STRUCTURAL EQUATION MODELING
2025, VOL. 32, NO. 4
ISSN: 1070-5511 | E-ISSN: 1532-8007

OPEN ACCESS

f_{MACS}: Generalizing d_{MACS} Effect Size for Measurement Noninvariance with Multiple Groups and Multiple Grouping Variables

Mark H. C. Lai, Yichi Zhang, Meltem Ozcan, Winnie Wing-Yee Tse, and Alexander Miles

University of Southern California

ABSTRACT

With the increase in empirical studies evaluating measurement invariance of psychological measures, psychometricians and methodologists have called for more attention to effect size when invariance is violated—that is, the practical significance of noninvariance. However, most existing effect sizes are limited to two-group designs. As researchers increasingly explore invariance across more than two groups (e.g., race and ethnicity) and multiple grouping variables (e.g., race and gender), we propose the f_{MACS} (and f^2_{MACS}) statistic as a natural extension to the d_{MACS} effect size by Nye and Drasgow (2011) for multiple groups, similar to how Cohen's f is an extension of Cohen's d for standardized mean differences of more than two groups. Using two empirical examples, we illustrate how f_{MACS} can be computed using parameter estimates of a partial invariance model, and show how f_{MACS} can quantify noninvariance due to both main effects and interactions when there are multiple grouping variables. We also provide suggestions for improving reporting practices of measurement invariance analyses to facilitate the computation of effect sizes in secondary research, making it easier to evaluate the replicability of invariance studies and synthesize findings from such studies.

KEYWORDS

Contrast; effect size; measurement invariance; factor analysis; differential item functioning

1. Introduction

Measurement invariance (MI) is achieved when an instrument measures a construct in an equivalent way across people or objects, settings, and time. In science, MI is a critical condition for quantitative analysis, as any statistical conclusions may be erroneous when confounded with nonequivalent measurement. For example, previous studies have found that female participants tended to report more crying symptoms in a depression screening test than male participants, even when participants were matched on other depressive symptoms (e.g., Cole et al., 2000). Kush et al. (2023), in a synthesis of five studies, found the item "Fights" in a teacher observation instrument of aggressive-disruptive behavior was endorsed more often for males than for females, when the latent trait was held constant. Dong and Dumas (2020) conducted a systematic review of 75 articles studying MI of personality measures across cultural/ethnic groups, and found that none of the articles achieved MI in the strict sense, highlighting the need for caution when interpreting cross-cultural comparisons of personality measures.

When MI is violated, observed scores of an instrument may not be directly comparable across groups, and the differences in observed scores may not reflect true differences in the latent construct of interest. This could lead to not only spurious group mean differences (e.g., Borsboom, 2006) or interaction effects (e.g., Hsiao & Lai, 2018), but also reduced statistical power (e.g., Tse et al., 2024).

While there has been tremendous growth in social and behavioral research assessing MI of psychological measures (e.g., Putnick & Bornstein, 2016), traditional invariance testing only permits binary conclusions of whether MI holds, and provides little information regarding the magnitude of noninvariance (i.e., violation of MI). As with any other statistical procedure, in MI analysis, effect size measures provide information regarding the *practical significance* of inferential results (American Educational Research Association, 2006; Appelbaum et al., 2018). For example, when the sample size is large, a statistically significant violation of invariance may correspond to negligible differences in the obtained scores (e.g., Lai et al., 2019; van Dijk et al., 2022), whereas in small samples, sizable violations of MI may not be statistically significant.

Previous systematic reviews have highlighted the lack of replicability of MI findings, and the inattention to effect size for measurement noninvariance is a potential cause (Dong & Dumas, 2020; Zhang, 2022). For example, in a systematic review, Dong and Dumas (2020) reported extremely large cross-study variations in the invariance conclusions of personality measures across culture, gender and age. In another meta-analysis of 32 articles evaluating gender invariance of the Center for Epidemiological Studies Depression (CES-D) scale, Zhang (2022) found that statistical conclusions were drastically different across studies, and each of the 20 items in CES-D was found noninvariant in at least one study. Both Dong and Dumas (2020) and Zhang (2022) reported that only a very small portion of the studies reported effect sizes or contained sufficient information to compute effect sizes, an observation consistent with the systematic review by Putnick and Bornstein (2016) and a more recent review by Maassen et al. (2023). The lack of effect size reporting in MI studies, therefore, makes it hard to assess the severity of the apparent "replication crisis" in those studies.

1.1. Measurement Invariance Under the Common Factor Model

Before the discussion of effect sizes for MI, we briefly review the common factor model, which is the basis for most MI applications. Readers can consult Millsap (2011) for a more detailed discussion. The common factor model assumes that, for a set of p indicators $Y_1, Y_2, \ldots, Y_p$ measuring q latent variables $\eta_1, \ldots, \eta_q$, where usually $q < p$, most of the shared variance among the indicators can be explained by the latent variables, while each indicator also has a unique component. The local independence assumption states that the unique components of different indicators are not interdependent, so that the latent variables are the only sources of shared variances among the items. For the current manuscript, we focus on unidimensional models with $q = 1$, although the concepts can be extended to multidimensional models.

Each indicator Y_j is linked to η by a measurement model, $P(Y_j) = f_Y(\eta; \boldsymbol{\omega}_j)$, where $\boldsymbol{\omega}_j$ is a vector of measurement parameters. The measurement model can be linear (e.g., linear factor analysis) or nonlinear (e.g., ordinal factor analysis and item response theory), and different procedures have been developed for assessing MI with different measurement models (Leitgöb et al., 2023; Somaraju et al., 2022; Vandenberg, 2002).

