


1 **Classification Accuracy of Multidimensional Tests: Quantifying the Impact of Noninvariance**

2 Mark H. C. Lai¹ & Yichi Zhang¹

3 ¹ Department of Psychology, University of Southern California

4 **Author Note**

5
6 Mark H. C. Lai  <https://orcid.org/0000-0002-9196-7406>

7 Yichi Zhang  <https://orcid.org/0000-0002-4112-2106>

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

12 *This is an Accepted Manuscript of an article to be published by Taylor & Francis in Structural*
13 *Equation Modeling: A Multidisciplinary Journal on September 3, 2021, to be available online:*
14 *<http://www.tandfonline.com/10.1080/10705511.2021.1977936>*

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

18

Abstract

19 There has been tremendous growth in research on measurement invariance over the past two decades.
20 However, given that psychological tests are commonly used for making classification decisions such as
21 personnel selections or diagnoses, surprisingly, there has been little research on how noninvariance impacts
22 classification accuracy. Millsap and Kwok (2004) proposed a selection accuracy framework for that
23 purpose, which has been recently extended to categorical data. Their framework, however, only deals with
24 classification using a unidimensional test. In contrast, classification in practice usually involves
25 multidimensional tests (e.g., personality) or multiple tests, with different weights assigned to each
26 dimension. In the current paper, we extend Millsap and Kwok's framework for examining the impact of
27 noninvariance to a multidimensional test on classification. We also provide an R script for the proposed
28 method and illustrate it with a personnel selection example using data from a published report featuring a
29 five-factor personality inventory.

30

Keywords: measurement invariance, sensitivity, specificity, classification accuracy, test bias

31 **Classification Accuracy of Multidimensional Tests: Quantifying the Impact of Noninvariance**

32 Scores on psychological tests are widely used when making selections, diagnostic, and admission
33 decisions. These tests are used to quantify people's relative standings on certain psychological constructs,
34 such as conscientiousness, self-esteem, vocational aptitude, or depression. However, the use of a
35 psychological test is only valid when *measurement invariance* holds, meaning that the test is free of
36 measurement bias such that it measures one or more latent construct in an equivalent and comparable way
37 across demographic subgroups (e.g., race, gender, age, disability status), modes of test administration (e.g.,
38 paper-and-pencil vs. computer-based), or any construct-irrelevant differences (Meredith, 1993; Millsap,
39 2011; Stark et al., 2004; Vandenberg, 2002). Given its importance, there has been exponential growth in
40 the number of studies detecting violations of measurement invariance of existing and newly developed
41 psychological tests. For example, Putnick and Bornstein (2016) identified 126 such articles published in
42 peer-reviewed journals in just one year, 2013.

43 On the other hand, there has been little research investigating how noninvariance affects the
44 quality of classification decisions based on these tests, such as in personnel selection, diagnosis, and
45 admission (Putnick & Bornstein, 2016; Schmitt & Kuljanin, 2008), which is often of great interest to users
46 of psychological tests. For example, personality assessment is commonly used in personnel selection (e.g.,
47 Schmit & Ryan, 1993); questionnaire and behavioral checklist are commonly used as screening tools for
48 mental health conditions. Most existing measurement invariance research on psychological tests, however,
49 have focused only on identifying noninvariant items, with little guidance on how to translate those research
50 findings for interpreting test scores. For instance, readers are usually only told, that two items in a test
51 were found noninvariant across ethnic groups, and test users are left wondering whether they should
52 remove those two items when administering the test. Even when effect size indices are reported, those are
53 usually presented in terms of the difference in loadings (e.g., Millsap, 2011) or test statistics (e.g., the
54 Mantel-Haenszel statistic; Zwick et al., 1999), which do not directly show whether the noninvariance makes
55 selection less effective or creates an unjustified barrier for certain subpopulations.

56 A useful framework to quantify the impact of measurement bias on selection or classification
57 accuracy was proposed by Millsap and Kwok (2004), which compares classification accuracy indices—such
58 as sensitivity and specificity—of a test with and without measurement invariance (see also Stark et al.,
59 2004). It allows researchers and test administrators to directly see the practical impact of measurement
60 bias on the effectiveness of a test for classification purposes. As shown in Millsap and Kwok, violation of
61 measurement invariance at the item level may or may not lead to meaningful impacts on the accuracy of a
62 classification procedure. More recently, Lai et al. (2017) have provided an R program to implement Millsap

63 and Kwok’s classification accuracy framework; Lai et al. (2019) and Gonzalez and Pelham (2021) have
64 extended the framework for binary and ordinal items.

65 However, so far the classification accuracy framework is limited to a unidimensional test, where
66 participants are measured on only one latent construct. In reality, classification is likely a decision based on
67 multiple tests or subtests. For example, in personnel selection, organizations may use combinations of
68 cognitive ability tests and dimensions of personality to select employees (e.g., Schmidt & Hunter, 1998). In
69 college admission, administrators may give different weights to different components of aptitude tests (e.g.,
70 verbal, mathematics, reading), together with other criteria, to rank potential students (e.g., Aguinis et al.,
71 2016). In these examples, it is common to assign more weights to dimensions deemed more important or
72 found more predictive of some criterion variables, like conscientiousness among personality dimensions
73 (Barrick & Mount, 1991; Hurtz & Donovan, 2000). Nevertheless, the unidimensional framework by Millsap
74 and Kwok (2004) and the recent extensions only allow examining each dimension separately, and thus do
75 not allow incorporating the relative importance weight of each dimension. Furthermore, some items may
76 tap into more than one dimensions, and how biases in those items affect classification decisions depend on
77 the correlations and the relative importance of the latent dimensions, the intricacies of which can only be
78 evaluated by considering all items and the dimensions simultaneously. Therefore, in the present paper, we
79 extend the classification accuracy framework to a multidimensional setting so that it quantifies the overall
80 impact of measurement noninvariance on the fairness and effectiveness of a classification procedure. In
81 addition, we provide an R script for implementing the proposed analysis.

82 In the following, we first define the model notations for a multi-group multidimensional factor
83 model and review previous approaches for evaluating measurement invariance. We then present the details
84 of the multidimensional classification accuracy analysis (MCAA) framework as an extension of Millsap and
85 Kwok (2004)’s framework, which includes defining the classification accuracy indices. The framework will
86 then be applied in a hypothetical selection scenario where job applicants are selected based on a weighted
87 composite of their subscale scores on a Big-Five personality test, with a step-by-step tutorial on the
88 relevant analyses using the R script provided.

89 **Factor Model**

90 For psychological tests, the factor model (Thurstone, 1947) is commonly used to represent the
91 statistical relations between item scores and the underlying constructs measured. Consider a set of p items
92 measuring m psychological constructs. Let \mathbf{y}_i be the $p \times 1$ item response vector of person i ’s score on the
93 items, and $\boldsymbol{\eta}_i$ be a $m \times 1$ vector containing scores on the underlying latent (i.e., unobserved) constructs.

94 Under a multidimensional common factor model (Thurstone, 1947), $\boldsymbol{\eta}$ and \mathbf{y} are linked statistically as a
 95 linear system,

$$\mathbf{y}_i = \mathbf{v} + \mathbf{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \quad (1)$$

96 where \mathbf{v} is a $p \times 1$ vector of measurement intercepts, which is analogous to regression intercepts, with
 97 elements v_j ($j = 1, \dots, p$) indicating the expected item scores for a person with zero scores on all latent
 98 variables; $\mathbf{\Lambda}$ is a $p \times m$ matrix of factor loadings, which is analogous to regression slopes, with elements λ_{jk}
 99 ($j = 1, \dots, p; k = 1, \dots, m$) indicating the strength of associations of item i with the k th latent construct;
 100 and $\boldsymbol{\varepsilon}_i$ is a $p \times 1$ column vector of the unique factor random variables, which captures the influence of
 101 factors that are irrelevant to $\boldsymbol{\eta}$ on \mathbf{y}_i . In other words, the factor model expressed in equation (1) says that
 102 a person's item scores are linear functions of their standings on the latent constructs ($\boldsymbol{\eta}$), plus some
 103 construct-irrelevant measurement errors ($\boldsymbol{\varepsilon}$).

