Exploring the Impact of Deleting (or Retaining) a Biased Item on Classification Accuracy

Meltem Ozcan and Mark H. C. Lai

University of Southern California

Author Note

Meltem Ozcan https://orcid.org/https://orcid.org/0000-0002-5054-8264

Mark H. C. Lai https://orcid.org/https://orcid.org/0000-0002-9196-7406

This work was sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) and was accomplished under Grant #W911NF-20-1-0282. The views, opinions, and/or findings contained in this paper are those of the authors and shall not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents.

Correspondence concerning this article should be addressed to Mark H. C. Lai, Department of Psychology, University of Southern California, 3620 S McClintock Ave., Los Angeles, CA 90089-1061, United States. E-mail: hokchiol@usc.edu

Ozcan, M., & Lai, M. H. C. Exploring the Impact of Deleting (or Retaining) a Biased Item: A Procedure Based on Classification Accuracy. *Assessment*, Advance online publication, pp. 1-15. Copyright © 2024 (The Authors). DOI: 10.1177/10731911241298081.

Abstract

Psychological test scores are commonly used in high-stakes settings to classify individuals. While measurement invariance across groups is necessary for valid and meaningful inferences of group differences, full measurement invariance rarely holds in practice. The classification accuracy analysis framework (Lai & Zhang, 2022; Millsap & Kwok, 2004) aims to quantify the degree and practical impact of noninvariance. However, how to best navigate the next steps remains unclear, and methods devised to account for noninvariance at the group level may be insufficient when the goal is classification. Furthermore, deleting a biased item may improve fairness but negatively affect performance, and replacing the test can be costly. We propose item-level effect size indices that allow test users to make more informed decisions by quantifying the impact of deleting (or retaining) an item on test performance and fairness, provide an illustrative example, and introduce *unbiasr*, an R package implementing the proposed methods.

Keywords: measurement invariance, item bias, classification accuracy, fairness, R package

Exploring the Impact of Deleting (or Retaining) a Biased Item on Classification Accuracy

Psychological tests are commonly used for selection and classification purposes. Medical professionals, government agencies, licensing boards, and employers alike use tests to measure and make comparisons between individuals' relative standings on constructs of interest (e.g., depression, aptitude), which are often key components for high-stakes decisions such as diagnosis, personnel selection, placement, licensing, and school admission (Reynolds et al., 2021). In health care, psychological tests are used to screen and assess treatment eligibility for conditions including depression, substance abuse, and sleep disorders, and may determine which patient gains access to or is denied certain resources and medical services. For example, screening tests are administered during primary care visits or as part of community screening initiatives for the early detection and treatment of depression (Arias de la Torre et al., 2024), and can help clinicians efficiently identify the individuals at greater risk and prioritize these individuals for further assessment. Accurate identification of probable cases of depression via screenings leads to improved health outcomes, expedites treatment delivery, and facilitates optimal allocation of limited resources, while inaccurate decisions may result in heavier burdens on the healthcare system and delays in treatment (Arias de la Torre et al., 2024; US Preventive Services Task Force, 2023).

Test scores contain random and systematic errors, which means that there is a chance that medical conditions may be misdiagnosed, a deserving applicant may be denied admission, or an unqualified employee may receive a promotion. If there are systematic differences in error rates across groups such that individuals belonging to one group (characterized by, for instance, racial identity) disproportionately lose access to opportunities, situations of adverse impact (Biddle, 2006) may arise. Clearly, the validity and fairness of any test is integral to its value and utility as a decision-making tool.

Implicit in the use of tests in such high-stakes contexts is an assumption that the tests measure the same construct the same way regardless of group membership or other construct-irrelevant conditions. For instance, the gender, SES, or ethnicity of test takers should have no bearing on scores on a test measuring risk of developing depression. If two individuals have the same underlying true risk of depression, their propensity distribution (Lord et al., 1968) for the test should be the same. This idea of equivalence of measurement operations across groups and conditions is termed *measurement invariance* (MI; Drasgow, 1984; Mellenbergh, 1989; Meredith, 1993). MI is considered a prerequisite of valid inference and interpretation in scientific inquiries (Horn & McArdle, 1992). However, the rigorous criteria for MI are rarely met in practice. More commonly, test users establish partial measurement invariance (PMI; Byrne et al., 1989), which exists when only a subset of the items are measurement invariant. For a test used for classification in high-stakes settings, violations of MI at the test or item levels may harm the prospects of some individuals by reflecting group-level differences when none exist. Such spurious inferences may have grave consequences, from psychiatric conditions being misdiagnosed disproportionately for individuals from disadvantaged groups to delays in treatment and misallocation of limited resources.

Most existing literature on MI has focused on inferences at the group level, but not on classification, which is a major purpose of psychological tests. While one can model PMI to obtain valid group difference estimates, modeling PMI may not be a feasible solution when the goal is the classification of individuals as (a) scoring is usually based on unweighted sums (or weighted sums with the same weights across groups), which leads to bias with biased items, and (b) if using factor scores based on PMI, different scoring formulas are used for different populations, which compounds fairness concerns.

Thus, after discovering PMI, test users are tasked with finding the best course of action going forward, which often entails answering some crucial questions: is the impact of bias negligible enough that the biased items can be retained? If not, should the test be discarded entirely in favor of a measurement invariant test? Should biased items be deleted, and if so, which ones? What is the practical impact of removing a biased item: does the performance of the test improve, deteriorate, or remain unaffected if a specific item is removed? At which point is the improvement in test performance big enough to justify deleting an item? While research on the importance of and methods for establishing MI is abundant, methods and guidelines for navigating the next steps after the detection of biased items remain sparse in comparison, and the decision to retain or remove items, or discard the test in favor of another (if such an alternative exists) ultimately depends on the researchers' professional judgment, existing literature, and the application context (Hammack-Brown et al., 2021; Millsap & Kwok, 2004).

Furthermore, a focus on MI in the context of classification is warranted as MI is implicated in the quality and practical impact of the decisions made using test scores, which is not necessarily a consideration when the test purpose is to describe group differences in latent means. The current research aims to remedy these gaps by developing item-level effect size indices that quantify the impact of deleting (or retaining) an item on test performance. We advocate for an impact-oriented lens for evaluating MI, which brings test purpose to the forefront, and introduce methods and guidelines for exploring and mitigating the practical impact of measurement bias on classification decisions.

This paper is structured as follows. We first introduce MI and review previous work on how PMI impacts classification, which constitute the building blocks of the current research. Then, we introduce the item deletion operations h and Δh which are based on Cohen's h effect size (1988), describe the item deletion indices that allow test users to assess how item-level bias impacts metrics such as sensitivity and specificity, and provide an illustrative example of the methods and functions from the R package *unbiasr* using parameter estimates from a previous invariance study involving the Center for Epidemiological Studies Depression (CES-D) Scale (Radloff, 1977; Zhang et al., 2011). We conclude with a discussion of the results, guidelines of interpretation, and future directions. All accompanying code is available as part of the *unbiasr* package, and function calls and parameter values for the illustrative example can be found in the supplementary materials.

Measurement Invariance

Measurement invariance (MI) is achieved when latent construct(s) (e.g., cognitive functioning, depression) are measured equivalently and comparably across groups (e.g., ethnicity, SES), test modes (e.g., paper, computer), or time points (Drasgow, 1984; Mellenbergh, 1989; Somaraju et al., 2021). The focus on the relationship between a test and the latent construct it purports to measure sets MI apart from prediction invariance, which concerns the relationship between test scores and criterion performance (Cleary, 1968).

While there is no universally accepted definition of fairness, here we define fairness to encapsulate freedom of scores from the effects of construct-irrelevant characteristics, and equivalence in meaning across individuals and groups in line with standards set jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA, APA, & NCME; 2014).

MI facilitates valid and meaningful comparisons of test scores across groups or conditions by ruling out construct-irrelevant group level attributes as potential sources of observed group differences (Maassen et al., 2023; Meredith, 1993). Especially in high-stakes contexts where inaccurate decisions may have far-reaching negative consequences, it is vital that researchers and practitioners using tests determine if PMI is present, and if so, assess its practical impact on test outcomes and take steps to mitigate any adverse impact caused by measurement bias.

