Evaluation of Five Effect Size Measures of

Measurement Non-Invariance for Continuous Outcomes

by

Heather J. Gunn

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved January 2019 by the
Graduate Supervisory Committee:

Kevin J. Grimm, Co-Chair
Michael C. Edwards, Co-Chair
Jenn-Yun Tein
Samantha F. Anderson

ARIZONA STATE UNIVERSITY

May 2019

ABSTRACT

To make meaningful comparisons on a construct of interest across groups or over time, measurement invariance needs to exist for at least a subset of the observed variables that define the construct. Often, chi-square difference tests are used to test for measurement invariance. However, these statistics are affected by sample size such that larger sample sizes are associated with a greater prevalence of significant tests. Thus, using other measures of non-invariance to aid in the decision process would be beneficial. For this dissertation project, I proposed four new effect size measures of measurement non-invariance and analyzed a Monte Carlo simulation study to evaluate their properties and behavior in addition to the properties and behavior of an already existing effect size measure of non-invariance. The effect size measures were evaluated based on bias, variability, and consistency. Additionally, the factors that affected the value of the effect size measures were analyzed. All studied effect sizes were consistent, but three were biased under certain conditions. Further work is needed to establish benchmarks for the unbiased effect sizes.

DEDICATION

I dedicate this dissertation to the many people who convinced me directly or indirectly to not drop out of graduate school.

To the late Roger Millsap: thank you for inspiring my interest in the area of psychometrics. I hope you would be proud of this work and my growth.

To Kevin Grimm: thank you for your emotional, intellectual, and financial support of me throughout graduate school. I am eternally grateful for your kindness, patience, and intellect throughout my graduate school experience.

To Mike Edwards: thank you for your humor, generosity, and respect. I left each of our meetings with more confidence. I am so grateful to have met you.

To Jenn-Yun Tein: thank you for standing up for me and being willing to meet with me whenever I asked. I learned so much from you.

To Alice Young and Marianne Evola: thank you for introducing me to the field of quantitative psychology. I would not be on my current path without your guidance.

To Gina Mazza and Matt Valente: I could not have asked for a better cohort. Both of you supported me in so many ways throughout this journey. Thank you for cheering me on during the good times and the bad times.

To Andie Parazo, Andrew Gaul, Kristin Bielling, and Liz Guerrero: the Four Horsemen of friendship. Thank you for your endless support and encouragement. Y’all kept me grounded and helped me rediscover myself when I was lost.

To my family: thank you for supporting and helping me achieve my goals.

To Michael Johnson: thank you for helping me during the hardest time of my life and for helping me create meaning amongst chaos.

ACKNOWLEDGMENTS

I thank my co-chairs Kevin Grimm and Mike Edwards as well as my committee members Jenn Tein and Samantha Anderson for their help in creating and executing this project. I also thank the other quantitative faculty members at Arizona State – Roger Millsap, Leona Aiken, Steve West, Craig Enders, Hye Won Suk, Dave MacKinnon, Roy Levy – who helped me develop and hone my quantitative skills. I would also like to express gratitude to the librarians who helped me access the materials I needed to write this dissertation. Lastly, I greatly appreciate all of my fellow graduate students in the quantitative psychology program at Arizona State for their support, intellectual discussions, and friendship.

TABLE OF CONTENTS

Page

CHAPTER Page

APPENDIX

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

There are outcomes of interest, like math ability or extraversion, which cannot be directly measured. One way to indirectly measure these outcomes is to measure a set of variables that are related to the construct of interest. Theoretically, we believe there is an underlying construct, called a latent variable, which directly influences the observed variables (Thurstone, 1947). If group comparisons on the latent variable are of interest, such as comparing males and females on parenting or assessing familism over time, then the relationship between the latent variable and the probability of obtaining a particular score on the observed variables needs to be equal across groups or time. Specifically, a majority of the observed variables need to be invariant. As the number of non-invariant observed variables increases, the more difficult it becomes to defend the position that the latent variable has the same meaning and metric across groups, making mean group comparisons on the latent variable tenuous.

Measurement invariance exists in the factor analytic framework if the following property holds:

$$P_g(y|\eta) = P(y|\eta), \tag{1}$$

where y is the score on an observed variable, η is the latent variable score, g is group membership, and $P_g(y|\eta)$ is the measured response function for the observed variable y and group g. If true, two people with the same level on the latent construct would be expected to have the same observed score, regardless of group membership. If this property does not hold, then the measurement properties of the observed variables in relation to the construct differ across groups and measurement invariance has been violated (Millsap, 2011).

Typically, in factor analysis, measurement invariance (also known as factorial invariance when using factor analytic models) is established by testing a series of nested hierarchical models with chi-square difference tests. However, these tests are highly affected by sample size such that statistically significant differences can be found for negligible group differences when the sample size is large. Thus, effect sizes should be used to quantify the magnitude of the non-invariance. For this dissertation project, I propose four new effect size measures for measurement non-invariance, and study the properties of these proposed effect size measures as well as one existing effect size of measurement non-invariance. I begin with a formal overview of measurement invariance testing and effect size measures. I then transition to discussing the creation of four new effect sizes of measurement non-invariance. Next, I describe the method and results of a simulation study that evaluates the behavior and properties of the four new effect size measures as well as a current effect size measure of non-invariance. Finally, I discuss the implication of the results.

Measurement Invariance

A measurement model expresses how unobserved latent variables relate to observed variables (Millsap, 2011). A measurement model that is often used when testing measurement invariance for continuous observed variables is the linear common factor model (Meredith, 1993; Thurstone, 1947). In this model, one or more common factors (i.e., latent variables) account for the covariances among a set of observed variables. For p observed variables and q common factors, the common factor model can be written as:

$$\mathbf{y}_{\mathrm{j}} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta}_{\mathrm{j}} + \boldsymbol{\varepsilon}_{\mathrm{j}}. \tag{2}$$

In Equation 2, $\mathbf{y}_{\mathrm{j}}$ is a $p \times 1$ vector of person j's scores on p observed variables, $\boldsymbol{\tau}$ is a $p \times 1$ vector of measurement intercepts, $\boldsymbol{\eta}_{\mathrm{j}}$ is a $q \times 1$ vector of factor scores (i.e., latent variable scores) for person j, $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor loadings (analogous to regression coefficients) that relate the factor scores to the observed scores, and $\boldsymbol{\varepsilon}_{\mathrm{j}}$ is a $p \times 1$ vector of unique factor scores for person j. Equation 2 has a similar structure to a regression equation (e.g., predictor, outcome, intercept, regression coefficient, residual); however, the predictors (i.e., the common factors) are unobserved variables. The common factors are assumed to follow a multivariate normal distribution, such that $\boldsymbol{\eta}_{\mathrm{j}} \sim MVN(\boldsymbol{\kappa}, \boldsymbol{\Psi})$, where $\boldsymbol{\kappa}$ is a $q \times 1$ vector of factor means and $\boldsymbol{\Psi}$ is a $q \times q$ matrix of common factor variances and covariances. The unique factors are assumed to follow a multivariate normal distribution, such that $\boldsymbol{\varepsilon}_{\mathrm{j}} \sim MVN(\mathbf{0}, \boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ is a $p \times p$ matrix of unique factor variances and covariances. $\boldsymbol{\Theta}$ is typically assumed to be a diagonal matrix (i.e., the unique factors are uncorrelated with one another); however, this assumption can be relaxed.

The common factor model leads to a set of expectations for the means, variances, and covariances of the observed variables. The expected covariance structure of the factor analysis model is:

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{3}$$

where $\boldsymbol{\Sigma}_{\mathbf{Y}}$ is a $p \times p$ model-implied covariance matrix for the observed variables, and the expected mean structure of the factor analysis model is:

$$\boldsymbol{\mu}_{\mathbf{Y}} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa}, \tag{4}$$

where $\boldsymbol{\mu}_{\mathbf{Y}}$ is a $p \times 1$ vector of model-implied means for the observed variables. In the single-population case, the mean structure is not often of importance; however, it is necessary for examining multiple populations and examining change over time – two situations where measurement invariance testing is necessary.

To find a unique solution to Equation 2, we need to impose constraints in the common factor model to solve the rotational uniqueness problem and to achieve global identification (e.g., define the scale and zero point for the latent variables; Bollen & Jöreskog, 1985; Millsap, 2011). There are many ways to identify a unidimensional common factor model for continuous variables so I focus on the two most common approaches: standardizing the common factor and using a reference indicator. Standardizing the common factor, η_{j}, is done by constraining its mean to be 0 and its variance to be 1. The reference indicator approach is implemented by constraining the factor loading and measurement intercept of one observed variable (an indicator) to 1 and 0, respectively.

We can expand the factor analysis model for a single population (as defined in Equation 2) to the multiple-population case by estimating the factor analysis model separately for each group, indexed by g, such that:

$$\boldsymbol{y}_{jg} = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \boldsymbol{\eta}_{jg} + \boldsymbol{\varepsilon}_{jg}, \tag{5}$$

with the expectation:

$$Cov(\boldsymbol{y}_{jg} | \boldsymbol{\eta}_{jg}) = \boldsymbol{\Theta}_g. \tag{6}$$

For factorial invariance to hold, the factor model parameters ($\boldsymbol{\tau}_g$, $\boldsymbol{\Lambda}_g$, $\boldsymbol{\Theta}_g$) need to have the same values for all of the groups or time points being compared. In other words, the factor model parameters do not have values that differ by group (i.e., $\boldsymbol{\tau}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Theta}$). There are four steps to testing for measurement invariance: 1) identifying a common baseline factor model for all groups, 2) choosing at least one reference variable to link the metric of the latent variable(s) across groups, 3) analyzing chi-square difference tests to identify non-invariance, and 4) estimating the final model.

Common baseline model. Factor models should be fit separately in each group to determine the best fitting model. If the groups are found to have the same number of factors with the same observed variables that define the factors, then a multiple-group model can be fit to the data where the only group equality constraints are those needed for identification and to link the metric of the latent variable across groups. This model is referred to as the configural invariance model (Meredith, 1993; Millsap, 2011). If the groups do not have the same number of factors or have different variables that define the factors, either testing stops and making valid group mean comparisons on the construct is doubtful or a partial invariance model can be analyzed, which I detail later in the document (Steenkamp & Baumgartner, 1998).

There are many ways to identify a multiple-group confirmatory factor analysis (CFA) model and link the metrics of the factors across groups for continuous measured variables. Here, I recreate one potential set of identification constraints when there is independent cluster structure (i.e., each indicator has only one non-zero loading and each factor is defined by at least three indicators with non-zero loadings). (Refer to Millsap (2011) to learn about identification for other scenarios.)

1. For one group (e.g., Group 1), fix the common factor means ($\boldsymbol{\kappa}_g$) to 0 and the factor variances, $\boldsymbol{\Psi}_{rr}$, to 1.

2. Choose a reference variable for each common factor and constrain its loading and its intercept to be equal across groups.

The common factor means and variances for the other group (or groups) are freely estimated. Standardizing the factors for both groups can lead to inaccurate invariance testing (Yoon & Millsap, 2007). The reference variable is known as an anchor item in the item response theory (IRT) framework. Even though the reference variable was previously defined as the variable which has a loading of one and an intercept of zero, in the context of measurement invariance in factor analysis, the reference variable or reference indicator refers to the measured variable that has group equality constraints on its parameters (see French & Finch, 2008; Johnson, Meade, & DuVernet, 2009; Jung & Yoon, 2017; Yoon & Millsap, 2007).

If the configural invariance model fits well, invariance testing can continue. While the χ^2 fit statistic is available to test exact fit, fit should not be assessed based on the χ^2 statistic (Jöreskog, 1971), partly because it is sensitive to even slight departures from multivariate normality (Jöreskog & Sörbom, 1983, p. I.39; West, Finch, & Curran, 1995) and sample size (Kelloway, 1995). Some of the global approximate fit statistics available for continuous indicators are the root mean square error of approximation (RMSEA; Browne & Cudeck, 1993; Steiger, 1989; Steiger & Lind, 1980), the comparative fit index (CFI; Bentler, 1990), and the standardized root mean square residual (SRMR; Jöreskog, & Sörbom, 1983). At the local level, residuals can aid in determining where in the model the misfit is occurring. Additionally, “substantive, theoretical and conceptual considerations” should be used when assessing fit (Jöreskog, 1971, p. 421). It is acceptable to sacrifice fit to increase interpretability (Cudeck & Browne, 1983).

Reference variable. As stated above, a reference variable needs to be chosen to link the metrics of the latent variable(s) by group. The reference variable must be invariant or the accuracy of the invariance testing is distorted (Bollen, 1989; Cheung & Rensvold, 1999; Yoon & Millsap, 2007) and the results are misleading (Johnson et al., 2009). Thus, an invariant observed variable needs to be used as the reference variable; however, invariance is rarely known a priori. As stated by French and Finch (2008), this leads to a circular situation where the reference variable needs to be invariant, invariance of parameters are established by estimating a model, and we cannot estimate a model for invariance without an invariant reference variable. There are many methods in factor analysis to identify which observed variable to use as a reference variable such as using modification indices (Yoon & Millsap, 2007), the factor-ratio test (Cheung & Rensvold, 1999), and the list-and-delete method (Rensvold & Cheung, 2001). Here, I describe a two-step approach to selecting a reference variable.

An empirical method that uses modification indices to identify which observed variable to use as the reference variable is the smallest modification index procedure with a partial invariance model (Jung & Yoon, 2017). First, a full scalar invariance model (i.e., a factor model where the loadings and intercepts are constrained to be equal across groups) is fit to the data. Then, the modification indices of just the invariance constraints (loadings and intercepts) are examined. A cutoff value is chosen for the modification indices. One choice is to use the value of 3.84, which is the value of the χ^2 distribution with one degree of freedom associated with a *p*-value of .05. Others have used a cutoff value of 5.0 (Byrne, Shavelson, & Muthén, 1989). However, the modification indices are affected by sample size such that larger sample sizes are associated with larger values of the modification indices. If there are any modification indices above the chosen cutoff value, then the parameter with the highest modification index is freed to vary across groups. This new model is estimated and the modification indices are once again examined. This procedure stops until there are no modification indices associated with invariance constraints above the chosen cutoff value left in the model. This model is considered to be the baseline model. Finally, the observed variable with the smallest modification index in the baseline model is chosen as the reference variable. In Jung and Yoon's (2017) simulation study, there was 99.4% accuracy with identifying an invariant reference variable across all conditions (sample size, location of non-invariance [loading or intercept], and size and pattern of non-invariance) where four out of the six observed variables were invariant; however, they did not simulate non-invariance in models with any model misspecification.

Testing. A forward approach of sequentially adding more model constraints is used to establish measurement invariance. Specifically, four hierarchically nested models are tested and compared: the configural invariance model, the metric invariance model, the scalar invariance model, and the strict invariance model (Meredith, 1993; Millsap, 2011; Vandenberg & Lance, 2000; Widaman & Reise, 1997). In the configural invariance model, as modeled in Equation 5, the dimensions of the loading matrices $\boldsymbol{\Lambda}_g$ are constrained to be equal (i.e., the groups have the same number of factors) and all zero loadings are in the same location across factor loading matrices (i.e., the same observed variables load on the same factors across groups). All remaining parameters (nonzero λs,

$\boldsymbol{\tau}_g$, and $\boldsymbol{\Theta}_g$) are free to vary across groups with the exception of the parameters constrained to invariance for identification purposes. The metric invariance model is identical to the configural invariance model except that the factor loading matrices $\boldsymbol{\Lambda}_g$ are constrained to be equal across groups ($\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$). The scalar invariance model is identical to the metric invariance model except that the $\boldsymbol{\tau}_g$ vectors are constrained to be equal across groups ($\boldsymbol{\tau}_g = \boldsymbol{\tau}$). Finally, the strict invariance model is identical to the scalar invariance model except the $\boldsymbol{\Theta}_g$ matrices are constrained to be equal across groups ($\boldsymbol{\Theta}_g = \boldsymbol{\Theta}$).

The four invariance models are nested and can be statistically compared using uncorrected χ^2 difference tests (e.g., the difference in χ^2 between the scalar invariance model and the metric invariance model; Bentler & Bonett, 1980; Bollen, 1989) or corrected χ^2 difference tests (Brace & Savalei, 2017; Satorra & Bentler, 2010). If the less-constrained (i.e., less invariant) model has good fit, the difference in the χ^2 statistics is distributed as a χ^2 with degrees of freedom equal to the number of estimated factor model parameters that differ between the models (Millsap, 2011, p. 194; Steiger, Shapiro, & Browne, 1985). If the difference test is significant, then the added invariance restrictions significantly worsen the fit of the model. If the difference test is not significant, the more parsimonious (i.e., the more invariant) model is chosen and the next invariance model is tested. The strict invariance model, in addition to not having significantly different fit from the scalar invariance model, should have good fit overall (Millsap, 2011).

Partial invariance. At any level of invariance, invariance may not hold (based on the significance of the χ^2 difference test) and thus the sequence of testing the hierarchical invariance models stops. However, it may be unrealistic to expect full measurement invariance to hold (Horn, 1991; Horn, McArdle, & Mason, 1983; Steenkamp & Baumgartner, 1998). It is possible that only one or a few parameters that were constrained to invariance are causing the misfit. Thus, partial invariance can be tested for where only some of the parameters are constrained to invariance (Byrne et al., 1989). For instance, if the test of scalar invariance does not hold, one can test a model where only a subset of the intercepts is constrained to invariance.

A specification search is needed to identify the location of the non-invariance. To determine which parameters should be non-invariant, a forward or backward approach can be adopted. If most of the parameters tested are invariant, it is more efficient to use a backward approach by releasing constraints from the more invariant model than to use a forward approach, which adds constraints to the less invariant model (Meade & Lautenschlager, 2004). If using the backward approach, modification indices and expected parameter change statistics can be used to identify which specific parameter is the source of the most misfit (Jöreskog & Sörbom, 1983, p. I.40-I.42; Reise, Widaman, & Pugh, 1993). However, using modification indices typically leads to incorrect conclusions (MacCallum, Roznowski, & Necowitz, 1992; Millsap, 2005). To minimize capitalizing on chance and increase generalizability, only parameters with severe violations should be freed to be non-invariant (MacCallum et al., 1992; Steenkamp & Baumgartner, 1998). Additionally, substantive theory should help guide decisions with regards to identifying non-invariant parameters, although this is rarely available (Steenkamp & Baumgartner, 1998).

Rather than testing the sequence of hierarchically nested models, we can test each parameter for invariance individually by calculating confidence intervals (Meade &

Bauer, 2007) or bias-corrected bootstrap confidence intervals (Cheung & Lau, 2012) of the difference in factor loadings or intercepts. In the latter approach, the configural invariance model is estimated and the bias-corrected bootstrap confidence intervals of the difference in factor loadings for the non-reference variables are calculated. If the confidence interval contains zero, then the parameter is considered to be invariant. Otherwise, it is categorized as non-invariant. Then, a metric or partial metric invariance model is estimated based on the previous results. Bias-corrected bootstrap confidence intervals of the difference in measurement intercepts are then calculated. This procedure is easily implemented in M*plus* (Muthén & Muthén, 1998-2014) via the MODEL CONSTRAINT command. This procedure can be used as a way to investigate partial invariance when a level of invariance does not hold. However, the power of these tests is influenced by many factors including sample size, factor overdetermination, item communality, and size of the factor loadings (Cheung & Lau, 2012).

There are different options on how to handle non-invariant observed variables in factor analysis (Cheung & Rensvold, 1998; Millsap & Kwok, 2004; Sass, 2011). The five options detailed by Sass (2011) are: 1) do not use the factor models or factor scores, 2) interpret scores independently and do not make any group comparisons, 3) delete the non-invariant variables from the model, 4) constrain non-invariant variables to invariance anyways, and 5) use a partial invariance model. The third option, while common in practice, is not desirable because it is possible that researchers would be using different variables to represent the same construct or scale (Millsap & Kwok, 2004). The justification for using the fourth option is the assumption that the population differences on the parameters are minimal even though they may not be in the specific sample (Horn et al., 1983). Instead of choosing only one of these options, multiple options can be analyzed. If the conclusions are drastically different, then further work needs to be done to determine which option is more valid. The number of non-invariant parameters and the size of the non-invariance impacts how different the options are (Sass, 2011).

The fifth option, using a partial invariant model, requires more detailed steps. Once the non-invariant parameter or parameters are identified, they are freed to vary across groups in the partial invariance model. Then, the partial invariance model is statistically compared to the baseline model. If the χ^2 difference test is not significant (or the change in another fit statistic (e.g., RMSEA) is below a recommended benchmark), then the sequence of invariance testing can continue (Reise et al., 1993). This is because full invariance at one level (e.g., metric invariance) is not needed to test invariance at the next level (e.g., scalar invariance; Byrne et al., 1989).