104 Let $E(\boldsymbol{\eta}) = \boldsymbol{\alpha}$ and $\text{Cov}(\boldsymbol{\eta}) = \boldsymbol{\Psi}$ be the mean vector and the variance-covariance matrix of the latent
 105 variables, respectively. Further, let the variance-covariance matrix among the unique factor variables be
 106 $\text{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Theta}$, and assume that each unique factor variable has a zero mean, $E(\boldsymbol{\varepsilon}) = 0$. In practice,
 107 researchers usually impose the local independence assumption such that $\boldsymbol{\Theta}$ is a diagonal matrix, meaning
 108 that the inter-item correlations are attributed solely to the variance of the underlying latent factor;
 109 however, the proposed framework can be applied when the local independence assumption is violated. It is
 110 assumed that $\boldsymbol{\eta}$ and $\boldsymbol{\varepsilon}$ are independent with $\text{Cov}(\boldsymbol{\eta}, \boldsymbol{\varepsilon}) = 0$, and together the model implies that
 111 $E(\mathbf{y}) = \mathbf{v} + \mathbf{\Lambda}\boldsymbol{\alpha}$ and $\text{Var}(\mathbf{y}) = \mathbf{\Lambda}\boldsymbol{\Psi}\mathbf{\Lambda}' + \boldsymbol{\Theta}$.

112 Factorial Invariance

113 Measurement invariance, or lack of item bias, is the condition where individuals from different
 114 subpopulations (e.g., age, gender, ethnicity, socioeconomic status), with the same standings on the latent
 115 constructs (e.g., cognitive ability or conscientiousness), demonstrate the same propensities in responding to
 116 all the items measuring these constructs. As such, measurement invariance is key to the "ideal of fairness"
 117 in workplace testing as described in the *Standards for Educational and Psychological Testing* (American
 118 Educational Research Association et al., 2014), according to which, fairness "is achieved if a given test
 119 score has the same meaning for all individuals and is not substantially influenced by construct-irrelevant
 120 barriers to individuals' performance." (p. 169). Therefore, the importance of identifying bias in
 121 psychological tests and evaluating measurement invariance cannot be understated.

122 Formally, measurement invariance holds when the conditional distribution of the observed item
 123 scores is the same across subpopulations that are not part of the construct domain (Mellenbergh, 1989).

124 That is, for the subpopulation membership variable W with levels $g = 1, \dots, G$, like gender and ethnicity,

$$P(\mathbf{y}|\boldsymbol{\eta}, W = g) = P(\mathbf{y}|\boldsymbol{\eta}), \quad \forall g.$$

125 Under the factor model defined in (1) and assuming multivariate normality of $(\boldsymbol{\eta}, \boldsymbol{\varepsilon})$, measurement
 126 invariance is also called strict factorial invariance (Meredith, 1993), or strict invariance, which holds when
 127 the measurement parameters: loadings ($\boldsymbol{\Lambda}$), intercepts (\mathbf{v}), and unique factor covariances ($\boldsymbol{\Theta}$), are equal
 128 across all subgroups. In math notations, factorial invariance implies

$$\mathbf{v}_g = \mathbf{v}, \boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}, \boldsymbol{\Theta}_g = \boldsymbol{\Theta}, \quad \forall g.$$

129 In practice, however, strict invariance does not commonly hold. Previous researchers have
 130 distinguished four stages of factorial invariance (Millsap, 2007), each with different implications for the use
 131 of test scores. The first stage is configural invariance, which requires that the factor structures be the same
 132 across subgroups, including the same number of factors and the same composition of items for each factor.
 133 An example violation of configural invariance is that an item is an indicator of math ability in one group,
 134 but is an indicator of both math ability and English proficiency in another group. The second stage is
 135 metric invariance (Horn & Mcardle, 1992), which, in addition to configural invariance, requires equal factor
 136 loadings (i.e., $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ for all g). As such, metric invariance ensures that a unit difference in the latent
 137 construct is comparable across subgroups. The third stage is scalar invariance, which, in addition to metric
 138 invariance, requires equal measurement intercepts across subgroups (i.e., $\mathbf{v}_g = \mathbf{v}$ for all g). Scalar invariance
 139 ensures that a given measure has the same origin or zero point. The final stage is strict invariance as
 140 previously discussed, where the unique factor variances and covariances are also identical (i.e., $\boldsymbol{\Theta}_g = \boldsymbol{\Theta}$ for
 141 all g).

142 *Partial Factorial Invariance/Item Bias*

143 In contrast to full factorial invariance, item bias, also called partial factorial invariance (Millsap &
 144 Kwok, 2004), is present when two individuals with exactly the same standings on the latent constructs
 145 demonstrate different propensities to respond to one or more items. When measurement invariance does
 146 not hold for some items, meaning that item bias is present, the comparison of test scores across
 147 subpopulations is not valid and can be highly misleading. Consider the hypothetical example in Figure 1,
 148 where scalar invariance is violated for a test of emotional intelligence with respect to paper-and-pencil and
 149 Internet-based administrations. The overall bias, due to differences in the intercepts of the items,
 150 systematically leads to lower scores for persons using the Internet-based test. Therefore, Person 2, who has

151 a higher true emotional intelligence level than Person 1 and takes the Internet-based test, gets a lower
152 observed test score than Person 1, who takes the paper-and-pencil-based test.

153 An abundance of the previous literature has focused on statistical methods for detecting violations
154 of factorial invariance, the most popular ones among which are the likelihood ratio test (LRT or χ^2 ;
155 Millsap, 2011) and the change in goodness-of-fit indices in structural equation modeling (Cheung &
156 Rensvold, 2002). Using maximum likelihood estimation, a likelihood ratio χ^2 statistic is obtained by
157 comparing the maximized log-likelihoods of a model with invariance constraints (e.g., the metric invariance
158 model with equality constraints on the factor loadings) and a model without such constraints (e.g., the
159 configural invariance model without factor loading constraints). A significant test statistic then indicates
160 that a particular step of measurement invariance is violated. However, the likelihood ratio test is very
161 sensitive to large sample sizes such that items are flagged as noninvariant even when the degree of bias is
162 trivial (Putnick & Bornstein, 2016). As an alternative, researchers rely on goodness-of-fit indices that are
163 less sensitive to sample sizes, such as the comparative fit index (CFI; Bentler, 1990) and the root mean
164 squared error of approximation (RMSEA; Steiger, 1980), and deem a psychological test practically
165 invariant when the change in these indices is within a certain threshold (e.g., $\Delta\text{CFI} < .01$ by Cheung &
166 Rensvold, 2002; $\Delta\text{RMSEA} < .005$ by Chen, 2007). These indices, however, are not meaningful metrics
167 when it comes to communicating the degree of noninvariance of tests, as a ΔCFI of $-.03$, for example, does
168 not indicate how using the test will be problematic in any concrete way.