The growing interest in measurement invariance has furnished researchers with a wealth of tools and procedures for the detection of noninvariance, which have been discussed extensively elsewhere (Schmitt & Kuljanin, 2008; Somaraju et al., 2021; Vandenberg & Lance, 2000). Many of these operate within the confirmatory factor analysis paradigm (CFA; Jöreskog, 1969). Of particular interest to the present research is the selection accuracy analysis framework by Millsap and Kwok (2004), which evaluates the practical impact of measurement bias on classification outcomes by comparing selection accuracy indices under MI and PMI. This framework was initially developed for a unidimensional test with continuous items, and has since been extended to work with binary (Lai et al., 2019) and ordinal (Gonzalez & Pelham, 2021) items, and multidimensional tests with continuous items and varying weights (the multidimensional classification accuracy analysis or the MCAA; Lai and Zhang, 2022). A similar framework is the Adverse Impact (AI) ratio (Nye & Drasgow, 2011), or the Ratio of Selection Ratios Index (Stark et al., 2004), which is a ratio of observed and expected selection proportions at a particular cut-off score that helps identify which of the two groups, if any, would be under or over-selected due to bias. The AI ratio compares the observed score distribution for one group against the expected distribution of scores for this group if the groups were matched on the latent trait(s).

These methods, along with many other innovative developments in MI research that fall outside of the current scope, reflect an exponential growth in literature on the importance of and methods for establishing MI. However, the next steps after detecting MI have not received as much attention, especially in the context of classification decisions, and there is a critical need for methods and guidelines for mitigating the practical impact of bias on classification decisions.

The Common Factor Model

The common factor model (Thurstone, 1947) is a statistical model of the relationship between an unobserved (latent) construct (e.g., depression) and observed (manifest) variables (e.g., item responses on a depression screening test) such that an individual's true standing on the latent construct governs the probability of observed responses through a system of linear equations. The relationship between items and the latent construct(s) is characterized by the loading, intercept, and uniqueness parameters, which refer to the correlation between the item and the factor, the expected item responses when the latent score equals zero, and the construct-irrelevant variance of the sum of measurement error and systematic error assumed to be distributed independently with mean zero, respectively (Thurstone, 1947). Confirmatory Factor Analysis (CFA; Jöreskog, 1969) can be used to estimate and test the equivalence of the parameters of this system (see Appendix A for a more comprehensive overview and technical details). If estimates are identical across groups, the test is factorially invariant (Byrne et al., 1989).

Factorial invariance (FI) has been shown to be equivalent to MI under the common factor model (Horn & McArdle, 1992; Thurstone, 1947); under MI, response probabilities of individuals with the same latent standing are expected to be invariant across groups. Depending on which parameters are the same across groups, the level of FI can be classified as, from the least to most stringent, configural, metric, scalar, and strict (Byrne et al., 2007; Horn & McArdle, 1992; Meredith, 1993). Configural invariance requires that the configuration of items and factors (the factor structure) is the same across groups. All measurement parameters are freely estimated under configural invariance. Metric invariance holds if, additionally, unstandardized factor loadings are equal across groups. If measurement intercepts are also the same across groups, it can be said that scalar invariance holds. Finally, strict factorial invariance (SFI) exists when measurement intercepts, factor loadings, and unique factor variance-covariances (i.e., uniqueness) are equal across groups or conditions, and is the most stringent level of invariance. More often, partial factorial invariance (PFI, Byrne et al., 1989) is met, meaning that invariance holds only for a subset of the items. Under the common factor model, MI is satisfied when SFI holds, and PMI is equivalent to PFI.

The Classification Accuracy Analysis Framework

Consider an example where the 20-item Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977) is used as an initial screener for risk of depression. Letting η denote an individual's true risk of depression, and Z denote observed scores on CES-D items, we can aggregate observed scores on the CES-D into a composite using some scoring rule, and classify individuals as at risk or not at risk based on a cut-off score Z_c (e.g., 16 points; Radloff, 1977)[1].

Given the probabilistic nature of inferences based on psychological tests (Borsboom, Romeijn, & Wicherts, 2008), these classifications are error-prone. The relationship between observed scale sums Z and theoretical factor scores η can be represented as a bivariate normal distribution and visualized as an ellipse, as in Figure 1. The latent and observed thresholds divide up the area of this ellipse into four quadrants, and depending on which quadrant a decision falls, it may be qualified as true positive (TP), true negative (TN), false positive (FP), and false negative (FN). For example, an individual who screened positive on the CES-D and who is truly at risk of depression ($Z > Z_c$ and $\eta > \eta_c$) is denoted a TP. Conversely, an individual who screened positive on the CES-D but is not at risk of depression ($Z > Z_c$ and $\eta < \eta_c$) reflects a FP. An individual who screened negative who is truly not at risk of depression is denoted a TN ($Z < Z_c$ and $\eta < \eta_c$) and an individual who is screened out but is truly at risk is denoted a FN ($Z < Z_c$ and $\eta > \eta_c$).

The proportion of decisions in each category (i.e., TP, FP, TN, and FN) may then be used to

[1] In other contexts, classifications may be made using a percentile (e.g., applicants performing in the top 10% on an entrance exam may be identified as the candidate pool).

compute summary classification accuracy indices[2] (CAI): proportion selected (PS), success ratio (SR), sensitivity (SE), and specificity (SP; Millsap & Kwok, 2004). Proportion selected,

$$PS = P(TP) + P(FP), \tag{1}$$

refers to the ratio of individuals who screened positive over the number of individuals assessed. Success ratio,

$$SR = P(TP) / (P[TP] + P[FP]), \tag{2}$$

(also termed *positive predictive value* or the *precision* of a test; Mohan et al., 2021) indicates the proportion of positive screens who are truly at risk of depression. Sensitivity,

$$SE = P(TP) / (P[TP] + P[FN]), \tag{3}$$

is also known as *true positive rate, hit rate*, or *recall* (Mohan et al., 2021), and refers to the success of the test in capturing individuals who meet the criteria: out of all the individuals who should be identified as at risk, how many of them actually screened positive? Finally, specificity,

$$SP = P(TN) / (P[TN] + P[FP]), \tag{4}$$

(*selectivity* or *true negative rate*), corresponds to the ability of the test in screening out the individuals who should have been excluded.

Under the simplifying assumption that individuals belong to one of two distinct populations (termed the *focal* and *reference* groups, where the reference group often corresponds to the majority group), the classification accuracy analysis framework entails the computation and comparison of CAI for the reference and focal groups under MI versus PMI to better understand the extent and practical impact of bias on test performance. If the negative impact of noninvariant items is deemed large enough by the test user, Millsap and Kwok (2004) suggest solutions such as removing noninvariant items or using a different test, and state that such decisions should be made with the usage of the test and the cost of each type of misclassification in mind. For instance, FPs and therefore SR and SP might be of greater concern if the test will be used to give

[2] These indices were originally termed *selection accuracy indices* in Millsap and Kwok (2004). We opted for *classification accuracy indices* to encompass a wider range of scenarios.

access to limited and costly resources (Millsap & Kwok, 2004).

When MI holds and the latent distributions are equal, we expect equal TP, FP, FN, and TN proportions for the reference and focal groups. However, proportions may be drastically different across groups under PMI (see Figure 1). Further, if the latent distributions are not equal across groups, it is not possible to compare the indices across groups even under MI. In order to address this concern, an additional set of indices termed ‘expected focal’ (Efocal) can be computed as the proportions we would expect to observe for the focal group if its latent distribution matched that of the reference group. One index of note based on this idea is the Adverse Impact (AI) ratio (Nye & Drasgow, 2011; Stark et al., 2004), which refers to the ratio of the expected proportion selected for the focal group and the observed proportion selected for the reference group. The AI ratio was developed to quantify the impact of differential item functioning on selection outcomes, and can be computed within Millsap and Kwok’s (2004) original framework.

The main idea behind the AI ratio is that if the latent trait level is equal across groups, the proportions of individuals scoring above the threshold should be equal in each group, which allows us to attribute any differences between selection proportions to measurement bias. Conditioning on the latent trait level η and using the group means and standard deviations from the two groups with the reference group’s ability density function means that any differences captured between the expected proportion selected in the focal group ($P_{Ef}\,[Z_f > Z_c]$; i.e., if the focal group has the same distribution of depression risk as the reference group) and the observed proportion selected (i.e., the proportion who screened positive) in the reference group $P_r\,(Z_r > Z_c)$ are not related to the construct being measured (see Appendix A for additional details).

