Classification Accuracy of Multidimensional Tests: Quantifying the Impact of Noninvariance

Yichi Zhang[1], Mark H. C. Lai[1]

[1] University of Southern California

Abstract

There has been tremendous growth in research on measurement invariance over the past two decades. However, given that psychological tests are commonly used for making personnel selection decisions, surprisingly there has been little research on how noninvariance impacts selection accuracy. Millsap & Kwok (2004) proposed a selection accuracy framework for that purpose for unidimensional tests. However, selection is usually based on multidimensional tests (e.g., personality) or multiple tests, with different weights assigned to each dimension. In the current project, we extend Millsap & Kwok's framework for examining the impact of noninvariance to a multidimensional test on selection. This multidimensional framework is implemented in R and illustrated with an example of selection using data from a published report featuring a five-factor personality inventory.

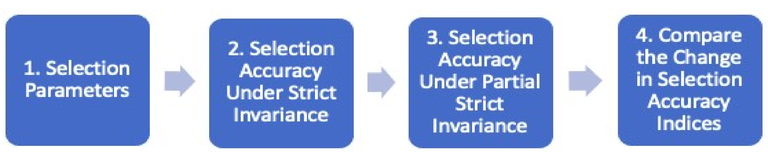
Introduction

- Psychological tests are commonly used for making selection decisions. However, tests are far from perfect, and items may suffer from systematic bias (i.e., *noninvariance*). Item bias = noninvariance = differential item functioning (DIF) can lead to devastating consequences.
- Previous literature mainly focuses on identifying non-invariant items and there is a dearth of studies on the practical impact of non-invariance to group differences in the latent construct level or score.
- Millsap & Kwok (2004) proposed the selection accuracy analysis framework, which allows researchers to evaluate the impact of item bias on selection accuracy indices, such as sensitivity and specificity.
- However, real-life selection is likely a decision based on multiple tests or subtests, such as personality tests which often based on multiple dimensions. Thus, there is a need to extend this single dimension framework to multiple dimensions.

Methods

- We extend the selection accuracy analysis framework to multidimensions and defined it as *Multidimensional Classification Accuracy Analysis* (MCAA) Framework. We also incorporate the adverse impact (AI) ratio (Nye & Drasgow, 2011), and implemented this framework to R.
- Under standard assumptions, the observed composite score and the weighted latent composite score follow a bivariate normal distribution. Following the derivation in Millsap & Kwok (2004), the selection accuracy indices can be calculated based on this joint distribution.

Steps for MCAA Framework



Empirical Example

- Goal: We reanalyzed a study done by Ock (2020), which evaluated the measurement invariance of the Mini-International Personality Item Pool (Mini-IPIP; Donnellan et al, 2006) across gender. Our goal is to apply the MCAA framework and examine the impact of noninvariance on selection.

Dataset descriptions

- The mini-IPIP is a personality measure based on the Five-Factor model (Donnellan et al., 2006). This scale has 20 items in total, with four items for each factor. Questions were descriptive statements answered on a 5-point Likert-type scale from 1 (very inaccurate) to 5 (very accurate). The sample consisted of 564 participants (239 males, 325 females), who were 20 to 85 years old (M = 51.7, SD = 12.5), and nearly all of them being Caucasian (97.7%).
- The conventional measurement invariance analysis showed:
  - all loadings were invariant
  - four items with noninvariant intercepts (female – male)
    - A2 (Agreeableness): $\Delta\nu = 0.157$
    - E6 (Extraversion): $\Delta\nu = 0.415$
    - N1 (Neuroticism): $\Delta\nu = 0.308$
    - N2 (Neuroticism): $\Delta\nu = 0.240$
  - three items with noninvariant uniqueness
    - N1(Neuroticism), $\Delta\theta = 0.281$
    - N2(Neuroticism), $\Delta\theta = 0.389$
    - C8(Conscientiousness), $\Delta\theta = 0.210$

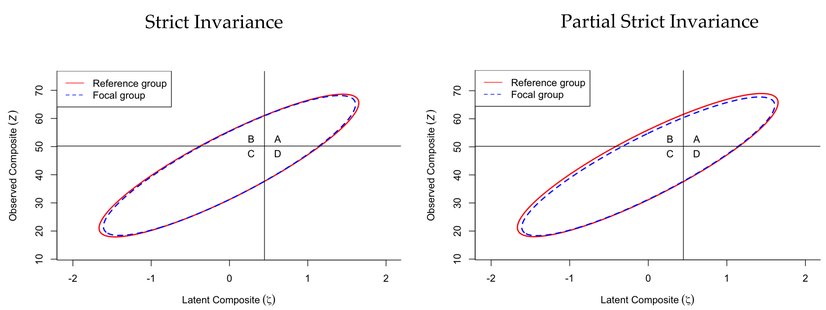
Step 1: Selection Parameters

- We set mixing proportions at 0.5 and assume that the mini-IPIP is used to select/screen top 25%
- Latent factor weights based on the prediction regression weights reported by Drasgow et al. (2012)

Table 1:
Weights for the Big Five Dimensions

	Latent weight	Item weight
Agreeableness	0.033	0.813
Conscientiousness	0.180	4.488
Extraversion	0.470	11.733
Neuroticism	-0.195	-4.878
Openness	0.123	3.090

Step 2 & 3: Selection Accuracy under Strict Invariance and Partial Strict Invariance



- Proportion selected (PS) = A + B
- Success Ratio (SR) = A/(A + B)
- Sensitivity (SE) = A/(A + D)
- Specificity (SP) = C/(B + C)

Table 2:
Impact of Item Bias on Selection Accuracy Indices

	Female	**Male**	E_F(Male)	**Female**	**Male**	E_F(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

- Female candidates would be selected in a slightly higher proportion compared to male candidates, and this gender difference is larger when partial strict invariance holds.

Step 4: Compare the Change in Selection Accuracy indices

- 0.8% more females being selected due to noninvariance (more false positives)
- Lower sensitivity for males
- Adverse impact ratio = 0.93, indicating slight disadvantage for male candidates

Discussion

- MCAA Framework
  - Extend the selection accuracy framework to multiple dimensions
  - Evaluate the impact of item bias on a practically meaningful metric
  - Incorporate recently developed effect size index
- Future Directions
  - Extend to categorical items
  - Account for sampling variability

References

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9, 93–115

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

Ock, J., McAbee, S. T., Mulfinger, E., & Oswald, F. L. (2020). The Practical Effects of Measurement Invariance: Gender Invariance in Two Big Five Personality Measures. *Assessment*, *27*(4), 657–674. https://doi.org/10.1177/1073191119885018

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. ***Correspondence concerning this presentation should be address to Yichi Zhang and Mark Lai, yzhang97@usc.edu, hokchiol@usc.edu.***