169 Recently, there have been increased research efforts to define interpretable effect size indices for
170 noninvariance at the item level. For example, Nye and Drasgow (2011) proposed the d_{MACS} effect size,
171 which corresponds to the expected standardized difference in observed item scores due to noninvariance;
172 Nye et al. (2019) further provided benchmark values for d_{MACS} based on a systematic review of the
173 organizational literature. Gunn et al. (2020) proposed and evaluated several indices that are conceptually
174 similar to d_{MACS} . However, because these indices focus on the impact of noninvariance on the item mean,
175 they do not directly inform the impact on classification—a common usage of psychological tests—for two
176 reasons. First, previous research usually assumes that biased items are automatically worse than unbiased
177 items and, therefore, should be removed in order to achieve valid cross-group comparisons. However, while
178 reducing bias in cross-group comparisons, removing biased items may make the test less reliable due to
179 reduced test length, resulting in less precise inferences. When using test results for decision-making, both
180 bias (systematic error) and precision (random error) should be taken into account. A slightly biased but
181 highly effective item may contribute more information than an unbiased but ineffective item, but existing
182 approaches for detecting item bias pays less attention to the role of unique variances and covariances (i.e.,

183 Θ), which is related to score reliability and is relevant to classification.¹

184 Second, as noted by Millsap and Kwok (2004), the evaluation of item bias should be made “in
185 relation to the purpose of the measure” (pp. 94–95). In the behavioral sciences, a common purpose of a
186 psychological test is to select or identify individuals based on their relative standings or absolute scores on
187 the test (Crocker, 2006). Surprisingly, and unfortunately, very little attention has been paid to how
188 noninvariance impacts selection. In the following section, we briefly review the relevant literature on item
189 bias and classification, after which we define the MCAA framework as an extension to the approach by
190 Millsap and Kwok (2004).

191 **Factorial Invariance in the Context of Selection**

192 Psychological and behavioral measures are commonly used for various classification purposes:
193 identifying people with depressive symptoms, selecting or promoting employees, and providing support for
194 college admissions decisions. Often employers and test administrators compute a scale score or a composite
195 score, denoted as Z , by applying a scoring rule on the item scores, such as by summing the items. The
196 classification decisions are then based on Z .

197 As a hypothetical personnel selection example, imagine that two subgroups of applicants respond
198 to a battery of assessment items (e.g., personality and cognitive tests). Without loss of generality, denote
199 the two groups as the *reference* and the *focal* groups (Millsap & Kwok, 2004), where the focal group is
200 considered to have a disadvantage due to potential measurement bias. Assume that the two groups are of
201 equal sizes and have identical distributions on their actual, latent competency level. Based on their
202 responses, each applicant receives a Z score, and a manager wants to use the battery to select the top 10%
203 of the combined pool of applicants. If the tests are bias-free, the final pool should consist of roughly 10% of
204 participants from the reference group and 10% from the focal group, as shown in Figure 2a (i.e., the
205 combined area of quadrants A and B). However, it is possible that due to noninvariance, or item bias, the
206 reference group on average gets higher scores than the focal group. As a result and as shown in Figure 2b,
207 13.4% of the reference group but only 6.7% of the focal group are selected. In other words, the selection
208 ratio changes from 1:1 to 2:1 between applicants in the reference and the focal groups.

209 The example is simplified because it assumes an equal number of applicants from the focal and the
210 reference groups with matched qualifications. Also, unless tests are perfectly reliable, some selected

¹ An example of this can be found in the supplemental material, where selection drops after deleting five noninvariant items in a 20-item test.

211 individuals in each subgroup will be “false positives;” a selection process that selects equal proportions of
 212 applicants in each subgroup is still problematic if it results in more false positives in some subgroups than
 213 others. A systematic approach to evaluating selection accuracy, as discussed in the remainder of the current
 214 paper, is needed to assess how factors such as group sizes, differences in the distribution of qualifications,
 215 and reliability of the test scores may influence the effect of item bias on classification accuracy.

216 *Classification Accuracy Analysis*

217 Millsap and Kwok (2004) proposed a framework to quantify how noninvariance affects
 218 classification accuracy. Specifically, based on the probabilities of true positives (qualified and selected),
 219 false positives (unqualified but selected), true negatives (unqualified and not selected), and false negatives
 220 (qualified but not selected) (i.e., quadrants A, B, C, D in Figure 2), one can summarize classification
 221 accuracy by the following indices:

$$\begin{aligned} \text{Proportion selected (PS)} &= P(\text{true positive}) + P(\text{false positive}), \\ \text{Success Ratio (SR)} &= \frac{P(\text{true positive})}{P(\text{true positive}) + P(\text{false positive})}, \\ \text{Sensitivity (SE)} &= \frac{P(\text{true positive})}{P(\text{true positive}) + P(\text{false negative})}, \\ \text{Specificity (SP)} &= \frac{P(\text{true negative})}{P(\text{true negative}) + P(\text{false positive})}, \end{aligned}$$

222 where $P(\cdot)$ denotes the probability of an outcome.

223 For example, in personnel selection, proportion selected refers to the proportion of candidates
 224 selected for the job based on the biodata items. Success ratio is the proportion of candidates who are truly
 225 qualified among all the selected candidates; thus, a low success ratio means that the selection procedure
 226 selects many unqualified candidates for the job. Sensitivity is the proportion of candidates who are selected
 227 among all qualified candidates; thus, a low sensitivity means that only a small proportion of truly qualified
 228 candidates are selected. Finally, specificity is the proportion of candidates who are not selected among all
 229 the candidates who do not meet the cutoff, so a low specificity means that only a small proportion of truly
 230 unqualified candidates are correctly screened out. In this hypothetical example, one can argue that success
 231 ratio and sensitivity are more important if the goal is to select the best candidates, so proportion selected,
 232 success ratio, and sensitivity are important factors concerning the fairness of the selection procedure across
 233 subgroups. In practice, however, different combinations of these indices may matter most, depending on
 234 the purpose of the selection procedure.

235 The analyses originally proposed by Millsap and Kwok (2004) only computed PS, SR, SE, and SP.
 236 In the context of personnel selection, one additional index that is of interest in personnel selection is the

237 adverse impact (AI) ratio, defined as (Nye & Drasgow, 2011)

$$\text{AI ratio} = \frac{E_R(\text{PS}_F)}{\text{PS}_R},$$

238 where PS_R is the proportion selected for the reference group (usually the majority group), and $E_R(\text{PS}_F)$ is
 239 the *expected proportion selected for the focal group based on the latent score distributions of the reference*
 240 *group*. In other words, $E_R(\text{PS}_F)$ is the proportion that would be selected from the focal group when the
 241 focal and the reference groups were matched in latent trait levels. When strict invariance holds, the AI
 242 ratio is 1, meaning that two candidates with equal latent trait levels—one from the focal group and the
 243 other from the reference group—are equally likely to be selected. When $\text{AI ratio} < 1$, it indicates that those
 244 in the focal group would be less likely to be selected due to factors not related to the target latent traits.

245 To evaluate the impact of noninvariance on selection, researchers compute the classification
 246 accuracy indices based on the parameter estimates from a partial strict invariance model and a strict
 247 invariance model, respectively. They can then compare the two sets of classification accuracy indices, and
 248 holistically evaluate the impact of noninvariance on selection for each subpopulation.

249 **Multidimensional Classification Accuracy Analysis (MCAA) Framework**

250 Millsap and Kwok (2004)'s framework and follow-up research assumed unidimensionality of the
 251 items, meaning that all items used for selection measure one single latent trait. In actual personnel
 252 selection as well as in many classification tasks, the items may tap into multiple constructs, or constructs
 253 with multiple dimensions. For example, the Mini-International Personality Item Pool (Mini-IPIP;
 254 Donnellan et al., 2006), a personality inventory commonly used as part of personnel selection, has five
 255 dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism). Also,
 256 classification in practice may assign different weights to different dimensions, which means that individuals
 257 are selected based on a weighted composite score on the latent variables. Therefore, in this article, we
 258 propose a more general, multidimensional selection accuracy framework, and illustrate it using a secondary
 259 data analytic example.