The AI ratio is defined as

$$AI\,ratio = \frac{P_{Ef}\left(Z_f \geq Z_c\right)}{P_r\left(Z_r \geq Z_c\right)} \tag{5}$$

(Nye & Drasgow, 2011; Stark et al., 2004) which we express as

$$AI = PS_{Ef} / PS_r \tag{6}$$

where PS_r denotes PS for the reference group, and PS_{Ef} denotes the expected PS for the focal group if both groups were matched to have the latent score distribution of the reference group. If SFI holds, the expected PS for the focal group will be equal to the PS for the reference group; hence, the AI ratio will equal 1. Deviations from 1 indicate the presence of measurement bias. A commonly used rule is the 'four-fifths' rule, which suggests that the focal group has suffered adverse impact if the AI ratio falls below 0.80 (Biddle, 2006; Nye & Drasgow, 2011). In an adverse impact situation, the item with the removal of which brings the AI ratio the closest to 1 would be our candidate for deletion.

The Multidimensional Classification Accuracy Analysis Framework

Noting that selection and classification decisions are rarely based on psychological tests measuring a single, unidimensional latent construct, and that different weights may be assigned to different dimensions in practice, Lai and Zhang (2022) expanded the selection accuracy analysis framework (Millsap & Kwok, 2004) to work with tests aimed to measure multiple latent constructs with different weights. Assuming the multivariate normality of the latent factor scores and the unique factor variables, the observed composite scores Z_g and the latent composite factor scores η_g (where the latent composite is a weighted combination of the latent dimensions and *g* denotes group membership) were shown to follow a bivariate normal distribution (see Appendix A; Lai & Zhang, 2022). Furthermore, the marginal distribution of (Z, η) was demonstrated to be a finite mixture of bivariate normal distributions with mixing proportion π_g, and the latent composite cut-off η_c can be computed as the quantile in the mixture corresponding to PS_{total} (Lai & Zhang, 2022; Millsap & Kwok, 2004). The researcher may choose to pre-specify PS_{total} (e.g., to select the top X% of candidates) or specify a cut-off Z_c (e.g., in a diagnostic screening setting), which will then be used to compute the proportion of individuals selected using the cut-off.

While this framework help test users to link measurement noninvariance to the practical impact on classification, it does not provide clear methods for or guidance on how test accuracy and fairness may be improved, for example, by dropping biased items. Our goal is to remedy this gap by providing test users with item deletion indices that allow for the assessment of improvements (or decreases) in test accuracy and fairness when a biased item is dropped.

Methods: Item Deletion

The methods discussed here concern the case where a psychological test used for classification decisions contains measurement bias, and the researcher aims to investigate which of the test items, if any, may be deleted to reduce the negative impact of this bias on the performance and fairness of the scale. The deletion of an item may not be necessary or beneficial in some scenarios, and may in fact harm the validity and reliability of the test as will be discussed later. The methods outlined here are provided to facilitate researchers' exploration of their data and to lead to more informed decisions about deleting or retaining an item.

The test instrument can consist of a single factor (e.g., depressive affect) or multiple factors (e.g., a scale of depression measuring different facets of depression such as positive affect, negative affect, and somatic symptoms). In this paper and in the accompanying *unbiasr* package, deletion is considered in a step-wise manner such that no more than one item is to be dropped at one time. Unless otherwise indicated by a subscript (e.g., SE_{sfi}), we assume that CAI are computed under PFI. The current method assumes that each item loads onto a single factor (i.e., no cross-loadings).

We can examine the impact of dropping an item on the difference in CAI from three distinct but complementary angles. The first approach entails an examination of an overall measure of classification accuracy, termed *aggregate* CAI ($\overline{CAI}$), which is a weighted average of CAI across the reference and focal groups. The second approach consists of a comparison of the AI ratio computed using the full item set (AI) with the AI ratio computed using an item set excluding the *j*-th item ($AI^{\setminus j}$). The third approach entails a comparison of CAI for the reference group (CAI_r) and the *expected* CAI for focal group (CAI_{Ef}) for a given set of items.

We now introduce h and Δh, operations used to compute item deletion indices that quantify differences in CAI and $\overline{CAI}$.

Operation: Cohen's *h* (Cohen, 1988)

Cohen's h (1988) is an effect size measure of the difference in two proportions or probabilities that was designed to account for the fact that probabilities can only range from 0 to 1, and uses the arcsine transformation so that the values better resemble an interval scale. Cohen's *h* effect size (Cohen, 1988) of the difference between proportions p_1 and p_2 is defined as

$$h = 2 \arcsin(\sqrt{p_1}) - 2 \arcsin(\sqrt{p_2}). \quad (7)$$

Resulting *h* values can be interpreted as indicators of small, medium, or large differences between proportions using the conventionally used benchmarks of 0.2, 0.5, and 0.8 (Cohen, 1988). For example, if p_1 = .65 and p_2 = .50, we have *h*(.65, .50) = 0.30, which corresponds to a small-medium effect size, and *h*(.95, .80) = 0.48.

Operation: Delta *h* (Δ*h*)

The change in the effect size *h* when a noninvariant item *j* is deleted is also of interest. Using * as a placeholder for the comparison *h* was computed for, the operation Δ*h* is defined as

$$\Delta h^{|j}\mathrm{CAI} = |\,\mathrm{h}^{*}\mathrm{CAI}\,| - |\,\mathrm{h}^{*}\mathrm{CAI}^{|j}\,|. \quad (8)$$

Delta *h* can be used to quantify the change in the difference between *h* values comparing CAI across groups or invariance conditions when the *j*-th item is dropped. As an example, consider a scenario where we are interested in the change in the effect size *h* associated with the difference between SE_r and the SE_{Ef} computed when item 2 is deleted (SE^{2}_{Ef}). First, the effect size *h* for the difference between SE_r versus SE_{Ef} (using the full item set) is computed:

$$h^{\text{r-Ef}}\mathrm{SE} = 2\arcsin(\sqrt{SE_r}) - 2\arcsin(\sqrt{SE_{Ef}}).$$

Second, *h* for the difference between SE^{2}_{r} versus SE^{2}_{Ef}) (on the item set excluding item 2) is computed:

$$h^{\text{r-Ef}}SE^{2} = 2\arcsin(\sqrt{SE^{2}_{r}}) - 2\arcsin(\sqrt{SE^{2}_{Ef}}).$$

Finally, these values are compared using

$$\Delta h^{|2}\mathrm{SE} = |h^{\text{r-Ef}}\mathrm{SE}| - |h^{\text{r-Ef}}\mathrm{SE}^{|2}|.$$

Note that Delta *h* is only computed on *h* values, in contrast to Cohen's *h* which can be computed for 'raw' proportions. Having defined these two operations, we now describe the first three approaches to item deletion in more detail.

Approach 1: Examining Changes in Aggregate Classification Accuracy Indices ($\overline{CAI}$)

We obtain aggregate classification accuracy indices as weighted averages across groups (see Appendix B for computation details) and $\overline{CAI}^{|j}$ (aggregate classification accuracy indices when a potentially biased item *j* is deleted). Comparing $\overline{CAI}$ and $\overline{CAI}^{|j}$ and examining the effect size *h* of any discrepancy helps us determine the impact of deleting a biased item. Increases in $\overline{CAI}$ when an item is deleted may point to one of the following scenarios: CAI may have increased for both groups, or CAI may have increased for one group but stayed constant or decreased for the other[3].

We suggest that the item j leading to the largest increase in $\overline{CAI}$ and resulting in negative $h^j\,\overline{CAI}$ when deleted may be considered a candidate for deletion. If $h^j\,\overline{CAI}$ is positive, deleting item *j* would lead to a decrease in CAI so researchers should be careful with deletion when $h^j\,\overline{CAI}$ is large.

Approach 2: Examining the AI Ratio

We then compare the AI ratio computed using the full item set (*AI*) to the one computed using the item set excluding biased item j ($AI^{|j}$). If the deletion of j does not lead to an $AI^{|j}$ closer to 1 than *AI* for any *j*, or leads to an AI ratio that is lower than the one computed using the full item set, all items should be retained as the deletion of items has no impact or leads to more adverse impact. If, on the other hand, the deletion of item j brings the AI ratio closer to 1, the discrepancy between PS_r and PS_{Ef} has decreased, signaling an improvement. If in fact $AI^{|j} = 1$, we can say that the difference between the groups in PS that is due to measurement bias is eliminated as the deletion of item j achieves a PS_f that is equivalent to that of the PS_r if these two groups were matched on their latent trait level. If there are multiple items the deletion of which lead to an improvement, the researcher is advised to consider the deletion of the item that

[3] We may consider an increase in $\overline{CAI}$ when the *j*-th item is deleted such that $\overline{CAI} < \overline{CAI}^{j}$ an overall improvement in all cases except when the increase in $\overline{CAI}$ is driven by improvements in CAI_r being given greater weight in computation due to a larger π_r that masks decreases in CAI_f. This case concerns the scenario where the removal of the item actually leads to greater discrepancy. Note that if there is an imbalance between CAI_r and CAI_f such that CAI is higher for one group than the other, $\overline{CAI}$ will take a value between CAI_r and CAI_f that is closer to CAI_r if $\pi_r > 0.5$, and equal to the midpoint between CAI_r and CAI_f if $\pi_r = 0.5$.

brings the AI ratio the closest to 1. If multiple items lead to a similar improvement in the AI ratio if deleted, the researcher may continue their exploration of the other indices and make a judgment call as to which item, if any, should be deleted.