Typically, the steps for assessing MI with continuous indicators include sequentially testing the four MI levels of configural, metric, scalar, and strict invariance. For ordered categorical items, similar steps are usually considered, although the specific ordering may not be as straightforward as in the continuous case (as discussed in Wu & Estabrook, 2016). Because latent variables generally do not have intrinsic units, in the traditional approach (e.g., Example 2 in the current paper), one or more items need to be *anchors*, with cross-group equality constraints on the measurement parameters, so that the latent variables are on the same metric across groups and the degree of noninvariance can be assessed for the remaining items. After that, equality constraints are placed on invariant parameters, while noinvariant parameters are freely estimated across groups, resulting in a *partial invariance* model. Based on the partial invariance model, effect sizes for noninvariant items can be computed. On the other hand, recent methods such as alignment optimization (Asparouhov & Muthén, 2014) and regularization (Belzak & Bauer, 2020) reframe MI assessment as an optimization problem with the goal of finding an approximate invariance solution, which makes the process less labor intensive and avoids the need for anchor items (see Example 1 in the current paper).

1.2. Effect Size Measures for Violations of MI

Discussion of effect size measures for MI can already be found in early literature in the context of item response theory (e.g., the area measures by Raju, 1988). We refer readers to Meade (2010) for a comprehensive taxonomy of MI effect sizes. In this article, we focus on extending the d_{MACS} effect size by Nye and Drasgow (2011), which quantifies the standardized mean difference in item scores due to noninvariance, for three reasons. First, d_{MACS}—one of the few standardized effect sizes proposed (see also Zwick et al., 1993)—can, arguably, be compared even when raw item scores are on different units (e.g., binary vs. 5-point scales). Second, d_{MACS} is analogous to the popular Cohen's d effect size for comparisons of two groups, making it more familiar and easily interpretable to researchers. Third, d_{MACS} is relatively well-researched: Nye et al. (2019) has provided benchmark values of small, medium, and large d_{MACS}, and Gunn et al. (2020) has discussed several extensions using signed differences and assigning different weights to the samples.

Consider an indicator Y_j measuring a latent variable η with probability distribution $f(\eta)$ in two groups: g_1 and g_2, and let the sample size of group g be n_g. Let $\hat{Y}_j(\eta) = E(Y_j|\eta)$ be the conditional expected item score given the latent score η and the measurement parameters. Nye and colleagues (Nye et al., 2019; Nye & Drasgow, 2011; see also Meade, 2010) proposed

$$d_{\text{MACS}j,(g_1,g_2)} = \sqrt{\frac{\int (\hat{Y}_{jg_1} - \hat{Y}_{jg_2}|\eta)^2 f(\eta) d\eta}{\text{Var}(Y_j)}} \tag{1}$$

to quantify the effect of measurement noninvariance or bias on the expected score of item j. In the above equation, $\text{Var}(Y_j) = \sum_g n_g \text{Var}(Y_{jg})/N$ is the pooled variance of item j, with $N = \sum_g n_g$. Note that if an item is found or assumed invariant such that the measurement parameters are constrainted equal across groups, d_{MACS} will be zero. Lai (2023) showed that d_{MACS} is effective in gauging the degree of noninvariance following approximate invariance analysis with alignment optimization (Asparouhov & Muthén, 2014). Gunn et al. (2020) further discussed modifications to d_{MACS} to use the variance of the reference group instead of the pooled variance across groups as the unit of standardization.

While d_{MACS} only concerns two groups, many grouping variables (e.g., ethnic and cultural groups, countries) naturally have more than two categories, and researchers are often interested in combinations and intersections of grouping variables (e.g., sex and race). While it is possible to handle multicategory grouping variables by (a) doing $G \times (G-1)/2$ pairwise comparisons and reporting many d_{MACS} statistics or (b) splitting the groups into two meaningful sets for comparisons,[1] approach (a) is tedious, and approach (b) represents some loss of information. Also, researchers may be interested in a summative statistic indicating the degree of noninvariance across $G > 2$ groups.

In this manuscript, we introduce f_{MACS} as an extension to d_{MACS}, using the same logic that the Cohen's f effect size is a natural extension to the Cohen's d effect size for mean differences across multiple groups. We then discuss how one can compute f_{MACS} for specific sources of noninvariance (e.g., interactions between gender and ethnicity) using contrasts, similar to the partitioning of sum of squares in analysis of variance (ANOVA). Later, we present two examples of calculating the f_{MACS} effect size using data reported in previous studies: one from a cross-cultural survey with a large number of countries and with the use of the relatively new technique of alignment optimization for approximate invariance (Asparouhov & Muthén, 2014), and the other from a validation study involving more than one grouping variable. The R code for reproducing the analyses has been made publicly available on GitHub (https://github.com/marklhc/fmacs-supp/).

2. The Proposed f_{MACS} Effect Size

Given that d_{MACS} is analogous to Cohen's d for two groups, we propose f_{MACS} as an analog to Cohen's f (Cohen, 1988) for the degree of noninvariance with two or more groups:

$$f^2_{\text{MACS}j} = \frac{1}{NG_j\text{Var}(Y_j)} \sum_{g=1}^{G_j} n_g \int_{-\infty}^{\infty} \left(\hat{Y}_{jg} - \bar{\hat{Y}}_j|\eta\right)^2 f(\eta)d\eta, \quad (2)$$

where G_j is the number of groups for item j (to accommodate missing data or items). Note that $\frac{1}{G_j}\sum_{g=1}^{G_j}(\hat{Y}_{jg} - \bar{\hat{Y}}_j|\eta)^2$ is the average squared deviation from the expected item mean, due to noninvariance. Therefore, f^2_{MACS} quantifies the ratio of (a) the variance due to noninvariance of the expected item score across groups to (b) the pooled within-group variance of the item score. Taking the square root, and treating the pooled item standard deviation as a unit of standardization, we can see that f_{MACS} is conceptually