260 For a selection test with q dimensions, let η_k be the true score for dimension k , and let \mathbf{w} be a
 261 $m \times 1$ vector of weights. In other words, if we know the true, error-free score $\boldsymbol{\eta}$ for every person, the
 262 selection should be based on $\boldsymbol{\zeta} = \mathbf{w}\boldsymbol{\eta}$. However, we only have the error-prone scores on p items, \mathbf{y} . Usually,
 263 the items can be similarly partitioned into m subsets $[\mathbf{y}'_1, \mathbf{y}'_2, \dots, \mathbf{y}'_m]'$, where each \mathbf{y}_k component consists of
 264 p_k items with $\sum_{k=1}^m p_k = p$. Let $\mathbf{c} = [c_1, c_2, \dots, c_p]$ be the vector of weights for the items. So with only the
 265 item scores, the selection is based on $Z = \mathbf{c}\mathbf{y}$. Following the derivation in Millsap and Kwok (2004), under

266 the multivariate normal assumption of $(\boldsymbol{\eta}, \boldsymbol{\varepsilon})$, within each subpopulation g , (Z_g, ζ_g) follows a bivariate
 267 normal distribution:

$$\begin{pmatrix} Z_g \\ \zeta_g \end{pmatrix} = N \left(\begin{bmatrix} \mathbf{c}\mathbf{v}_g + \mathbf{c}\boldsymbol{\Lambda}_g\boldsymbol{\alpha}_g \\ \mathbf{w}\boldsymbol{\alpha}_g \end{bmatrix}, \begin{bmatrix} \mathbf{c}\boldsymbol{\Lambda}_g\boldsymbol{\Psi}_g\boldsymbol{\Lambda}'_g\mathbf{c}' + \mathbf{c}\boldsymbol{\Theta}_g\mathbf{c}' & \\ & \mathbf{w}\boldsymbol{\Psi}_g\mathbf{w}' \end{bmatrix} \right), \quad (2)$$

268 and the marginal distribution of (Z, ζ) is a finite mixture of bivariate normal distributions, with mixing
 269 proportions π_1, \dots, π_G based on the relative sizes of the subpopulations.

270 With a given cutscore on the observed composite, Z_c , and the total proportion selected,
 271 $\text{PS}_T = P(Z > Z_c)$, Millsap and Kwok (2004) showed that the cutscore on the latent composite, ζ_c , can be
 272 determined as the quantile in the mixture bivariate normal distribution corresponding to probability PS_T .
 273 Once Z_c and ζ_c are set, the classification accuracy indices for group g can be easily obtained as cumulative
 274 probabilities in a bivariate normal distribution. We have created an R script with the major function
 275 `PartInvMulti_we()` (see the supplemental material) that automates the computation, so that users simply
 276 need to input the parameter values for each subpopulation, together with the mixing proportions and Z_c ,
 277 to get the selection accuracy indices for each subpopulation. Below, we demonstrate the MCAA using real
 278 data of a personality inventory.

279 *Comparing MCAA With the Unidimensional Counterpart*

280 To demonstrate the need for a multidimensional framework, we conducted a simulation in which
 281 classification is based on two mildly correlated latent variables, and compared the classification accuracy
 282 indices based on MCAA as opposed to the unidimensional framework by Millsap and Kwok (2004).
 283 Specifically, we simulated 1,000 data sets, each with 10 items, where the first five items loaded on the first
 284 factor and the next five items loaded (primarily) on the second factor, with all loadings = .70; we also
 285 made item 10 to cross-load on the first factor with loading = .30. The items were fully metric invariant but
 286 three items were scalar noninvariant, with intercepts = 0 for all items in the first group and were 0.3, -0.1,
 287 and 0.5 for items 4, 5, and 10. Both latent factors had unity variance with a .2 correlation in both groups,
 288 and the unique variances were .51 for all items. We simulated 1,000 data sets, and for each data set, we
 289 obtained classification accuracy indices using MCAA with weights of $\mathbf{w} = [.7, .3]$ given to the two factors,
 290 respectively. The classification accuracy indices were based on selecting the top 25% on the latent
 291 composite. We also applied the unidimensional framework (as implemented in Lai et al., 2017) to obtain
 292 classification accuracy indices based on the first five items and the last five items separately, and obtained
 293 the weighted averages of the two sets of indices with the same weights of .7 and .3.

294 Table 1 compares the true population-level classification accuracy indices and the mean values

295 across replications from MCAA and the unidimensional approach, with the first group as the reference
296 group under partial strict invariance models. In summary, whereas MCAA recovered the true values of the
297 indices well for both groups, using the unidimensional approach resulted in biased values of proportions
298 selected and values of other indices being smaller than the true values.

299 **Illustrative Example**

300 To illustrate the application of the multidimensional classification accuracy analysis (MCAA), we
301 used data from Ock et al. (2020), which examined measurement invariance of the mini-IPIP across gender.
302 The data was a subset of the Eugene-Springfield Community Sample collected from 1994 Spring to 1996
303 Fall (Goldberg, 2018), a well-studied community sample who completed a mail survey, including the
304 mini-IPIP. Ock et al. (2020) performed listwise deletion and provided complete data in their supplemental
305 material. The sample consisted of 564 participants (239 males, 325 females), who were 20 to 85 years old
306 ($M = 51.7$, $SD = 12.5$), and nearly all of them being Caucasian (97.7%).

307 The mini-IPIP is a short form of the International Personality Item Pool, a personality measure
308 based on the Five-Factor model (Donnellan et al., 2006; Goldberg, 1999). The mini-IPIP had 20 items in
309 total, with four items for each factor. Specifically, items A2, A5, A7, A9 measure the factor Agreeableness;
310 items C3, C4 C6, C8 measure the factor Conscientiousness; items E1, E4, E6, E7 measure the factor
311 Extraversion; items N1, N2, N6, N8 measure the factor Neuroticism; and items O2, O8, O9, O10 measure
312 the factor Openness to Experience. Further details about these items can be found in the Appendix in
313 Donnellan et al. (2006). Questions were descriptive statements answered on a 5-point Likert-type scale
314 from 1 (*very inaccurate*) to 5 (*very accurate*). Table 2 shows the means, standard deviations, and the
315 correlations of the mini-IPIP items by gender.

316 To identify noninvariant parameters, we used the *lavaan* R package (Rosseel, 2012) and the
317 forward specification search procedure (Yoon & Kim, 2014) using likelihood ratio tests ($\Delta\chi^2$).² Given the
318 categorical nature of the items, we followed Ock et al. (2020) to use the robust maximum likelihood (MLR)
319 estimator, with the scaled $\Delta\chi^2$ test by Satorra and Bentler (2001). We first fitted a configural invariance
320 model, which showed poor fit, $\chi^2(320) = 662.94$, $p < .001$, RMSEA = 0.06, 95%CI [0.06, 0.07], CFI = 0.84,
321 SRMR = 0.06. Based on the modification indices, we decided to free eight pairs of unique factor
322 covariances: A2 and A5, E4 and E7, I2 and I10, I8 and I9, A9 and I9, C3 and E6, A2 and E7, E7 and N2.
323 The modified configural invariance model with five factors (see Figure 3) showed acceptable fit,

² Jung and Yoon (2016) and Jung and Yoon (2017) are accessible introductions to other methods for identifying non-invariant parameters.