Approach 3: Examining Differences in CAI for Reference and Efocal Group

Comparisons can then be made between the observed scores for the reference group (CAI_r) and the scores we would expect to see for the focal group if the focal group followed the same distribution as the reference group (CAI_{Ef}) by conditioning on the matched latent trait. Unlike the AI ratio, which focuses solely on PS, this approach allows the researcher to quantify discrepancies between SE_r, SR_r, and SP_r, and SE_{Ef}, SR_{Ef}, and SP_{Ef}, and to interpret any observed difference between CAI_r and CAI_{Ef} as being truly due to measurement bias, that is, as a difference that is not due to true group-level differences in the trait being measured.

After computing Cohen's *h* values for the difference between CAI_r and CAI_{Ef} for the full item set ($h^{\text{r-Ef}}\text{CAI}$) and an item set excluding biased item *j* ($\text{h}^{\text{h-Ef}}CAI^{j}$, the change in this difference can be computed using equation (8). Item *j* that leads to the smallest $|h^{\text{r-Ef}}\text{CAI}|$ and the largest $\Delta h^{|j}\text{CAI}$ introduces the most bias and its deletion has the largest effect size may be considered the candidate for deletion. In contrast, items that lead to a larger $|h^{\text{r-Ef}}\text{CAI}|$ or result in an insubstantial improvement (as indicated by a very small $\Delta h^{|j}\text{CAI}$) should be retained.

The three approaches are intended to be examined in conjunction, and test users are advised to compare and contrast results from each approach before making a final decision about item deletion. If there is unanimity across the approaches supporting the deletion of an item (and assuming that its deletion does not have a major impact on the conceptual breadth of the test), the item may be dropped. If the three approaches agree, but the improvement as indicated by the indices is minimal, the test user may opt to retain the item in order to preserve the statistical properties and construct coverage of the scale. If there is disagreement between the approaches such that, for example, one approach indicates an improvement and one approach indicates a decrease in accuracy and fairness if an item is deleted, the user is advised to proceed with caution and examine raw classification accuracy indices. We suggest that items should be retained unless there is clear indication that the deletion of an item would lead to a concrete improvement in and would not harm accuracy and fairness.

Illustrative Example

We now illustrate the use and interpretation of the item-level deletion indices in a diagnostic application context using CFA estimates from a previous study investigating the measurement invariance of the CES-D (Radloff, 1977) across Chinese and Dutch elderly populations (Zhang et al., 2011). CES-D is made up of 20 items and four factors: positive affect (*good, hopeful, happy, enjoyed*), depressive affect (*blues, depressed, failure, fearful, lonely, crying, sad*), somatic complaints (*bothered, appetite, mind, effort, sleep, talk, get going*), and interpersonal problems (*unfriendly, dislike*). Participants are asked to rate each item on a scale of 0 to 3 based on how they felt in the past week. The maximum score is 60 on the full scale.

In their examination of data collected from 4903 elderly adults from China and 1903 elderly adults from the Netherlands, Zhang and colleagues (2011) found that configural and metric invariance held, and demonstrated partial scalar and partial strict invariance such that while the same construct was being measured across groups, there were differences in intercepts (*failure*, *good*) and uniqueness (*depressed*, *fearful*, and *dislike*). Depending on the size and direction of these differences, more individuals from the Chinese elderly (reference) group may be flagged for depression, resulting in a potential waste of valuable and limited resources. Likewise, fewer individuals from the Dutch elderly (focal) group who are truly at risk for depression may screen positive, which may mean that their treatment is delayed, or they lose access to resources or interventions. These observed differences may also be mistaken for true group-level differences, leading to spurious conclusions in theory building which may have unforeseeable downstream consequences. Taking informed steps to delete the item introducing the most bias to the scale may allow practitioners and researchers mitigate unfair disadvantages caused by measurement bias.

We demonstrate the item deletion framework assuming that the CES-D scale is used as an initial screener for risk of depression, where selected individuals would be further assessed by a clinician who may leverage multiple additional information sources (e.g., a diagnostic interview) to determine whether the individual qualifies for some treatment or intervention program for depression. We use unstandardized factor loading, uniqueness, intercept, factor mean and factor standard deviation estimates from Zhang et al. (2011) and a latent factor variance-covariance matrix computed using factor correlation estimates from a previous study by Miller et al. (1997) as our input parameters[4,5].

We use a cut-off score of 16 on the full CES-D scale following the example of Radloff (1977), and hold the proportions selected using the full set of items constant in the item deletion scenarios considered. Note that researchers can instead choose to provide a new post-deletion cut-off to be used.

The mixing proportion π_r is set to 4903/(1903 + 4903) ≈ 0.72. As the depressive affect and somatic affect factors have 7 items each, the lack of positive affect factor has 4 items, and the interpersonal problems factor has 2 items, this allocation of weights results in 35%, 35%, 20% and 10% weighting for the aforementioned latent dimensions. All relevant parameter values, function calls, and outputs can be found in the code excerpts included in the supplementary materials[6].

Item deletion on the 20-item, four-factor CES-D scale

Under partial factorial invariance, PS = 0.457 of the Chinese elderly group and PS = 0.144 of the Dutch elderly group scored above the cut-off score of Z_c = 16, which corresponds to an aggregated $\overline{PS}$ of

---

[4] The factor correlation estimates from Miller et al. (1997) were used as a proxy as estimates for the latent factor variance-covariance matrix or factor correlations were not provided in Zhang et al. (2011). As items in the positive affect subscale were reversed in Zhang et al. (2011) to achieve a 'lack of positive affect' interpretation, we reversed the signs of Miller et al.'s (1997) correlation estimates in our computations of the variance-covariance matrix.

[5] The parameters reported in Zhang et al. (2011) were obtained via maximum likelihood (ML) while Miller et al. (1997) used the asymptotically distribution free weighted least squares (WLS) estimator. The assumptions of continuous, normally distributed data for the ML estimator are unlikely to hold in the case of CES-D, which has four response options. It is recommended that ordinal methods are employed when dealing with data with less than five response options (Rhemtulla et al., 2012), and ordinal data should be handled differently than continuous data while testing for MI (Wu & Estabrook, 2016). We use parameter estimates for the CES-D scale for illustrative purposes only, and the item deletion methods were developed to facilitate researchers' exploration of the impact of item-level bias after fitting their model.

[6] An extension to the illustrative example in which the analyses are repeated for each of the four subscales of CES-D, assuming for the purposes of illustration that the subscales will be used independently to select individuals can be found in the supplementary materials.

0.387. PS_r and PS_f are held constant to achieve an aggregate PS of $\overline{PS}$ = 0.387 across item deletion scenarios.

Items 4, 9, 10, 11, 15, and 20 (*effort, depressed, failure, fearful, good, dislike*) were identified as biased in Zhang (2011). Table 1 illustrates the $\overline{CAI}$ and Cohen's h values associated with the deletion of each of these biased items. The h values here range between 0.003 and 0.013. The higher $\overline{SP}$ = 0.934 compared to $\overline{SR} = \overline{SE}$ = 0.887 suggests that, overall, the scale performs somewhat better at not selecting individuals who are not at risk for depression. The effect size of removing any of the biased items on $\overline{CAI}$ is quite low, as seen in the Cohen's h values provided in Table 1, and no item's removal leads to an improvement in $\overline{CAI}$~~)~~.