Because testing for partial invariance is a post-hoc fitting procedure, there are many criticisms against using it due to its exploratory nature (MacCallum, 1986). First, if there are many non-invariant parameters, then multiple χ^2 difference tests are analyzed, but the Type I error rate is not controlled for (Green, Thompson, & Babyak, 1998; Kaplan, 1989). To remedy this, we can use a Type I error correction such as the false discovery rate procedure (Benjamini & Hochberg, 1995). Additionally, the partial invariance models are modified to increase model fit to a particular data set. Because of sampling error, the sample may not be representative of the population and that could be the cause of non-invariance (Horn et al., 1983). To combat this issue, the model should be cross-validated with an independent sample (Anderson & Gerbing, 1988). One method to cross-validate the model is to split the data set into a calibration sample and a validation sample (Bentler, 1980). Rather than using the full sample to establish measurement invariance, researchers can use a calibration sample and make modifications (e.g., allow a parameter to freely vary across groups) to improve model fit. The final model with the empirical modifications is then tested using the validation sample. If the model has good fit in the validation sample, then the modifications were appropriate and the model is generalizable.

There is debate as to how much partial invariance is too much. Steenkamp and Baumgartner (1998) argue that only one other indicator other than the reference variable needs to be invariant to have meaningful group mean comparisons; however, they acknowledge that having more invariant parameters is desired. Dimitrov (2010) suggested that no more than 20% of the factor model parameters should be freed to vary across groups; however, there is no empirical support for this suggestion. Mean comparisons on the common factor are valid with a partial invariant model (Byrne et al., 1989) and power to detect mean differences on the common factor is minimally affected when there are non-invariant parameters (Kaplan & George, 1995; Whittaker, 2013). However, if the majority of the indicators are non-invariant, is the same construct being measured in both groups? Sometimes, it is rational to believe the measured constructs differ between groups (including the same people measured at different times). For instance, controlling for the latent variable, Dutch soldiers pre- and post-deployment perceive the symptoms of post-traumatic stress disorder differently presumably due to the trauma of war (Lommen, van de Schoot, & Engelhard, 2014). In this case, mean comparisons of this measure should not be conducted. Millsap and Kwok (2004) note that the purpose of the measure drives how much partial invariance should be tolerated.

Statistically comparing measurement invariance models. One of the issues with relying on χ^2 difference tests to determine invariance is that the χ^2 difference test is sensitive to sample size (Brannick, 1995; Kelloway, 1995) such that larger sample sizes are associated with more significant statistical tests even when the difference in parameters is trivial (Wu, Li, & Zumbo, 2007). Rather than solely relying on the χ^2 difference test, changes in other fit statistics can be used to evaluate invariance (Chen, 2007; Cheung & Rensvold, 2002; Little, 1997; Meade, Johnson, & Braddy, 2008). For instance, if the change in CFI (ΔCFI) between nested models is less than .01, then the null hypothesis (that the fit is the same for both models) should not be rejected and the more parsimonious (i.e., invariant) model should be used (Cheung & Rensvold, 2002). (Note: the change in fit indices are calculated by subtracting the fit index of the more invariant model from the fit index of the less invariant model.) Meade et al. (2008), on the other hand, recommended using a ΔCFI < .002 to establish metric or scalar invariance. However, the CFI is not very sensitive to changes in the mean structure, which is important for invariance testing (Chen, Sousa, & West, 2005). Chen (2007) recommended different cutoffs based on whether metric invariance, scalar invariance, or strict invariance was evaluated. If the total sample size is greater than 300 and the sample sizes are equal across group (i.e., 150 participants in each group), then metric invariance is established if ΔCFI < .010 and if either ΔRMSEA < 0.015 or ΔSRMR < 0.030. Scalar or strict invariance is established if ΔCFI < .010 and if either ΔRMSEA < 0.015 or ΔSRMR < 0.010. Ideally, the changes of these goodness-of-fit statistics should not be affected by sample size; however, there are conflicting results concerning how sample size impacts ΔCFI. Some researchers have found the ΔCFI to not be impacted by sample size (Chen, 2007; Cheung & Rensvold, 2002), others found that as sample size increased, ΔCFI increased (Meade & Bauer, 2007), and still others have found that as sample size decreased, ΔCFI decreased (Kang, McNeish, & Hancock, 2016).

While sample size is one of the factors that influence detection of non-invariance (Meade & Lautenschlager, 2004), the χ^2 difference tests and change in fit statistics can be affected by many other factors. The ratio of group sample sizes impacts the detection of non-invariance, such that if the sample sizes are equal, ΔCFI, ΔRMSEA, and ΔSRMR tend to be larger than if the sample sizes are unequal and Type II error (i.e., concluding invariance when there is non-invariance) increases as the sample sizes become more equal (Chen, 2007). Additionally, the communalities of the variables (proportion of variance in the variable accounted for by the common factors) affect detection of non-invariance such that χ^2 difference tests perform better with higher communalities (increased power and more accurate; Meade & Bauer, 2007; Meade & Lautenschlager, 2004) and as the communalities of variables with non-invariance increased, ΔCFI decreased (Meade & Bauer, 2007). χ^2 difference tests have more power to detect measurement invariance when factor overdetermination (i.e., the ratio of indicators to factors) is high and the non-invariance pattern was mixed (e.g., one group had higher loadings for some observed variables and lower loadings for other observed variables compared to the other group; Meade & Bauer, 2007). Finally, as the magnitude of the loadings increased (i.e., improved measurement quality), the ΔCFI decreased (Kang et al., 2016).

In conclusion, while there are recommended cutoffs for the changes in fit statistics (including the χ^2 difference test), these should be used with caution because the changes in fit statistics are affected by many factors. Wu et al. (2007) and Kline (2011) point out that the cutoffs for ΔCFI, ΔRMSEA, etc. are based on limited simulation conditions and may not generalize to all data sets and models. For instance, Kang and colleagues (2016) do not recommend using ΔCFI because they found it to be affected by measurement quality, which has not often been manipulated in simulation studies. In addition, others argue that approximate fit statistics like RMSEA and CFI should never be used because, by their nature, they are not precise (Barrett, 2007). These complicated issues with using fit statistics to detect non-invariance highlight the need for using effect sizes to better understand the group parameter differences and the measure itself.

Effect Size Measures

Effect size measures are "a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (Kelley & Preacher, 2012, p. 140) and "provide information about the magnitude" of the effect being studied (Durlak, 2009, p. 917). For instance, an effect size can quantify the difference between two parameters. There are three purposes for using effect sizes: power analysis, research synthesis, and research reporting (Steinberg & Thissen, 2006). In the case of assessing non-invariance, effect sizes are used for research reporting purposes. The American Psychological Association (2001) recommends reporting effect size estimates for all effects studied in addition to the significance of those effects. This is because the *p*-value does not provide information on the magnitude of the effect. Statistical significance and effect size measures supplement each other and should both be used when making research decisions (Fan, 2001). Thus, researchers should use effect size measures to aid in the invariance testing process.

While effect size measures can provide additional information about the statistical test, they can only be interpreted in a specific research context. The size of the effect does not determine its practical or clinical value (Durlak, 2009). For instance, changing behaviors is more difficult than changing attitudes (Durlak, 2009). Thus, an effect size value of 0.5 may be meaningful when studying behavioral changes, but negligible when studying changes in attitudes. Additionally, just because an effect has a small magnitude does not mean that the effect does not have practical implications and importance. A biomedical study investigating the effects of taking aspirin on number of heart attacks found the magnitude of the effect ($r^2 = .0012$) to be “so small as to be considered quantitatively unimpressive by methodological convention” (Rosnow & Rosenthal, 1989, p. 1279). However, this effect was considered to be so impactful on the health of the participants that the Steering Committee of the Physicians’ Health Study Research Group (1988) told participants in the control group to start taking aspirin. Because the outcome was life or death, and the cost of treatment was small, the effect was considered to be practically significant. Effect size measures cannot solely determine practical importance, but can provide additional information about the effect being studied above and beyond the significance test.

Useful and high-quality effect size measures should have four properties: an appropriate scale, calculable confidence intervals, independence from sample size, and good estimation properties (Preacher & Kelley, 2011). The scale of the effect size measure should be appropriate for the research question at hand to increase interpretability. If the outcome of interest is in interpretable units, unstandardized effect size measures are preferable over standardized measures (Baguley, 2009).

Another property of a good effect size is the availability of confidence intervals (Wilkinson & Task Force on Statistical Inference, 1999). When used for research reporting, it is important to note that the calculated effect sizes are estimates and thus are subject to sampling error just like any other sample statistic. It is possible to find a large effect even though the effect size in the population is small due to sampling error, especially if the sample size is small (Fan, 2001).

Even though the effect size is subject to sampling error, which decreases as the sample size increases, the point estimate of the effect size should not be affected by sample size (Preacher & Kelley, 2011). In other words, two researchers studying the same phenomenon should come to the same conclusion about the effect size regardless of their sample sizes.

Finally, the effect size measure should have good estimation properties. Specifically, the effect size should be consistent (as sample size increases, the sample estimate converges to the population value), unbiased (the sample value in expectation equals the population value), and efficient (low sampling variability).

Effect size measures of non-invariance in IRT. An item response theory (IRT) model can be used instead of a factor analysis model when the indicators are categorical. Invariance testing in the IRT framework is known as differential item functioning (DIF). Detailing the models and testing process in IRT is beyond the scope of this paper; however, in this framework, many effect size measures of DIF have been developed and thus it is worth mentioning the importance of the work done in this area. Interested readers should refer to Meade (2010), who developed a taxonomy for categorizing effect size measures used to measure DIF in an IRT framework.

Effect size measures of non-invariance in factor analysis. One of the unresolved problems of measurement invariance is using an effect size measure to assess degree of non-invariance (Millsap, 2005). There have been attempts to create an effect size measure of non-invariance for continuous indicators; however, they are not widely used and their properties (e.g., consistency, bias) have not been studied.

One effect size that measures non-invariance for continuous indicators in mean and covariance structure (MACS) analysis is d_{MACS} (Nye & Drasgow, 2011). The formula for this effect size is:

$$d_{\text{MACS}} = \frac{1}{SD_{iPooled}} \sqrt{\int \left(\hat{Y}_{i1} - \hat{Y}_{i2} | \eta \right)^2 f_2(\eta) d\eta}\,, \tag{7}$$

where $SD_{jPooled}$ is the pooled within-group standard deviation of indicator *i* for Group 1 and Group 2, $\hat{Y}_{i1}$ is the expected observed score on indicator *i* using factor model parameters for Group 1, $\hat{Y}_{i2}$ is the expected observed score on indicator *i* using factor model parameters for Group 2, and $f_2(\eta)$ is the distribution of the latent variable for Group 2 only. Nye and Drasgow (2011) define the pooled within-group standard deviation as:

$$SD_{iPooled} = \frac{(N_1 - 1)SD_1 + (N_2 - 1)SD_2}{(N_1 - 1) + (N_2 - 1)}\,, \tag{8}$$

where N_1 is the sample size of Group 1, N_2 is the sample size of Group 2, SD_1 is the standard deviation of the indicator for Group 1, and SD_2 is the standard deviation of the indicator for Group 2. It should be noted that this formula is not the standard way of calculating a pooled standard deviation. Typically, the pooled variance is estimated and then the square root of that pooled variance is calculated to estimate the pooled standard deviation. The expected observed score for indicator *i* and Group 1, $\hat{Y}_{i1}$, at a particular value of the latent variable is calculated as:

$$\hat{Y}_{i1} = \tau_{i1} + \lambda_{i1}\eta, \tag{9}$$

where τ_{i1} is the measurement intercept for indicator *i* and Group 1, λ_{i1} is the factor loading for indicator *i* and Group 1, and η is the value of the latent variable being evaluated. We can apply the same formula to calculate the expected observed score for Group 2, $\hat{Y}_{i2}$, by using Group 2 parameters (i.e., τ_{i2} and λ_{i2}). In IRT, if we plot the expected item score (in actuality, we are plotting the probability of endorsing an item or answering the item correctly in the binary case) against the latent variable, the resulting curve is called an item response function or trace line. There is no corresponding terminology in factor analysis. Thus, I refer to the regression line produced by Equation 9 as an indicator response function (IRF). This effect size measure, d_{MACS}, is interpreted as the standardized average difference in expected indicator scores across a normal latent variable distribution for Group 2 assuming the differences were uniform. The larger the value of d_{MACS}, the greater the magnitude of non-invariance. Thus, we prefer smaller values of d_{MACS} compared to larger values.

Even though the intended use for d_{MACS} is in the factor analytic framework, I describe how it would be categorized in the four dimensions of Meade’s (2010) IRT taxonomy to describe the components of this effect size. First, d_{MACS} is measured at the indicator level. In other words, d_{MACS} can be calculated for each observed variable. Second, an assumed distribution (with estimated parameters) is used rather than sample estimates of latent variable scores to calculate the effect size. Specifically, a normal latent variable distribution is assumed for Group 2. Most effect sizes in the IRT framework only use Group 2 participants to determine the adverse impact of item or test scores (Flowers, Oshima, & Raju, 1999; Raju, van der Linden, & Fleer, 1995; Stark, Chernyshenko, & Drasgow, 2004). Group 2 is typically the group suspected of being penalized by bias (e.g., females, African-Americans) or the lowest-scoring group and is typically referred to as the focal group in IRT. Third, this effect size does not allow for cancellation. At the indicator level, cancellation can occur when evaluating across the latent variable distribution. If the factor loadings for an indicator are different across groups, then it is possible that at some levels of the latent variable distribution one group is expected to score higher than the other group whereas at other levels of the latent variable distribution the reverse is true. Thus, positive differences and negative differences can be summed together and cancel each other out. Therefore, it is possible to have an effect size value of zero even when there is non-invariance because of cancellation. Because the group difference in expected indicator scores in d_{MACS} is squared, the sign of the squared difference will always be positive and thus cancellation cannot occur for that effect size. Fourth and finally, d_{MACS} is in a standardized metric. Standardizing the effect size is important for comparing the effect size across continuous indicators because the indicators can have vastly different scales.

Millsap and Olivera-Aguilar (2012) developed an effect size measure of indicator non-invariance when there are two groups and metric invariance holds, but scalar invariance does not. In this scenario, the mean group difference in observed indicators can be expressed as:

$$\mu_{i1} - \mu_{i2} = (\tau_{i1} - \tau_{i2}) + \boldsymbol{\lambda}_i(\boldsymbol{\kappa}_1 - \boldsymbol{\kappa}_2), \tag{10}$$

where μ_{i1} is the observed mean of indicator *i* for Group 1, τ_{i1} is the measurement intercept of indicator *i* for Group 1, $\boldsymbol{\lambda}_i$ is the *i*th row of the factor loading matrix, and $\boldsymbol{\kappa}_1$ is the vector of common factor means for Group 1. The remaining parameters in the equation have the same interpretation except applied to Group 2. This equation illustrates that the difference in observed indicator means $(\mu_{i1} - \mu_{i2})$ is affected by the group difference in intercepts and the group difference in factor means. The portion of the observed mean difference on indicator *i* due to the group difference on the intercepts can be calculated via the ratio of the intercept difference as:

$$\frac{\tau_{i1} - \tau_{i2}}{\mu_{i1} - \mu_{i2}}. \tag{11}$$

This ratio is the portion of the group difference in the observed means that can be explained by non-invariance. The remaining portion is the group difference that can be explained by group differences in the latent variable means. A drawback to this effect size is that the observed group difference in indicator means can be zero, leading to an undefined number of the effect size. Additionally, the effect size measure can be negative if the difference in observed means is of the opposite sign to the difference in measurement intercepts, making interpretation difficult. Millsap and Olivera-Aguilar (2012) did not name this effect size nor provide benchmarks; however, Millsap and Kim (2018) argue that an intercept difference to observed indicator mean difference ratio of 1:2 or larger prevents researchers from making valid group mean comparisons on that indicator.

Millsap and Olivera-Aguilar (2012) also developed an effect size measure of indicator non-invariance when scalar invariance holds, but strict invariance does not. In this scenario, the variance of indicator *i* for group *g* can be given as:

$$\sigma_{ig}^2 = \boldsymbol{\lambda}_i' \boldsymbol{\Psi}_g \boldsymbol{\lambda}_i + \Theta_{ig} \,. \tag{12}$$

In a similar vein to the effect size above, the portion of group difference in the observed variances due to non-invariance can be calculated as:

$$\frac{\Theta_{i1} - \Theta_{i2}}{\sigma_{i1}^2 - \sigma_{i2}^2} \,. \tag{13}$$

The remaining portion is due to group difference on the factor distribution ($\boldsymbol{\Psi}_g$). Again, this effect size can be negative, which makes interpretation difficult.

Proposed effect size measures of non-invariance in factor analysis. Inspired by the effect sizes of DIF in IRT and by d_{MACS}, I propose four new effect size measures of measurement non-invariance for continuous outcomes in the factor analytic framework. First, the signed difference in expected indicator scores for Group 2 (SDI_2) is defined as:

$$SDI_2 = \frac{\int_{-\infty}^{\infty} \left[\hat{Y}_{i1} - \hat{Y}_{i2} | \eta\right] \cdot f_2(\eta) d\eta}{SD(indicator)_2} \,, \tag{14}$$

where the parameters have the same interpretation as before in the formula for d_{MACS} except $SD(indicator)_2$ is the standard deviation of the observed indicator scores for Group 2. This measure is similar to d_{MACS}, but it allows for cancellation across the latent variable distribution (the sign of the difference in expected indicator scores $[\hat{Y}_{i1} - \hat{Y}_{i2}]$ is preserved and is thus called a signed measure). Additionally, the denominator differs from d_{MACS}. The numerator measures the impact of non-invariance on Group 2 participants. Thus, the denominator should refer to Group 2 only as well.[1]

To create an unsigned version of SDI_2, the absolute value of the difference in expected indicator scores can be calculated instead of the raw difference. Specifically, the unsigned difference in expected indicator scores for Group 2 (UDI_2) is calculated as:

$$UDI_2 = \frac{\int_{-\infty}^{\infty} \left| \hat{Y}_{i1} - \hat{Y}_{i2} | \eta \right| \cdot f_2(\eta) d\eta}{SD(indicator)_2}. \quad (15)$$

The parameters have the same interpretation as they do for SDI_2. Here, the absolute value of the group differences in expected indicator scores are calculated to put UDI_2 on the same metric as SDI_2, which is a different metric than the metric of d_{MACS}.

I intend for the SDI_2 and UDI_2 to be used together when assessing non-invariance because they provide information independent of each other. Chalmers, Counsell, and Flora (2016) discussed the different combinations of values for signed and unsigned measures, which I modified in Table 1. Elaborating on the table, there are predictable ways we know the two measures will behave even though there is not necessarily a deterministic relationship. As illustrated in Figure 1, first, unsigned effect size measures will always be positive since either the negative differences are squared or the absolute value of the difference is calculated. Second, if they are on the same scale, the unsigned effect size measure will always be greater than or equal to the signed effect size measure since the signed effect size allows for cancellation. If the indicator is invariant, then both

[1] I ran all analyses using a (corrected) pooled standard deviation in the denominator for all five studied effect sizes and the conclusions did not change (e.g., standardized bias conclusions were the same). Additionally, the pooled and not pooled versions of the effect sizes were correlated .99 for all effect sizes. Thus, I chose the denominator that was more interpretable.

effect size values will be zero, as illustrated by the filled-in circle in Figure 1. If the loadings are invariant, but the intercepts are non-invariant, then the signed effect size will be equal to the unsigned effect size, which is shown by the point-up triangles on the 45-degree angles. The pattern of the point-down triangles represents when there is complete cancellation (i.e., the loadings are non-invariant and the cross of the IRFs occurs at the mean of the symmetrical factor distribution). In conclusion, both the signed and unsigned measures provide independent pieces of information and thus both should be calculated in conjunction with one another to evaluate the magnitude of non-invariance.