$$\frac{\text{SD}_{\text{noninvariance}}}{\text{SD}_{\text{item score}}}.$$

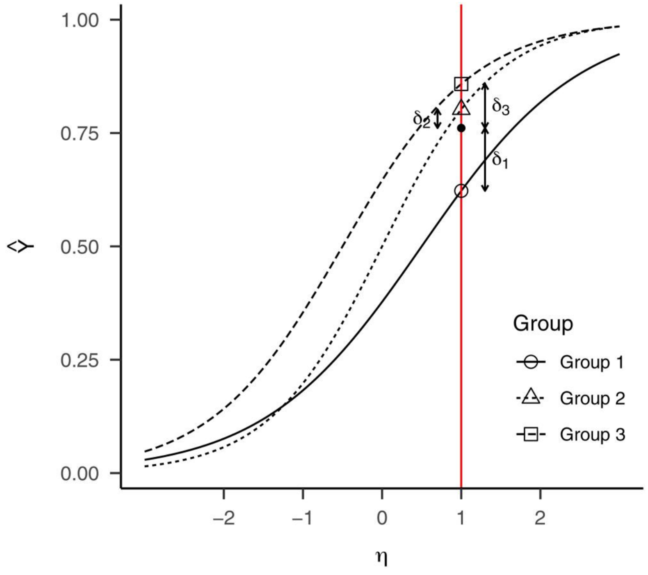

Figure 1. Variability of expected item score ($\hat{Y}$) at a given η.
Note: The item characteristic curves are based on a two-parameter logistic model. Discrepancies in the three curves (indicated by the δ s) show measurement noninvariance across groups.

Figure 1 illustrates between-group variations in expected item score ($\hat{Y}_{jg}$) at a given level of η, based on a two-parameter logistic item response model. The three curves show the relation between $\hat{Y}_{jg}$ and η across three groups, and the discrepancy in the curves indicates that the item is not invariant. For example, at $\eta = 1$, the expected item score $\hat{Y}_{jg}$ is 0.62, 0.80, and 0.86, respectively for $g = 1$, 2, and 3. Assuming the three groups are of equal sizes, the grand mean $\bar{\hat{Y}}_j$ is 0.76. The deviation of $\hat{Y}_{jg}|\eta = 1$ from the grand mean is -0.14 for $g = 1$, 0.04 for $g = 2$, and 0.10 for $g = 3$, so the mean squared deviation across the three groups, $E[(\hat{Y}_{jg} - \bar{\hat{Y}}_j)^2|\eta = 1]$ is 0.01. The computation is analogous to that of the between-group sum of squares in one-way ANOVA. To determine the variance due to noninvariance, one can similarly compute $E[(\hat{Y}_{jg} - \bar{\hat{Y}}_j)^2|\eta]$ for all values of η, and then obtain the weighted average by integrating out the latent score η.

In a linear factor model, it can be shown that f^2_{MACS} simplifies to

$$\begin{aligned} f^2_{\text{MACS}j} = & \frac{1}{NG_j\text{Var}(Y_j)} \sum_{g=1}^{G_j} n_g \\ & \Big[(\nu_{jg} - \bar{\nu}_j)^2 + 2(\nu_{jg} - \bar{\nu}_j)(\lambda_{jg} - \bar{\lambda}_j)E(\eta) \\ & +(\lambda_{jg} - \bar{\lambda}_j)^2 E(\eta^2)\Big], \end{aligned} \quad (3)$$

where ν and λ represent measurement intercepts and loadings, respectively, and $\bar{\nu}$ and $\bar{\lambda}$ are the sample size weighted averages of intercepts and loadings. Therefore, if a researcher studying invariance across ethnic groups finds $f_{\text{MACS}} = 0.5$, it means that the differences due to noninvariance in item scores *across* ethnic groups are half the size of the differences in scores *within* each ethnic group. In other words, even when there were no difference in the latent construct across groups, each ethnic group would differ from the grand mean by, on average, 0.5 standard deviations, which is generally considered substantial.

2.1. Different Weighting Schemes

While the above equation assumes that the parameters of each group are weighted by the respective sample size, sometimes researchers may want to weight the groups differently (e.g., Gunn et al., 2020), especially when the population sizes are very different from the sample sizes. The following is a more general version of equation (2) that incorporates group weights, w_{gj}:

$$f^2_{\text{MACS}j} = \frac{1}{(\sum_g w_{gj})G_j\text{Var}(Y_j)} \sum_{g=1}^{G_j} w_{gj} \int_{-\infty}^{\infty} \left(\hat{Y}_{jg} - \bar{\hat{Y}}_j|\eta\right)^2 f(\eta)d\eta. \quad (4)$$

A special case is when $w_{gj} = w_{gj'}$ for all g s and $j \neq j'$, meaning that every group is weighted equally. As a hypothetical example, consider a researcher studying noninvariance of an item across a sample of White, Black, and Asian Americans with a sample size of 100 for each group. The researcher finds the item intercept to be 0.6, 0.7, and 0.9, respectively. If the groups are weighted by their sample sizes, the f_{MACS} effect size would be 0.12. However, if the researcher is interested in the item variation due to noninvariance at the population level, the researcher may choose to weight the groups based on the population sizes, with 75.3% White, 13.7% Black, and 6.5% Asian at the U.S. national level according to the 2024 Census (U.S. Census Bureau, n.d.), which would yield a different f_{MACS} effect size of 0.08.

Table 1. Information needed to compute effect sizes for noninvariant items in secondary analyses.