324 $\chi^2(304) = 408.96$, $p < .001$, RMSEA = 0.03, 95%CI [0.03, 0.04], CFI = 0.95, SRMR = 0.05. Equality
 325 constraints in the loadings did not result in poorer model fit, scaled $\Delta\chi^2(15) = 10.83$, $p = .764$. We then
 326 added the constraints to the intercepts, which resulted in poorer model fit, scaled $\Delta\chi^2(15) = 49.38$,
 327 $p < .001$. One item in Agreeableness (A2, “Sympathize with others’ feelings”), one item in Extraversion
 328 (E6 “Don’t talk a lot”) and two items in Neuroticism (N1, “Am relaxed most of the time”; N2, “Seldom
 329 feel blue”) showed noninvariant intercepts across groups ($\Delta\nu_{F-M} = 0.16, 0.42, 0.31, 0.24$). After freeing
 330 these items, the scalar model showed acceptable fit, $\chi^2(330) = 426.75$, $p < .001$, RMSEA = 0.03, 95%CI
 331 [0.02, 0.04], CFI = 0.95, SRMR = 0.05. A strict invariance model was further fitted to the data, which
 332 fitted the data worse than the partial scalar invariance model, scaled $\Delta\chi^2(20) = 40.65$, $p = .004$. One item
 333 in Conscientiousness (C8, “Make a mess of things”) and two items in Neuroticism (N1, “Am relaxed most
 334 of the time”; N2, “Seldom feel blue”) showed noninvariant unique factor variance across groups ($\Delta\theta_{F-M} =$
 335 $0.21, 0.28, 0.39$). The final model is a partial strict invariance model, $\chi^2(347) = 446.25$, $p < .001$, RMSEA
 336 = 0.03, 95%CI [0.02, 0.04], CFI = 0.95, SRMR = 0.06. The parameter estimates from the partial strict
 337 invariance model can be found in the supplemental material.

338 While the conventional invariance testing identified four items with noninvariant intercepts, the
 339 results did not provide information on how these noninvariant parameters may impact personnel selection
 340 using the mini-IPIP. For example, do the noninvariant intercepts give a substantial or a negligible
 341 advantage to females? Does dropping the noninvariant items improve the selection procedure? To answer
 342 these questions, we show how MCAA can be applied in a step-by-step fashion.

343 *Step 1: Selection Parameters*

344 As a first step of doing MCAA, we need to consider several parameters related to selection: (a) the
 345 mixing proportion (π_g), (b) the relative weights given to each dimension (w_k), (c) the weights given to each
 346 item (c_j), and (d) the selection cutoff, either in terms of an absolute cutoff score (Z_c) or a relative cutoff
 347 proportion (i.e., proportion selected). In this example, because the population sizes for females and for
 348 males are roughly equal, we use $\pi_1 = \pi_2 = .5$. The weights given to the items and to the different
 349 dimensions require some more considerations. If the test items are summed together to get one single scale
 350 score for selection purposes, and each dimension contains an equal number of items, then we can specify
 351 $\mathbf{w} = \mathbf{1}$ (i.e., a vector of ones). However, it is well documented in previous research (e.g., Barrick & Mount,
 352 1991) that different personality dimensions had different associations with job performance. Instead, we
 353 used the regression weights reported by Drasgow et al. (2012), which conducted a meta-analysis to examine
 354 the predictive validity of five personality dimensions in eight criteria of job performance (e.g., the

355 predictive validity of conscientiousness ranges from -0.23 to 0.20). After averaging the regression weights
356 for each dimension, we set $\mathbf{w} = [.0325, .1795, .4693, -.1951, .1236]$ for agreeableness, conscientiousness,
357 extraversion, neuroticism (with a negative weight), and openness. Note that the sum of the absolute values
358 of the weights is one. On the item side, because each dimension has the same number of items, we set the
359 item weights to be proportional to the latent weights, while keeping the maximum weighted score for each
360 participant to 100 (same as the unweighted score); specifically, $\mathbf{c} = 5 \times [w_1, w_1, w_1, w_1, \dots, w_5, w_5, w_5, w_5]$
361 for 20 items. The codes for obtaining the weights can be found in the supplemental material. For the
362 selection cutoff, we assume that the mini-IPIP is used to select the top 25% of the candidates.

363 *Step 2: Classification Accuracy Under Strict Invariance*

364 To establish the baseline information of using the mini-IPIP in selecting males and females, we
365 first obtained the parameter estimates under full strict invariance. The supplemental material contains
366 codes for extracting parameter estimates from a fitted *lavaan* model object as inputs for the MCAA;
367 however, researchers can also manually input the parameter estimates into the provided R function,
368 `PartInvMulti_we()`. Following Millsap and Kwok (2004), for the four noninvariant intercepts and the
369 three noninvariant unique factor variances, we obtained the average parameter estimates weighted by the
370 mixing proportions as parameters for the strict invariance model. We used female candidates as the
371 reference group and male candidates as the focal group. Using the selection parameters in Step 1 and the
372 R script in the supplemental material, one can obtain the selection indices when strict invariance holds, as
373 shown in Table 3. Specifically, the selection is expected to comprise slightly more female candidates
374 (25.2%) than male candidates. The other classification accuracy indices (success ratio, sensitivity, and
375 specificity) were similar for the two groups.

376 *Step 3: Classification Accuracy Under Partial Strict Invariance*

377 The selection accuracy of mini-IPIP under partial strict invariance can be obtained in the same
378 way as in Step 2, except that the intercept parameters were different for males and females, as well as the
379 unique variances and covariances. The results are again shown in Table 3. In the presence of test bias,
380 male candidates are selected in a lower proportion than female candidates (0.260 for female and 0.240 for
381 male). The selection procedure has a higher sensitivity for female than male candidates (0.758 for female
382 and 0.733 for male). However, female candidates have a lower success ratio (0.732 for female and 0.759 for
383 male) and specificity (0.907 for female and 0.923 for male) than male candidates.

384 *Step 4: Compare the Change in Classification Accuracy indices*

385 Comparing the results in Steps 2 and 3, we see male candidates are selected in a lower proportion
 386 (24.0%), whereas female candidates are selected in a higher proportion (26.0%). The increased proportion
 387 selected for female candidates due to item bias, however, results in a lower success ratio (0.732 as opposed
 388 to 0.748 under strict invariance), meaning that there are more false positives among qualified female
 389 candidates. Item bias also results in a higher sensitivity (0.758 as opposed to 0.749) and a lower specificity
 390 (0.907 as opposed to 0.915) for females. On the contrary, the lower proportion selected for male candidates
 391 results in a higher success ratio (0.759 as opposed to 0.743), lower sensitivity (0.733 as opposed to 0.742),
 392 and higher specificity (0.923 as opposed to 0.915).

393 The columns labelled $E_F(\text{Male})$ in Table 3 represent the expected classification performance for
 394 male candidates based on the latent score distributions of the female candidates. The differences between
 395 columns Female and $E_F(\text{Male})$ show the impact of item bias on classification accuracy, as they are identical
 396 when strict invariance holds. Our R function also computed the AI ratio for male candidates to be 0.935,
 397 which is the ratio of proportions selected for females and the proportions selected for $E_F(\text{Male})$ under
 398 strict invariance, that is, .243 / .260. The computed AI ratio indicates that, due to item bias, for every
 399 1,000 female candidates selected, only 935 equally qualified male candidates will be selected. Thus, it
 400 demonstrates a disadvantage for male candidates when using the mini-IPIP for selection.