Most effect sizes of non-invariance in the IRT framework as well as d_{MACS}, SDI_2, and UDI_2 measure the impact of non-invariance for Group 2 only. Specifically, the effect sizes compare the expected indicator scores of people in Group 2 compared to the expected indicator scores of people in Group 2 if they were instead in Group 1. However, what is typically of practical interest is comparing how the expected indicator scores change when modeling invariance versus allowing the groups to have different measurement model parameters. More accurately, if the amount of non-invariance is not problematic, rather than analyzing a model separately in each group and constraining parameters to be invariant, many researchers combine the groups and analyze the factor analytic model for the entire sample (see Sandler, Wolchik, Mazza, Gunn, Tein, Berkel, Jones, & Porter (2019) for an example). Braun and Holland (1982) referred to the combination of populations as a synthetic population. In this case, the comparison of interest is between the expected indicator score when non-invariance is modeled in a multiple-group model to the expected indicator score when the measurement model is estimated for the synthetic sample. Additionally, there is utility in considering how Group

1 is affected by non-invariance in addition to how Group 2 is affected. Thus, the third effect size measure of non-invariance I propose is the weighted signed difference in expected indicator scores (*WSDI*), which is defined as:

$$WSDI = p_1 \frac{\int_{-\infty}^{\infty} \left[\hat{Y}_{i1} - \hat{Y}_{i\mathrm{S}} | \eta\right] \cdot f_1(\eta) d\eta}{SD(indicator)_1} + p_2 \frac{\int_{-\infty}^{\infty} \left[\hat{Y}_{i\mathrm{S}} - \hat{Y}_{i2} | \eta\right] \cdot f_2(\eta) d\eta}{SD(indicator)_2}, \quad (16)$$

where p_1 is the proportion of people in Group 1, p_2 is the proportion of people in Group 2, $\hat{Y}_{iS}$ is the expected indicator score on indicator *i* using parameters from a single-population model (which needs to be equated to the multiple-group model), and the other parameters have the same meaning as before. The proportions are calculated using the observed sample sizes and thus assume the observed proportions match the population proportions. Rather than comparing the expected indicator scores using the parameters for the two groups like for the *SDI*$_2$, the comparison for the *WSDI* is between the expected indicator score using separate group parameters to the expected indicator score using the synthetic group parameters.

The unsigned version of *WSDI* is the weighted unsigned difference in expected indicator scores (*WUDI*), which is defined as:

$$WUDI = p_1 \frac{\int_{-\infty}^{\infty} \left|\hat{Y}_{i1} - \hat{Y}_{iS} | \eta\right| \cdot f_1(\eta) d\eta}{SD(indicator)_1} + p_2 \frac{\int_{-\infty}^{\infty} \left|\hat{Y}_{i\mathrm{S}} - \hat{Y}_{i2} | \eta\right| \cdot f_2(\eta) d\eta}{SD(indicator)_2}. \quad (17)$$

This effect size is identical to the effect size in Equation 16 except the absolute values of the expected indicator score differences are calculated and analyzed.

Present Study and Hypotheses

I conducted Monte Carlo simulations to study the properties of the four proposed effect size measures of measurement non-invariance (i.e., *SDI*$_2$, *UDI*$_2$, *WSDI*, and *WUDI*) and of d_{MACS} to answer three research questions. First, I tested if the five effect size measures were unbiased and consistent. I anticipated that the three unsigned measures would exhibit bias for the truly invariant indicators. For invariant indicators, the value of the unsigned effect size measures is zero in the population (there are no group differences in expected indicator scores). The sample estimate of the unsigned effect size will always be zero or a positive number. Because the sample value of the effect size is not expected to be zero in every simulated data set, this will result in positive bias of the sample values of the unsigned effect sizes. I also expected the five effect sizes to be consistent. Congruence between population values and sample estimates in a factor analysis model increases as sample size increases (MacCallum, Widaman, Zhang, & Hong, 1999). Because the effect sizes are calculated using parameter estimates (e.g., loadings), I expected this to cause the effect sizes to be consistent.

Second, I tested if the total sample size, ratio of Group 1 sample size to Group 2 sample size, magnitude of non-invariance, location of non-invariance (e.g., loadings, intercepts), or the latent variable distribution of Group 2 affected the value of the effect sizes. I anticipated that the magnitude of non-invariance would be an important predictor (significant and explains a large portion of the variance of the outcome) of the value of the effect size measure for all five effect sizes. I hypothesized that the latent variable distribution of Group 2 affected the value of the effect sizes for the indicators that had a non-invariant loading, but not for the invariant indicator or the indicator with just a non-invariant intercept. If the IRFs are identical or parallel, then the difference in expected indicator scores is uniform at each level of the latent variable. If, however, the IRFs cross, then the difference in expected indicator scores depends on the value of the latent variable. Where the latent variable distribution of the groups is centered in relation to where the IRFs cross affects how much cancellation occurs and how much weight is given to the bigger group differences in expected indicator scores. I expected sample size and balance of the group sample sizes to not be meaningful predictors because I expected the point estimate of the effect sizes to not be affected by sample size, which is a property of high-quality effect size measures (Preacher & Kelley, 2011).

Finally, I investigated the relationships among the five measures. I expected d_{MACS} to be highly related to UDI_2. While there are differences in how these effect sizes are calculated, I did not have reason to believe that they would be differentially affected by my simulation factors and thus I expected there to be a monotonic relationship between their population values. I also expected those two effect sizes to have a high relationship to *WUDI* but that the relationship would be weaker because of the different conceptualization for *WUDI*. I anticipated the two signed effect sizes to have a moderate to high relationship, similar to the relationship between UDI_2 and *WUDI*. Finally, I expected the signed and unsigned versions of the same effect size (e.g., *WSDI* and *WUDI*) to have a weaker relationship compared to the like-signed effect size measures (e.g., *WSDI* and SDI_2) because the signed and unsigned versions of the same effect size provide independent pieces of information.

CHAPTER 2

METHOD

Manipulated Simulation Factors

A full factorial Monte Carlo simulation with five factors was implemented. The five manipulated factors were: (1) total sample size (three levels), (2) balance of group sample sizes (two levels), (3) magnitude of non-invariance (three levels), (4) location of non-invariance (four levels), and (5) latent variable distribution of Group 2 (three levels). In total, there were $3 \times 2 \times 3 \times 4 \times 3 = 216$ design cells. The four levels of location of non-invariance, which are described in detail below, were captured in one replication. Thus, even though there were 216 design cells, only $216 \div 4 = 54$ different types of models were simulated. For each type of model, 1,000 replications[2] were simulated for a total of 54,000 samples. Previous simulation studies of measurement invariance have used this number of replications per design cell (Cheung & Rensvold, 2002; Fan & Sivo, 2009; French & Finch, 2008). Because each sample contains four different locations of non-invariance, there were a total of 216,000 records that were used in the analyses.

The total sample sizes investigated were 300, 500, and 1,000. These sample sizes, or similar sample sizes, were used in previous simulation studies of measurement invariance (e.g., Chen, 2007; Stark, Chernyshenko, & Drasgow, 2006; Yoon & Millsap, 2007) and are representative of the sample sizes used in empirical studies. For example, 244 children participated in the efficacy trial of the Family Bereavement Program (Sandler et al., 2003) and 749 children participated in the La Familia study (Gonzales,

[2] The Monte Carlo standard error (a measure of between-simulation variability) was calculated for each effect size within each design cell. The largest value was 0.0008. Thus, the number of replications was deemed acceptable.

Knight, Gunn, Tein, Tanaka, & White, 2018). The balance of the two group sample sizes was simulated to be a ratio of either 1:1 or 2:1. For instance, if the total sample size was 300 and the 1:1 sample size ratio was used, then Group 1 was simulated to have a sample size of 150 and Group 2 was simulated to have a sample size of 150. If the 2:1 sample size ratio was used, then Group 1 was simulated to have a sample size of 200 and Group 2 was simulated to have a sample size of 100. These ratios reflect ratios commonly seen in invariance analyses (e.g., treatment to control ratio in a study is typically 1:1 whereas racial or ethnic ratios can be closer to 2:1 depending on the specific categorization). For example, 886 parents participated in the effectiveness trial of the New Beginnings Program with 409 parents assigned to the two-session control condition and 477 parents assigned to the ten-session treatment condition (Sandler et al., 2019). Of those 886 parents, 526 were non-Hispanic white and 280 were Hispanic (80 parents were categorized as another race or ethnicity).

The magnitude of non-invariance was simulated to be small, medium, or large. A small magnitude of non-invariance was defined as a raw difference of 0.10 in the loadings and 0.20 in the intercepts. A medium magnitude of non-invariance was defined as a difference of 0.25 in the loadings and 0.40 in the intercepts. Finally, a large magnitude of non-invariance was defined as a difference of 0.40 in the loadings and 0.60 in the intercepts. In all cases, Group 2 was simulated to have a smaller (or more negative) loading or intercept compared to Group 1. These values were chosen based on previous research that defined these differences when the factor was standardized for one group (Kim, 2011; Stark, Chernyshenko, & Drasgow, 2006; Yoon & Millsap, 2007). Because the factor was standardized and the indicators were simulated to have an expected variance of 1 (explained later) for Group 1, the loadings for Group 1 were in a standardized metric.

The four levels of location of non-invariance were: (1) the loading of the indicator was non-invariant, (2) the intercept of the indicator was non-invariant, (3) the loading and the intercept of the indicator were non-invariant, and (4) neither the loading nor the intercept of the indicator was non-invariant (i.e., the parameters of the indicator were invariant). While these effect size measures are not expected to be calculated for invariant indicators in practice, I did so to compare the magnitude of the effect size due to non-invariant parameters to the magnitude of the effect size due to sampling error of invariant parameters. Lastly, impact (i.e., the group mean difference of the latent variable) was manipulated by varying the population latent variable distribution for Group 2 across simulations. For all design cells, latent variable scores for participants in Group 1 were randomly drawn from a standard normal distribution. Latent variable scores for participants in Group 2 were randomly drawn from one of three normal distributions: $N(0, 1)$, $N(-0.5, 1.3)$, or $N(-0.5, 0.7)$. The first two were chosen based on previous research that used the same distributions (Kim, 2011; Millsap & Kwok, 2004). The last distribution was chosen to tease apart if the differences in effect size values were due to a factor mean difference or a factor variance difference.

Data Generation

For both groups, a one-factor model with eight indicators was simulated using R v. 3.1.2. Data were simulated according to the following equation

$$\boldsymbol{y}_{jg} = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \eta_{jg} + \boldsymbol{\varepsilon}_{jg}, \tag{18}$$

where $\boldsymbol{y}_{jg}$ is an 8×1 vector of indicator scores for person j in group g. Group 1 always had the same factor loading and measurement intercept values for each replication such that $\boldsymbol{\Lambda}_1 = [0.8, 0.9, 0.6, 0.7, 0.8, 0.8, 0.8, 0.8]$ and $\boldsymbol{\tau}_1 = [0.2, 0.4, -0.2, -0.1, 0.0, 0.0, 0.0, 0.0]$. Group 2 was simulated to have the same loading and intercept values as Group 1 for the first five indicators (i.e., the first five indicators were simulated to be invariant across group). The loading of the sixth indicator was simulated to be lower in Group 2, but the intercept for this indicator was invariant. For Group 2, the intercept of the seventh indicator was simulated to be more negative, but the loading for the indicator was invariant. Finally, the loading and the intercept of the eighth indicator were simulated to be lower or more negative in Group 2. The group difference of the loadings and intercepts varied depending on the magnitude of measurement non-invariance. Specifically, for the small magnitude condition, the population loadings for Group 2 were $\boldsymbol{\Lambda}_2 = [0.8, 0.9, 0.6, 0.7, 0.8, 0.7, 0.8, 0.7]$ and the population intercepts were $\boldsymbol{\tau}_2 = [0.2, 0.4, -0.2, -0.1, 0.0, 0.0, -0.2, -0.2]$. For the medium magnitude condition, the population loadings for Group 2 were $\boldsymbol{\Lambda}_2 = [0.8, 0.9, 0.6, 0.7, 0.8, 0.55, 0.8, 0.55]$ and the population intercepts were $\boldsymbol{\tau}_2 = [0.2, 0.4, -0.2, -0.1, 0.0, 0.0, -0.4, -0.4]$. Finally, for the large magnitude condition, the population loadings for Group 2 were $\boldsymbol{\Lambda}_2 = [0.8, 0.9, 0.6, 0.7, 0.8, 0.4, 0.8, 0.4]$ and the population intercepts were $\boldsymbol{\tau}_2 = [0.2, 0.4, -0.2, -0.1, 0.0, 0.0, -0.6, -0.6]$.

Latent variable scores for participants in Group 1, η_{j1}, were randomly drawn from a standard normal distribution (mean = 0, variance = 1). The normal distribution used to generate latent variables scores for Group 2 varied across conditions as described above. The expected variance of an indicator is based on the following equation:

$$\mathrm{var}(indicator) = \lambda \cdot \mathrm{var}(factor) \cdot \lambda + \mathrm{var}(unique). \qquad (19)$$

Thus, for the participants in Group 1 to have an expected variance equal to one for each indicator, the unique factor scores, ε_{ij1}, were randomly drawn from a normal distribution (mean = 0, variance = $1 - \lambda_{i1} \cdot \Psi_1 \cdot \lambda_{i1}$). The unique factor scores for Group 2, ε_{ij2}, were also randomly drawn from the same normal distribution, meaning the unique variances were invariant and that the expected variance of the indicators for Group 2 varied depending on the values of the factor loadings and variance of the common factor for Group 2.

Calculation of Effect Sizes

Group 1 and Group 2 parameters used in the calculation of the effect sizes were taken from an estimated multiple-group one-factor CFA model where the measurement model parameters of the first four indicators were constrained to be invariant across groups and the measurement model parameters for the last four indicators were freed to vary across groups. The model was further identified by standardizing the factor for Group 1.[3] This scenario is similar to the scenario envisioned for applied researchers when they calculate these effect size measures (e.g., after testing for invariance and concluding a partial invariance model is the best-fitting model). For the sample estimate of the effect size measures, the sample estimate of the standard deviation of the indicator was used in the denominator. For the population effect size calculations, the expected standard deviation of the indicators was used in the denominator (see Equation 19). Because of the complexities of integrating across a normal distribution, quadrature was used to

[3] Using different identification constraints (e.g., standardizing the factor for one group or using a reference variable) does not affect the value of the effect size.

approximate the integral in the calculation of the five effect size measures. The group differences in expected indicator scores were evaluated across the range of $-5 \leq \eta \leq 5$ using 101 quadrature nodes spaced 0.1 apart.[4]

For the two weighted effect size measures, the synthetic parameters are from a single-population measurement model. However, the synthetic loadings and intercepts need to be on the same scale as the measurement model parameters from the multiple-group model to make the expected indicator scores comparable. In other words, the same reference group needs to be used to estimate both models or the parameters need to be equated. The former option was used for this study. The sample estimate of the synthetic parameter was calculated by duplicating the entire generated data set for the current replication and labeling the duplicated data as the data for the synthetic group. Then, a three-group measurement model was estimated with Group 1 as the reference group (i.e., the common factor was standardized for this group), Group 2 as the second group, and the synthetic group as the third group. The first five indicators were constrained to be invariant across the three groups. The loadings and intercepts for the last three indicators (i.e., the indicators simulated to be non-invariant) were freely estimated in all groups. The sample estimate of the synthetic parameters were the estimated loadings and intercepts in the third group (i.e., the synthetic group) for the last four indicators (i.e., the studied indicators). The population value of the parameters of non-invariant indicators for the synthetic population cannot be easily calculated because the amount and type of non-

[4] I simulated 12 design cells (4 location × 3 magnitude) with 10 replications using 1,001 quadrature nodes spaced 0.01 apart and there was no appreciable difference (largest difference was 1×10^{-6}) in the effect size sample estimates or population values compared to using 101 quadrature nodes. To reduce computational burden, the latter option was used.

invariance, the relative size of the groups, and the latent variable distribution of the groups affect it. Thus, they were estimated via simulated data. Because four of the five simulated factors (all but total sample size) affect the value of the synthetic parameters, 72 (2×3×4×3) different population synthetic loadings and synthetic intercepts were estimated and used in the calculations of the weighted effect sizes. For each of the 72 design cells, a data set with a total sample size of 100,000 was generated. Similar to the process to obtain the sample estimate, this data set was duplicated to create the data for the synthetic group and a three-group measurement model was estimated with Group 1 as the reference group, Group 2 as the second group, and the synthetic group as the third group. The population synthetic parameters were the estimated loadings and intercepts in the third group (i.e., the synthetic group) for the last four indicators (i.e., the studied indicators).

Evaluation Criteria

The effect size measures were evaluated based on average raw bias, standardized bias, the root mean square error, consistency, and the relation of the effect size measures to one another. Average raw bias refers to the difference between the estimated effect size in the sample and the population value of the effect size divided by the total number of records in a design cell and is defined as:

$$ARB(\hat{\theta}) = \frac{\sum_{r=1}^{R}(\hat{\theta}_r - \theta)}{R}, \tag{20}$$

where $\hat{\theta}_r$ is the parameter estimate for the r^{th} record, θ is the parameter, and R is the number of records. Standardized bias refers to the difference between the average estimated effect size and the average population value of the effect size divided by the standard deviation of the estimated effect size within a design cell, also known as the empirical standard error of that parameter. Specifically, standardized bias is defined as:

$$SB(\hat{\theta}) = \frac{ARB(\hat{\theta})}{SD(\hat{\theta})}. \tag{21}$$

Unacceptable bias was defined as the absolute value of standardized bias being greater than 0.40 (Collins, Schafer, & Kam, 2001; Lai & Kwok, 2016). Parameter variability was assessed by calculating the root mean square error (RMSE) of the parameter estimates across all records in a design cell. The formula for the RMSE of parameter estimate $\hat{\theta}$ is:

$$RMSE(\hat{\theta}) = \sqrt{\frac{\sum_{r=1}^{R}\left(\hat{\theta}_r - \theta\right)^2}{R-1}}. \tag{22}$$

Thus, each design cell had one estimate of average raw bias, standardized bias, and RMSE.

The consistency of the effect size measures was examined analytically by averaging the RMSE values for each level of the total sample size and for each sample size of individual groups. If the marginalized RMSE values decreased as sample size increased, then the effect size was deemed consistent.

To determine the amount of overlapping information of the effect sizes, I calculated the correlations between the like-signed effect sizes (e.g., SDI_2 and *WSDI*). Additionally, I calculated the correlations between the population values of the signed and unsigned versions of the same effect size (e.g., SDI_2 and UDI_2).

Data Analyses

A five-way between-subjects analysis of variance (ANOVA) was conducted to determine if any of the design factors affected the values of the effect size measures.

There were 216 conditions × 1,000 replications = 216,000 records used in the analysis. Because of this large sample size, the power to detect effects with trivial effect sizes was high. Thus, partial η^2 (no relation to the common factor variable) was used. Partial η^2 was calculated via the following formula

$$partial\,\eta^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}} \tag{23}$$

where *SS* refers to the sums of squares. Only effects with an effect size of partial $\eta^2 \geq .01$ (a small effect by Cohen's (1988) standards) are reported and described, which is a criterion that has been used in previous simulation studies (for an example, see Krull & MacKinnon, 1999). Additionally, only pairwise or simple pairwise comparisons that had a Cohen's *d* value ≥ 0.2 are reported and described.

CHAPTER 3

RESULTS

All factor models converged to a proper solution (e.g., no negative variances). Results are organized in terms of bias, consistency, values, and relationship of the effect sizes.

Bias

Table 2 presents the average raw bias of the five effect size measures for each design cell. An ANOVA on the differences between the sample estimate in each record and the population value of the design cell revealed that there was a meaningful effect of sample size and location of non-invariance for the three unsigned effect sizes (i.e., d_{MACS}, UDI_2, and *WUDI*). This is seen in Table 2. The invariant indicator had greater average raw bias values compared to the other three studied indicators for the three unsigned effect sizes. Additionally, the average raw bias decreased as sample size increased.

Table 3 presents the standardized bias values for each design cell for the five effect size measures. Sample estimates of the signed effect size measures (i.e., SDI_2 and *WSDI*) were unbiased in all 216 conditions (average raw bias range for SDI_2 = -0.007 to 0.008, standardized bias range for SDI_2 = -0.078 to 0.095, average raw bias range for *WSDI* = -0.004 to 0.003, standardized bias range for *WSDI* = -0.133 to 0.110). The three unsigned effect size measures had problematic bias (i.e., |standardized bias| greater than 0.40) for all conditions involving the indicator that was simulated to be invariant. For these conditions, the population value of the effect size was 0, or close to 0 in the case of *WUDI* where the synthetic parameters did not always match the multiple-group parameters. (This is because the value of the synthetic parameters are affected by latent variable distributions of the two groups in addition to the values of the non-invariant parameters.) The sample estimates of the unsigned effect size measures can only be positive. Thus, the unsigned effect sizes are positively biased for the invariant indicator. Additionally, the three unsigned effect size measures had problematic bias for some of the conditions involving the indicator with a non-invariant loading but an invariant intercept where the non-invariance was small in magnitude. The standardized bias decreased as the total sample size increased and the bias was less for the Group 2 latent variable distribution with a mean of -0.5 and a variance of 1.3.

Consistency

Table 4 presents the RMSE values for each design cell for the five effect size measures. Table 5 presents the marginal RMSE values by Group 2 sample size for d_{MACS}, SDI_2, and UDI_2. As the Group 2 sample size increased, the three effect sizes became more efficient. Table 6 presents the marginal RMSE values by Group 1 sample size for *WSDI* and *WUDI*. As the Group 1 sample size increased, the two effect sizes became more efficient.