Option 1: Raw Data
- Raw data with clear variable descriptions
- Analytic scripts for reproducing invariance analyses

Option 2: Group-Specific Summary Statistics
- Group sizes
- For continuous indicators, means and covariance matrices
- For categorical items, thresholds and polychoric correlations (or tetrachoric correlations for binary items)

Option 3: Group-Specific Parameter Estimates
- Group sizes
- Item standard deviations
- For continuous indicators, parameter estimates of loadings, intercepts, and unique variances and covariances
- For categorical items, parameter estimates of loadings, thresholds, and unique variances and covariances

2.2. Relationship Between f_{MACS} and d_{MACS}

In mean comparisons with two groups, Cohen's d and f effect sizes are related such that $d = 2f$ (Cohen, 1988). For d_{MACS} and f_{MACS}, $2f_{\text{MACS}} = d_{\text{MACS}}$ also holds exactly for continuous indicators. For example, for a noninvariant item, if $\lambda_f = 0.7$ and $\nu_f = 0.2$ for the focal group, and $\lambda_r = 0.4$ and $\nu_r = 0$ for the reference group, when $\eta \sim N(0, 1)$ for the reference group, and $\text{Var}(Y_j) = 1$, we have $d_{\text{MACS}} = 0.64$, and $f_{\text{MACS}} = 0.32$.

2.3. Bias of Sample f_{MACS} Estimates

Given that f_{MACS} is conceptually linked to Cohen's d and f effect sizes, and the sample estimates of d and f are generally biased in small samples (e.g., Hedges, 1981), we performed a small-scale simulation to investigate the bias of sample f_{MACS} estimates. The details of the simulation are provided in the supplemental materials (https://github.com/marklhc/fmacs-supp/). As a brief summary, the results showed that the sample f_{MACS} estimates were generally unbiased with a sample size of $n =$ 250 per group, while the use of the bootstrap method for bias correction (Efron, 1987; see details in the supplemental materials) generally reduced the bias to an acceptable level with $n = 100$. The need for bias correction was stronger with categorical indicators than with continuous indicators.

2.4. Information Needed for Computing d_{MACS} and f_{MACS}

In addition to effect sizes for violations of MI, we also advocate for the reporting of the information necessary to compute these effect sizes, so that other researchers can reproduce the results, and secondary researchers and meta-analysts can have the option to compute effect sizes differently, if needed. As a result, effect sizes from multiple studies can be compared and synthesized.

Like most effect size statistics, computing point estimates of d_{MACS} and f_{MACS} does not require the raw data, but only some secondary information about the data and the invariance analyses. Unfortunately, information such as group-specific loadings and intercepts, was rarely reported in previous invariance studies, as found in Nye et al. (2019), Dong and Dumas (2020), and Zhang (2022) when attempting to compute effect sizes for noninvariance from the primary studies. While guidelines for reporting invariance analyses exist (e.g., Maassen et al., 2023; Putnick & Bornstein, 2016), they are geared towards assessing whether different levels of invariance are tenable, not the magnitude of noninvariance. Therefore, we summarize below (and also in Table 1) the information needed for computing d_{MACS} and f_{MACS} effect sizes for researchers conducting secondary analyses and meta-analyses, with the hope of improving the replicability of invariance analyses and making them part of cumulative science (Lakens, 2013).

2.4.1. Raw Data

If possible, the best way to allow future researchers conducting secondary or meta-analyses computing effect sizes is to share the raw data with clear documentation for conducting the invariance analyses. This gives maximum flexibility for other researchers as they can use alternative methods for parameter estimation, determination of noninvariant items, and effect size computation, and is in the spirit of the open science movement (e.g., Shrout & Rodgers, 2018).

2.4.2. Means and Covariance Matrices

If it is not feasible to share raw data, researchers could share summary statistics that are (practically) sufficient to reproduce the invariance analyses. For analyses with normal-theory maximum likelihood estimation, the sufficient statistics are the means and covariance matrices of all items for each group, as well as the group sample sizes. For analyses with categorical data and least squares estimation, one should share the item thresholds, the tetrachoric (for binary items) or the polychoric (for ordered categorical items) correlations for each group, and the group sample sizes.[2] Access to these summary statistics would allow secondary studies to reproduce and reconduct the invariance analyses, except for potential differences due to, for example, missing data and robust standard errors.

2.4.3. Parameter Estimates

From equations (1) and (2), one can see that d_{MACS} and f_{MACS} use the following information: (a) sample sizes for each item in each group, (b) parameter estimates of the final partial invariance model, which include loadings, intercepts (or thresholds and unique variances and covariances for categorical items; Tse et al., 2024), and latent means and variances, and (c) item standard deviations for each group. If (c) is missing, one can also compute the model-implied item standard deviations when the unique variances (and any covariances) are also available.

Note that the "Journal Article Reporting Standards" by the American Psychological Association (Appelbaum et al., 2018) suggested that sample means, covariance matrices, and the parameter estimates should be routinely reported when doing structural equation modeling, which subsumes invariance analyses. Also, the reporting guidance above is specific to computing d_{MACS} or f_{MACS} effect sizes; researchers should still consult the guidelines by, for example, Putnick and Bornstein (2016), Luong and Flake (2023), and Maassen et al. (2023) for recommendations for other information to be reported in invariance analyses.

Next, we present two empirical examples to illustrate how f_{MACS} can inform the magnitude of item noninvariance. The R package *pinsearch* can compute f_{MACS} with either a fitted partial invariance model from *lavaan* (Rosseel, 2012) or with parameter estimates as discussed above. The full R code for the examples can be found in the online supplemental materials.

3. Empirical Example 1

In the first example, we used data analyzed in Asparouhov and Muthén (2014), which illustrated the alignment optimization method for approximate invariance. The data included 49,894 participants from 26 European countries answering four items measuring the construct "Tradition and Conformity." Each of the four items was on a 6-point scale, with an example item "It is important for him to be humble and modest. He tries not to draw attention to himself." More details can be found in Asparouhov and Muthén (2014) and Beierlein et al. (2012).