Table 2 contains the item deletion indices quantifying the discrepancy between CAI_r and CAI_{Ef}. The AI ratio for the full scale is AI = 0.908, which is greater than the 80% threshold for adverse impact. The deletion of item 9 (*depressed*) or 11 (*fearful*) leads to AI ratios further away from the optimal ratio of 1, whereas deleting item 4 or (*effort*) or 20 (*dislike*) leads to no change in the AI ratio. The greatest improvement in the AI ratio is observed for the removal of item 15 (*good*), followed by item 10 (*failure*) as the removal of either item brings the AI ratio closer to 1: $AI^{|15}$ = 0.977 and $AI^{|10}$ = 0.930. The rightmost three columns of Table 2 illustrate the effect size of the discrepancies between CAI_r and CAI_{Ef} attributable to measurement bias. $h^{r\text{-}Ef}SR = -0.157$ and $h^{r\text{-}Ef}SP = -0.164$ on the full CES-D scale suggests higher SR and SP values for the focal group (Dutch elderly) had the focal group been matched with the reference group (Chinese elderly) on the latent traits. Similarly, $h^{r\text{-}Ef}SE = 0.141$ suggests a lower SE for the expected focal group. Not only does the deletion of item 15 (*good*) attenuate the discrepancy between CAI_r and CAI_{Ef}, bringing the h values closer to 0 ($h^{|15}SR = -0.047$, $h^{|15}SE = 0.026$, $h^{|15}SP = -0.048$), improvements caused by the deletion of this item also have the largest effect sizes out of all item deletion scenarios ($\Delta h^{|15}SR = 0.110$, $\Delta h^{|15}SE = 0.116$, $\Delta h^{|15}SP = 0.116$).

In light of these findings, and barring any domain specific reasons to retain this item, we can conclude that deleting item 15 would help mitigate the impact of measurement bias on classification accuracy and fairness, and render the diagnostic accuracy of the CES-D for the Chinese and Dutch elderly groups more comparable (see the supplementary materials for an illustration of the distributions of latent and observed scores for depression for the Dutch and Chinese elderly groups before and after the deletion of item 15).

We would like to emphasize that the suggestion to delete item 15 (*good*) only applies to the findings discussed here by Zhang et al. (2011) regarding the comparison between Chinese and Dutch elderly individuals, and does not necessarily generalize to item or test performance in other contexts. For instance, if a clinician is adapting the Chinese version of the CES-D to screen depression risk for their clients, we would recommend repeating the analyses here with their data and carefully examining the performance of all items including item 15 before proceeding with any deletion.

Four-factor CES-D Scale after deleting item 15

Continuing with the diagnostic example, we perform item deletion on the remaining 19 items of the CES-D to see whether the deletion of a second item may further reduce the impact of measurement bias. The cut-off score is recomputed as $Z_c = 16/60 \times (60-3) = 15.2$ to account for the deleted item. Results are illustrated in Tables 3 and 4.

In Table 3, we see that the deletion of any of the remaining biased items (4, 9, 10, 11, 20, or effort, depressed, failure, fearful, dislike) leads to decreases from $\overline{SR} = 0.886$, $\overline{SE} = 0.886$ and $\overline{SP} = 0.925$, harming overall classification accuracy.

In the first column of Table 4, only item 10 (*failure*) leads to an AI ratio that is closer to 1 if deleted, with $AI^{|10} = 0.999$ from $AI = 0.979$[7]. In the next three columns of Table 4, we see that deleting item 10

[7] Note that the AI ratio is 0.979 on the 19-item CES-D scale, which is slightly different than the previously reported delete-one AI ratio of $AI^{|15} = 0.977$ (see Table 2). Any such difference in the row labeled '|15' in Table 2 and in the row labeled Full in Table 4 is due to the difference in providing a cut-off value versus a proportion to be selected for the computations. In the computations for the AI value that would be achieved by deleting item 15, the proportions selected using the provided cut-off score on the 20-item scale were held constant in the 19-item scenarios. As such, $AI^{|15} = 0.979$ was achieved using a proportion of selection. On the other hand, once we dropped item 15 and repeated our computations to consider the deletion of a second item, $AI = 0.999$ was computed based on the cut-off score for the 19-item scale. Accordingly, the delete-one statistics reported for the 19-item scale (i.e., for an 18-item subset) were computed based on the proportions selected when using the full 19-item scale.

may also slightly reduce the discrepancy between CAI_r and CAI_{Ef}, with $h^{\text{r-Ef}}\text{SR}^{|10} = -0.011$, $h^{\text{r-Ef}}\text{SE}^{|10} = -0.007$, and $h^{\text{r-Ef}}\text{SP}^{|10} = -0.011$. The effect sizes of these changes in discrepancy between CAI_r and CAI_{Ef} are $\Delta h^{|10}\text{SR} = 0.032$, $\Delta h^{|10}\text{SE} = 0.017$, and $\Delta h^{|10}\text{SP} = 0.036$. The deletion of any other item either leads to an insubstantial improvement, or exacerbates the discrepancy between CAI_r and expected CAI_f, increasing bias.

While these results show that item 10 introduces the most bias after item 15, the potential improvement achieved from dropping this item is not as clear-cut as that from the deletion of item 15. Given the ambiguity of these results and the lower magnitude of the improvements in AI and $h^{\text{r-Ef}}$ CAI compared to when item 15 was the candidate for deletion, we would recommend retaining item 10 and proceeding with the 19-item CES-D scale unless further, theory-based justification supporting the deletion of item 10 is established. It may be worthwhile examining the raw classification accuracy indices as, depending on the application context, whether an increase in SR is caused by a decrease in FP or an increase in TP may give additional insight into the best course of action if the scale will be used for allocating limited resources such as access to a treatment program. We believe that the methods and guidelines outlined here equip test users to make more informed decisions about whether improvements in AI and $h^{\text{r-Ef}}$ CAI are large enough to warrant item deletion.

Implementation using R package *unbiasr*

The R package *unbiasr* implements the item deletion methods proposed in the current paper. The main function in *unbiasr* is *PartInv()*, which allows users to evaluate the practical impact of classification accuracy across groups and requires only the CFA parameter estimates as input. *item_deletion_h()* computes effect size indices quantifying the impact of deleting biased item(s) on classification accuracy indices. *unbiasr* incorporates the R scripts from Lai et al. (2017) and Lai and Zhang (2022).

First, CAI are computed under SFI and PFI for the full set of items using the user-specified item weights. Then, summary statistics are computed for the item set excluding item *j* using an adjusted item weight vector where an item weight of zero is assigned to the *j*-th item. In the calculation of the new item weights, the weight that had been allocated to the j-th item is redistributed across the remaining test items proportionally to the current weights of these items. If the test is multi-dimensional, the weighting is redistributed only across the items that belong to the same subscale as item j. Once relevant delete-one classification accuracy indices are computed for the reference, focal, and expected focal groups under strict and partial factorial invariance, operations h and Delta h are used to compute the deletion indices ($\overline{CAI}$, $h^{|j}\,\overline{CAI}$, $AI^{|j}$, $h^{\text{r-Ef}}\text{CAI}$, $\Delta h^{|j}\text{CAI}$).

Depending on the purpose and application context of the test, users may indicate a cut-off score (Z_c; e.g., to identify patients scoring above a clinically meaningful cut-off for treatment referral), or input a proportion for selection (*propsel*; e.g., to hire the candidates scoring in the top 10% of the applicant pool). If the user specifies a cut-off Z_c as well as a delete-one cut-off score adjusting for the decrease in the maximum total score when an item is dropped from the scale, the second cut-off score is used as the new Z_c in item deletion scenarios. If the user specifies a proportion for selection, this value is held constant in item deletion scenarios. If a delete-one cut-off score is not provided by the user, the PS_{sfi} and PS_{pfi} using Z_c on the full item set are held constant in the computations of CAI in item deletion scenarios. For example, if Z_c = 16 on the full scale corresponds to PS_{sfi} = 0.30 and PS_{pfi} = 0.28, summary statistics will be computed with *propsel* = 0.30 and *propsel* = 0.28 so that the highest scoring 30% and 28% of individuals in each item deletion scenario will be selected under strict and partial invariance conditions respectively.

Discussion

Psychological tests provide decision-making bodies and scientists alike with a relatively time-efficient and objective tool for the assessment and comparison of individuals' relative standings on constructs of interest and are used in a range of applications from theory construction and advancement to decision-making. As such tests are commonly used in high-stakes contexts and may have wide-reaching consequences beyond the immediate application of the test, it is critical that test scores are valid and free of bias. A notion inextricably linked to validity and bias is measurement invariance, which holds when a test measures the same construct in the same way across grouping variables that are irrelevant to the construct under study (e.g., race). The current framework provides decision-makers with tools and guidelines to better navigate the seldom-discussed next-steps following the discovery of noninvariant items.