Values of Effect Sizes

Table 7 reports the minimum, first quartile, mean, third quartile, and maximum value of each effect size by location of non-invariance for the small magnitude condition. This summary allows for easy comparison of the value of the effect size due to non-invariant parameters of small magnitude to the value of the effect size due to sampling error of invariant parameters. On average, the values of the unsigned effect sizes for the invariant indicator are smaller than the values of the unsigned effect sizes for the three non-invariant indicators. Additionally, the average values of the unsigned effect sizes for each of the three non-invariant indicators are greater than the third quartile of the unsigned effect sizes for the invariant indicator, indicating that, on average, the effect size values for indicators that are invariant are much smaller than the effect size values for indicators that are not invariant.

To determine which simulation factors affected the value of the effect size, I analyzed a five-way between-subjects ANOVA separately for each effect size. All possible interactions between predictors were included in the analyses.

***d_{MACS}*.** The highest order effect that was impactful (i.e., partial $\eta^2 \geq .01$) on the value of d_{MACS} was the three-way interaction of location × magnitude × latent variable distribution for Group 2 (henceforth labeled as LVD2; partial $\eta^2 = .05$). In other words, the two-way interaction of magnitude × LVD2 differed by location of non-invariance. Figure 2 illustrates the four simple two-way interactions of magnitude × LVD2 for each level of location of non-invariance.

For the invariant indicator, the level of magnitude and level of LVD2 did not meaningfully affect the value of d_{MACS} (i.e., partial $\eta^2 < .01$ for interaction and main effects). This is shown in the top left panel in Figure 2. The lines representing the three levels of the LVD2 factor are horizontal and overlapping.

The simple two-way interaction of magnitude × LVD2 was meaningful for the indicator with a non-invariant loading (partial $\eta^2 = .05$). In the top right panel of Figure 2, the line for the lower mean, higher variance condition (i.e., mean = -0.5, variance = 1.3) diverged from the other two LVD2 conditions as magnitude of non-invariance increased. The simple main effect of LVD2 was impactful at all three levels of magnitude. For all three levels of magnitude, the lower mean, lower variance condition (i.e., mean = -0.5, variance = 0.7) and the same mean, same variance condition (i.e., mean = 0, variance = 1) did not differ from one another based on the value of d_{MACS} (Cohen's $d < 0.2$). The average value of d_{MACS} for the lower mean, higher variance condition was meaningfully larger than the other two conditions at all magnitudes, with the difference increasing as magnitude of non-invariance increased. Table 8 presents the Cohen's *d* values and average value of d_{MACS} for the studied comparisons that had a meaningful pairwise or simple pairwise comparison.

The simple two-way interaction of magnitude × LVD2 was not impactful on the value of d_{MACS} for the indicator with a non-invariant intercept. This is illustrated in the bottom left panel in Figure 2. The lines are roughly parallel, indicating there is no interaction. The main effect of magnitude was impactful (partial $\eta^2 = .85$) as was the main effect of LVD2 (partial $\eta^2 = .04$). As magnitude of non-invariance increased, the value of d_{MACS} increased. None of the pairwise comparisons between the three LVD2 groups were meaningful (i.e., Cohen's $d < 0.2$ for all three pairwise comparisons).

The simple two-way interaction of magnitude × LVD2 was meaningful for the indicator with a non-invariant loading and non-invariant intercept (partial $\eta^2 = .14$). As shown in the bottom right panel of Figure 2, the LVD2 conditions diverged from one another as magnitude of non-invariance increased. The simple main effect of LVD2 was meaningful at each level of magnitude of non-invariance. The same mean, same variance condition had a higher average value of d_{MACS} compared to the other two conditions at all levels of magnitude, with the difference increasing as magnitude of non-invariance increased. The lower mean, higher variance condition had a larger average value of d_{MACS} when the magnitude of non-invariance was medium and when it was large.

In addition to the three-way interaction, there was also an impactful two-way interaction of sample size × location (partial $\eta^2 = .14$). The simple main effect of sample size was not impactful for the three indicators that had non-invariance; however, the simple main effect of sample size was impactful for the invariant indicator (partial $\eta^2 = .16$). As sample size increased, the average value of d_{MACS} decreased due to the bias decreasing (i.e., the sample estimate approached the population value of 0).

***SDI_2*.** Table 9 presents the Cohen's *d* values and average value of SDI_2 for each condition that had a meaningful pairwise or simple pairwise comparison. The highest order effect that was impactful on the value of SDI_2 was the three-way interaction of location × magnitude × LVD2 (partial $\eta^2 = .08$). Figure 3 illustrates the four simple two-way interactions of magnitude × LVD2 for each level of location of non-invariance. For the invariant indicator, the magnitude of non-invariance and LVD2 factors did not meaningfully affect the value of SDI_2 (i.e., partial $\eta^2 < .01$ for interaction and main effects).

The simple two-way interaction of magnitude × LVD2 was meaningful for the indicator with a non-invariant loading (partial $\eta^2 = .19$). The simple main effect of LVD2 was meaningful at each level of magnitude of non-invariance. For the same mean, same variance condition, the average value of SDI_2 was roughly 0, regardless of the magnitude of non-invariance, due to cancellation of positive and negative differences. This value was higher than the average value of SDI_2 for the two lower mean LVD2 conditions, with the difference increasing as magnitude of non-invariance increased. The lower mean, lower variance condition had lower (i.e., more negative) values of SDI_2 than the lower mean, higher variance condition when the magnitude of non-invariance was medium and large.

The simple two-way interaction of magnitude × LVD2 was meaningful for the indicator with a non-invariant intercept (partial η^2 = .03). The simple main effect of LVD2 was meaningful for the small magnitude of non-invariance condition (partial η^2 = .06), the medium magnitude condition (partial η^2 = .17), and the large magnitude condition (partial η^2 = .30). In all conditions, the lower mean, lower variance condition always had a higher average value of SDI_2 than the other two conditions and the lower mean, higher variance condition always had a lower average value of SDI_2 than the other two conditions. The difference between conditions increased as magnitude of non-invariance increased.

The simple two-way interaction of magnitude × LVD2 was meaningful for the indicator with a non-invariant loading and non-invariant intercept (partial η^2 = .19). The simple main effect of LVD2 was meaningful for the small magnitude of non-invariance condition (partial η^2 = .12), the medium magnitude condition (partial η^2 = .42), and the large magnitude condition (partial η^2 = .60). The same mean, same variance condition always had a higher average value of SDI_2 than the other two conditions and the lower mean, higher variance condition always had a lower average value of SDI_2 than the other two conditions. The difference between conditions increased as magnitude of non-invariance increased.

***UDI*₂.** Table 10 presents the Cohen's *d* values and average value of UDI_2 for each condition that had a meaningful pairwise or simple pairwise comparison. The highest order effect that was impactful on the value of UDI_2 was the three-way interaction of location × magnitude × LVD2 (partial $\eta^2 = .07$). Figure 4 illustrates the four simple two-way interactions of magnitude × LVD2 for each level of location of non-invariance.

The simple two-way interaction of magnitude × LVD2 was not meaningful for the invariant indicator (partial $\eta^2 < .01$); however, the main effect of LVD2 was impactful (partial $\eta^2 = .011$). The lower mean, lower variance condition had, on average, higher values of UDI_2 than the lower mean, higher variance condition.

The simple two-way interaction of magnitude × LVD2 was meaningful for the indicator with a non-invariant loading (partial $\eta^2 = .04$). The simple main effect of LVD2 was not impactful for the small magnitude condition, but was impactful for the medium magnitude condition (partial $\eta^2 = .04$) and the large magnitude condition (partial $\eta^2 = .15$). In the medium magnitude condition, the lower mean, higher variance condition had a meaningfully higher mean than the lower mean, lower variance condition and the same mean, same variance condition. Additionally, the lower mean, lower variance condition had a meaningfully higher mean than the same mean, same variance condition. In the large magnitude condition, the lower mean, higher variance condition had a meaningfully higher mean than other two LVD2 conditions.

The simple two-way interaction of magnitude × LVD2 was meaningful for the indicator with a non-invariant intercept (partial $\eta^2 = .03$). The simple main effect of LVD2 was meaningful for the small magnitude of non-invariance condition (partial $\eta^2 = .06$), the medium magnitude condition (partial $\eta^2 = .17$), and the large magnitude condition (partial $\eta^2 = .30$). In all conditions, the lower mean, lower variance condition had the highest average value of UDI_2 followed by the same mean, same variance condition and then the lower mean, higher variance condition with the differences becoming more pronounced as magnitude of non-invariance increased.

The simple two-way interaction of magnitude × LVD2 was meaningful for the indicator with a non-invariant loading and intercept (partial $\eta^2 = .16$). The simple main effect of LVD2 was meaningful for the small magnitude of non-invariance condition (partial $\eta^2 = .10$), the medium magnitude condition (partial $\eta^2 = .36$), and the large magnitude condition (partial $\eta^2 = .53$). In all conditions, the same mean, same variance condition had higher values of UDI_2, on average, than the other two LVD2 conditions. The lower mean, lower variance condition had higher values of UDI_2 than the lower mean, higher variance condition when the magnitude was small, essentially the same average values when the magnitude was medium (Cohen's $d < .2$), and lower average values when the magnitude was large.

The main effect of sample size was impactful (partial $\eta^2 = .012$). However, none of the pairwise comparisons had an effect size greater than or equal to a small effect.

***WSDI*.** Table 11 presents the Cohen's *d* values and average value of *WSDI* for each condition that had a meaningful pairwise or simple pairwise comparison. The highest order effect that was impactful on the value of *WSDI* was the three-way interaction of location × magnitude × LVD2 (partial $\eta^2 = .03$). Figure 5 illustrates the four simple two-way interactions of magnitude × LVD2 for each level of location of non-invariance. For the invariant indicator, the magnitude of non-invariance and LVD2 factors did not meaningfully affect the value of *WSDI* (i.e., partial $\eta^2 < .01$ for interaction and main effects).

The simple two-way interaction of magnitude × LVD2 was meaningful for the indicator with a non-invariant loading (partial η^2 = .08). The simple main effect of LVD2 was impactful for the small magnitude condition (partial η^2 = .04), the medium magnitude condition (partial η^2 = .24), and the large magnitude condition (partial η^2 = .42). For the same mean, same variance condition, the average value of *WSDI* was roughly 0, regardless of the magnitude of non-invariance, due to complete cancellation. This value was higher than the average value of *WSDI* for the lower mean, lower variance condition and the lower mean, higher variance condition at all levels of magnitude. Additionally, the lower mean, lower variance condition had lower *WSDI* values, on average, than the lower mean, higher variance condition for medium and large magnitudes.

The simple two-way interaction of magnitude × LVD2 was not impactful on the value of *WSDI* for the indicator with intercept non-invariance, but the main effect of magnitude was impactful (partial η^2 = .83) as was the main effect of LVD2 (partial η^2 = .05). The same mean, same variance condition had higher values of *WSDI*, on average, than the lower mean, higher variance condition. The values of *WSDI* increased as magnitude of non-invariance increased.

The simple two-way interaction of magnitude × LVD2 was meaningful for the indicator with a non-invariant loading and non-invariant intercept (partial η^2 = .10). The simple main effect of LVD2 was impactful for the small magnitude condition (partial η^2 = .07), the medium magnitude condition (partial η^2 = .29), and the large magnitude condition (partial η^2 = .45). The same mean, same variance condition had higher values of *WSDI*, on average, than the other two LVD2 conditions.

The two-way interaction of location × balance was impactful on the value of *WSDI* (partial $\eta^2 = .03$). The simple main effect of balance was not meaningful for the invariant indicator nor for the indicator with a non-invariant loading. The simple main effect of balance was meaningful for the indicator with a non-invariant intercept and for the indicator with a non-invariant loading and intercept, such that the balanced condition had a higher average than the unbalanced condition.

WUDI. Table 12 presents the Cohen's *d* values and average value of *WUDI* for each condition that had a meaningful pairwise or simple pairwise comparison. The highest order effect that was impactful on the value of *WUDI* was the three-way interaction of location × magnitude × LVD2 (partial $\eta^2 = .02$). Figure 6 illustrates the four simple two-way interactions of magnitude × LVD2 for each level of location of non-invariance. For the invariant indicator, the magnitude of non-invariance and LVD2 factors did not meaningfully affect the value of *WUDI* (i.e., partial $\eta^2 < .01$ for interaction and main effects).

The simple two-way interaction of magnitude × LVD2 was meaningful for the indicator with a non-invariant loading (partial $\eta^2 = .01$). The lower mean, higher variance condition had meaningfully higher values of *WUDI*, on average, than the other two LVD2 conditions when the magnitude of non-invariance was medium or large.

The simple two-way interaction of magnitude × LVD2 was not impactful on the value of *WUDI* for the indicator with intercept non-invariance, but the main effect of magnitude was impactful (partial $\eta^2 = .84$) as was the main effect of LVD2 (partial $\eta^2 = .05$). For the former effect, as magnitude of non-invariance increased, the value of *WUDI* increased. For the latter effect, the same mean, same variance condition had meaningfully higher values than the lower mean, higher variance condition.

The simple two-way interaction of magnitude × LVD2 was meaningful for the indicator with a non-invariant loading (partial $\eta^2 = .06$). The same mean, same variance condition had meaningfully higher values than the other two LVD2 conditions, with the difference increasing as magnitude of non-invariance increased. Additionally, the lower mean, higher variance condition had higher values than the lower mean, lower variance condition when the magnitude of non-invariance was large.

In addition to the just-described three-way interaction, there was a second impactful three-way interaction of location × magnitude × balance (partial $\eta^2 = .01$) on the value of *WUDI*. Figure 7 illustrates the four simple two-way interactions of magnitude × balance for each level of location of non-invariance. For the invariant indicator, the magnitude of non-invariance and balance of sample sizes did not meaningfully affect the value of *WUDI* (i.e., partial $\eta^2 < .01$ for interaction and main effects).

The simple two-way interaction of magnitude × balance was not impactful on the value of *WUDI* for the indicator with a non-invariant loading, but the main effect of magnitude was impactful (partial $\eta^2 = .80$) as was the main effect of balance (partial $\eta^2 = .04$). As magnitude increased, the value of *WUDI* increased. The balanced condition had higher values of *WUDI*, on average, compared to the unbalanced condition; however, Cohen's *d* was less than 0.2.

The simple two-way interaction of magnitude × balance was meaningful for the indicator with a non-invariant intercept (partial $\eta^2 = .02$). The balanced conditions had higher values of *WUDI*, on average, compared to the unbalanced conditions for all levels of magnitude of non-invariance with the difference increasing as magnitude increased.

The simple two-way interaction of magnitude × balance was meaningful for the indicator with a non-invariant loading and intercept (partial $\eta^2 = .04$). The balanced conditions had higher values of *WUDI*, on average, compared to the unbalanced conditions for all levels of magnitude of non-invariance with the difference increasing as magnitude increased.

In addition to the two three-way interactions, the main effect of sample size was impactful (partial $\eta^2 = .013$). However, none of the pairwise comparisons of the three sample size conditions had an effect size greater than or equal to a small effect.

Relationship of Effect Sizes

The population values of the signed effect sizes were almost perfectly related ($r = .988$). The population values of the unsigned effect sizes were also almost perfectly related ($r_{dmacs,UDI2} = .996$, $r_{dmacs,WUDI} = .988$, $r_{UDI2,WUDI} = .988$). Finally, the population values of the signed and unsigned versions of the same effect size were highly related ($r_{SDI2,UDI2} = .694$, $r_{WSDI,WUDI} = .807$).

CHAPTER 4

DISCUSSION

Relying solely on *p*-values of a statistical test to make decisions is not good practice because the *p*-value is highly affected by sample size. Instead, effect size measures should be used in conjunction with *p*-values to understand the magnitude of the effect being studied in addition to examining statistical significance. However, in the area of measurement invariance, only a few effect size measures of non-invariance exist; however, they are not widely used and their properties have not been studied. Creating an unbiased and consistent effect size of measurement non-invariance is important to help researchers understand the impact of non-invariance on their models. This study examined the statistical properties of an existing effect size measure and of four proposed effect size measures under different simulated conditions. I discuss the results in terms of the findings, limitations of the study, future directions, and recommendations.

Overview and Implications of Results

Two of the estimation properties that Preacher and Kelley (2011) state high-quality effect size measures should have are that they are unbiased and consistent. All five effect size measures were consistent. The two signed effect size measures were unbiased across all simulated conditions. The three unsigned effect size measures were generally unbiased; however, they exhibited bias in some of the simulated conditions. These effect sizes were positively biased when a truly invariant indicator was not constrained to be invariant in the estimated model. Additionally, the unsigned effect size measures were biased in most of the conditions where the loading of the indicator was non-invariant and the magnitude of non-invariance was small. The bias decreased as the total sample size increased. Given these results, when the sample estimate of the unsigned effect size is small (e.g., UDI_2 = 0.07) and the value of the corresponding signed effect size is close to 0, the unsigned effect size is most likely overestimated.

Another property of high-quality effect size measures is that the value of the effect size should be independent of sample size (Preacher & Kelley, 2011). The ANOVA results illustrated that sample size did not affect the value of the two signed effect sizes as there were no meaningful main or simple main effects involving sample size as a predictor. Sample size did predict the value of d_{MACS} for the invariant indicator, such that as sample size increased, the value of d_{MACS} decreased. This is because bias decreased as sample size increased and the sample estimate converged to the population value of 0. Sample size was flagged as an important predictor of the values of UDI_2 and *WUDI*, but none of the pairwise comparisons were meaningful. In both cases, as sample size increased, the average value of the effect size decreased.

For all effect size values, there was a meaningful three-way interaction of location of non-invariance, the latent variable distribution for Group 2, and magnitude of non-invariance. This occurred because magnitude of non-invariance did not affect the sample estimates of the invariant indicator, but did affect the values of the non-invariant indicators, such that as magnitude of non-invariance increased, the absolute value of the effect size increased. One notable exception to this pattern occurred when the latent variable distribution for Group 2 (LVD2) was centered where the indicator response functions (IRFs) for the two groups crossed. In this case, there was complete cancellation and the expected value of the signed effect sizes was zero regardless of the magnitude of non-invariance.

The location of LVD2 is important when there is loading non-invariance. If the IRFs are parallel, then LVD2 does not affect the value of any of the effect sizes. If the loadings appreciably differ between groups, where the latent variable distributions of the two groups are in relation to where the IRFs cross affects the value of all five effect sizes. For instance, for the indicator with a non-invariant loading, the IRFs crossed when $\eta = 0$. However, when the indicator had a non-invariant loading and a non-invariant intercept and the magnitude of non-invariance was medium, the IRFs crossed when $\eta = -1.6$. Where the latent variable distributions of the two groups were centered in relation to that crossing affected how much cancellation occurred and how much weight was given to the bigger group differences in expected indicator scores.

The balance of group sample sizes did not affect the values of d_{MACS}, SDI_2, or UDI_2, but was impactful on the values of the two weighted effect sizes. When the difference was meaningful, the average value of the effect size in the balanced condition was always greater than the average value of the effect size in the unbalanced condition. However, this may not generalize to all possible conditions (e.g., Group 2 has a larger loading than Group 1, Group 2 has a larger sample size than Group 1) and further work is needed before establishing a pattern with regards to the effect of the balance of sample sizes on the value of the weighted effect sizes.

The five effect size measures were highly related to one another. This is to be expected for d_{MACS} and UDI_2 because the formulas for these two effect sizes are very similar. The extremely high relationship between *WUDI* and the two other unsigned effect sizes was surprising. While a strong relationship was expected, a greater difference between the effect sizes was anticipated. It is important to note that the effect sizes were highly related given the simulated conditions studied. Including other simulated conditions may affect the correlations. For instance, exploring more disparate sample sizes, more disparate latent variable distributions, and different IRFs may lead to different conclusions regarding how closely related *WUDI* is with the other two unsigned effect sizes. Additionally, the correlations between signed and unsigned versions of the same effect size measure should be interpreted cautiously because the full range of possibilities was not simulated. More specifically, non-invariance was simulated such that Group 1 always had higher loadings and/or intercepts and a higher or the same mean on the latent variable continuum. If the full range of possibilities was simulated as shown in Figure 1, I would expect the correlations to attenuate.

Limitations

As with any simulation study, this study was limited in scope. For instance, there were simulation factors and conditions within studied factors that were not examined that are important to investigate in a future study. The remainder of this section addresses five limitations of the current study.

First, the population communalities for the studied indicators were all equal to .64 for Group 1. However, the communalities of indicators affect the recovery of parameters (MacCallum et al., 1999). As communalities decrease, the sampling variability of parameter estimates increases, thus affecting the effect size calculations. Anything that affects the recovery of parameters is expected to affect the effect size estimates.