401 *Comparison With Separate Unidimensional Analyses*

402 We also applied separate unidimensional analysis to each dimension of the mini-IPIP. Because
 403 Agreeableness, Conscientiousness, and Openness were not found to show partial invariance when evaluated
 404 separately, only the results of Extraversion and Neuroticism are reported in Table 4.³ Note that the results
 405 for Neuroticism were based on reversely coded items, as lower neuroticism is usually preferred in personnel
 406 selection. It can be seen that the impact of item bias on selection accuracy is larger when considering each
 407 dimension separately than the combined impacts obtained from MCAA. Consider selecting individuals
 408 solely based on extraversion. Under strict invariance, female candidates will be selected at a similar
 409 proportion (24.7%) as male candidates (25.3%), but more females will be selected (26.7%) under partial
 410 strict invariance. The reverse is true for Neuroticism, as more males will be selected under partial strict
 411 invariance. The item biases showed a particularly large impact on sensitivity. However, when considering

³ Evaluating measurement invariance for each dimension, we found one item for Extraversion (E6 “Don’t talk a lot”) had noninvariant intercepts, and two items in Neuroticism (N1, “Am relaxed most of the time”; N2, “Seldom feel blue”) had noninvariant intercepts and unique variances across gender.

412 the combined effect of item biases across all dimensions, as shown by MCAA, the impact was much
413 smaller, as the biases in Extraversion and in Neuroticism somewhat cancelled out. The sensitivity and
414 specificity also had higher values under MCAA, as the selection was generally more accurate with 20 items
415 from five dimensions as opposed to only 4 items in one dimension. Therefore, when a multidimensional test
416 is used for selection or classification purposes, using the MCAA framework provides results more closely
417 aligned to how item biases affect the actual classification decisions overall.

418 **Discussion**

419 Despite tremendous growth in the measurement invariance literature, there has been a disconnect
420 between how results of invariance testing are presented—usually in the form of statistical significance and
421 change in fit indices—and how psychological measures are used in practice for making classification
422 decisions. The work by Millsap and Kwok (2004) and Stark et al. (2004) provided a foundation for
423 understanding the impact of noninvariance on classification. Nevertheless, the previous work only
424 considered unidimensional items, but in practice, classification is usually done based on multiple
425 dimensions.

426 The current work proposes a multidimensional classification accuracy framework (MCAA), which
427 examines the change in classification accuracy indices attributable to noninvariance across demographic
428 subgroups. We conceptualize the multidimensional problem by considering the joint distribution of the
429 weighted observed composite score and the weighted latent composite score, and provide software code that
430 computes the change in classification accuracy indices due to noninvariance. We illustrate MCAA with a
431 step-by-step example of a five-dimensional personality inventory commonly used for personnel selection.

432 Readers should note that the MCAA and the foundational work by Millsap and Kwok (2004) only
433 represent one option for understanding the practical significance of noninvariance. Other effect size indices
434 exist in the literature, some of which were nicely summarized in Meade (2010), and some recent
435 development was made by Nye et al. (2019) and Gunn et al. (2020). In our opinion, information on changes
436 in classification accuracy indices is most relevant for measures that are potentially used for classification
437 purposes, such as tests of cognitive and noncognitive abilities, as well as screening and diagnostic tools.
438 When conducting invariance analysis, we encourage researchers to carefully consider the intended usage of
439 the measure being studied and report the magnitude of noninvariance (if any) in a metric that is easily
440 interpretable and fits the context in which the measure will be used.

441 While we think the MCAA represents a step closer to linking invariance research and actual
442 practices in the context of classification, we also recognize some limitations of the current work and

443 encourage future research efforts to address them. First, the current framework assumes that item
444 responses are approximately continuous; given that binary and ordinal items are commonly used in
445 psychological measures, future research can combine MCAA and the recent extension by Gonzalez and
446 Pelham (2021) and Lai et al. (2019). Second, as with previous literature, the implied classification accuracy
447 indices under the strict and the partial strict invariance models are only point estimates and are subject to
448 sampling error. Given the recommendations on reporting uncertainty estimates for measures on practical
449 significance (e.g., American Psychological Association, 2020), future work is needed to develop methods for
450 obtaining standard errors and confidence intervals for the classification accuracy indices.

References

- 451
- 452 Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2016). Differential prediction generalization in college
453 admissions testing. *Journal of Educational Psychology, 108*(7), 1045–1059.
454 <https://doi.org/10.1037/edu0000104>
- 455 American Educational Research Association, American Psychological Association, & National Council on
456 Measurement in Education. (2014). *Standards for educational and psychological testing*.
457 <https://www.testingstandards.net/open-access-files.html>
- 458 American Psychological Association. (2020). *Publication manual of the American Psychological Association*
459 (7th ed.). <https://doi.org/10.1037/000016S-000>
- 460 Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A
461 meta-analysis. *Personnel Psychology, 44*(1), 1–26.
462 <https://doi.org/https://doi.org/10.1111/j.1744-6570.1991.tb00688.x>
- 463 Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2),
464 238–246. <https://doi.org/10.1037/0033-2909.107.2.238>
- 465 Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural*
466 *Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504.
467 <https://doi.org/10.1080/10705510701301834>
- 468 Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement
469 invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233–255.
470 https://doi.org/10.1207/S15328007SEM0902_5
- 471 Crocker, L. (2006). *Introduction to classical and modern test theory*. Cengage Learning.
- 472 Donnellan, M., Oswald, F., Baird, B., & Lucas, R. (2006). The Mini-IPIP scales: Tiny-yet-effective
473 measures of the big five factors of personality. *Psychological assessment, 18*(2), 192–203.
474 <https://doi.org/10.1037/1040-3590.18.2.192>
- 475 Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development*
476 *of the Tailored Adaptive Personality Assessment System (TAPAS) to support Army personnel*
477 *selection and classification decisions* (Technical Report 1311). U.S. Army Research Institute for the
478 Behavioral and Social Sciences. <https://apps.dtic.mil/sti/citations/ADA564422>
- 479 Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the
480 lower-level facets of several five-factor models. *Personality psychology in Europe, 7*(1), 7–28.
- 481 Goldberg, L. R. (2018). (*2,8,10 & others*) *International Personality Item Pool (IPIP)* (Version V1) [Data
482 set]. Harvard Dataverse. <https://doi.org/10.7910/DVN/UF52WY>

- 483 Gonzalez, O., & Pelham, W. E. (2021). When does differential item functioning matter for screening? A
484 method for empirical evaluation. *Assessment*, *28*(2), 446–456.
485 <https://doi.org/10.1177/1073191120913618>
- 486 Gunn, H. J., Grimm, K. J., & Edwards, M. C. (2020). Evaluation of six effect size measures of
487 measurement non-Invariance for continuous outcomes. *Structural Equation Modeling: A
488 Multidisciplinary Journal*, *27*(4), 503–514. <https://doi.org/10.1080/10705511.2019.1689507>
- 489 Horn, J. L., & Mcardle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging
490 research. *Experimental Aging Research*, *18*(3), 117–144.
491 <https://doi.org/10.1080/03610739208253916>
- 492 Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of
493 Applied Psychology*, *85*(6), 869–879. <https://doi.org/10.1037/0021-9010.85.6.869>
- 494 Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance:
495 Forward, backward, and factor-ratio tests. *Structural Equation Modeling: A Multidisciplinary
496 Journal*, *23*(4), 567–584. <https://doi.org/10.1080/10705511.2015.1138092>
- 497 Jung, E., & Yoon, M. (2017). Two-step approach to partial factorial invariance: Selecting a reference
498 variable and identifying the source of noninvariance. *Structural Equation Modeling: A
499 Multidisciplinary Journal*, *24*(1), 65–79. <https://doi.org/10.1080/10705511.2016.1251845>
- 500 Lai, M. H. C., Kwok, O.-M., Yoon, M., & Hsiao, Y.-Y. (2017). Understanding the impact of partial
501 factorial invariance on selection accuracy: An R script. *Structural Equation Modeling: A
502 Multidisciplinary Journal*, *24*(5), 783–799. <https://doi.org/10.1080/10705511.2017.1318703>
- 503 Lai, M. H., Richardson, G. B., & Mak, H. W. (2019). Quantifying the impact of partial measurement
504 invariance in diagnostic research: An application to addiction research. *Addictive Behaviors*, *94*,
505 50–56. <https://doi.org/10.1016/j.addbeh.2018.11.029>
- 506 Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and
507 scales. *The Journal of Applied Psychology*, *95*(4), 728–743. <https://doi.org/10.1037/a0018966>
- 508 Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational
509 Research*, *13*(2), 127–143. [https://doi.org/10.1016/0883-0355\(89\)90002-5](https://doi.org/10.1016/0883-0355(89)90002-5)
- 510 Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*,
511 *58*(4), 525–543. <https://doi.org/10.1007/BF02294825>
- 512 Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*(4), 461–473.
513 <https://doi.org/10.1007/s11336-007-9039-7>
- 514 Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

