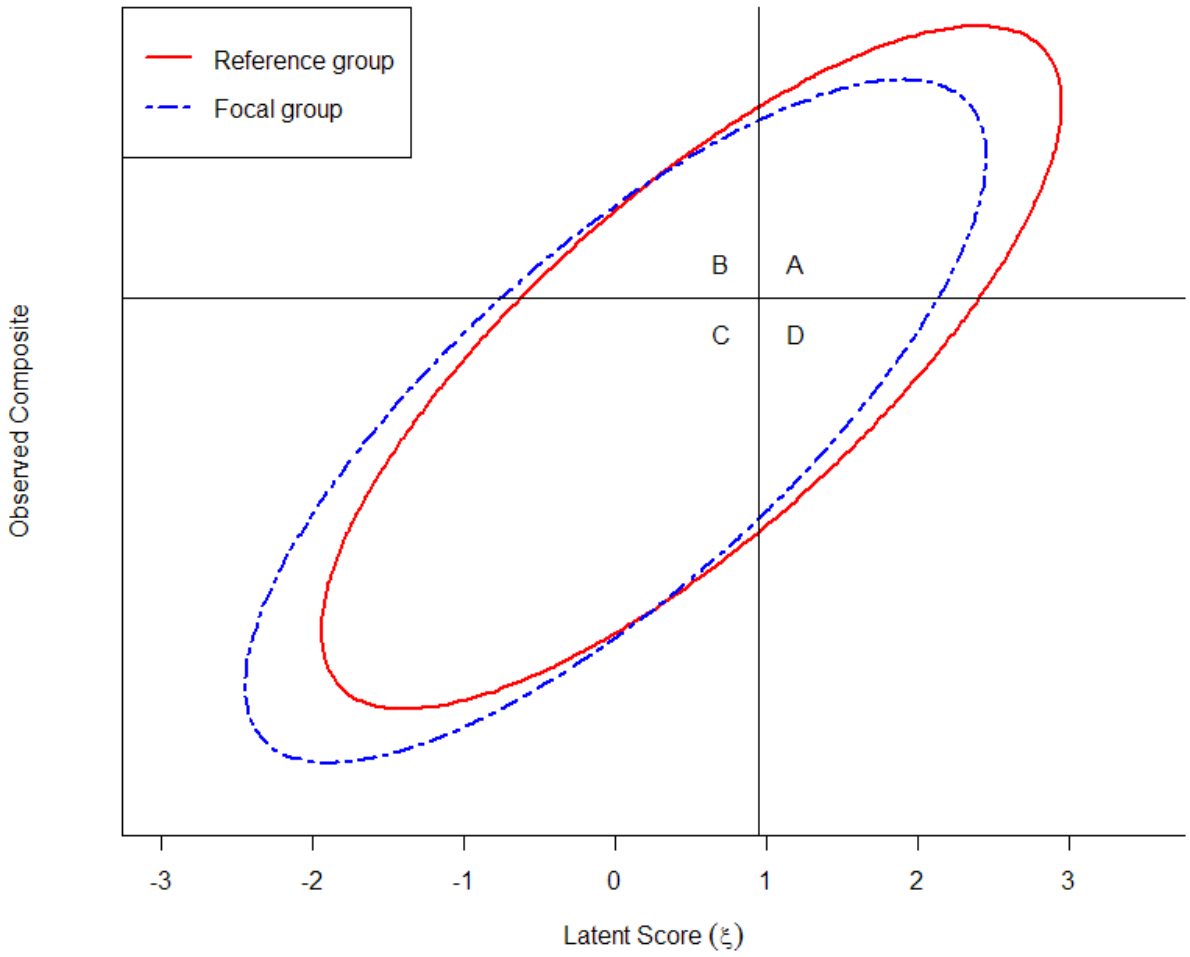

Figure 1. Recreation of Figure 2 from Millsap and Kwok (2004) showing the bivariate distribution of latent score and observed sum score with respect to two subpopulations. Note that the graph is used mainly for conceptual understanding; when the indicators were binary, the contour is not smooth and continuous.

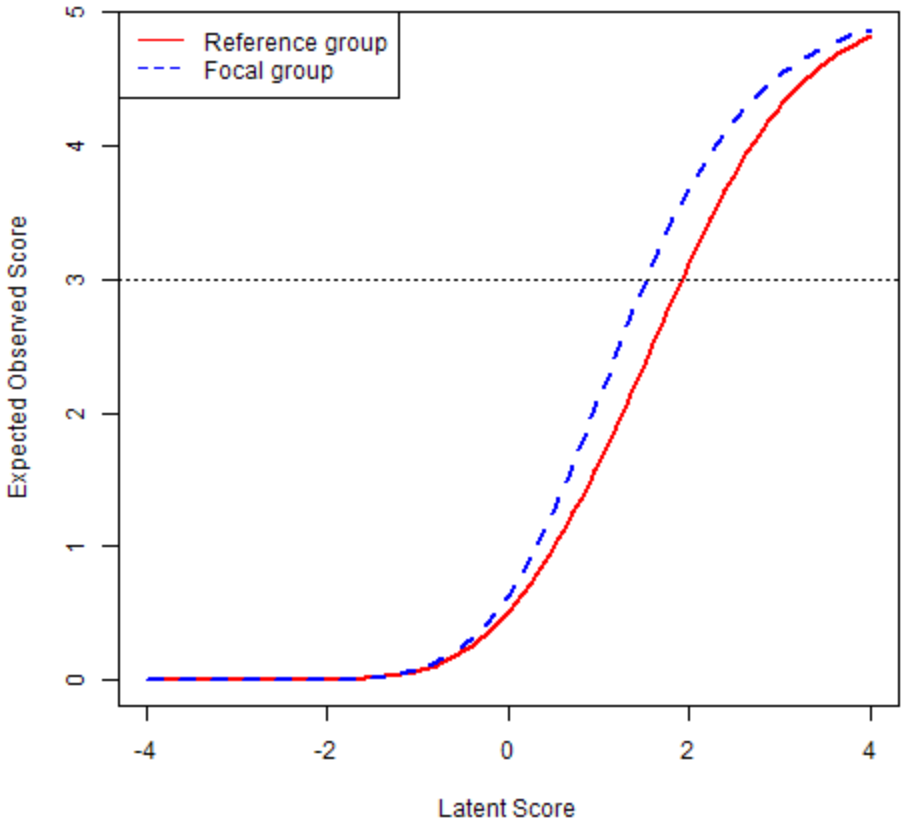

Figure 2. Expected number of endorsed items as a function of the latent factor score for the heuristic example.