Second, a condition that was not studied that is important to vary is level and type of misspecification. In this study, the tested model was essentially the same as the generating model. (It was not exactly because a few invariant parameters were not constrained to invariance.) Fitting a variety of incorrect models to the data should be conducted to see how model misspecification affects the value and bias of the effect sizes. For instance, if a non-invariant, non-studied indicator was modeled to be invariant, then I would expect the values of the effect sizes of the studied indicators to be impacted. As stated previously, if the reference variable (or variables) that links the metrics between the groups is not invariant, then the accuracy of the invariance testing is biased and the other parameters in the models are not accurately recovered (Bollen, 1989; Cheung & Rensvold, 1999; Johnson et al., 2009; Yoon & Millsap, 2007). In addition to analyzing models with specific misspecifications, it would be important to model passive misfit as well (Cudeck & Browne, 1992) and see how that impacts the value, consistency, and bias of the effect size measures.

Third, the indicators for Group 1 were simulated to have the same expected observed variances. While the denominator of the effect sizes were designed to make the effect size values comparable across indicators with different variances, it would be important to establish if and how different scales affect the value of the effect sizes. If scaling does affect the value of the effect sizes, then this would impact what effect size values are considered small or large.

Fourth, there was not an analytical calculation of the synthetic parameters (measurement parameters for the synthetic population). The proposed method for calculating the population value of the synthetic parameters was an empirical one and thus just an estimate of the true value. This affects the population value of the weighted effect sizes as well as the sample estimate and thus affects measures of bias. The population value of the synthetic parameters for the invariant indicator can be calculated analytically (they are equal to the multiple-group parameters); however, they were calculated empirically in this study to match the procedure for calculating the synthetic parameters of the non-invariant indicators.

Fifth and finally, high quality effect size measures should be efficient (Preacher & Kelley, 2011). Efficiency was not evaluated in this study because different estimators of the same effect size were not compared. Thus, it cannot be concluded that these effect size measures were efficient.

Future Directions

Beyond addressing the limitations of the current study, there are many avenues for future research regarding these effect size measures. In this section, I address six future directions.

First, in addition to the properties stated by Preacher and Kelley (2011), another important quality of a good effect size measure is the use of benchmarks and guidelines, which aid in interpretability (Kirk, 1996). For instance, Cohen (1988) developed benchmarks for Cohen's *d* where a value of 0.2 indicated a small effect, 0.5 a medium effect, and 0.8 a large effect. d_{MACS} is in the same metric of Cohen's *d*. The advantage of putting an effect size on the same scale as Cohen's *d* is that researchers have a good understanding of the metric, can easily make comparisons across studies, and can use the same benchmarks. A disadvantage of applying Cohen's *d* benchmarks to effect sizes of non-invariance is that the benchmarks were developed by looking at different effect sizes across many studies on psychological effects and may not be generalizable to non-invariance studies. However, researchers are using Cohen's benchmarks to interpret magnitude of non-invariance for d_{MACS} (Clark, Listro, Lo, Durbin, Donnellan, & Neppl, 2016). A future study should calculate the five effect size measures for applied measurement invariance studies in the extant literature to get a sense of the range of typical values. It will be difficult to calculate the weighted effect sizes based on published research because researchers rarely publish the results from both a multiple-group model and a single-population model.

Second, as mentioned before, Preacher and Kelley (2011) stated that good effect size measures have calculable confidence intervals. Future research should investigate the proper way to calculate confidence intervals for these effect size measures either analytically or empirically. Chalmers and colleagues (2016) described an empirical way to calculate confidence intervals for an effect size of non-invariance in the IRT framework. First, using the sample-obtained point estimate of the parameters and the estimated variation of those parameters, impute plausible parameter values for the parameters used in the effect size calculations (e.g., loadings and intercepts). Second, compute the effect size using the imputed parameter estimates. Third, repeat steps one and two *M* times (e.g., 1,000). After this process is over, there will be *M* effect size estimates that can be used to calculate empirical confidence intervals.

Third, developing corrections for the bias seen in the unsigned effect size measures would help the sample estimates be more trustworthy. However, because the unsigned effect size measures were biased only in specific conditions, we would need to be cautious about introducing bias to the unbiased conditions if we try to correct for the bias in the biased conditions. The conditions that were biased were the ones that had small or null values for the population effect size. Given this, it would be helpful to determine a cutoff population value where the sample estimates are not biased. If a sample estimate falls below that cutoff value, then the results are less trustworthy. (Though it is important to note that a sample estimate might fall below that cutoff value due to sampling error even though the population value is larger than the cutoff value.)

Rather than adjusting the formula to correct for bias, another possibility is to adjust the model. The sample estimates of the effect sizes were calculated by allowing the loading and intercept to differ by group. However, invariance testing may lead researchers to conclude that one or both of those parameters are invariant (e.g., the loading is invariant, but the intercept is not). This would affect the sample estimate of the effect sizes and thus would affect conclusions of bias. For instance, if a truly invariant indicator is modeled to be invariant, then the sample estimate of the effect size will match the population value. In this study, however, the parameters of the truly invariant indicator were not constrained to invariance and this caused the unsigned effect sizes for that indicator to exhibit positive bias.

Fourth, understanding how the proposed effect sizes should be calculated and how to interpret them in complex models is an important next step. This study looked at a simple case of one factor with eight indicators. In reality, researchers work with multiple factors, which can lead to complexities such as cross-loadings. As noted by Nye and Drasgow (2011), more complex effect size formulas are needed for indicators that load on multiple latent variables.

Fifth, there is potential to create more effect size measures of non-invariance in the factor analytic framework. More effect sizes of non-invariance are needed because effect sizes are purpose-specific and, thus, some are more appropriate for certain inferences (e.g., latent mean estimation, inferences at the person level, validity of a cut score). The five effect size measures analyzed in this study were measured at the indicator level. A scale-level effect size measure would be useful to understand the impact of non-invariance on scale scores or factor scores. Additionally, an effect size that accompanies the tests of levels of invariance (e.g., metric, scalar) can be used jointly with the *p*-values to detect non-invariance and measure the magnitude of misfit.

These effect size measures are designed for continuous outcomes; however, they can be easily translated for categorical/ordinal outcomes. In the IRT framework, the five effect sizes can be calculated as is, except the expected indicator scores would be calculated using the item response functions based on the IRT parameters. Even though there are many effect size measures in the IRT framework, none of the effect size measures in existence compare expected item scores using multiple-group parameters to expected item scores using the synthetic group parameters. Thus, the *WSDI* and *WUDI* would be useful to translate into the IRT framework.

Finally, once more studies have been conducted on these effect sizes to better understand their properties and behavior, an important future direction is to convince researchers to use these effect sizes. The best way to do so is to make the effect size measures and their confidence intervals easily available in popular statistical software programs and to create interpretable benchmarks.

Recommendations

I recommend that applied researchers calculate and report at least one of these effect sizes based on a partial invariant model that has been finalized through invariance testing. The intended use of the four proposed effect sizes is that they should be calculated once invariance testing is completed and the final model is settled on. They are not designed to be used to detect non-invariance, but to measure the impact of non-invariance that has been detected with statistical tests.

The specific effect size a researcher should use is determined by the substantive question at hand. Because the unsigned effect size measures were biased under certain conditions, cautions should be taken when using those effect sizes and thus should not be the only effect size calculated. More caution should be taken when the value of the unsigned effect size is relatively small (e.g., UDI_2 is 0.07) because that is when the effect size is most likely to be over-estimated. Given that the like-signed effect sizes were highly related to one another, it is best to choose the effect size that is most interpretable.

More simulation and empirical work is needed before making valid recommendations regarding interpretation of the magnitude of the effect sizes (e.g., determining the cutoff value for a small effect of non-invariance). Until then, the benchmarks of Cohen's *d* can be used as a rough proxy for d_{MACS}, SDI_2, and UDI_2, but should not be taken as hard cut-offs.

REFERENCES

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411-423. doi:10.1037/0033-2909.103.3.411

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100,* 603–617. doi:10.1348/000712608X377117

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42,* 815–824. doi:10.1016/j.paid.2006.09.018

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B*, *57,* 289-300. doi:10.2307/2346101

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, *31,* 419-456. doi:10.1146/annurev.ps.31.020180.002223

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246. doi:10.1037/0033-2909.107.2.238

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606. doi:10.1037/0033-2909.88.3.588

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons. doi:10.1002/9781118619179

Bollen, K. A., & Jöreskog, K. G. (1985). Uniqueness does not imply identification. *Sociological Methods and Research, 14,* 155-163. doi:10.1177/0049124185014002003

Brace, J. C., & Savalei, V. (2017). Type I error rates and power of several versions of scaled chi-square difference tests in investigations of measurement invariance. *Psychological Methods*, *22*(3), 467-485. doi:10.1037/met0000097

Brannick, M. T. (1995). Critical comments on applying covariance structure modelling. *Journal of Organizational Behavior*, *16*, 201-213. doi:10.1002/job.4030160303

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equations models* (pp. 136–162). Newbury Park, CA: Sage.

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

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*(1), 114-140. doi:10.1177/0013164415584576

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 464-504. doi:10.1080/10705510701301834

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 471-492. doi:10.1207/s15328007sem1203_7

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organization Research Methods*, *15*(2), 167-198. doi:10.1177/1094428111421987

Cheung, G. W., & Rensvold, R. B. (1998). Cross-cultural comparisons using non-invariant measurement items. *Applied Behavioral Science Review*, *6*, 93–110. doi:10.1016/S1068-8595(99)80006-3

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1–27. doi:10.1177/014920639902500101

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233-255. doi:10.1207/S15328007SEM0902_5

Clark, D. A., Listro, C. J., Lo, S. L., Durbin, C. E., Donnellan, M. B., & Neppl, T. K. (2016). Measurement invariance and child temperament: An evaluation of sex and informant differences on the child behavior questionnaire. *Psychological Assessment*, *28*(12), 1646-1662. doi:10.1037/pas0000299

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330-351. doi:10.1037/1082-989X.6.4.330

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18,* 147-167. doi:10.1207/s15327906mbr1802_2

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika*, *57*(3), 357-369. doi:10.1007/BF02295424

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*, 121–149. doi:10.1177/0748175610373459

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917-928. doi:10.1093/jpepsy/jsp004

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *The Journal of Educational Research*, *94,* 275-282. doi:10.1080/00220670109598763

Fan, X., & Sivo, S. A. (2009). Using ∆goodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(1), 54-69. doi:10.1080/10705510802561311

Flowers, C. P., Oshima, C., & Raju, N. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement*, *23*(4), 309-326. doi:10.1177/01466219922031437

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 96-113. doi:10.1080/10705510701758349

Gonzales, N. A., Knight, G. P., Gunn, H. J., Tein, J.-Y., Tanaka, R., & White, R. M. B. (2018). Intergenerational gaps in Mexican American values trajectories: Associations with parent-adolescent conflict and adolescent psychopathology. *Developmental Psychopathology*, *30*(5), 1611-1627. doi:10.1017/S0954579418001256

Green, S. B., Thompson, M. S., & Babyak, M. A. (1998). A Monte Carlo investigation of methods for controlling Type I errors with specification searches in structural equation modeling. *Multivariate Behavioral Research, 33,* 365-383. doi:10.1207/s15327906mbr3303_3

Horn, J. L. (1991). Comments on “Issues in Factorial Invariance.” In L. M. Collins, & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 114-125). Washington, DC: American Psychological Association. doi:10.1037/10099-000

Horn, J. L., McArdle, J., & Mason, R. (1983). When is invariance not invariant: A practical scientist’s look at the ethereal concept of factor invariance. *Southern Psychologist*, *1*(4), 179-188.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 642-657. doi:10.1080/10705510903206014

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409-426. doi:10.1007/BF02291366

Jöreskog, K. G., & Sörbom, D. (1983). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood*. Chicago, IL: International Educational Services.

Jung, E., & Yoon, M. (2017). Two-step approach to partial factorial invariance: Selecting a reference variable and identifying the source of noninvariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 65-79. doi:10.1080/10705511.2016.1251845

Kang, Y., McNeish, D. M., & Hancock, G. R. (2016). The role of measurement quality on practical guidelines for assessing measurement and structural invariance. *Educational and Psychological Measurement*, *76*(4), 533-561. doi:10.1177/0013164415603764

Kaplan, D. (1989). The problem of error rate inflation in covariance structure models. *Educational and Psychological Measurement*, *49*, 333-337. doi:10.1177/0013164489492005

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*(2), 101-118. doi:10.1080/10705519509539999

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*(2), 127-152. doi:10.1037/a0028086

Kelloway, E. K. (1995). Structural equation modelling in perspective. *Journal of Organizational Behavior, 16,* 215-224. doi:10.1002/job.4030160304

Kim, E. S. (2011). Testing measurement invariance using MIMIC: Likelihood ratio test and modification indices with a critical value adjustment. Unpublished doctoral dissertation, Texas A&M University.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759. doi:10.1177/0013164496056005002

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review*, *23*(4), 418-444. doi:10.1177/0193841X9902300404

Lai, M. H. C., & Kwok, O. (2016). Estimating standardized effect sizes for two- and three-level partially nested data. *Multivariate Behavioral Research, 51*(6), 740–756. doi:10.1080/00273171.2016.1231606

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32,* 53–76. doi:10.1207/s15327906mbr3201_3

Lommen, M. J. J., van de Schoot, R., & Engelhard, I. M. (2014). The experience of traumatic events disrupts the stability of a posttraumatic stress scale. *Frontiers in Psychology, 5*, 1-7. doi:10.3389/fpsyg.2014.01304

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*, 107-120. doi:10. 1037/0033-2909.100.1.107

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490-504. doi:10.1037/0033-2909.111.3.490

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84-99. doi:10.1037/1082-989X.4.1.84

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

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 611-635. doi:10.1080/10705510701575461

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568-592. doi:10.1037/0021-9010.93.3.568

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(1), 60-72. doi:10.1207/S15328007SEM1101_5

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525-543. doi:10.1007/BF02294825

Millsap, R. E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Multivariate applications book series. Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 153-171). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

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

Millsap, R. E., & Kim, H. (2018). Factorial invariance across multiple populations in discrete and continuous data. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 847-884). Hoboken, NJ: Wiley-Blackwell. doi:10.1002/9781118489772.ch26

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

Millsap, R. E., & Olivera-Aguilar (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380-392). New York, NY: Guildford Press.
Muthén, L. K., & Muthén, B. O. (1998-2014). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

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*(5), 966-980. doi:10.1037/a0022955

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*(2), 93-115. doi:10.1037/a0022658

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*, 353-368. doi:10.1177/014662169501900405

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychology Bulletin*, *114*(3), 552-566. doi:10.1037/0033-2909.114.3.552

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management: Vol. 1. Equivalence in measurement* (pp. 21-50). Greenwich, CT: Information Age.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276-1284. doi:10.1037/0003-066X.44.10.1276

Sandler, I. N., Ayers, T. S., Wolchik, S. A., Tein, J.-Y., Kwok, O.-M., Haine, R. A., Twohey, J. L., Suter, J., Lin, K., Padgett-Jones, S., Weyer, J. L., Cole, E., Kriege, G., & Griffin, W. A. (2003). The Family Bereavement Program: Efficacy evaluation of a theory-based prevention program for parentally-bereaved children and adolescents. *Journal of Consulting and Clinical Psychology*, *71*(3), 587–600. doi:10.1037/0022-006X.71.3.587

Sandler, I., Wolchik, S., Mazza, G., Gunn, H., Tein, J.-Y., Berkel, C., Jones, S. & Porter, M. (2019, online publication). Randomized effectiveness trial of the New Beginnings Program for divorced families with children and adolescents. *Journal of Clinical Child and Adolescent Psychology*. doi:10.1080/15374416.2018.1540008

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*(4), 347-363. doi:10.1177/0734282911406661

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*, 243–248. doi:10.1007/s11336-009-9135-y

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of a differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, *89*, 497-508. doi:10.1037/0021-9010.89.3.497

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292-1306. doi:10.1037/0021-9010.91.6.1292

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78-90. doi:10.1086/209528

Steering Committee of the Physicians’ Health Study Research Group (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine*, *318*, 262-264. doi:10.1056/NEJM198801283180431
Steiger, J. H. (1989). *EzPATH: Causal modeling.* Evanston, IL: SYSTAT.

Steiger, J. H., & Lind, J. M. (1980). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, *50*, 253-264. doi:10.1007/BF02294104

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11,* 402-415. doi:10.1037/1082-989X.11.4.402

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: The University of Chicago Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4-70. doi:10.1177/109442810031002

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56-75). Thousand Oaks, CA, US: Sage Publications, Inc.

Whittaker, T. A. (2013). The impact of noninvariant intercepts in latent means models. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 108-130. doi:10.1080/10705511.2013.742397

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association. doi:10.1037/10222-009

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604. doi:10.1037/0003-066X.54.8.594

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation*, *12*(3), 1-26.

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. doi:10.1080/10705510701301677

APPENDIX A

TABLES AND FIGURES

Table 1

Possible outcomes for signed and unsigned effect size combinations.

	Small signed effect size	Large signed effect size
Small unsigned effect size	Indicator is invariant or close to invariant.	This outcome is not possible to observe because the signed effect size ≤ unsigned effect size property will always hold.
Large unsigned effect size	The expected indicator score lines cross to create a balanced overall scoring on the indicator. However, there is non-ignorable non-invariance at particular levels of the latent variable.	Non-ignorable non-invariance. Indicators with large intercept differences but no loading differences will lead to this scenario.

Note. This table is a modification of Table 1 from Chalmers, Counsell, and Flora (2016).

Table 2

Average raw bias of the five effect size measures for each design cell.

Loc[1]	N[2]	Bal[3]	Mag[4]	LVD2[5]	d_{MACS}	UDI_2	*WUDI*	SDI_2	*WSDI*
I	300	1:1	Small	(0,1)	0.099	0.085	0.035	0.001	0.001
I	300	1:1	Small	(-0.5,1.3)	0.103	0.085	0.039	0.002	0.001
I	300	1:1	Small	(-0.5,0.7)	0.104	0.095	0.036	-0.006	-0.003
I	300	1:1	Medium	(0,1)	0.097	0.083	0.037	0.000	0.000
I	300	1:1	Medium	(-0.5,1.3)	0.100	0.082	0.036	0.005	0.001
I	300	1:1	Medium	(-0.5,0.7)	0.104	0.095	0.037	-0.004	-0.004
I	300	1:1	Large	(0,1)	0.100	0.086	0.039	0.000	0.000
I	300	1:1	Large	(-0.5,1.3)	0.102	0.085	0.037	0.003	0.001
I	300	1:1	Large	(-0.5,0.7)	0.105	0.096	0.036	0.001	0.000
I	300	2:1	Small	(0,1)	0.108	0.094	0.039	-0.001	0.000
I	300	2:1	Small	(-0.5,1.3)	0.109	0.090	0.035	0.000	-0.001
I	300	2:1	Small	(-0.5,0.7)	0.107	0.100	0.038	0.002	0.000
I	300	2:1	Medium	(0,1)	0.106	0.092	0.032	0.002	-0.001
I	300	2:1	Medium	(-0.5,1.3)	0.108	0.088	0.035	-0.001	-0.001
I	300	2:1	Medium	(-0.5,0.7)	0.109	0.102	0.037	0.000	0.001
I	300	2:1	Large	(0,1)	0.108	0.093	0.036	0.001	0.002
I	300	2:1	Large	(-0.5,1.3)	0.107	0.087	0.035	0.003	0.001
I	300	2:1	Large	(-0.5,0.7)	0.113	0.106	0.040	-0.001	0.001
I	500	1:1	Small	(0,1)	0.078	0.068	0.027	0.001	0.000
I	500	1:1	Small	(-0.5,1.3)	0.077	0.064	0.029	0.001	0.000
I	500	1:1	Small	(-0.5,0.7)	0.082	0.076	0.027	-0.002	-0.001
I	500	1:1	Medium	(0,1)	0.080	0.069	0.031	-0.003	-0.002
I	500	1:1	Medium	(-0.5,1.3)	0.079	0.065	0.027	0.000	-0.001
I	500	1:1	Medium	(-0.5,0.7)	0.082	0.075	0.027	0.000	-0.003
I	500	1:1	Large	(0,1)	0.078	0.067	0.030	-0.002	-0.001
I	500	1:1	Large	(-0.5,1.3)	0.078	0.064	0.027	-0.001	-0.001
I	500	1:1	Large	(-0.5,0.7)	0.082	0.075	0.026	0.003	0.001
I	500	2:1	Small	(0,1)	0.082	0.070	0.029	-0.002	-0.001
I	500	2:1	Small	(-0.5,1.3)	0.085	0.069	0.027	-0.001	-0.001
I	500	2:1	Small	(-0.5,0.7)	0.085	0.078	0.029	0.002	0.000
I	500	2:1	Medium	(0,1)	0.082	0.070	0.023	0.003	-0.001
I	500	2:1	Medium	(-0.5,1.3)	0.083	0.067	0.027	-0.002	-0.001
I	500	2:1	Medium	(-0.5,0.7)	0.084	0.078	0.028	0.000	0.001
I	500	2:1	Large	(0,1)	0.080	0.069	0.026	-0.002	0.000
I	500	2:1	Large	(-0.5,1.3)	0.081	0.066	0.026	0.001	0.001
I	500	2:1	Large	(-0.5,0.7)	0.086	0.080	0.030	-0.002	0.000
I	1,000	1:1	Small	(0,1)	0.055	0.047	0.016	-0.002	-0.001
I	1,000	1:1	Small	(-0.5,1.3)	0.056	0.046	0.020	0.001	0.000
I	1,000	1:1	Small	(-0.5,0.7)	0.057	0.052	0.016	-0.004	-0.002
I	1,000	1:1	Medium	(0,1)	0.055	0.047	0.019	0.004	0.002
I	1,000	1:1	Medium	(-0.5,1.3)	0.055	0.045	0.018	0.001	0.000
I	1,000	1:1	Medium	(-0.5,0.7)	0.058	0.053	0.017	0.000	-0.003
I	1,000	1:1	Large	(0,1)	0.055	0.047	0.020	-0.001	-0.001
I	1,000	1:1	Large	(-0.5,1.3)	0.056	0.046	0.018	-0.003	-0.002
I	1,000	1:1	Large	(-0.5,0.7)	0.058	0.053	0.015	0.000	-0.001