- 515 Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in
516 two populations. *Psychological Methods, 9*(1), 93–115. <https://doi.org/10.1037/1082-989X.9.1.93>
- 517 Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? examining
518 the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research*
519 *Methods, 22*(3), 678–709. <https://doi.org/10.1177/1094428118761122>
- 520 Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence:
521 Understanding the practical importance of differences between groups. *Journal of Applied*
522 *Psychology, 96*(5), 966–980. <https://doi.org/10.1037/a0022955>
- 523 Ock, J., McAbee, S. T., Mulfinger, E., & Oswald, F. L. (2020). The practical effects of measurement
524 invariance: Gender invariance in two Big Five personality measures. *Assessment, 27*(4), 657–674.
525 <https://doi.org/10.1177/1073191119885018>
- 526 Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state
527 of the art and future directions for psychological research. *Developmental Review, 41*, 71–90.
528 <https://doi.org/10.1016/j.dr.2016.06.004>
- 529 Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software,*
530 *48*(2), 1–36. <http://www.jstatsoft.org/v48/i02/>
- 531 Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure
532 analysis. *Psychometrika, 66*(4), 507–514. <https://doi.org/10.1007/BF02296192>
- 533 Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel
534 psychology: Practical and theoretical implications of 85 years of research findings. *Psychological*
535 *Bulletin, 262*–274.
- 536 Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant
537 and nonapplicant populations. *Journal of Applied Psychology, 78*(6), 966–974.
538 <https://doi.org/10.1037/0021-9010.78.6.966>
- 539 Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human*
540 *Resource Management Review, 18*(4), 210–222. <https://doi.org/10.1016/j.hrmr.2008.03.003>
- 541 Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item
542 (functioning and differential) test functioning on selection decisions: When are statistically
543 significant effects practically important? *Journal of Applied Psychology, 89*(3), 497–508.
544 <https://doi.org/10.1037/0021-9010.89.3.497>
- 545 Steiger, J. H. (1980). Statistically based tests for the number of common factors. *the annual meeting of the*
546 *Psychometric Society. Iowa City, IA. 1980*. Retrieved March 8, 2021, from
547 <https://ci.nii.ac.jp/naid/10012870999/>

- 548 Thurstone, L. (1947). *Multiple factor analysis*. University of Chicago Press.
- 549 Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance
550 methods and procedures. *Organizational Research Methods*, 5(2), 139–158.
551 <https://doi.org/10.1177/1094428102005002001>
- 552 Yoon, M., & Kim, E. S. (2014). A comparison of sequential and nonsequential specification searches in
553 testing factorial invariance. *Behavior Research Methods*, 46(4), 1199–1206.
554 <https://doi.org/10.3758/s13428-013-0430-2>
- 555 Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF
556 analysis. *Journal of Educational Measurement*, 36(1), 1–28.
557 <https://doi.org/10.1111/j.1745-3984.1999.tb00543.x>

Table 1*Simulation Results*

	Group 1				Group 2			
	PS	SR	SE	SP	PS	SR	SE	SP
Population value	0.236	0.822	0.774	0.944	0.264	0.781	0.826	0.923
MCAA	0.236	0.821	0.774	0.944	0.264	0.780	0.826	0.923
UCAA	0.239	0.800	0.762	0.936	0.261	0.767	0.803	0.919

Note. UCAA = Unidimensional classification accuracy analysis applied separately to the two dimensions. PS = proportion selected. SR = success ration. SR = success ratio SE = sensitivity SP = specificity

Table 2
Mean, Standard Deviations, and Item-Level Correlations of the Mini-IPIP Scales by Gender.

	Female		Male																					
	Mean	SD	Mean	SD	a2	a5	a7	a9	c3	c4	c8	c9	e1	e4	e6	e7	n1	n2	n6	n8	i2	i8	i9	i10
a2	4.34	0.73	3.93	0.87	1	.29	.18	.20	-.02	-.03	.10	.07	-.00	.06	.03	-.14	-.02	-.01	.06	-.03	-.01	.01	-.04	.01
a5	4.07	0.84	3.53	1.02	.51	1	.33	.39	.04	-.10	.02	-.06	.12	.20	.13	.05	-.08	.00	.09	-.02	.19	.11	.14	.14
a7	4.45	0.83	4.04	0.90	.30	.45	1	.44	-.02	-.12	.14	.06	.08	.19	.08	.04	-.07	-.14	-.08	-.13	.13	.11	.13	.15
a9	4.18	0.82	3.61	0.96	.32	.51	.56	1	-.01	-.13	.03	-.03	.01	.09	.00	.01	-.12	-.06	-.08	-.13	.06	.05	.07	.09
c3	3.37	1.22	3.28	1.09	.06	.08	.10	.01	1	.31	.31	.39	.05	.05	-.02	-.01	-.07	-.12	-.04	-.17	-.11	-.01	-.10	-.00
c4	4.32	0.77	4.26	0.71	.04	-.01	.04	-.06	.22	1	.20	.35	.02	-.05	.00	.06	.10	-.07	.05	-.03	-.10	-.10	-.08	-.12
c8	4.23	0.97	4.18	0.81	.05	-.03	.09	.05	.21	.27	1	.33	.13	.13	.07	.13	-.02	-.20	-.10	-.26	-.12	.01	-.04	-.05
c9	3.70	1.32	3.61	1.22	.00	.01	-.01	-.13	.32	.41	.30	1	.04	.05	-.01	.01	-.03	-.13	-.04	-.10	-.05	.07	.02	-.03
e1	2.22	1.15	2.31	1.03	.11	.16	.10	.11	.01	-.08	-.10	.03	1	.38	.37	.40	-.11	-.11	.04	-.03	.14	.10	.14	.15
e4	2.94	1.30	2.88	1.24	.04	.18	.25	.16	.13	.02	.02	.10	.41	1	.38	.34	-.07	-.11	.00	-.09	.10	.08	.06	.17
e6	3.41	1.24	3.01	1.21	.12	.19	.34	.29	-.12	-.08	-.01	-.01	.35	.40	1	.39	.01	-.03	.05	-.09	.14	.11	.18	.19
e7	3.08	1.02	3.09	1.08	.07	.06	.18	.12	.07	-.03	.10	-.00	.42	.35	.48	1	-.06	-.22	-.02	-.16	.07	.07	.12	.17
n1	2.54	1.11	2.26	0.97	-.15	-.05	.00	.05	-.04	-.02	-.10	.06	.09	-.03	.11	.04	1	.27	.37	.33	-.07	.09	.02	-.10
n2	2.66	1.32	2.46	1.12	-.24	-.08	-.15	-.09	-.09	.02	-.17	-.03	-.02	-.12	-.01	-.06	.41	1	.41	.49	-.04	.04	-.02	-.10
n6	2.31	1.08	2.30	1.05	-.19	-.09	-.14	-.14	-.13	.04	-.18	-.02	.08	-.03	.12	-.01	.31	.40	1	.41	.00	-.04	-.04	-.03
n8	2.23	1.20	2.29	1.15	-.14	-.03	-.09	-.12	-.12	-.00	-.23	.01	-.05	-.11	.01	-.04	.39	.54	.41	1	.08	.06	.02	-.02
i2	3.76	1.13	3.94	0.97	.08	.15	.21	.13	-.02	.02	-.06	.00	.26	.17	.31	.25	-.00	.02	.03	.09	1	.26	.32	.63
i8	3.35	1.26	3.65	1.13	.10	.02	.08	.11	.01	-.07	.06	-.06	.09	.19	.22	.28	.02	-.03	-.14	-.08	.23	1	.65	.24
i9	3.38	1.23	3.62	1.09	.17	.17	.19	.32	-.06	-.02	-.01	-.13	.08	.14	.21	.16	-.00	.01	-.15	-.00	.26	.50	1	.34
i10	3.81	1.18	4.07	1.05	.06	.08	.20	.14	.05	.07	.02	.03	.22	.17	.26	.27	.00	-.04	-.12	-.02	.57	.32	.26	1

Note. The item-level correlations of males (females) are shown in the lower (upper) triangle.