Using the alignment algorithm in the R package *sirt* (Robitzsch, 2024),[3] the aligned loadings were, across the 26 countries, 0.351 to 0.825 for Item 1, 0.358 to 0.859 for Item 2, 0.538 to 0.905 for Item 3, and 0.764 to 1.002 for Item 4, respectively. The aligned intercepts were 2.27 to 3.47 for Item 1, 2.04 to 3.10 for Item 2, 2.62 to 3.83 for Item 3, and 2.50 to 2.73 for Item 4. The aligned loadings, aligned intercepts, and the item uniqueness are needed to compute the f_{MACS} effect size statistics and can be found in the online supplemental materials.[4] The fit of the aligned model, which was the same as the configural invariance model, was acceptable, χ^2 $(df = 52) = 317.1$, $p < .001$, RMSEA = .052, CFI = .99, SRMR = .013. The ranges of the loadings and intercepts suggest that noninvariance is larger for Items 1 to 3 than for Item 4, which is consistent with the f_{MACS} values of .260, .186, .244, .053 for Items 1 to 4. The f_{MACS} values for Item 1 indicate that the expected item score for each country differs from the grand mean by .26 in standardized units, which is substantial. On the other hand, for Item 4, the expected item score differs from the grand mean by .053 in standardized units, which is negligible.

According to Nye et al. (2019), $d_{\text{MACS}} \leq 0.20$ can be considered negligible, and using the conversion $d_{\text{MACS}} = 2\ f_{\text{MACS}}$, we can consider $f_{\text{MACS}} \leq 0.10$ to indicate negligible noninvariance. Given that three out of the four items showed f_{MACS} much higher than .10, it appears that participants from different countries interpret those three items somewhat differently. Previous literature on alignment optimization (e.g., Asparouhov & Muthén, 2014; Lai, 2023) recommended no more than one-third of the items to have substantial noninvariance for the results to be trusted, so researchers should exert extreme cautions when trying to compare the Tradition and Conformity construct across countries.

4. Empirical Example 2

As a second example, we re-analyzed the data set made publicly available by Lui (2019), which aimed to evaluate invariance of a measure of alcohol beliefs among 1,148 undergraduate students from a private university in the United States. Alcohol beliefs were measured by the College Life Alcohol Salience Scale (CLASS; Osberg et al., 2010) consisting of 15 Likert items (1 = *strongly disagree* to 5 = *strongly agree*). The author conducted an invariance analysis of CLASS across gender (65% female) and ethnicity (44.9% Caucasian, 19.9% Asian, 10.3% African American, 16.7% Latinx/Hispanic, and 8.3% mixed/other ethnic backgrounds), as well as a few other demographic background variables. They determined noninvariance by the likelihood ratio test (at .01 level), the change in goodness of fit indices (ΔCFI and ΔRMSEA; Cheung & Rensvold, 2002), and the bootstrap confidence interval test (Cheung & Lau, 2012). They found six items to have noninvariant intercepts across ethnic groups. While the author investigated invariance separately on different demographic groupings, in our reanalysis, we combined gender and ethnic groups to investigate their intersectionality. In addition, given the relatively small number of African American participants and the results from Lui (2019) that the one-factor model for CLASS did not fit this subgroup, we focused our illustration on Caucasian, Asian, and Latinx/Hispanic participants, resulting in an analytic sample size of 937. Following Lui (2019), we used maximum likelihood estimation with sandwich estimates of standard errors and scaled test statistics (i.e., the `estimator = "MLR"` option in *lavaan*).

4.1. Specification Search

We first evaluated different levels of invariance of all 15 items across the six Gender × Ethnicity combinations. Specifically, we followed the specification search approach by Yoon and Millsap (2007) to sequentially identify noninvariant parameters using modification indices (MI; Sörbom, 1989). The algorithm proceeds in stages of loading, intercept, and uniqueness invariances. In each stage, all parameters at that stage (e.g., loadings) are first assumed invariant, and the invariance constraint with the largest MI is repeatedly freed in the next model, until the largest MI is below a pre-specified cutoff. Here we chose a cutoff of 3.84, which corresponds to the 95th percentile of the χ^2 distribution with one degree of freedom.

Caucasian males were the reference group, and the latent Alcohol Beliefs variable was identified by fixing the mean to zero and the variance to one for the reference group. We freed the unique covariances between Items 1 and 2 for item wording similarity, as suggested by modification indices. The configural invariance model had an acceptable fit, scaled χ^2 $(df = 534)$ = 935.1, $p <$ .001, RMSEA = .073, CFI = .93, SRMR = .051. The specification search for noninvariant parameters was automated using the *pinsearch* R package (Lai, 2024), and the final partial invariance model had scaled χ^2 $(df = 735)$ = 1,180, $p <$ .001, RMSEA = .065, CFI = .92, SRMR = .077. As shown in Table 2, while metric invariance was found tenable, a total of 12 intercepts from 8 items were found to be noninvariant for at least one group in the final partial invariance model. The parameter estimates of the partial invariance model can be found in the supplemental material.

4.2. f_{MACS} Effect Size

As shown in Table 3, f_{MACS} ranged between 0.06 and 0.15 for these 8 items. Therefore, the impact of noninvariance on the observed item score was generally small, but was larger than or close to 0.10 for Items 1, 2, 4, and 14. These items overlap with those found by Lui (2019): Items 2, 4, 7, 14, and the discrepancy could be due to the exclusion of African American participants in our analysis, as well as the fact that Lui (2019) did not explore interactions between gender and ethnicity.