I	1,000	2:1	Small	(0,1)	0.059	0.050	0.020	-0.001	0.000
I	1,000	2:1	Small	(-0.5,1.3)	0.058	0.047	0.017	-0.003	-0.002
I	1,000	2:1	Small	(-0.5,0.7)	0.060	0.056	0.020	0.001	0.000
I	1,000	2:1	Medium	(0,1)	0.058	0.049	0.014	0.002	-0.002
I	1,000	2:1	Medium	(-0.5,1.3)	0.060	0.048	0.019	-0.002	-0.001
I	1,000	2:1	Medium	(-0.5,0.7)	0.060	0.056	0.019	0.000	0.001
I	1,000	2:1	Large	(0,1)	0.059	0.051	0.019	-0.001	0.001
I	1,000	2:1	Large	(-0.5,1.3)	0.058	0.047	0.018	0.000	0.000
I	1,000	2:1	Large	(-0.5,0.7)	0.060	0.055	0.020	0.001	0.001
NIL	300	1:1	Small	(0,1)	0.034	0.033	0.014	0.001	0.001
NIL	300	1:1	Small	(-0.5,1.3)	0.028	0.029	0.014	-0.004	-0.003
NIL	300	1:1	Small	(-0.5,0.7)	0.037	0.039	0.018	0.004	0.002
NIL	300	1:1	Medium	(0,1)	0.017	0.019	0.007	0.006	0.003
NIL	300	1:1	Medium	(-0.5,1.3)	0.011	0.014	0.005	-0.006	-0.003
NIL	300	1:1	Medium	(-0.5,0.7)	0.017	0.023	0.008	-0.005	-0.001
NIL	300	1:1	Large	(0,1)	0.008	0.011	0.003	-0.003	-0.003
NIL	300	1:1	Large	(-0.5,1.3)	0.005	0.011	0.003	-0.006	-0.002
NIL	300	1:1	Large	(-0.5,0.7)	0.008	0.014	0.003	-0.002	-0.001
NIL	300	2:1	Small	(0,1)	0.040	0.041	0.017	-0.005	-0.003
NIL	300	2:1	Small	(-0.5,1.3)	0.038	0.038	0.016	-0.002	0.000
NIL	300	2:1	Small	(-0.5,0.7)	0.038	0.043	0.017	0.001	0.001
NIL	300	2:1	Medium	(0,1)	0.014	0.017	0.006	-0.001	-0.001
NIL	300	2:1	Medium	(-0.5,1.3)	0.012	0.016	0.005	-0.002	0.001
NIL	300	2:1	Medium	(-0.5,0.7)	0.014	0.023	0.007	-0.002	0.000
NIL	300	2:1	Large	(0,1)	0.008	0.014	0.003	-0.004	-0.001
NIL	300	2:1	Large	(-0.5,1.3)	0.010	0.017	0.004	-0.005	0.000
NIL	300	2:1	Large	(-0.5,0.7)	0.004	0.011	0.002	0.003	0.002
NIL	500	1:1	Small	(0,1)	0.021	0.021	0.009	0.000	0.000
NIL	500	1:1	Small	(-0.5,1.3)	0.015	0.016	0.008	0.000	-0.001
NIL	500	1:1	Small	(-0.5,0.7)	0.024	0.027	0.012	0.000	0.001
NIL	500	1:1	Medium	(0,1)	0.011	0.012	0.005	-0.002	-0.001
NIL	500	1:1	Medium	(-0.5,1.3)	0.007	0.009	0.004	-0.003	-0.002
NIL	500	1:1	Medium	(-0.5,0.7)	0.013	0.017	0.007	-0.004	-0.001
NIL	500	1:1	Large	(0,1)	0.008	0.009	0.002	-0.004	-0.002
NIL	500	1:1	Large	(-0.5,1.3)	0.005	0.008	0.002	-0.002	-0.001
NIL	500	1:1	Large	(-0.5,0.7)	0.002	0.005	0.001	0.002	0.000
NIL	500	2:1	Small	(0,1)	0.025	0.026	0.011	-0.002	-0.001
NIL	500	2:1	Small	(-0.5,1.3)	0.017	0.018	0.008	0.003	0.001
NIL	500	2:1	Small	(-0.5,0.7)	0.021	0.025	0.010	0.003	0.002
NIL	500	2:1	Medium	(0,1)	0.008	0.009	0.003	-0.002	-0.002
NIL	500	2:1	Medium	(-0.5,1.3)	0.004	0.008	0.002	0.001	0.002
NIL	500	2:1	Medium	(-0.5,0.7)	0.007	0.011	0.004	0.002	0.000
NIL	500	2:1	Large	(0,1)	0.003	0.006	0.001	-0.007	-0.003
NIL	500	2:1	Large	(-0.5,1.3)	0.000	0.005	0.001	-0.001	0.001
NIL	500	2:1	Large	(-0.5,0.7)	0.000	0.004	0.001	0.003	0.002
NIL	1,000	1:1	Small	(0,1)	0.013	0.013	0.005	0.001	0.001
NIL	1,000	1:1	Small	(-0.5,1.3)	0.010	0.010	0.005	-0.001	-0.001
NIL	1,000	1:1	Small	(-0.5,0.7)	0.010	0.012	0.005	0.001	0.001
NIL	1,000	1:1	Medium	(0,1)	0.003	0.004	0.002	-0.002	0.000
NIL	1,000	1:1	Medium	(-0.5,1.3)	0.001	0.002	0.001	0.002	0.001

NIL	1,000	1:1	Medium	(-0.5,0.7)	0.005	0.007	0.003	-0.002	-0.001
NIL	1,000	1:1	Large	(0,1)	0.002	0.003	0.001	0.000	-0.001
NIL	1,000	1:1	Large	(-0.5,1.3)	0.003	0.005	0.001	0.000	0.000
NIL	1,000	1:1	Large	(-0.5,0.7)	0.003	0.005	0.001	-0.003	-0.002
NIL	1,000	2:1	Small	(0,1)	0.009	0.010	0.005	-0.001	-0.001
NIL	1,000	2:1	Small	(-0.5,1.3)	0.009	0.010	0.004	0.000	0.000
NIL	1,000	2:1	Small	(-0.5,0.7)	0.011	0.014	0.005	0.000	0.001
NIL	1,000	2:1	Medium	(0,1)	0.006	0.007	0.002	0.000	-0.001
NIL	1,000	2:1	Medium	(-0.5,1.3)	0.003	0.005	0.001	-0.002	0.000
NIL	1,000	2:1	Medium	(-0.5,0.7)	0.004	0.007	0.002	-0.003	-0.001
NIL	1,000	2:1	Large	(0,1)	0.002	0.003	0.000	-0.003	-0.001
NIL	1,000	2:1	Large	(-0.5,1.3)	0.003	0.005	0.001	-0.001	0.001
NIL	1,000	2:1	Large	(-0.5,0.7)	0.003	0.006	0.001	0.000	0.001
NII	300	1:1	Small	(0,1)	0.018	0.006	0.003	0.002	0.000
NII	300	1:1	Small	(-0.5,1.3)	0.019	0.006	0.003	0.000	-0.001
NII	300	1:1	Small	(-0.5,0.7)	0.019	0.009	0.006	0.003	0.002
NII	300	1:1	Medium	(0,1)	0.009	0.002	0.000	0.001	0.000
NII	300	1:1	Medium	(-0.5,1.3)	0.009	0.000	0.000	0.000	-0.001
NII	300	1:1	Medium	(-0.5,0.7)	0.009	0.001	0.001	0.001	-0.001
NII	300	1:1	Large	(0,1)	0.012	0.007	0.002	0.007	0.002
NII	300	1:1	Large	(-0.5,1.3)	0.010	0.005	0.001	0.005	0.000
NII	300	1:1	Large	(-0.5,0.7)	0.005	-0.001	0.000	-0.001	-0.002
NII	300	2:1	Small	(0,1)	0.015	0.003	0.000	-0.002	-0.002
NII	300	2:1	Small	(-0.5,1.3)	0.022	0.008	0.004	0.001	0.000
NII	300	2:1	Small	(-0.5,0.7)	0.024	0.010	0.005	0.003	0.000
NII	300	2:1	Medium	(0,1)	0.007	-0.001	-0.003	-0.001	-0.004
NII	300	2:1	Medium	(-0.5,1.3)	0.015	0.005	0.004	0.004	0.003
NII	300	2:1	Medium	(-0.5,0.7)	0.009	0.001	0.002	0.001	0.000
NII	300	2:1	Large	(0,1)	0.008	0.006	0.001	0.006	0.001
NII	300	2:1	Large	(-0.5,1.3)	0.012	0.006	0.003	0.006	0.002
NII	300	2:1	Large	(-0.5,0.7)	0.010	0.007	0.002	0.007	0.001
NII	500	1:1	Small	(0,1)	0.014	0.006	0.003	0.004	0.002
NII	500	1:1	Small	(-0.5,1.3)	0.012	0.002	0.002	0.000	0.000
NII	500	1:1	Small	(-0.5,0.7)	0.011	0.002	0.003	0.000	0.000
NII	500	1:1	Medium	(0,1)	0.005	0.000	0.000	0.000	0.000
NII	500	1:1	Medium	(-0.5,1.3)	0.006	0.001	0.000	0.001	-0.001
NII	500	1:1	Medium	(-0.5,0.7)	0.006	0.002	0.001	0.002	0.000
NII	500	1:1	Large	(0,1)	0.002	-0.001	-0.001	-0.001	-0.001
NII	500	1:1	Large	(-0.5,1.3)	0.007	0.003	0.000	0.003	0.000
NII	500	1:1	Large	(-0.5,0.7)	0.004	0.002	0.000	0.002	-0.001
NII	500	2:1	Small	(0,1)	0.008	0.000	-0.001	-0.002	-0.002
NII	500	2:1	Small	(-0.5,1.3)	0.014	0.004	0.002	0.001	0.000
NII	500	2:1	Small	(-0.5,0.7)	0.013	0.005	0.003	0.003	0.001
NII	500	2:1	Medium	(0,1)	0.008	0.003	-0.002	0.002	-0.002
NII	500	2:1	Medium	(-0.5,1.3)	0.011	0.006	0.004	0.006	0.003
NII	500	2:1	Medium	(-0.5,0.7)	0.005	-0.001	0.001	-0.002	0.000
NII	500	2:1	Large	(0,1)	0.002	-0.001	0.000	-0.001	0.000
NII	500	2:1	Large	(-0.5,1.3)	0.004	0.002	0.001	0.002	0.001
NII	500	2:1	Large	(-0.5,0.7)	0.008	0.007	0.002	0.007	0.001
NII	1,000	1:1	Small	(0,1)	0.004	0.000	0.000	0.000	0.000

NII	1,000	1:1	Small	(-0.5,1.3)	0.005	-0.001	0.000	-0.001	0.000
NII	1,000	1:1	Small	(-0.5,0.7)	0.005	0.001	0.001	0.001	0.000
NII	1,000	1:1	Medium	(0,1)	0.003	0.000	0.000	0.000	0.000
NII	1,000	1:1	Medium	(-0.5,1.3)	0.002	0.000	-0.001	0.000	-0.001
NII	1,000	1:1	Medium	(-0.5,0.7)	0.004	0.003	0.000	0.003	0.000
NII	1,000	1:1	Large	(0,1)	0.003	0.002	0.000	0.002	0.000
NII	1,000	1:1	Large	(-0.5,1.3)	0.003	0.001	-0.001	0.001	-0.001
NII	1,000	1:1	Large	(-0.5,0.7)	0.003	0.001	0.001	0.001	0.000
NII	1,000	2:1	Small	(0,1)	0.006	0.001	-0.001	0.001	-0.001
NII	1,000	2:1	Small	(-0.5,1.3)	0.004	-0.001	0.000	-0.002	-0.001
NII	1,000	2:1	Small	(-0.5,0.7)	0.008	0.003	0.001	0.003	0.001
NII	1,000	2:1	Medium	(0,1)	0.003	0.001	-0.003	0.001	-0.003
NII	1,000	2:1	Medium	(-0.5,1.3)	0.003	0.000	0.001	0.000	0.001
NII	1,000	2:1	Medium	(-0.5,0.7)	0.001	-0.001	0.000	-0.001	0.000
NII	1,000	2:1	Large	(0,1)	0.000	-0.001	0.000	-0.001	0.000
NII	1,000	2:1	Large	(-0.5,1.3)	0.004	0.002	0.002	0.002	0.001
NII	1,000	2:1	Large	(-0.5,0.7)	0.007	0.005	0.002	0.005	0.001
NILI	300	1:1	Small	(0,1)	0.018	0.018	0.008	0.001	0.000
NILI	300	1:1	Small	(-0.5,1.3)	0.020	0.023	0.006	-0.003	-0.002
NILI	300	1:1	Small	(-0.5,0.7)	0.021	0.023	0.008	-0.004	-0.002
NILI	300	1:1	Medium	(0,1)	0.012	0.017	0.007	0.004	0.001
NILI	300	1:1	Medium	(-0.5,1.3)	0.011	0.018	0.005	0.001	0.000
NILI	300	1:1	Medium	(-0.5,0.7)	0.012	0.020	0.005	-0.001	-0.002
NILI	300	1:1	Large	(0,1)	0.006	0.014	0.003	0.004	-0.001
NILI	300	1:1	Large	(-0.5,1.3)	0.010	0.019	0.006	0.008	0.001
NILI	300	1:1	Large	(-0.5,0.7)	0.009	0.018	0.004	-0.003	-0.002
NILI	300	2:1	Small	(0,1)	0.017	0.018	0.007	0.001	0.000
NILI	300	2:1	Small	(-0.5,1.3)	0.020	0.024	0.006	0.000	0.000
NILI	300	2:1	Small	(-0.5,0.7)	0.026	0.029	0.008	0.001	-0.001
NILI	300	2:1	Medium	(0,1)	0.010	0.018	0.005	0.004	0.000
NILI	300	2:1	Medium	(-0.5,1.3)	0.013	0.022	0.007	0.004	0.001
NILI	300	2:1	Medium	(-0.5,0.7)	0.012	0.022	0.004	0.000	0.000
NILI	300	2:1	Large	(0,1)	0.003	0.014	0.004	0.001	0.000
NILI	300	2:1	Large	(-0.5,1.3)	0.009	0.020	0.007	0.008	0.003
NILI	300	2:1	Large	(-0.5,0.7)	0.012	0.023	0.006	0.001	0.000
NILI	500	1:1	Small	(0,1)	0.010	0.011	0.005	0.002	0.001
NILI	500	1:1	Small	(-0.5,1.3)	0.014	0.017	0.005	0.001	0.000
NILI	500	1:1	Small	(-0.5,0.7)	0.016	0.017	0.006	0.001	0.000
NILI	500	1:1	Medium	(0,1)	0.005	0.009	0.003	0.001	0.000
NILI	500	1:1	Medium	(-0.5,1.3)	0.006	0.010	0.002	-0.004	-0.002
NILI	500	1:1	Medium	(-0.5,0.7)	0.005	0.010	0.002	-0.004	-0.003
NILI	500	1:1	Large	(0,1)	0.002	0.005	0.001	-0.002	-0.002
NILI	500	1:1	Large	(-0.5,1.3)	0.009	0.014	0.004	0.003	-0.001
NILI	500	1:1	Large	(-0.5,0.7)	0.002	0.010	0.001	0.002	-0.001
NILI	500	2:1	Small	(0,1)	0.006	0.007	0.003	-0.003	-0.002
NILI	500	2:1	Small	(-0.5,1.3)	0.017	0.020	0.006	0.003	0.001
NILI	500	2:1	Small	(-0.5,0.7)	0.016	0.018	0.004	0.000	-0.001
NILI	500	2:1	Medium	(0,1)	0.004	0.010	0.002	0.002	-0.001
NILI	500	2:1	Medium	(-0.5,1.3)	0.008	0.014	0.004	0.003	0.001
NILI	500	2:1	Medium	(-0.5,0.7)	0.010	0.018	0.004	0.006	0.003

NILI	500	2:1	Large	(0,1)	0.002	0.008	0.003	0.002	0.001
NILI	500	2:1	Large	(-0.5,1.3)	0.006	0.013	0.005	0.007	0.003
NILI	500	2:1	Large	(-0.5,0.7)	0.003	0.009	0.002	0.000	-0.001
NILI	1,000	1:1	Small	(0,1)	0.004	0.004	0.002	-0.001	0.000
NILI	1,000	1:1	Small	(-0.5,1.3)	0.008	0.010	0.003	0.000	0.000
NILI	1,000	1:1	Small	(-0.5,0.7)	0.008	0.010	0.003	0.002	0.000
NILI	1,000	1:1	Medium	(0,1)	0.003	0.004	0.002	0.000	0.000
NILI	1,000	1:1	Medium	(-0.5,1.3)	0.004	0.006	0.002	0.001	0.000
NILI	1,000	1:1	Medium	(-0.5,0.7)	0.007	0.011	0.003	0.005	0.000
NILI	1,000	1:1	Large	(0,1)	0.003	0.006	0.001	0.002	-0.001
NILI	1,000	1:1	Large	(-0.5,1.3)	0.004	0.006	0.002	0.002	-0.001
NILI	1,000	1:1	Large	(-0.5,0.7)	0.003	0.006	0.001	0.001	0.000
NILI	1,000	2:1	Small	(0,1)	0.003	0.004	0.001	-0.001	-0.001
NILI	1,000	2:1	Small	(-0.5,1.3)	0.008	0.011	0.002	0.001	0.000
NILI	1,000	2:1	Small	(-0.5,0.7)	0.008	0.010	0.002	0.002	-0.001
NILI	1,000	2:1	Medium	(0,1)	0.004	0.007	0.001	0.002	-0.001
NILI	1,000	2:1	Medium	(-0.5,1.3)	0.002	0.005	0.001	0.000	0.000
NILI	1,000	2:1	Medium	(-0.5,0.7)	0.001	0.005	0.000	-0.002	-0.001
NILI	1,000	2:1	Large	(0,1)	0.004	0.008	0.003	0.004	0.001
NILI	1,000	2:1	Large	(-0.5,1.3)	0.002	0.006	0.003	0.001	0.001
NILI	1,000	2:1	Large	(-0.5,0.7)	0.002	0.005	0.001	0.000	-0.001

Notes. [1]Location of non-invariance (I = invariant indicator, NIL = non-invariant loading, NII = non-invariant intercept, NILI = non-invariant loading and intercept), [2]Total sample size, [3]Balance of sample sizes, [4]Magnitude of non-invariance, [5]Latent variable distribution of Group 2 (mean, variance).

Table 3

Standardized bias of five effect size measures for each design cell.

