Methodological and Statistical Advances in the Consideration of

Cultural Diversity in Assessment:   
A Critical Review of Group Classification and Measurement Invariance Testing

**Online Supplements**

**S1: True score**

A psychological construct is an unobservable theoretical variable that must be inferred from observable behaviors; an individual’s standing on a psychological construct is often estimated from responses to test items that assess specific thoughts, feelings, behaviors, or symptoms (Hopwood & Donnellan, 2010). An observed score comprises a true score and an error score. The true score reflects an individual’s true standing on the attribute that the test is intended to assess; it is an unobserved score that can only be estimated. The error score can reflect random or systematic error: random error is usually the product of a single testing situation (e.g., sickness, guessing, lack of sleep), whereas systematic error can reflect attributes of the person (e.g., test anxiety, reading speed, acquiescence, social desirability) or the test (e.g., poorly worded questions) that are present across administrations.

There are several ways to estimate the true score. Use of some advanced psychometric techniques, such as confirmatory factor analysis (CFA) is premised on the idea that an individual’s standing on a latent construct can be estimated from multiple indicators (sometimes called manifest variables) of the construct. Although each indicator may alone be a flawed measure of the construct, an aggregated set of such indicators can produce a useful assessment of the latent construct (Davidov et al., 2014). In a confirmatory factor analysis framework, an individual’s score on each indicator (“observed score”) is assumed to be a combination of a latent factor score (“true score”) and a measurement error score. The true score is estimated based on the relationships among the indicators.

**S2: Definition of test fairness**

According to the *Standards* (2014) set by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME),

“A test that is fair within the meaning of the *Standards* reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct. To the degree possible, characteristics of all individuals in the intended test population, including those associated with race, ethnicity, gender, age, socioeconomic status, or linguistic or cultural background, must be considered throughout all stages of development, administration, scoring, interpretation, and use so that barriers to fair assessment can be reduced. At the same time, test scores must yield valid interpretations for intended uses, and different test contexts and uses may call for different approaches to fairness. (p. 50).”

**S3**: **Definition of race and ethnicity**

There are numerous definitions of race and ethnicity. We present the definitions by Markus (2008), which are based on theoretical and empirical integration.

*“Race* is a dynamic set of historically derived and institutionalized ideas and practices that (1) sorts people into ethnic groups according to perceived physical and behavioral human characteristics; (2) associates differential value, power, and privilege with these characteristics and establishes a social status ranking among the different groups; and (3) emerges (a) when groups are perceived to pose a threat (political, economic, or cultural) to each other’s world view or way of life; and/or (b) to justify the denigration and exploitation (past, current, or future) of, and prejudice toward, other groups. *Ethnicity* is a dynamic set of historically derived and institutionalized ideas and practices that (1) allows people to identify or to be identified with groupings of people on the basis of presumed (and usually claimed) commonalities including language, history, nation or region of origin, customs, ways of being, religion, names, physical appearance, and/or genealogy or ancestry; (2) can be a source of meaning, action, and identity; and (3) confers a sense of belonging, pride, and motivation. (p. 654).”

**S4**: **Definition of sex and gender**

“*Sex* refers to a person’s biological status and is typically categorized as male, female, or intersex (i.e., atypical combinations of features that usually distinguish male from female). There are a number of indicators of biological sex, including sex chromosomes, gonads, internal reproductive organs, and external genitalia. *Gender* refers to the attitudes, feelings, and behaviors that a given culture associates with a person’s biological sex. Behavior that is compatible with cultural expectations is referred to as gender normative; behaviors that are viewed as incompatible with these expectations constitute gender nonconformity.” (APA, 2012, p. 11).

**S5: Mean difference and test bias issue**

Laypeople and some researchers have sometimes viewed mean score differences as evidence of test bias (e.g., Warne et al., 2014). This erroneous view is sometimes termed the *egalitarian fallacy*, which entails the assumption that all groups are equal in the characteristics measured by a test, and measured group differences must therefore result from bias. This interpretation of test bias is unanimously rejected by experts on testing (e.g., AERA, APA, & NCME, 2014; SIOP, 2018) because score differences between groups may reflect true differences (e.g., Clarizio, 1979).

**S6:** **Predictive invariance issues**

Much of the research examining predictive bias has employed the technique of moderated multiple regression (Mattern & Patterson, 2013), in which predictor variables and their interaction term are entered into the prediction model sequentially. At each step, change in *R2* (*ΔR2*) is examined to make decisions about the presence and magnitude of the two types of predictive bias (i.e., intercept and slope bias; Clearly, 1968). For example, an outcome measure (e.g., supervisor ratings of employee risk-taking behaviors at work) is regressed onto a predictor measure (e.g., scores on a self-report measure of antisocial behavior) in step 1, and onto a group membership variable (e.g., gender or ethnicity) in step 2. If group membership accounts for significant predictive variance over and above the predictor measure (*ΔR2*), test scores are said to exhibit intercept bias. The product of the group membership variable and the predictor variable (which carries the effect of their interaction) is then entered into the regression equation in step 3. If this new term explains a significant amount of incremental variance in the criterion variable, the test scores are said to exhibit slope bias (Clearly, 1968).