We have fully automated the three complementary approaches to item deletion outlined in this paper, and the functions for the computation of item deletion indices are available in our open-source R package *unbiasr*. The outlined methods expedite and give structure to the otherwise laborious and error-prone process of determining the best course of action to handle item bias by converting differences in classification accuracy indices to comparable and easily interpretable units. As such, test users can make more informed decisions about item deletion (or retention) more efficiently, prevent the misallocation of limited resources, expedite the time it takes for patients to receive the care they need, and reduce the influence of construct-irrelevant factors on classification decisions, promoting fairness. We hope that the detailed examination and discussion of the item deletion indices in the illustrative example helps elucidate the process of determining whether a biased item can, or should, be deleted to improve accuracy and fairness in classification decisions.

There are a number of limitations to the current work. First, the methods outlined here concern binary classification decisions, such as selection versus rejection or diagnosis versus no diagnosis. Future work is planned to extend to classification into multiple categories (e.g., classification of an individual's depression level into severity categories; class placement of students based on levels of language proficiency). Second, we only considered noninvariance across two groups, whereas many demographic characteristics have multiple subgroups (e.g., ethnicity, race, SES). We hope to extend the framework to the classification of individuals across multiple groups. Third, we assumed that the test items were measured on an interval scale. We have proposed and illustrated the current framework in the context of interval level data[8], but we plan to extend the framework to ordered categorical data in future research. Moreover, the current methods do not quantify the uncertainty around the estimates. Additional tasks for our package

[8] Note that the CES-D items are measured on a 0-3 scale and would ideally be treated as ordinal. The illustrative example assumed interval level data for the sake of simplicity.

therefore include extending the current methods to performing item deletion for multiple groups as well as for when test items may be measured on a binary or ordinal scale, and computing uncertainty estimates (e.g., Bayesian credible intervals).

Any item deletion decision should be made with the context and application of the test in mind (Millsap & Kwok, 2004), as one potential consequence of deleting items is reduced construct coverage (Kruyen et al., 2013). While the deletion of an item may lead to better classification accuracy and increased fairness, the item may nevertheless be important to retain, particularly in application contexts where inference and interpretability take precedence over prediction. It may be more important in a research context to get a holistic picture that taps into all facets of the construct for theory-building purposes as opposed to in more applied contexts where the goal is to make a decision[9]. For example, imagine the item *loss of interest and pleasure*, which measures an aspect of depression that is integral to the construct definition of depression, is found noninvariant across groups and that the deletion of this item leads to better classification accuracy and higher fairness. If the goal is to determine the individuals that qualify for a treatment program, the improvement in performance and fairness in outcomes may justify the deletion of the item as the predictive validity of the scale as a diagnostic tool may be of greater interest. However, if the scores on the depression scale are, for example, used to gain a better understanding the manifestation of the symptoms of depression in different cultural contexts, we recommend consulting existing literature as well as domain experts to clarify the potential reductions in construct coverage. It may also be worthwhile to explore alternative approaches, such as going back to the drawing table and piloting modified versions of the noninvariant item with samples from the different groups to rebuild the scale with an unbiased replacement item, assuming that resource and time constraints allow for such a detour.

Furthermore, test users should exercise great caution while considering deleting multiple items at a time from a scale, and note the close relationship between the test length and its internal reliability (Brown, 1910; Kruyen et al., 2012; Spearman, 1910). We stress that the shorter the test, the riskier it may be to drop

[9] See Chapter 4 of AERA, APA & NCME (2014) and Bandalos (2018) Chapter 16 for additional discussions of MI applications for theory building and item revision.

items.

The item deletion indices, methods, and guidelines introduced here function as exploratory tools to scrutinize the ‘what-if’ scenarios concerning biased items. It is ultimately up to the decision-maker to judge whether the magnitude of an improvement is large enough to warrant deletion, and determine whether one or more items, if any, can (and should) be deleted in a given application context.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Arias de la Torre, J., Ronaldson, A., Vilagut, G., Martínez-Alés, G., Dregan, A., Bakolis, I., Valderas, J. M., Molina, A. J., Martín, V., Bellón, J. Á., & Alonso, J. (2024). Implementation of community screening strategies for depression. *Nature Medicine*, 10.1038/s41591-024-02821-1. Advance online publication. https://doi.org/10.1038/s41591-024-02821-1

Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences* (pp. 478-518). Guilford Publications.

Biddle, D. (2006). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing* (2nd) [doi: https://doi.org/10.4324/9781315263298]. Routledge.

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

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *1904-1920*, *3* (3), 296–322. https://doi.org/ 10.1111/j.2044-8295.1910.tb00207.x

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

Byrne, B. M., Shavelson, R. J., & Muthén, B. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*, 461–473. https://doi.org/10.1007/s11336-007-9039-7

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*, 115–124.

Cohen, J. (1988). *Multiple Factor Analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95(1)*. https://doi.org/10.1037/0033-2909.95.1.134

Gonzalez, O., & Pelham, W. E. (2021). When does differential item functioning matter for screening? A method for empirical evaluation. *Assessment*, *28*, 446–456. https://doi.org/https://doi.org/10.1177/1073191120913618

Hammack-Brown, B., Fulmore, J., Keiffer, G., & Nimon, K. (2021). Finding invariance when noninvariance is found: An illustrative example of conducting partial measurement invariance testing with the automation of the factor-ratio test and list-and-delete procedure. *Human Resource Development Quarterly*. https://doi.org/10.1002/hrdq. 21452

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the mantel-haenszel procedure. *ETS Research Report Series*, *1986* (2), i–24.

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

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34(2)*, 183–202. https://doi.org/10.1007/ BF02289343

Kline, P. M., Rose, E. K., & Walters, C. R. (2021). Systemic discrimination among large U.S. employers. (29053). https://doi.org/10.3386/w29053

Kruyen, P. M., Emons, W. H., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, *12* (4), 321–344.

Kruyen, P. M., Emons, W. H., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, *13* (3), 223–248.

Lai, M. H. C., Kwok, O., Yoon, M., & Hsiao, Y. (2017). Understanding the impact of partial factorial invariance on selection accuracy: An R script. *Structural Equation Modeling*, *24(5)*. https://doi.org/10.1080/10705511.2017.1318703

Lai, M. H. C., Richardson, G. B., & Mak, H. W. (2019). Quantifying the impact of partial measurement invariance in diagnostic research: An application to addiction research. *Addictive Behaviors*, *94*, 50–56. https://doi.org/10.1016/j.addbeh. 2018.11.029

Lai, M. H. C., & Zhang, Y. (2022). Classification accuracy of multidimensional tests: Quantifying the impact of noninvariance. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2021.1977936

Lord, F., Novick, M., & Birnbaum, A. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Lord, F. (1952). A Theory of Test Scores. *Psychometric Monographs*.

Maassen, E., D'Urso, E. D., Van Assen, M. A., Nuijten, M. B., De Roover, K., & Wicherts, J. M. (2023). The dire disregard of measurement invariance testing in psychological science. *Psychological Methods.* Advance online publication. https://dx.doi.org/10.1037/met0000624

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

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

Meredith, W., & Millsap, R. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, *57*, 289–311. https://doi.org/10.1007/BF02294510

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44 (11, Suppl 3)*. https://doi.org/10.1097/01.mlr.0000245438.73837.89

Miller, T. Q., Markides, K. S., & Black, S. A. (1997). The factor structure of the CES-D in two surveys of elderly Mexican Americans. *The Journals of Gerontology: Series B, 52B*, S259–S269. https://doi.org/10.1093/geronb/52B.5.S259

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. https://doi.org/10.1037/1082-989X.9.1.93

Mohan, R., Singla, Kaushal, P., & Kadry, S. (2021). *Artificial intelligence, machine learning, and data science technologies: Future impact and well-being for society 5.0*. https://doi.org/10.1201/9781003153405

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

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

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. https ://doi.org/10.1177/014662167700100306

Reynolds, C. R., Altmann, R., & Allen, D. N. (2021). *Mastering modern psychological testing theory and methods* (2nd) [ISBN 978-3-030-59454-1]. Springer, Cham.

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

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18(4)*. https://doi.org/10.1016/j.hrmr.2008.03.003

Somaraju, A. V., Nye, C. D., & Olenick, J. (2021). A review of measurement equivalence in organizational research: What's old, what's new, what's next? *Organizational Research Methods*. https://doi.org/10.1177/10944281211056524

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3* (3), 271.

Standards for Educational and Psychological Testing. (2014). American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.)

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://dx.doi.org/10.1037/0021-9010.89.3.497

Thurstone, L. L. (1947). *Multiple Factor Analysis*. University of Chicago Press.