4.3. Contrast

Much like in ANOVA, where researchers may be interested in specific contrasts in addition to the omnibus effect, the same can be done for f_{MACS}, such as (a) male compared to female, (b) difference among Caucasian, Asian, and Hispanic/Latinx, and (c) the interaction (intersectionality) of gender and ethnicity. Specifically, denote the variable for a constrat as C with K levels. To compute f_{MACS} for a contrast, we can again use equation (4), but use weighted averages of parameters at each contrast level (i.e., $\bar{\nu}_{jk} = \sum_{\{g:C_g=k\}} w_{jg}\nu_{jg} / \sum_{\{g:C_g=k\}} w_{jg}$ and $\bar{\lambda}_{jk} = \sum_{\{g:C_g=k\}} w_{jg}\lambda_{jg} / \sum_{\{g:C_g=k\}} w_{jg}$ for continuous indicators) instead of the group-specific parameters (i.e., ν_{jg} and λ_{jg}). The weight for each contrast level is the sum of the weights of the groups with that level.

In addition, one can also compute f_{MACS} for a set of contrast coefficients (e.g., Casella & Berger, 2002, Chapter 11), again analogous to the practice in ANOVA. Let $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M$ be a set of contrast coefficients, where $M < G$ and, for each $\mathbf{a}_l$, $\sum_g a_{lg} = 0$. Let $\mathbf{L} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_M]$ be the contrast matrix. For example, one can represent the overall difference among $G = 4$ groups using the following matrix for effect coding:

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{bmatrix}.$$

Then, for each level of η, we can replace $(\hat{Y}_{jg} - \overline{\hat{Y}}_j)^2$ in equation (2) with the variance due to the set of contrasts (see, e.g., Rencher & Schaalje, 2008, Chapter 15):

$$\left[\mathbf{L}^\top \hat{\mathbf{Y}}_j(\eta)\right]^\top \left[\mathbf{L}^\top \mathbf{W}^{-1} \mathbf{L}\right]^{-1} \left[\mathbf{L}^\top \hat{\mathbf{Y}}_j(\eta)\right],$$

where $\hat{\mathbf{Y}}_j = [\hat{Y}_{j1}, \hat{Y}_{j2}, \ldots, \hat{Y}_{jG}]^\top$ is the vector of group-specific expected item score at a given level of η, and $\mathbf{W}$ is a diagonal matrix of group weights.

As shown in Table 3, f_{MACS} was larger for ethnicity than for gender for most items, but f_{MACS} for the interaction was similar in size as the main effect for gender. It should be pointed out that, given the unbalanced sample sizes, the main effects of gender, ethnicity, and their interaction are correlated, so the sum of f^2_{MACS} for the main and interaction effects would not be equal to the overall f_{MACS}.

The results of Example 2 suggest that items 1, 2, 4, and 14 to have more than negligible violations of invariance, and the contrast analysis shows ethnicity to be more predictive of noninvariance than gender for item 4 ("The reward at the end of a hard week of studying should be a weekend of heavy drinking"). These results suggest possible directions for revising items so that they are more comparable across groups, and researchers should analyze the wordings and participant understanding of the items before deciding whether to drop or revise an item. Note that noninvariance from multiple items can compound or cancel each other out at the test score level, so further analyses could be performed, such as the selection accuracy analysis suggested by Millsap and Kwok (2004) and the item deletion analysis by Ozcan and Lai (2024).

Table 2. Noninvariant parameters from specification search.

Step	Item	Group	Noninvariant parameter
1	class14	Caucasian Female	intercepts
2	class1	Caucasian Female	intercepts
3	class2	Asian Male	intercepts
4	class2	Hispanic Male	intercepts
5	class4	Hispanic Female	intercepts
6	class4	Asian Female	intercepts
7	class8	Asian Female	intercepts
8	class4	Asian Male	intercepts
9	class7	Asian Female	intercepts
10	class3	Asian Female	intercepts
11	class7	Asian Male	intercepts
12	class5	Hispanic Female	intercepts

Table 3. f_{MACS} Effect sizes for empirical example 2.

Item	Overall	Gender	Ethnicity	Gender x ethnicity
class1	0.10	0.03	0.05	0.05
class2	0.10	0.08	0.06	0.06
class3	0.07	0.03	0.04	0.04
class4	0.11	0.04	0.09	0.05
class5	0.06	0.03	0.04	0.04
class7	0.08	0.00	0.07	0.00
class8	0.09	0.04	0.05	0.05
class14	0.15	0.04	0.07	0.07

5. Discussion

Despite the growth in empirical studies investigating measurement invariance of psychological instruments, most studies have focused only on significance tests and goodness-of-fit indices of the invariance models, and less attention has been paid to quantifying the magnitude of noninvariance on an interpretable metric. The d_{MACS} effect size proposed by Nye and Drasgow (2011) is a useful metric that informs the standardized mean difference in expected item scores due to noninvariance, and has gained popularity in recent years. However, it is only applicable when there are two groups, and applied researchers are usually interested in grouping variables with more than two levels or multiple grouping variables.

In this paper, we propose the f_{MACS} (and f^2_{MACS}) effect size to quantify the ratio of between-group item variance due to noninvariance to within-group item variance, which is a natural extension of d_{MACS} when there are more than two groups. We provide two empirical examples of how f_{MACS} can quantify the magnitude of noninvariance when using different invariance testing methods. Given that effect size has been part of the reporting standard for quantitative analyses in the past two decades, we recommend researchers, journal editors, and reviewers follow the same standard for invariance analyses. We also outline what information can facilitate effect size computations in secondary analyses and meta-analyses for measurement invariance, with the hope that future invariance analyses can be more transparent and replicable.