Loc[1]	Mag[2]	N[3]	Bal[4]	LVD2[5]	d_{MACS}	UDI_2	*WUDI*	SDI_2	*WSDI*
I	Small	300	1:1	(0,1)	**1.87**	**1.86**	**1.55**	0.02	0.02
I	Small	300	1:1	(-0.5,1.3)	**1.87**	**1.84**	**1.80**	0.02	0.02
I	Small	300	1:1	(-0.5,0.7)	**1.82**	**1.75**	**1.48**	-0.06	-0.06
I	Small	300	2:1	(0,1)	**1.92**	**1.86**	**1.85**	-0.01	-0.01
I	Small	300	2:1	(-0.5,1.3)	**1.84**	**1.79**	**1.71**	0.00	-0.03
I	Small	300	2:1	(-0.5,0.7)	**1.86**	**1.81**	**1.77**	0.02	0.01
I	Small	500	1:1	(0,1)	**1.88**	**1.81**	**1.43**	0.01	0.01
I	Small	500	1:1	(-0.5,1.3)	**1.89**	**1.89**	**1.78**	0.02	0.00
I	Small	500	1:1	(-0.5,0.7)	**1.90**	**1.85**	**1.45**	-0.02	-0.02
I	Small	500	2:1	(0,1)	**1.86**	**1.83**	**1.79**	-0.04	-0.04
I	Small	500	2:1	(-0.5,1.3)	**1.85**	**1.81**	**1.68**	-0.01	-0.03
I	Small	500	2:1	(-0.5,0.7)	**2.01**	**1.98**	**1.90**	0.02	0.00
I	Small	1,000	1:1	(0,1)	**1.88**	**1.86**	**1.30**	-0.05	-0.05
I	Small	1,000	1:1	(-0.5,1.3)	**1.97**	**1.96**	**1.78**	0.01	-0.01
I	Small	1,000	1:1	(-0.5,0.7)	**1.86**	**1.81**	**1.21**	-0.08	-0.09
I	Small	1,000	2:1	(0,1)	**1.91**	**1.88**	**1.81**	-0.02	-0.02
I	Small	1,000	2:1	(-0.5,1.3)	**1.89**	**1.88**	**1.66**	-0.08	-0.13
I	Small	1,000	2:1	(-0.5,0.7)	**1.93**	**1.88**	**1.76**	0.02	-0.01
I	Medium	300	1:1	(0,1)	**1.88**	**1.86**	**1.70**	0.00	0.00
I	Medium	300	1:1	(-0.5,1.3)	**1.82**	**1.81**	**1.64**	0.07	0.04
I	Medium	300	1:1	(-0.5,0.7)	**1.90**	**1.86**	**1.57**	-0.04	-0.11
I	Medium	300	2:1	(0,1)	**1.90**	**1.84**	**1.54**	0.02	-0.04
I	Medium	300	2:1	(-0.5,1.3)	**1.87**	**1.80**	**1.77**	-0.01	-0.03
I	Medium	300	2:1	(-0.5,0.7)	**1.86**	**1.81**	**1.73**	0.00	0.03
I	Medium	500	1:1	(0,1)	**1.91**	**1.89**	**1.68**	-0.05	-0.05
I	Medium	500	1:1	(-0.5,1.3)	**1.88**	**1.85**	**1.66**	-0.01	-0.03
I	Medium	500	1:1	(-0.5,0.7)	**1.86**	**1.81**	**1.45**	0.00	-0.09
I	Medium	500	2:1	(0,1)	**1.89**	**1.86**	**1.44**	0.04	-0.04
I	Medium	500	2:1	(-0.5,1.3)	**1.93**	**1.91**	**1.82**	-0.03	-0.05
I	Medium	500	2:1	(-0.5,0.7)	**1.93**	**1.89**	**1.76**	0.00	0.03
I	Medium	1,000	1:1	(0,1)	**1.95**	**1.94**	**1.60**	0.09	0.09
I	Medium	1,000	1:1	(-0.5,1.3)	**1.82**	**1.81**	**1.49**	0.03	-0.02
I	Medium	1,000	1:1	(-0.5,0.7)	**1.79**	**1.75**	**1.21**	0.01	-0.12
I	Medium	1,000	2:1	(0,1)	**1.90**	**1.87**	**1.27**	0.04	-0.08
I	Medium	1,000	2:1	(-0.5,1.3)	**1.87**	**1.86**	**1.75**	-0.05	-0.09
I	Medium	1,000	2:1	(-0.5,0.7)	**1.96**	**1.92**	**1.70**	0.00	0.03
I	Large	300	1:1	(0,1)	**1.88**	**1.85**	**1.71**	0.01	0.00
I	Large	300	1:1	(-0.5,1.3)	**1.90**	**1.88**	**1.72**	0.04	0.02
I	Large	300	1:1	(-0.5,0.7)	**1.84**	**1.77**	**1.46**	0.01	0.01
I	Large	300	2:1	(0,1)	**1.85**	**1.84**	**1.72**	0.02	0.04
I	Large	300	2:1	(-0.5,1.3)	**1.87**	**1.82**	**1.79**	0.03	0.04
I	Large	300	2:1	(-0.5,0.7)	**1.93**	**1.86**	**1.84**	-0.01	0.02
I	Large	500	1:1	(0,1)	**1.90**	**1.89**	**1.69**	-0.03	-0.03
I	Large	500	1:1	(-0.5,1.3)	**1.85**	**1.84**	**1.64**	-0.02	-0.04
I	Large	500	1:1	(-0.5,0.7)	**1.83**	**1.76**	**1.33**	0.04	0.02
I	Large	500	2:1	(0,1)	**1.89**	**1.87**	**1.69**	-0.04	0.00
I	Large	500	2:1	(-0.5,1.3)	**1.88**	**1.86**	**1.78**	0.02	0.03

I	Large	500	2:1	(-0.5,0.7)	**1.89**	**1.84**	**1.78**	-0.02	0.01
I	Large	1,000	1:1	(0,1)	**1.89**	**1.86**	**1.57**	-0.02	-0.02
I	Large	1,000	1:1	(-0.5,1.3)	**1.85**	**1.84**	**1.54**	-0.07	-0.12
I	Large	1,000	1:1	(-0.5,0.7)	**1.95**	**1.92**	**1.22**	0.00	-0.03
I	Large	1,000	2:1	(0,1)	**1.97**	**1.94**	**1.70**	-0.02	0.03
I	Large	1,000	2:1	(-0.5,1.3)	**1.83**	**1.81**	**1.69**	0.00	0.01
I	Large	1,000	2:1	(-0.5,0.7)	**1.93**	**1.89**	**1.78**	0.02	0.05
NIL	Small	300	1:1	(0,1)	**0.48**	**0.53**	**0.52**	0.02	0.03
NIL	Small	300	1:1	(-0.5,1.3)	0.36	**0.43**	**0.53**	-0.05	-0.07
NIL	Small	300	1:1	(-0.5,0.7)	**0.48**	**0.54**	**0.61**	0.04	0.05
NIL	Small	300	2:1	(0,1)	**0.57**	**0.63**	**0.68**	-0.06	-0.07
NIL	Small	300	2:1	(-0.5,1.3)	**0.49**	**0.55**	**0.62**	-0.03	0.00
NIL	Small	300	2:1	(-0.5,0.7)	**0.53**	**0.60**	**0.64**	0.01	0.03
NIL	Small	500	1:1	(0,1)	0.37	**0.42**	0.39	0.00	0.01
NIL	Small	500	1:1	(-0.5,1.3)	0.24	0.30	0.39	0.00	-0.02
NIL	Small	500	1:1	(-0.5,0.7)	0.40	**0.46**	**0.50**	0.00	0.02
NIL	Small	500	2:1	(0,1)	**0.43**	**0.48**	**0.52**	-0.02	-0.04
NIL	Small	500	2:1	(-0.5,1.3)	0.27	0.33	0.39	0.04	0.05
NIL	Small	500	2:1	(-0.5,0.7)	0.35	**0.42**	**0.45**	0.04	0.06
NIL	Small	1,000	1:1	(0,1)	0.32	0.35	0.28	0.03	0.04
NIL	Small	1,000	1:1	(-0.5,1.3)	0.21	0.25	0.30	-0.01	-0.03
NIL	Small	1,000	1:1	(-0.5,0.7)	0.23	0.29	0.31	0.03	0.03
NIL	Small	1,000	2:1	(0,1)	0.21	0.26	0.29	-0.02	-0.04
NIL	Small	1,000	2:1	(-0.5,1.3)	0.19	0.23	0.28	-0.01	0.00
NIL	Small	1,000	2:1	(-0.5,0.7)	0.24	0.30	0.31	0.01	0.03
NIL	Medium	300	1:1	(0,1)	0.19	0.22	0.21	0.06	0.07
NIL	Medium	300	1:1	(-0.5,1.3)	0.11	0.15	0.15	-0.06	-0.07
NIL	Medium	300	1:1	(-0.5,0.7)	0.18	0.24	0.23	-0.04	-0.03
NIL	Medium	300	2:1	(0,1)	0.15	0.18	0.17	-0.01	-0.03
NIL	Medium	300	2:1	(-0.5,1.3)	0.12	0.17	0.16	-0.02	0.02
NIL	Medium	300	2:1	(-0.5,0.7)	0.15	0.22	0.19	-0.02	-0.01
NIL	Medium	500	1:1	(0,1)	0.16	0.18	0.18	-0.02	-0.03
NIL	Medium	500	1:1	(-0.5,1.3)	0.10	0.13	0.14	-0.04	-0.05
NIL	Medium	500	1:1	(-0.5,0.7)	0.18	0.23	0.22	-0.04	-0.03
NIL	Medium	500	2:1	(0,1)	0.10	0.12	0.12	-0.02	-0.05
NIL	Medium	500	2:1	(-0.5,1.3)	0.06	0.11	0.09	0.02	0.06
NIL	Medium	500	2:1	(-0.5,0.7)	0.09	0.14	0.13	0.02	0.01
NIL	Medium	1,000	1:1	(0,1)	0.07	0.08	0.10	-0.03	-0.02
NIL	Medium	1,000	1:1	(-0.5,1.3)	0.02	0.05	0.06	0.03	0.02
NIL	Medium	1,000	1:1	(-0.5,0.7)	0.10	0.14	0.14	-0.03	-0.04
NIL	Medium	1,000	2:1	(0,1)	0.12	0.13	0.12	0.00	-0.04
NIL	Medium	1,000	2:1	(-0.5,1.3)	0.05	0.08	0.07	-0.04	0.02
NIL	Medium	1,000	2:1	(-0.5,0.7)	0.08	0.12	0.11	-0.04	-0.06
NIL	Large	300	1:1	(0,1)	0.08	0.10	0.07	-0.03	-0.06
NIL	Large	300	1:1	(-0.5,1.3)	0.05	0.09	0.07	-0.05	-0.05
NIL	Large	300	1:1	(-0.5,0.7)	0.08	0.12	0.08	-0.02	-0.01
NIL	Large	300	2:1	(0,1)	0.08	0.13	0.09	-0.03	-0.01
NIL	Large	300	2:1	(-0.5,1.3)	0.09	0.13	0.11	-0.04	0.01
NIL	Large	300	2:1	(-0.5,0.7)	0.04	0.09	0.05	0.02	0.06
NIL	Large	500	1:1	(0,1)	0.10	0.11	0.08	-0.04	-0.06
NIL	Large	500	1:1	(-0.5,1.3)	0.06	0.09	0.08	-0.02	-0.03
NIL	Large	500	1:1	(-0.5,0.7)	0.03	0.06	0.04	0.02	0.01

NIL	Large	500	2:1	(0,1)	0.04	0.07	0.04	-0.07	-0.08
NIL	Large	500	2:1	(-0.5,1.3)	0.01	0.05	0.03	-0.01	0.02
NIL	Large	500	2:1	(-0.5,0.7)	0.01	0.05	0.02	0.03	0.06
NIL	Large	1,000	1:1	(0,1)	0.03	0.05	0.04	-0.01	-0.04
NIL	Large	1,000	1:1	(-0.5,1.3)	0.06	0.08	0.07	-0.01	-0.01
NIL	Large	1,000	1:1	(-0.5,0.7)	0.06	0.08	0.05	-0.04	-0.09
NIL	Large	1,000	2:1	(0,1)	0.04	0.05	0.02	-0.04	-0.03
NIL	Large	1,000	2:1	(-0.5,1.3)	0.05	0.07	0.06	-0.02	0.04
NIL	Large	1,000	2:1	(-0.5,0.7)	0.05	0.09	0.07	0.00	0.03
NII	Small	300	1:1	(0,1)	0.24	0.08	0.08	0.02	0.01
NII	Small	300	1:1	(-0.5,1.3)	0.25	0.08	0.10	0.00	-0.01
NII	Small	300	1:1	(-0.5,0.7)	0.23	0.10	0.17	0.04	0.05
NII	Small	300	2:1	(0,1)	0.19	0.04	0.00	-0.03	-0.07
NII	Small	300	2:1	(-0.5,1.3)	0.27	0.11	0.13	0.02	0.01
NII	Small	300	2:1	(-0.5,0.7)	0.28	0.12	0.15	0.03	0.01
NII	Small	500	1:1	(0,1)	0.23	0.10	0.09	0.07	0.06
NII	Small	500	1:1	(-0.5,1.3)	0.19	0.03	0.06	0.00	-0.01
NII	Small	500	1:1	(-0.5,0.7)	0.17	0.03	0.10	0.00	0.01
NII	Small	500	2:1	(0,1)	0.14	0.00	-0.04	-0.04	-0.08
NII	Small	500	2:1	(-0.5,1.3)	0.21	0.06	0.08	0.01	0.00
NII	Small	500	2:1	(-0.5,0.7)	0.19	0.07	0.09	0.04	0.02
NII	Small	1,000	1:1	(0,1)	0.10	0.01	0.00	0.00	-0.01
NII	Small	1,000	1:1	(-0.5,1.3)	0.11	-0.01	0.01	-0.02	-0.02
NII	Small	1,000	1:1	(-0.5,0.7)	0.11	0.02	0.05	0.01	0.01
NII	Small	1,000	2:1	(0,1)	0.14	0.03	-0.03	0.02	-0.04
NII	Small	1,000	2:1	(-0.5,1.3)	0.09	-0.03	-0.01	-0.04	-0.05
NII	Small	1,000	2:1	(-0.5,0.7)	0.16	0.06	0.06	0.05	0.02
NII	Medium	300	1:1	(0,1)	0.11	0.02	0.00	0.01	-0.01
NII	Medium	300	1:1	(-0.5,1.3)	0.11	0.00	0.00	0.00	-0.03
NII	Medium	300	1:1	(-0.5,0.7)	0.11	0.01	0.02	0.01	-0.03
NII	Medium	300	2:1	(0,1)	0.09	-0.01	-0.09	-0.01	-0.10
NII	Medium	300	2:1	(-0.5,1.3)	0.17	0.06	0.11	0.05	0.07
NII	Medium	300	2:1	(-0.5,0.7)	0.10	0.01	0.04	0.01	0.01
NII	Medium	500	1:1	(0,1)	0.08	0.00	-0.01	0.00	-0.01
NII	Medium	500	1:1	(-0.5,1.3)	0.10	0.01	0.00	0.01	-0.02
NII	Medium	500	1:1	(-0.5,0.7)	0.09	0.03	0.02	0.03	0.00
NII	Medium	500	2:1	(0,1)	0.12	0.04	-0.06	0.04	-0.06
NII	Medium	500	2:1	(-0.5,1.3)	0.16	0.09	0.13	0.09	0.11
NII	Medium	500	2:1	(-0.5,0.7)	0.07	-0.02	0.02	-0.02	0.00
NII	Medium	1,000	1:1	(0,1)	0.06	0.01	0.00	0.01	0.00
NII	Medium	1,000	1:1	(-0.5,1.3)	0.05	-0.01	-0.04	-0.01	-0.05
NII	Medium	1,000	1:1	(-0.5,0.7)	0.07	0.05	0.01	0.05	0.00
NII	Medium	1,000	2:1	(0,1)	0.07	0.01	-0.13	0.01	-0.13
NII	Medium	1,000	2:1	(-0.5,1.3)	0.06	0.01	0.04	0.01	0.03
NII	Medium	1,000	2:1	(-0.5,0.7)	0.02	-0.02	0.00	-0.02	-0.01
NII	Large	300	1:1	(0,1)	0.15	0.08	0.06	0.08	0.06
NII	Large	300	1:1	(-0.5,1.3)	0.12	0.06	0.02	0.06	0.01
NII	Large	300	1:1	(-0.5,0.7)	0.05	-0.01	-0.01	-0.01	-0.03
NII	Large	300	2:1	(0,1)	0.09	0.06	0.03	0.06	0.03
NII	Large	300	2:1	(-0.5,1.3)	0.14	0.07	0.09	0.07	0.07
NII	Large	300	2:1	(-0.5,0.7)	0.11	0.07	0.05	0.07	0.03
NII	Large	500	1:1	(0,1)	0.03	-0.02	-0.03	-0.02	-0.03

NII	Large	500	1:1	(-0.5,1.3)	0.10	0.05	0.01	0.05	0.00
NII	Large	500	1:1	(-0.5,0.7)	0.06	0.03	0.00	0.03	-0.01
NII	Large	500	2:1	(0,1)	0.04	-0.02	-0.01	-0.02	-0.01
NII	Large	500	2:1	(-0.5,1.3)	0.06	0.03	0.05	0.03	0.03
NII	Large	500	2:1	(-0.5,0.7)	0.11	0.08	0.06	0.08	0.04
NII	Large	1,000	1:1	(0,1)	0.07	0.04	0.02	0.04	0.02
NII	Large	1,000	1:1	(-0.5,1.3)	0.06	0.02	-0.03	0.02	-0.03
NII	Large	1,000	1:1	(-0.5,0.7)	0.07	0.02	0.03	0.02	0.02
NII	Large	1,000	2:1	(0,1)	0.01	-0.02	-0.02	-0.02	-0.02
NII	Large	1,000	2:1	(-0.5,1.3)	0.08	0.04	0.08	0.04	0.07
NII	Large	1,000	2:1	(-0.5,0.7)	0.13	0.09	0.07	0.09	0.06
NILI	Small	300	1:1	(0,1)	0.22	0.22	0.21	0.01	0.00
NILI	Small	300	1:1	(-0.5,1.3)	0.29	0.35	0.19	-0.04	-0.06
NILI	Small	300	1:1	(-0.5,0.7)	0.28	0.29	0.21	-0.04	-0.06
NILI	Small	300	2:1	(0,1)	0.22	0.22	0.20	0.01	-0.01
NILI	Small	300	2:1	(-0.5,1.3)	0.26	0.32	0.20	0.00	-0.01
NILI	Small	300	2:1	(-0.5,0.7)	0.34	0.35	0.24	0.01	-0.02
NILI	Small	500	1:1	(0,1)	0.16	0.17	0.17	0.03	0.03
NILI	Small	500	1:1	(-0.5,1.3)	0.25	0.31	0.18	0.02	0.00
NILI	Small	500	1:1	(-0.5,0.7)	0.26	0.27	0.20	0.02	0.00
NILI	Small	500	2:1	(0,1)	0.09	0.11	0.09	-0.04	-0.06
NILI	Small	500	2:1	(-0.5,1.3)	0.28	0.35	0.22	0.04	0.05
NILI	Small	500	2:1	(-0.5,0.7)	0.25	0.27	0.16	0.00	-0.03
NILI	Small	1,000	1:1	(0,1)	0.08	0.09	0.09	-0.02	-0.02
NILI	Small	1,000	1:1	(-0.5,1.3)	0.20	0.27	0.14	0.01	-0.01
NILI	Small	1,000	1:1	(-0.5,0.7)	0.17	0.20	0.13	0.05	-0.01
NILI	Small	1,000	2:1	(0,1)	0.07	0.09	0.06	-0.02	-0.05
NILI	Small	1,000	2:1	(-0.5,1.3)	0.19	0.27	0.13	0.02	0.02
NILI	Small	1,000	2:1	(-0.5,0.7)	0.17	0.20	0.09	0.03	-0.02
NILI	Medium	300	1:1	(0,1)	0.14	0.18	0.18	0.04	0.03
NILI	Medium	300	1:1	(-0.5,1.3)	0.14	0.22	0.13	0.01	0.00
NILI	Medium	300	1:1	(-0.5,0.7)	0.14	0.21	0.12	-0.01	-0.03
NILI	Medium	300	2:1	(0,1)	0.11	0.17	0.14	0.04	0.00
NILI	Medium	300	2:1	(-0.5,1.3)	0.15	0.24	0.19	0.04	0.03
NILI	Medium	300	2:1	(-0.5,0.7)	0.13	0.21	0.11	0.00	0.00
NILI	Medium	500	1:1	(0,1)	0.07	0.11	0.11	0.01	0.00
NILI	Medium	500	1:1	(-0.5,1.3)	0.09	0.16	0.07	-0.05	-0.07
NILI	Medium	500	1:1	(-0.5,0.7)	0.08	0.13	0.06	-0.05	-0.09
NILI	Medium	500	2:1	(0,1)	0.06	0.12	0.08	0.02	-0.02
NILI	Medium	500	2:1	(-0.5,1.3)	0.11	0.20	0.15	0.04	0.03
NILI	Medium	500	2:1	(-0.5,0.7)	0.15	0.23	0.13	0.07	0.08
NILI	Medium	1,000	1:1	(0,1)	0.05	0.08	0.09	0.00	0.00
NILI	Medium	1,000	1:1	(-0.5,1.3)	0.08	0.14	0.09	0.02	0.00
NILI	Medium	1,000	1:1	(-0.5,0.7)	0.14	0.20	0.14	0.08	0.02
NILI	Medium	1,000	2:1	(0,1)	0.09	0.12	0.07	0.03	-0.02
NILI	Medium	1,000	2:1	(-0.5,1.3)	0.03	0.10	0.07	-0.01	-0.02
NILI	Medium	1,000	2:1	(-0.5,0.7)	0.03	0.09	0.00	-0.03	-0.03
NILI	Large	300	1:1	(0,1)	0.06	0.12	0.07	0.03	-0.03
NILI	Large	300	1:1	(-0.5,1.3)	0.11	0.18	0.14	0.06	0.02
NILI	Large	300	1:1	(-0.5,0.7)	0.10	0.16	0.09	-0.02	-0.04
NILI	Large	300	2:1	(0,1)	0.02	0.10	0.10	0.01	-0.01
NILI	Large	300	2:1	(-0.5,1.3)	0.10	0.17	0.16	0.06	0.06

NILI	Large	300	2:1	(-0.5,0.7)	0.12	0.19	0.13	0.01	-0.01
NILI	Large	500	1:1	(0,1)	0.03	0.06	0.03	-0.02	-0.06
NILI	Large	500	1:1	(-0.5,1.3)	0.12	0.17	0.12	0.03	-0.01
NILI	Large	500	1:1	(-0.5,0.7)	0.03	0.11	0.03	0.02	-0.01
NILI	Large	500	2:1	(0,1)	0.03	0.08	0.09	0.02	0.02
NILI	Large	500	2:1	(-0.5,1.3)	0.08	0.14	0.16	0.06	0.07
NILI	Large	500	2:1	(-0.5,0.7)	0.03	0.10	0.06	0.00	-0.03
NILI	Large	1,000	1:1	(0,1)	0.06	0.09	0.05	0.03	-0.02
NILI	Large	1,000	1:1	(-0.5,1.3)	0.08	0.11	0.09	0.04	-0.02
NILI	Large	1,000	1:1	(-0.5,0.7)	0.06	0.10	0.05	0.01	-0.01
NILI	Large	1,000	2:1	(0,1)	0.06	0.11	0.11	0.05	0.04
NILI	Large	1,000	2:1	(-0.5,1.3)	0.04	0.10	0.12	0.01	0.03
NILI	Large	1,000	2:1	(-0.5,0.7)	0.04	0.07	0.06	0.00	-0.03

Notes. [1]Location of non-invariance (I = invariant indicator, NIL = non-invariant loading, NII = non-invariant intercept, NILI = non-invariant loading and intercept), [2]Magnitude of non-invariance, [3]Total sample size, [4]Balance of sample sizes, [5]Latent variable distribution of Group 2 (mean, variance). Bolded values represent standardized bias values greater than 0.4.