US Preventive Services Task Force, Barry, M. J., Nicholson, W. K., Silverstein, M., Chelmow, D., Coker, T. R., Davidson, K. W., Davis, E. M., Donahue, K. E., Jaén, C. R., Li, L., Ogedegbe, G., Pbert, L., Rao, G., Ruiz, J. M., Stevermer, J. J., Tsevat, J., Underwood, S. M., & Wong, J. B. (2023). Screening for Depression and Suicide Risk in Adults: US Preventive Services Task Force Recommendation Statement. *JAMA*, *329*(23), 2057–2067. https://doi.org/10.1001/jama.2023.9297

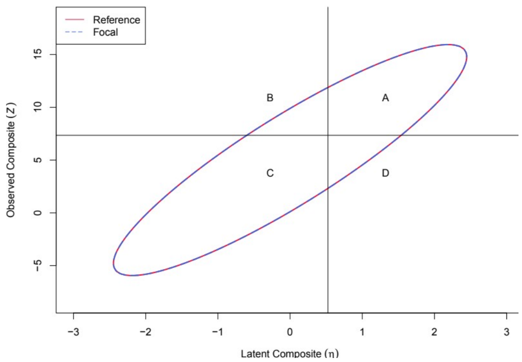
Vandenberg, R., & Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3(1)*. https://doi.org/10.1177/109442810031002

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

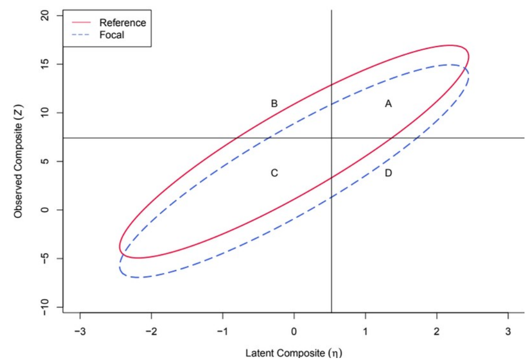
Zhang, B., Fokkema, M., Cuijpers, P., J., Li, Smits, N., & Beekman, A. (2011). Measurement invariance of the Center for Epidemiological Studies Depression Scale (CES-D) among Chinese and Dutch elderly. *BMC Medical Research Methodology*, *11*, 74–83. https://doi.org/10.1186/1471-2288-11-74.

Figure 1

Distribution of observed and latent scores by group and invariance condition.



(a) *Strict factorial invariance.*



(b) *Partial factorial invariance.*

Note. An illustration of the joint bivariate distributions of observed and latent scores for the cases where strict measurement invariance holds (a), and partial measurement invariance holds (b). The distributions are indicated separately for the reference and focal groups. Dotted lines denote thresholds on the observed and latent scores. The quadrants A, B, C, and D correspond to TP, FP, TN, FN rates.

Table 1

Aggregate Classification Accuracy Indices and h values computed for the 20-item CES-D scale

	Aggregate Classification Accuracy Indices					
	$\overline{SR}$	h	$\overline{SE}$	h	$\overline{SP}$	h
Full	0.885	-	0.885	-	0.933	-
|4	0.881	0.013	0.881	0.013	0.930	0.010
|9	0.882	0.009	0.882	0.009	0.931	0.007
|10	0.884	0.004	0.884	0.004	0.932	0.003
|11	0.884	0.005	0.884	0.005	0.932	0.003
|15	0.883	0.008	0.883	0.008	0.931	0.006
|20	0.883	0.006	0.883	0.006	0.932	0.004

Note. Columns $\overline{SR}$, $\overline{SE}$, and $\overline{SP}$ indicate aggregate classification accuracy indices computed for a given item set (either 20-items, "Full", or 19-items excluding item *j* indicated in the row). The columns titled *h* indicate the Cohen's *h* values for comparisons between $\overline{CAI}$ on the 20-item scale and possible 19-item scales excluding item *j*. The dashes in the second row indicate that there is no comparison of $\overline{CAI}$ on the 20-item scale with itself.

Table 2

Item deletion indices comparing the reference and the (expected) focal groups on the 20-item CES-D scale

		CAI$_r$ vs. Expected CAI$_f$					
	AI	$h^{\text{r-Ef}}{}_{\text{SR}}$	$\Delta h^{\text{r-Ef}}{}_{\text{SR}}$	$h^{\text{r-Ef}}{}_{\text{SE}}$	$\Delta h^{\text{r-Ef}}{}_{\text{SE}}$	$h^{\text{r-Ef}}{}_{\text{SP}}$	$\Delta h^{\text{r-Ef}}{}_{\text{SP}}$
Full	0.908	-0.151	-	0.141	-	-0.159	-
|4	0.908	-0.147	0.004	0.138	0.002	-0.156	0.003
|9	0.904	-0.154	-0.003	0.147	-0.006	-0.164	-0.004
|10	0.930	-0.115	0.036	0.104	0.037	-0.122	0.037
|11	0.905	-0.153	-0.002	0.145	-0.005	-0.162	-0.003
|15	0.977	-0.045	0.106	0.026	0.114	-0.047	0.113
|20	0.908	-0.145	0.006	0.141	-0.001	-0.154	0.005

Note. The first column contains the AI ratio for a given item set. $h^{\text{r-Ef}}$CAI columns indicate effect sizes for the discrepancy between classification accuracy indices computed for the reference group (CAI$_r$) and expected CAI computed for the focal group (CAI$_{Ef}$) for an item set (either 20-items, "Full", or 19-items excluding biased item *j* indicated in the row). $\Delta h^{\text{r-Ef}}$CAI columns denote the change in the discrepancy between CAI$_r$ and CAI$_{Ef}$ when item *j* is deleted. The dashes in the first row indicate that there is no comparison of an item deletion index on the 20-item scale with itself. As Cohen's *h* cannot be computed for non-proportions, there are no *h* values reported for AI values.

Table 3

Aggregate classification accuracy indices computed for the 19-item CES-D scale.

	Aggregate classification accuracy indices					
Item Set	$\overline{SR}$	h	$\overline{SE}$	h	$\overline{SP}$	h
Full	0.888	-	0.888	-	0.926	-
|4	0.884	0.011	0.884	0.011	0.924	0.009
|9	0.879	0.008	0.879	0.008	0.915	0.006
|10	0.881	0.003	0.881	0.003	0.916	0.002
|11	0.880	0.004	0.880	0.004	0.916	0.003
|20	0.875	0.020	0.875	0.020	0.912	0.016

Note. Columns $\overline{SR}$, $\overline{SE}$, and $\overline{SP}$ indicate aggregate classification accuracy indices computed for a given item set (either 19-items, "Full", or 18-items excluding item *j* indicated in the row). Columns titled *h* Cohen's *h* values for comparisons between $\overline{CAI}$ on the 19-item scale and 18-item scales excluding item *j*. Note that item numbers are the same after the deletion of item 15 (*good*). The dashes in the first row indicate that there is no comparison of $\overline{CAI}$ on the 19-item scale with itself.

Table 4

Item deletion indices comparing the reference and the (expected) focal groups on the 19-item CES-D scale

		CAI$_r$ vs. Expected CAI$_f$					
Item Set	AI	$h^{\text{r-Ef}}\text{SR}$	$\Delta h^{\text{r-Ef}}\text{SR}$	$h^{\text{r-Ef}}\text{SE}$	$\Delta h^{\text{r-Ef}}\text{SE}$	$h^{\text{r-Ef}}\text{SP}$	$\Delta h^{\text{r-Ef}}\text{SP}$
Full	0.979	-0.044	-	0.025	-	-0.048	-
|4	0.979	-0.042	0.002	0.024	0.001	-0.046	0.002
|9	0.973	-0.047	-0.004	0.032	-0.006	-0.053	-0.005
|10	0.999	-0.011	0.032	-0.007	0.018	-0.010	0.037
|11	0.973	-0.047	-0.004	0.031	-0.006	-0.052	-0.005
|20	0.977	-0.040	0.003	0.028	-0.002	-0.045	0.002

Note. The first column contains the AI ratio for a given item set. $h^{\text{r-Ef}}$CAI columns indicate effect sizes for the discrepancy between classification accuracy indices computed for the reference group (CAI$_r$) and expected CAI computed for the focal group (CAI$_{Ef}$) for an item set (either 19-items, "Full", or 18-items excluding biased item *j* indicated in the row). $\Delta h^{\text{r-Ef}}$CAI columns denote the change in the discrepancy between CAI$_r$ and CAI$_{Ef}$ when item *j* is deleted. Note that item numbers are the same after the deletion of item 15 (*good*). The dashes in the first row indicate that there is no comparison of an item deletion index on the 19-item scale with itself. As Cohen’s *h* cannot be computed for non-proportions, there are no *h* values reported for AI values.