While the current paper focuses on item-level noninvariance, test-level invariance may also be of interest when test scores, usually unweighted or weighted sums of item scores (such as predicted factor scores in factor analysis), are used (e.g., Chalmers et al., 2016; Stark et al., 2004). One can obtain test-level f_{MACS} by replacing $\hat{Y}$ in equation (2) by $\hat{Z}$, where $\hat{Z} = w_1\hat{Y}_1 + w_2\hat{Y}_2 + \ldots$ is the test score variable with weights $w_1, w_2, \ldots$. The test-level f_{MACS} is available in the *pinsearch* package by supplying the `item_weights` argument, as illustrated in the supplemental materials for both applied examples.

The current paper also only focuses on violation of MI with respect to discrete grouping variables. As Cohen's f can also be applied to regression analysis with a continuous predictor, it should be possible, in theory, to compute f_{MACS} for violation of MI with respect to a continuous variable, W, such as age. In Equation (2), the finite sum $\sum_{g=1}^{G_j} n_g(\hat{Y}_{jg} - \overline{\hat{Y}}_j|\eta)^2$ represents the between group variance due to noninvariance, which can be replaced by the variance accounted for by W. For example, for a multiple-indicator multiple-cause (MIMIC) model with noninvariant intercept for item j,

$$Y_j = (\nu_j + b_j W) + \lambda_j \eta + \varepsilon_j,$$

where $\eta = \gamma_0 + \gamma W$ and the intercept depends on W, the variance accounted for by W, conditioned on η, is $b_j^2 \text{Var}(W)$. Future research can further formalize and validate f_{MACS} for continous violators.

Like other effect size statistics, one main area of future research for d_{MACS} and f_{MACS} is to determine the distribution of effect sizes in empirical research, and use such information to determine possible benchmark values. In other words: what values of d_{MACS} and f_{MACS} are considered small or large? Currently, a major obstacle is that the majority of existing invariance research did not report sufficient information for effect size computation (Dong & Dumas, 2020; Nye et al., 2019). Nye et al. (2019) instead used simulations to give two sets of benchmark values for d_{MACS} in terms of the impact of item noninvariance for subsequent analyses: .20 or .40 for small noninvariance, .40 or .60 for medium noninvariance, and .70 or .80 for large noninvariance. Given the benchmark values suggested by Nye et al. (2019), one can derive benchmark values for f_{MACS} using the relation $d_{\text{MACS}} = 2f_{\text{MACS}}$ when there are two groups, which is also the approach Cohen (1988) took when suggesting benchmark values for f and some other effect size statistics. Based on our experiences, we suggest using $f_{\text{MACS}} < .10$ for negligible noninvariance, as a starting point. While these benchmark values could provide impetus for researchers to adopt effect size statistics for noninvariance, they should be used with caution as the interpretation of the magnitude of noninvariance should be based on many other factors, such as the construct of interest, the grouping variables, the main usage of the instrument, the context of the measurement, and so on. Future research should focus on refining the benchmark values and providing more fine-grained guidelines for using these effect size statistics.

There are several limitations in the current paper. First, we have only defined f_{MACS} at the population level and substituted in sample estimates of the model parameters to estimate f_{MACS}. As long as the parameter estimates are consistent, the sample f_{MACS} estimate should also be consistent, as demonstrated in our simulation studies. On the other hand, our simulations also showed that sample estimates of f_{MACS} could be biased in small samples, much like how Cohen's d is a biased estimator of standardized mean difference in the population level, and future research should fully investigate the properties of sample f_{MACS} estimates across a wider range of conditions like number of items and larger number of groups using simulation studies. Second and relatedly, there has not been discussion on obtaining uncertainty measures, such as standard errors and confidence intervals, for d_{MACS} and f_{MACS}. Some possible options are the delta method, bootstrapping, and the Monte Carlo method by simulating from the multivariate sampling distributions of the parameter estimates (e.g., Preacher & Selig, 2012), and future research can demonstrate and investigate the performance of these methods for noninvariance effect size, as well as that of alternative methods. Third, both d_{MACS} and f_{MACS} are model-dependent as they assume that the partial or approximate invariance model is correctly specified. However, in practice, such a model is usually obtained after some kind of specification search procedure, as in our Example 2. It is unclear to what extent uncertainty in model selection would bias both point and uncertainty estimates of d_{MACS} and f_{MACS}, and future research is needed to shed light on this issue. Future research may also look into ways of incorporating model uncertainty when estimating d_{MACS} and f_{MACS}, and effect size estimators of noninvariance that are less model-dependent.

Notes

1. For example, Zieger et al. (2019) performed pairwise comparisons to test measurement invariance of a teacher satisfaction measure between England and 17 other countries.
2. With weighted least squares estimation, one also needs the $p' \times p'$ weight matrix, which involves the fourth moments of the items and $p' = p(p-1)/2$. However, given the size of the matrix, such a matrix is rarely reported in practice except when the number of items is small. For secondary analyses, the polychoric correlations and thresholds can be used to perform unweighted least squares estimation by using an identity matrix for the weight matrix.
3. The alignment model is identified by the fixed = FALSE option in *sirt*, which constrains the product of latent *SD*s across groups to 1.
4. The item uniqueness is not affected by the alignment algorithm.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*, 33–40. https://doi.org/10.3102/0013189X035006033

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *The American Psychologist*, *73*, 3–25. https://doi.org/10.1037/amp0000191

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 495–508. https://doi.org/10.1080/10705511.2014.919210

Beierlein, C., Davidov, E., Schmidt, P., Schwartz, S. H., & Rammstedt, B. (2012). Testing the discriminant validity of Schwartz' Portrait Value Questionnaire items – A replication and extension of Knoppen and Saris (2009). *Survey Research Methods*, *6*, 25–36. Pages https://doi.org/10.18148/SRM/2012.V6I1.5092

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*, 673–690. https://doi.org/10.1037/met0000253

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*, S176–S181. https://doi.org/10.1097/01.mlr.0000245143.08679.cc

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Thomson Learning.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76*, 114–140. https://doi.org/10.1177/0013164415584576

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods*, *15*, 167–198. https://doi.org/10.1177/1094428111421987

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates.