The prediction bias of scores on cognitive ability tests has been widely investigated in educational and clinical assessment contexts (e.g., Berry, Clark, & McClure, 2011; Roth, Bevier, Bobko, Switzer, & Tyler, 2001). For example, a representative finding in the educational research literature is that the regression line derived on a combined sample or on a White subsample generally over-predicts college GPA somewhat for African American SAT examinees. As over-prediction is associated with *more favorable* outcomes for the relatively disadvantaged group (i.e., higher likelihood of college admission for African Americans compared to their White counterparts producing the same SAT score), researchers generally do not express much concern about the observed (modest) intercept bias (Clarizio, 1979; Kaplan & Saccuzzo, 2009). In a classic study of prediction bias in the clinical assessment literature, Arbisi, Ben-Porath, and McNulty (2002) used a modified moderated multiple regression with scores on the MMPI-2 (the Minnesota Multiphasic Personality Inventory–2, Butcher et al., 1989) from African American and Caucasian psychiatric inpatients predicting conceptually relevant clinical criteria. Results provided evidence of bias for scores on some scales that slightly under-predicted psychopathology in African Americans, but the authors interpreted the magnitude of this prediction bias as not clinically significant.

**S7:** See Chun, Stark, Kim, and Chernyshenko (2016) for a comparison between MGCFA and MIMIC methods.

**S8**: **Brief summary of Chen (2008)’s results**

When factor loadings for the predictor measure were higher for the reference group (culture where the measure was developed, e.g., U.S.) than in the focal group (culture where the measure was imported, e.g., China), the predictive relationship (e.g., stress predicting depression) was artificially stronger in the focal group but weaker in the reference group, creating a bogus interaction effect of predictor by group (e.g., stress by culture). This result would imply that although the stress measure is better for the U.S. group than for the Chinese group (factor loadings being higher for the U.S. group), the relationship between stress and depression is weaker for the U.S. group than for the Chinese group. When the reference group had higher loadings on a criterion measure, however, the opposite pattern was found: predictive validity was weaker for the Chinese group than for the U.S. group.

Table S1

*Key Publications on Practices and Guidelines and involving the Evaluation of Measurement Invariance (MI) via Multi-group Confirmatory Factor Analysis and Related Statistical Methods*

|  |  |
| --- | --- |
| Publication | Brief Description |
| Asparouhov & Muthén (2009) | Introduction to exploratory structural equation modeling (ESEM), a method that allows complex factor model (e.g., cross-loadings) |
| Asparouhov & Muthén (2014) | Introduction to alignment optimization, a method that is useful for testing MI with a large number of groups |
| Bauer (2017) | Introduction to moderated nonlinear factor analysis, a flexible method for evaluating MI and differential item functioning with categorical (e.g., sex) and/or continuous (e.g., age) group variables |
| Byrne, Shavelson, & Muthén (1989) | Introduction to partial measurement invariance |
| Byrne & Stewart (2006)  Chen, Sousa, & West (2005)  Cheung (2008) | Overview of MI testing of second-order factor models |
| Chen (2007, 2008) | Demonstration of the impact of degrees and patterns of invariance on fit indices, slopes, and means |
| Cheung & Lau (2012) | Introduction to “Bias-corrected bootstrap confidence intervals,” an efficient way to identify referent items and noninvariance using bootstrap confidence intervals |
| Cheung & Rensvold (1999) | Discussion of the importance of referent items in testing MI and introduction of the factor-ratio test to identify referent items |
| Cheung & Rensvold (2002) | Proposal for the use of change in fit indices (e.g., ΔCFA ≤ -.01) in MI testing as an alternative to Δχ2 (continuous indicators, maximum likelihood estimator) |
| Greiff & Scherer (2018) | Accessible and brief summary of MI testing by the editor of *European Journal of Psychological Assessment* |
| Hildebrandt, Lüdtke, Robitzsch, Sommer, & Wilhelm, (2016) | Introduction to local structural equation, a method to examine MI across continuous variables |
| Hirschfeld & von Brachel (2014) | Tutorial for conducting MGCFA using the R packages lavaan, semTools, and semPlot |
| Jung & Yoon (2017) | Proposal of alternate methods for identifying referent items |
| Khojasteh & Lo (2015) | Discussion of cut-off criteria for noninvariance in bifactor models |
| Kim, Cao, Wang, & Nguyen (2017) | Overview of five approaches to evaluation of MI when comparing many groups (see Table 1s in their supplementary materials) |
| Lai & Yoon (2015) | Proposal of a modified CFI index |
| Marsh, Guo, Parker, Nagengast, Asparouhov, Muthén, & Dicke (2017) | Introduction to the extended alignment method, “Alignment within CFA” (AwC) |
| Marsh, Morin, Parker, & Kaur (2014) | Introduction to the exploratory structural equation model within CFA (EwC) |
| Marsh, Muthén, Asparouhov, Lüdtke, Robitzsch, Morin, & Trautwein (2009) | Introduction of a 13 category taxonomy for evaluating measurement invariance using exploratory structural equation modeling (ESEM) |
| Mead & Bauer (2007) | Examination of factors that affect precision of factor loadings and power |
| Meredith (1993) | Discussion of magnitudes of invariance: weak, strong, and strict |
| Millsap (2011) | Heavily quantitative reference book summarizing and integrating various approaches to testing MI |
| Muthén & Asparouhov (2013) | Introduction to Bayesian structure equation modeling (BSEM), a method that allows researchers to incorporate prior knowledge of parameters |
| Muthén & Asparouhov (2017) | Introduction to multilevel factor analysis, allowing MI testing with many groups; groups considered as random |
| Nye & Drasgow (2011) | Discussion of effect size indices in MI testing |
| Pendergast, von der Embse, Kilgus, & Erlund (2017) | Demonstration of MI testing of ordered categorical indicators in a school psychology context |
| Putnick & Bornstein (2016) | Step by step description of MGCFA procedures in nontechnical language, with helpful suggestions for various issues involving MGCFA |
| Sass, Schmitt, & Marsh (2014) | Proposal of guidelines for invariance criteria for ordered categorical indicators |
| Vandenberg & Lance (2000)  Vandenberg (2002) | First major studies providing a comprehensive and integrative paradigm for conducting sequences of MI tests |
| Yuan & Chan (2016) | Discussion of misuse of chi-square difference tests |