Appendix A

1. The Common Factor Model

For a set of J items ($j = 1, \ldots, J$) aimed to measure M latent constructs ($m = 1, \ldots, M$), let $\boldsymbol{y}_{ig}$ denote a $J \times 1$ vector of observed item scores, and $\boldsymbol{\eta}_{ig}$ a $M \times 1$ vector of latent factor scores distributed with $M \times 1$ mean vector $\mathrm{E}(\boldsymbol{\eta}) = \boldsymbol{\alpha}$ and $M \times M$ variance-covariance matrix $\mathrm{Cov}(\boldsymbol{\eta}) = \boldsymbol{\Psi}$. Here, i denotes the individual ($i = 1, \ldots, N$), and g denotes group membership, time point, or test condition. The common factor model postulates that the relationship between the latent and observed variables is expressed by

$$\boldsymbol{y}_{ig} = \boldsymbol{\nu}_g + \Lambda_g \boldsymbol{\eta}_{ig} + \boldsymbol{\epsilon}_{ig}$$

where $\boldsymbol{\nu}_g$ is a $J \times 1$ vector of intercepts, Λ_g is a $J \times M$ matrix of factor loadings, and $\boldsymbol{\epsilon}_{ig}$ is a $J \times 1$ vector of unique factor variables (Lai & Zhang, 2022; Meredith & Teresi, 2006). Unique factor variables (ϵ) refer to the construct-irrelevant variance of the sum of measurement error and systematic error, and each is assumed to be distributed independently with mean $\mathrm{E}(\boldsymbol{\epsilon}) = 0$ and variance-covariance matrix $\mathrm{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Theta}$. Assuming additionally that the latent and unique factor variables are uncorrelated ($\mathrm{Cor}[\boldsymbol{\epsilon}, \boldsymbol{\eta}] = 0$), the observed variables are distributed with mean $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\nu} + \Lambda\, \boldsymbol{\alpha}$ and variance-covariance matrix $\boldsymbol{\Sigma} = \Lambda\, \boldsymbol{\Theta}\, \Lambda + \Psi$.

Depending on which parameters are the same across groups, the level of factorial invariance can be classified as, from the least to most stringent, configural, metric, scalar, and strict (Byrne et al., 2007; Horn & McArdle, 1992; Meredith, 1993). Configural invariance requires the same factor structure across groups, and freely estimates all parameters. Metric invariance additionally requires equal unstandardized factor loadings (Λ) across groups. Scalar invariance holds if measurement intercepts ($\boldsymbol{\nu}$) are also the same across groups. Finally, strict factorial invariance (SFI) exists when measurement intercepts, factor loadings, and unique factor variance-covariances ($\mathrm{Var}[\epsilon]$; uniqueness) are equal across groups or conditions ($\boldsymbol{\nu}_g = \boldsymbol{\nu}$, $\Lambda_g = \Lambda$, $\boldsymbol{\theta}_g = \boldsymbol{\theta}$, $\forall g$). While SFI is necessary for valid and meaningful comparison of factor scores across groups, it may be difficult for these demanding criteria to be met in practice. More often, partial factorial invariance

(PFI, Byrne et al., 1989) is met, meaning that invariance holds only for a subset of the items.

Factorial invariance has been shown to be equivalent to MI under the common factor model (Horn & McArdle, 1992; Thurstone, 1947). Then, response probabilities of individuals with the same latent standing are expected to be invariant across groups if MI holds. Mathematically, MI exists if conditioned on the latent construct, observed scores and group membership are independent such that

$$P\,(\boldsymbol{y}|\boldsymbol{\eta}, G = g) = P\,(\boldsymbol{y}|\boldsymbol{\eta}), \forall g$$

(Mellenbergh, 1989; Meredith & Millsap, 1992). Under the common factor model, MI is satisfied when SFI holds, and PMI is equivalent to PFI.

2. Adverse Impact Ratio

Letting $P_f\,(Z_f > Z_c)$ and $P_r(Z_r > Z_c)$ denote the proportion of selected individuals who scored above the cut-off point Z_c in the focal and reference groups and P_{Ef} denote the proportion of selected individuals expected for the focal group, the AI ratio is defined as

$$AI\,ratio = \frac{P_{Ef}\left(Z_f \geq Z_c\right)}{P_r\left(Z_r \geq Z_c\right)}$$

where

$$\mathrm{P_{Ef}}\,(\mathrm{Z_f} \geq \mathrm{Z_c}) = \int \mathrm{P_f}\,(\mathrm{Z_f} \geq \mathrm{Z_c}|\eta)\ \mathbf{f}_\mathrm{r}\,(\eta)\, d\eta,$$

$$\mathrm{P_r}\,(\mathrm{Z_r} \geq \mathrm{Z_c}) = \int \mathrm{P_r}\,(\mathrm{Z_r} \geq \mathrm{Z_c}|\eta)\ \mathbf{f}_\mathrm{r}\,(\eta)\, d\eta$$

(Nye & Drasgow, 2011; Stark et al., 2004).

3. The Multidimensional Classification Accuracy Analysis Framework

Let $\boldsymbol{c}$ be a $J \times 1$ vector of item weights. For a multidimensional test with J items measuring M latent constructs, assuming the multivatiate normality of $(\boldsymbol{\eta}, \boldsymbol{\epsilon})$, the observed scale sums Z_g and the latent factor scores η_g were shown to follow a bivariate normal distribution such that

$\begin{pmatrix} Z_g \\ \eta_g \end{pmatrix} = N\left(\begin{bmatrix} c\,v_g + c\,\Lambda_g \alpha_g \\ w\,\alpha_g \end{bmatrix}, \quad \left[c\,\Lambda_g \Psi_g \Lambda_g c + c\,\Theta_g c \quad ¿\,c\,\Lambda_g \Psi_g w \quad w\,\Psi_g w\right]\right)$ where $\boldsymbol{w}$ is a $1 \times M$ vector of latent factor weights (Lai & Zhang, 2022). Furthermore, the marginal distribution of (Z, η) was demonstrated to be a finite mixture of bivariate normal distributions with mixing proportion π_g, and the latent score cut-off η_c can be computed as the quantile in the mixture corresponding to PS_{total} (Lai & Zhang, 2022; Millsap & Kwok, 2004).

Appendix B

Computing Aggregate Classification Accuracy Indices ($\overline{CAI}$)

We compute aggregate TP ($\overline{TP}$), FP ($\overline{FP}$), TN ($\overline{TN}$), and FN ($\overline{FN}$) using the following formulas where π_r indicates the mixing proportion (the relative size) of the reference group:

$$\overline{TP} = \text{TP}_r \times \pi_r + \text{TP}_f \times (1 - \pi_r),$$

We then compute $\overline{CAI}$ on the full item set:

$$\overline{PS} = \overline{TP} + \overline{FP},$$

$$\overline{SR} = \overline{TP} / (\overline{TP} + \overline{FP}),$$

$$\overline{SE} = \overline{TP} / (\overline{TP} + \overline{FN}),$$

$$\overline{SP} = \overline{TN} / (\overline{TN} + \overline{FP}).$$

$\overline{PS}$ equals the user-specified proportion to be selected, or the quantile as identified by the user-specified cut-off. Then, $\overline{PS}^{\text{¿}j}$, $\overline{SR}^{\text{¿}j}$, $\overline{SE}^{\text{¿}j}$ and $\overline{SP}^{\text{¿}j}$ are computed and compared against $\overline{PS}$, $\overline{SR}$, $\overline{SE}$, and $\overline{SP}$ to determine the impact of deleting a biased item.

$h^{|j}\,\overline{CAI}$ effect size for the change in $\overline{CAI}$ when the $\boldsymbol{j}$-th item is deleted is computed using:

$$h^{\text{¿}j}\,\overline{CAI} = 2\arcsin(\sqrt{\overline{CAI}}) - 2\arcsin(\sqrt{\overline{CAI}^{\text{¿}j}})$$

For example, the improvement in $\overline{SE}$ if the first item is deleted is computed as

$$h^{\text{¿}1}\,\overline{SE} = 2\arcsin(\sqrt{\overline{SE}}) - 2\arcsin(\sqrt{\overline{SE}^{\text{¿}1}}).$$