Cole, S. R., Kawachi, I., Maller, S. J., & Berkman, L. F. (2000). Test of item-response bias in the CES-D scale. *Journal of Clinical Epidemiology*, *53*, 285–289. https://doi.org/10.1016/S0895-4356(99)00151-1

Dong, Y., & Dumas, D. (2020). Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. *Personality and Individual Differences*, *160*, 109956. https://doi.org/10.1016/j.paid.2020.109956

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*, 171–185. https://doi.org/10.1080/01621459.1987.10478410

Gunn, H. J., Grimm, K. J., & Edwards, M. C. (2020). Evaluation of six effect size measures of measurement non-Invariance for continuous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 503–514. https://doi.org/10.1080/10705511.2019.1689507

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128. https://doi.org/10.3102/10769986006002107

Hsiao, Y.-Y., & Lai, M. H. C. (2018). The impact of partial measurement invariance on testing moderation for single and multi-level data. *Frontiers in Psychology*, *9*, 740. https://doi.org/10.3389/fpsyg.2018.00740

Kush, J. M., Masyn, K. E., Amin-Esmaeili, M., Susukida, R., Wilcox, H. C., & Musci, R. J. (2023). Utilizing moderated non-linear factor analysis models for integrative data analysis: A tutorial. *Structural Equation Modeling: a Multidisciplinary Journal*, *30*, 149–164. https://doi.org/10.1080/10705511.2022.2070753

Lai, M. H. C. (2023). Adjusting for measurement noninvariance with alignment in growth modeling. *Multivariate Behavioral Research*, *58*, 30–47. https://doi.org/10.1080/00273171.2021.1941730

Lai, M. H. C., Richardson, G. B., & Wa Mak, H. (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. (2024). *Pinsearch: Specification search for partial factorial invariance* [Manual]. https://github.com/marklhc/pinsearch

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., De Roover, K., Jak, S., Meitinger, K., Menold, N., Muthén, B., Rudnev, M., Schmidt, P., & Van De Schoot, R. (2023). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research*, *110*, 102805. https://doi.org/10.1016/j.ssresearch.2022.102805

Lui, P. P. (2019). College alcohol beliefs: Measurement invariance, mean differences, and correlations with alcohol use outcomes across sociodemographic groups. *Journal of Counseling Psychology*, *66*, 487–495. https://doi.org/10.1037/cou0000338

Luong, R., & Flake, J. K. (2023). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods*, *28*, 905–924. https://doi.org/10.1037/met0000441

Maassen, E., D'Urso, E. D., van Assen, M. A. L. M., 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://doi.org/10.1037/met0000624

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *The Journal of Applied Psychology*, *95*, 728–743. https://doi.org/10.1037/a0018966

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge. https://doi.org/10.4324/9780203821961

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

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods*, *22*, 678–709. https://doi.org/10.1177/1094428118761122

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

Osberg, T. M., Atkins, L., Buchholz, L., Shirshova, V., Swiantek, A., Whitley, J., Hartman, S., & Oquendo, N. (2010). Development and validation of the College Life Alcohol Salience Scale: A measure of beliefs about the role of alcohol in college life. *Psychology of Addictive Behaviors: journal of the Society of Psychologists in Addictive Behaviors*, *24*, 1–12. https://doi.org/10.1037/a0018197

Ozcan, M., & Lai, M. H. C. (2024). Exploring the impact of deleting (or retaining) a biased Item: A procedure based on classification accuracy. *Assessment*. Advance online publication. https://doi.org/10.1177/10731911241298081

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*, 77–98. https://doi.org/10.1080/19312458.2012.679848

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

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502. https://doi.org/10.1007/BF02294403

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics* (2nd ed.). Wiley-Interscience.

Robitzsch, A. (2024). *Sirt: Supplementary item response theory models* (Version 4.0-32). https://CRAN.R-project.org/package=sirt

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://www.jstatsoft.org/v48/i02/ https://doi.org/10.18637/jss.v048.i02

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Somaraju, A. V., Nye, C. D., & Olenick, J. (2022). A review of measurement equivalence in organizational research: What's old, what's new, what's next? *Organizational Research Methods*, *25*, 741–785. https://doi.org/10.1177/10944281211056524

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*, 371–384. https://doi.org/10.1007/BF02294623

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? *The Journal of Applied Psychology*, *89*, 497–508. https://doi.org/10.1037/0021-9010.89.3.497

Tse, W. W.-Y., Lai, M. H. C., & Zhang, Y. (2024). Does strict invariance matter? Valid group mean comparisons with ordered-categorical items. *Behavior Research Methods*, *56*, 3117–3139. https://doi.org/10.3758/s13428-023-02247-6

U.S. Census Bureau. (n.d). *QuickFacts*. Retrieved January 1, 2025, from https://www.census.gov/quickfacts/

van Dijk, W., Schatschneider, C., Al Otaiba, S., & Hart, S. A. (2022). Assessing measurement invariance across multiple groups: When is fit good enough? *Educational and Psychological Measurement*, *82*, 482–505. https://doi.org/10.1177/00131644211023567

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

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

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

Zhang, G. (2022). *A systematic review of measurement invariance research of the CES-D scale across gender* [Unpublished master's thesis]. University of Southern California. https://doi.org/10.25549/usc-theses-oUC111375873

Zieger, L., Sims, S., & Jerrim, J. (2019). Comparing teachers' job satisfaction across countries: A multiple-pairwise measurement invariance approach. *Educational Measurement: Issues and Practice*, *38*, 75–85. https://doi.org/10.1111/emip.12254

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*, 233–251. https://doi.org/10.1111/j.1745-3984.1993.tb00425.x