Table 3

Impact of Item Bias on Selection Accuracy Indices From the Multidimensional Classification Accuracy Analysis

	Strict Invariance			Partial Strict Invariance		
	Female	Male	$E_F(\text{Male})$	Female	Male	$E_F(\text{Male})$
Proportion selected	0.252	0.248	0.252	0.260	0.240	0.243
Success ratio	0.748	0.743	0.748	0.732	0.759	0.764
Sensitivity	0.749	0.742	0.749	0.758	0.733	0.739
Specificity	0.915	0.915	0.915	0.907	0.923	0.923

Note. The column $E_F(\text{Male})$ shows the expected values for male candidates using the latent distributions of female candidates.

Table 4

Impact of Item Bias on Selection Accuracy Indices When the Dimensions Are Considered Separately.

	Strict Invariance		Partial Strict Invariance	
	Female	Male	Female	Male
Extraversion				
Proportion selected	0.247	0.253	0.267	0.233
Success ratio	0.714	0.715	0.691	0.738
Sensitivity	0.715	0.714	0.748	0.680
Specificity	0.906	0.903	0.891	0.918
Neuroticism (Reverse coded)				
Proportion selected	0.261	0.239	0.238	0.262
Success ratio	0.741	0.708	0.756	0.693
Sensitivity	0.734	0.717	0.683	0.768
Specificity	0.908	0.909	0.921	0.895

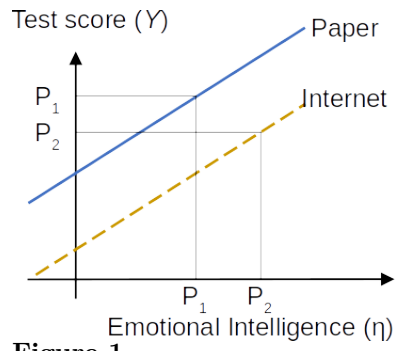


Figure 1

Example of scalar non-invariance where a participant taking a paper test is mistakenly given a lower score.

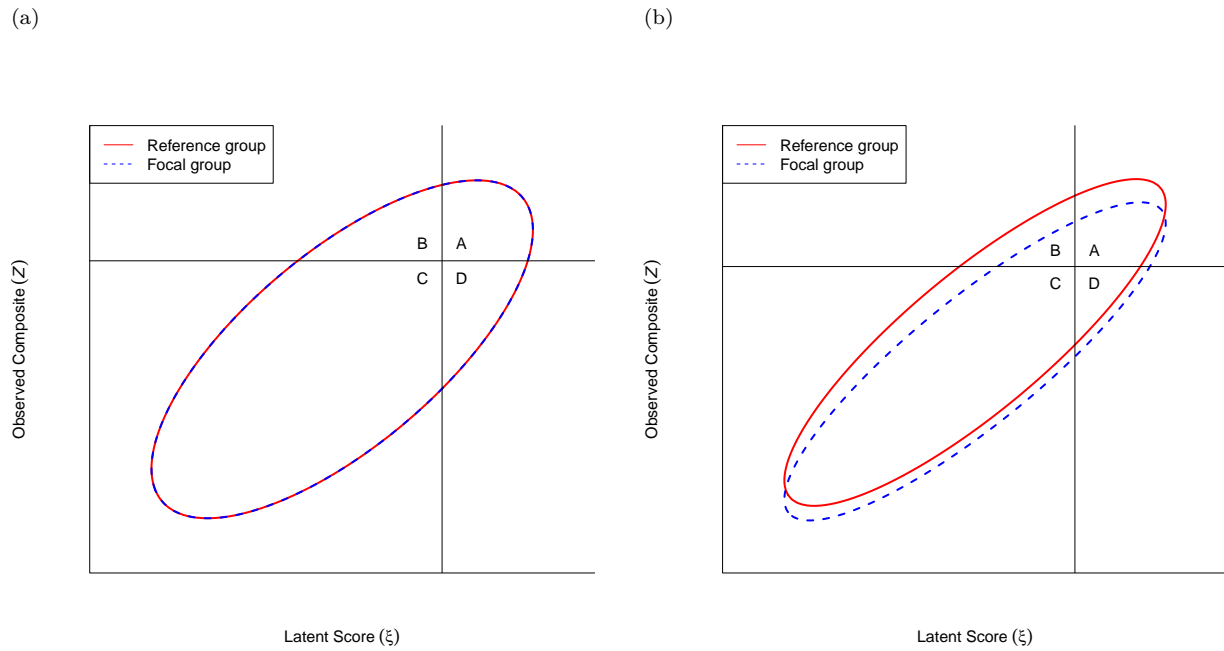


Figure 2

Relationship between true latent construct scores (x-axis) and observed test scores (y-axis) for (a) a test with no item bias and (b) a test with biases against one subgroup. Quadrants A, B, C, and D indicates the proportions of true positives, false positives, true negatives, and false negatives.

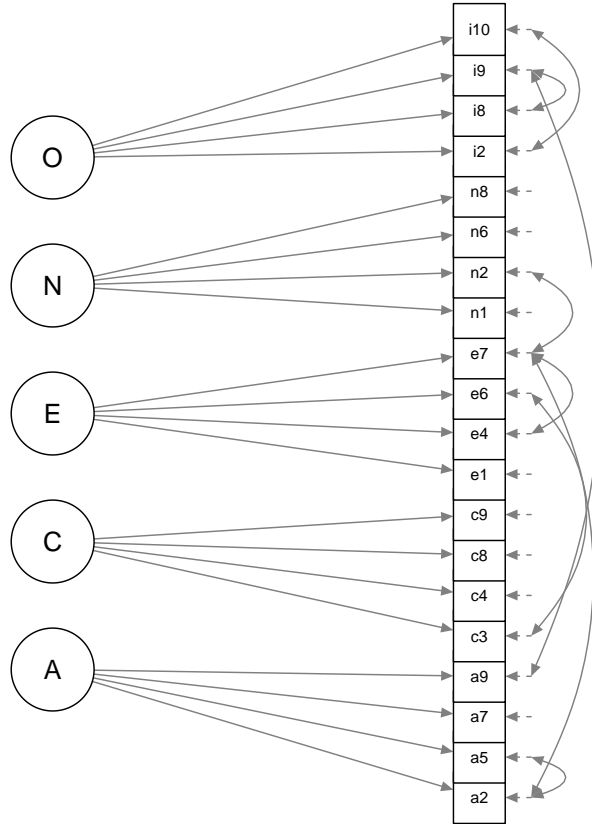


Figure 3

Path diagram of the factor model for the mini-IPIP items in the illustrative example.