Table 4

RMSE of the five effect size measures for each design cell.

Loc[1]	LVD2[2]	Mag[3]	N[4]	Bal[5]	d_{MACS}	*SDI_2*	*UDI_2*	*WSDI*	*WUDI*
I	(0,1)	Small	300	2:1	0.122	0.036	0.087	0.044	0.107
I	(0,1)	Small	300	1:1	0.112	0.038	0.077	0.042	0.097
I	(0,1)	Small	500	2:1	0.093	0.027	0.064	0.033	0.080
I	(0,1)	Small	500	1:1	0.089	0.032	0.064	0.032	0.078
I	(0,1)	Small	1,000	2:1	0.066	0.019	0.045	0.023	0.057
I	(0,1)	Small	1,000	1:1	0.062	0.021	0.043	0.021	0.054
I	(0,1)	Medium	300	2:1	0.120	0.035	0.084	0.038	0.104
I	(0,1)	Medium	300	1:1	0.109	0.038	0.076	0.043	0.094
I	(0,1)	Medium	500	2:1	0.092	0.027	0.065	0.028	0.080
I	(0,1)	Medium	500	1:1	0.090	0.032	0.063	0.036	0.078
I	(0,1)	Medium	1,000	2:1	0.065	0.019	0.045	0.018	0.056
I	(0,1)	Medium	1,000	1:1	0.061	0.021	0.042	0.023	0.053
I	(0,1)	Large	300	2:1	0.123	0.035	0.084	0.042	0.106
I	(0,1)	Large	300	1:1	0.113	0.039	0.078	0.046	0.098
I	(0,1)	Large	500	2:1	0.090	0.026	0.062	0.030	0.078
I	(0,1)	Large	500	1:1	0.088	0.030	0.060	0.034	0.076
I	(0,1)	Large	1,000	2:1	0.066	0.019	0.046	0.022	0.057
I	(0,1)	Large	1,000	1:1	0.062	0.021	0.043	0.023	0.053
I	(-0.5,1.3)	Small	300	2:1	0.124	0.035	0.084	0.041	0.103
I	(-0.5,1.3)	Small	300	1:1	0.117	0.037	0.078	0.045	0.096
I	(-0.5,1.3)	Small	500	2:1	0.097	0.026	0.065	0.031	0.079
I	(-0.5,1.3)	Small	500	1:1	0.087	0.028	0.058	0.033	0.072
I	(-0.5,1.3)	Small	1,000	2:1	0.066	0.017	0.042	0.020	0.053
I	(-0.5,1.3)	Small	1,000	1:1	0.063	0.020	0.042	0.023	0.052
I	(-0.5,1.3)	Medium	300	2:1	0.122	0.033	0.081	0.041	0.101
I	(-0.5,1.3)	Medium	300	1:1	0.114	0.036	0.075	0.042	0.094
I	(-0.5,1.3)	Medium	500	2:1	0.094	0.025	0.060	0.031	0.076
I	(-0.5,1.3)	Medium	500	1:1	0.089	0.028	0.061	0.032	0.074
I	(-0.5,1.3)	Medium	1,000	2:1	0.068	0.018	0.043	0.022	0.055
I	(-0.5,1.3)	Medium	1,000	1:1	0.063	0.020	0.042	0.022	0.052
I	(-0.5,1.3)	Large	300	2:1	0.121	0.034	0.082	0.041	0.100
I	(-0.5,1.3)	Large	300	1:1	0.116	0.037	0.078	0.043	0.096
I	(-0.5,1.3)	Large	500	2:1	0.092	0.025	0.060	0.030	0.075
I	(-0.5,1.3)	Large	500	1:1	0.089	0.028	0.058	0.032	0.073
I	(-0.5,1.3)	Large	1,000	2:1	0.066	0.018	0.044	0.021	0.053
I	(-0.5,1.3)	Large	1,000	1:1	0.063	0.020	0.042	0.022	0.052
I	(-0.5,0.7)	Small	300	2:1	0.122	0.035	0.093	0.043	0.114
I	(-0.5,0.7)	Small	300	1:1	0.119	0.041	0.094	0.044	0.110
I	(-0.5,0.7)	Small	500	2:1	0.095	0.026	0.069	0.033	0.088
I	(-0.5,0.7)	Small	500	1:1	0.093	0.031	0.072	0.033	0.086
I	(-0.5,0.7)	Small	1,000	2:1	0.068	0.020	0.053	0.024	0.064
I	(-0.5,0.7)	Small	1,000	1:1	0.065	0.022	0.050	0.021	0.060
I	(-0.5,0.7)	Medium	300	2:1	0.124	0.036	0.096	0.043	0.116
I	(-0.5,0.7)	Medium	300	1:1	0.118	0.040	0.092	0.044	0.108

I	(-0.5,0.7)	Medium	500	2:1	0.095	0.027	0.072	0.032	0.088
I	(-0.5,0.7)	Medium	500	1:1	0.094	0.031	0.072	0.033	0.086
I	(-0.5,0.7)	Medium	1,000	2:1	0.067	0.019	0.050	0.022	0.063
I	(-0.5,0.7)	Medium	1,000	1:1	0.066	0.023	0.052	0.022	0.061
I	(-0.5,0.7)	Large	300	2:1	0.127	0.037	0.099	0.045	0.120
I	(-0.5,0.7)	Large	300	1:1	0.119	0.041	0.094	0.043	0.110
I	(-0.5,0.7)	Large	500	2:1	0.097	0.028	0.075	0.034	0.091
I	(-0.5,0.7)	Large	500	1:1	0.094	0.032	0.073	0.033	0.087
I	(-0.5,0.7)	Large	1,000	2:1	0.067	0.019	0.051	0.023	0.062
I	(-0.5,0.7)	Large	1,000	1:1	0.065	0.022	0.050	0.020	0.060
NIL	(0,1)	Small	300	2:1	0.081	0.039	0.093	0.031	0.077
NIL	(0,1)	Small	300	1:1	0.078	0.040	0.085	0.031	0.071
NIL	(0,1)	Small	500	2:1	0.064	0.030	0.071	0.024	0.059
NIL	(0,1)	Small	500	1:1	0.061	0.031	0.065	0.024	0.054
NIL	(0,1)	Small	1,000	2:1	0.044	0.020	0.048	0.016	0.039
NIL	(0,1)	Small	1,000	1:1	0.045	0.022	0.046	0.017	0.039
NIL	(0,1)	Medium	300	2:1	0.092	0.040	0.106	0.034	0.094
NIL	(0,1)	Medium	300	1:1	0.092	0.042	0.098	0.036	0.090
NIL	(0,1)	Medium	500	2:1	0.074	0.032	0.083	0.027	0.075
NIL	(0,1)	Medium	500	1:1	0.071	0.032	0.076	0.028	0.069
NIL	(0,1)	Medium	1,000	2:1	0.051	0.022	0.058	0.019	0.052
NIL	(0,1)	Medium	1,000	1:1	0.049	0.023	0.053	0.019	0.047
NIL	(0,1)	Large	300	2:1	0.096	0.044	0.126	0.038	0.113
NIL	(0,1)	Large	300	1:1	0.098	0.046	0.113	0.038	0.105
NIL	(0,1)	Large	500	2:1	0.080	0.035	0.099	0.031	0.093
NIL	(0,1)	Large	500	1:1	0.075	0.036	0.089	0.030	0.081
NIL	(0,1)	Large	1,000	2:1	0.054	0.025	0.072	0.021	0.062
NIL	(0,1)	Large	1,000	1:1	0.054	0.025	0.062	0.022	0.058
NIL	(-0.5,1.3)	Small	300	2:1	0.086	0.036	0.090	0.030	0.077
NIL	(-0.5,1.3)	Small	300	1:1	0.083	0.037	0.082	0.030	0.072
NIL	(-0.5,1.3)	Small	500	2:1	0.067	0.027	0.067	0.023	0.058
NIL	(-0.5,1.3)	Small	500	1:1	0.063	0.029	0.066	0.022	0.054
NIL	(-0.5,1.3)	Small	1,000	2:1	0.049	0.020	0.050	0.016	0.042
NIL	(-0.5,1.3)	Small	1,000	1:1	0.047	0.021	0.046	0.016	0.040
NIL	(-0.5,1.3)	Medium	300	2:1	0.096	0.036	0.101	0.032	0.095
NIL	(-0.5,1.3)	Medium	300	1:1	0.097	0.040	0.100	0.032	0.091
NIL	(-0.5,1.3)	Medium	500	2:1	0.077	0.028	0.080	0.026	0.076
NIL	(-0.5,1.3)	Medium	500	1:1	0.077	0.031	0.078	0.025	0.072
NIL	(-0.5,1.3)	Medium	1,000	2:1	0.056	0.020	0.058	0.019	0.055
NIL	(-0.5,1.3)	Medium	1,000	1:1	0.055	0.022	0.056	0.018	0.051
NIL	(-0.5,1.3)	Large	300	2:1	0.108	0.040	0.127	0.037	0.126
NIL	(-0.5,1.3)	Large	300	1:1	0.110	0.043	0.123	0.037	0.119
NIL	(-0.5,1.3)	Large	500	2:1	0.083	0.032	0.103	0.029	0.096
NIL	(-0.5,1.3)	Large	500	1:1	0.084	0.034	0.094	0.028	0.089
NIL	(-0.5,1.3)	Large	1,000	2:1	0.056	0.023	0.070	0.019	0.064
NIL	(-0.5,1.3)	Large	1,000	1:1	0.060	0.023	0.065	0.020	0.064
NIL	(-0.5,0.7)	Small	300	2:1	0.081	0.037	0.100	0.031	0.084
NIL	(-0.5,0.7)	Small	300	1:1	0.084	0.040	0.099	0.035	0.082
NIL	(-0.5,0.7)	Small	500	2:1	0.064	0.029	0.078	0.024	0.065

NIL	(-0.5,0.7)	Small	500	1:1	0.066	0.031	0.078	0.027	0.064
NIL	(-0.5,0.7)	Small	1,000	2:1	0.049	0.021	0.056	0.018	0.048
NIL	(-0.5,0.7)	Small	1,000	1:1	0.046	0.023	0.055	0.018	0.044
NIL	(-0.5,0.7)	Medium	300	2:1	0.100	0.041	0.123	0.037	0.108
NIL	(-0.5,0.7)	Medium	300	1:1	0.094	0.043	0.111	0.037	0.097
NIL	(-0.5,0.7)	Medium	500	2:1	0.075	0.032	0.093	0.028	0.080
NIL	(-0.5,0.7)	Medium	500	1:1	0.076	0.032	0.084	0.030	0.078
NIL	(-0.5,0.7)	Medium	1,000	2:1	0.052	0.022	0.064	0.019	0.056
NIL	(-0.5,0.7)	Medium	1,000	1:1	0.051	0.023	0.060	0.020	0.052
NIL	(-0.5,0.7)	Large	300	2:1	0.101	0.042	0.134	0.039	0.124
NIL	(-0.5,0.7)	Large	300	1:1	0.105	0.044	0.130	0.040	0.117
NIL	(-0.5,0.7)	Large	500	2:1	0.078	0.034	0.106	0.030	0.093
NIL	(-0.5,0.7)	Large	500	1:1	0.081	0.034	0.098	0.032	0.090
NIL	(-0.5,0.7)	Large	1,000	2:1	0.055	0.023	0.073	0.021	0.066
NIL	(-0.5,0.7)	Large	1,000	1:1	0.054	0.024	0.068	0.021	0.060
NII	(0,1)	Small	300	2:1	0.079	0.036	0.083	0.034	0.078
NII	(0,1)	Small	300	1:1	0.074	0.038	0.077	0.036	0.073
NII	(0,1)	Small	500	2:1	0.061	0.028	0.063	0.027	0.061
NII	(0,1)	Small	500	1:1	0.061	0.031	0.063	0.030	0.061
NII	(0,1)	Small	1,000	2:1	0.045	0.020	0.045	0.019	0.045
NII	(0,1)	Small	1,000	1:1	0.042	0.021	0.043	0.021	0.043
NII	(0,1)	Medium	300	2:1	0.083	0.037	0.087	0.037	0.087
NII	(0,1)	Medium	300	1:1	0.080	0.040	0.083	0.040	0.083
NII	(0,1)	Medium	500	2:1	0.067	0.030	0.070	0.030	0.070
NII	(0,1)	Medium	500	1:1	0.061	0.031	0.064	0.031	0.064
NII	(0,1)	Medium	1,000	2:1	0.046	0.021	0.048	0.021	0.048
NII	(0,1)	Medium	1,000	1:1	0.044	0.022	0.044	0.022	0.044
NII	(0,1)	Large	300	2:1	0.088	0.039	0.097	0.039	0.097
NII	(0,1)	Large	300	1:1	0.081	0.041	0.086	0.041	0.086
NII	(0,1)	Large	500	2:1	0.069	0.031	0.073	0.031	0.073
NII	(0,1)	Large	500	1:1	0.062	0.031	0.065	0.031	0.065
NII	(0,1)	Large	1,000	2:1	0.048	0.022	0.053	0.022	0.053
NII	(0,1)	Large	1,000	1:1	0.042	0.021	0.045	0.021	0.045
NII	(-0.5,1.3)	Small	300	2:1	0.085	0.035	0.082	0.033	0.076
NII	(-0.5,1.3)	Small	300	1:1	0.079	0.037	0.076	0.035	0.071
NII	(-0.5,1.3)	Small	500	2:1	0.067	0.028	0.065	0.027	0.062
NII	(-0.5,1.3)	Small	500	1:1	0.062	0.028	0.059	0.028	0.057
NII	(-0.5,1.3)	Small	1,000	2:1	0.044	0.018	0.042	0.018	0.042
NII	(-0.5,1.3)	Small	1,000	1:1	0.043	0.020	0.042	0.020	0.041
NII	(-0.5,1.3)	Medium	300	2:1	0.089	0.037	0.086	0.036	0.085
NII	(-0.5,1.3)	Medium	300	1:1	0.082	0.039	0.079	0.038	0.079
NII	(-0.5,1.3)	Medium	500	2:1	0.068	0.029	0.066	0.028	0.066
NII	(-0.5,1.3)	Medium	500	1:1	0.064	0.030	0.061	0.030	0.061
NII	(-0.5,1.3)	Medium	1,000	2:1	0.047	0.019	0.045	0.019	0.045
NII	(-0.5,1.3)	Medium	1,000	1:1	0.044	0.021	0.042	0.021	0.042
NII	(-0.5,1.3)	Large	300	2:1	0.087	0.037	0.087	0.036	0.087
NII	(-0.5,1.3)	Large	300	1:1	0.086	0.041	0.084	0.041	0.084
NII	(-0.5,1.3)	Large	500	2:1	0.069	0.029	0.069	0.029	0.069
NII	(-0.5,1.3)	Large	500	1:1	0.068	0.032	0.065	0.032	0.065

NII	(-0.5,1.3)	Large	1,000	2:1	0.048	0.020	0.047	0.020	0.047
NII	(-0.5,1.3)	Large	1,000	1:1	0.047	0.022	0.045	0.022	0.045
NII	(-0.5,0.7)	Small	300	2:1	0.088	0.038	0.097	0.036	0.090
NII	(-0.5,0.7)	Small	300	1:1	0.086	0.041	0.092	0.039	0.087
NII	(-0.5,0.7)	Small	500	2:1	0.068	0.029	0.073	0.028	0.071
NII	(-0.5,0.7)	Small	500	1:1	0.067	0.032	0.072	0.031	0.070
NII	(-0.5,0.7)	Small	1,000	2:1	0.052	0.022	0.056	0.022	0.055
NII	(-0.5,0.7)	Small	1,000	1:1	0.047	0.022	0.050	0.022	0.050
NII	(-0.5,0.7)	Medium	300	2:1	0.096	0.042	0.105	0.041	0.105
NII	(-0.5,0.7)	Medium	300	1:1	0.088	0.042	0.094	0.041	0.094
NII	(-0.5,0.7)	Medium	500	2:1	0.071	0.031	0.078	0.031	0.078
NII	(-0.5,0.7)	Medium	500	1:1	0.070	0.033	0.073	0.033	0.073
NII	(-0.5,0.7)	Medium	1,000	2:1	0.051	0.022	0.055	0.022	0.055
NII	(-0.5,0.7)	Medium	1,000	1:1	0.050	0.023	0.053	0.023	0.053
NII	(-0.5,0.7)	Large	300	2:1	0.094	0.041	0.105	0.040	0.105
NII	(-0.5,0.7)	Large	300	1:1	0.096	0.046	0.102	0.045	0.102
NII	(-0.5,0.7)	Large	500	2:1	0.075	0.032	0.083	0.032	0.083
NII	(-0.5,0.7)	Large	500	1:1	0.074	0.035	0.078	0.035	0.078
NII	(-0.5,0.7)	Large	1,000	2:1	0.052	0.023	0.058	0.023	0.058
NII	(-0.5,0.7)	Large	1,000	1:1	0.052	0.025	0.055	0.025	0.055
NILI	(0,1)	Small	300	2:1	0.082	0.037	0.090	0.034	0.084
NILI	(0,1)	Small	300	1:1	0.084	0.041	0.088	0.038	0.083
NILI	(0,1)	Small	500	2:1	0.065	0.030	0.073	0.028	0.067
NILI	(0,1)	Small	500	1:1	0.065	0.031	0.066	0.030	0.063
NILI	(0,1)	Small	1,000	2:1	0.046	0.020	0.049	0.019	0.046
NILI	(0,1)	Small	1,000	1:1	0.044	0.021	0.046	0.020	0.044
NILI	(0,1)	Medium	300	2:1	0.092	0.042	0.111	0.039	0.107
NILI	(0,1)	Medium	300	1:1	0.090	0.044	0.102	0.041	0.097
NILI	(0,1)	Medium	500	2:1	0.072	0.032	0.086	0.030	0.083
NILI	(0,1)	Medium	500	1:1	0.072	0.035	0.083	0.032	0.078
NILI	(0,1)	Medium	1,000	2:1	0.051	0.023	0.060	0.021	0.058
NILI	(0,1)	Medium	1,000	1:1	0.050	0.024	0.056	0.022	0.054
NILI	(0,1)	Large	300	2:1	0.101	0.049	0.144	0.045	0.137
NILI	(0,1)	Large	300	1:1	0.098	0.048	0.123	0.044	0.118
NILI	(0,1)	Large	500	2:1	0.077	0.036	0.105	0.034	0.101
NILI	(0,1)	Large	500	1:1	0.077	0.039	0.100	0.036	0.094
NILI	(0,1)	Large	1,000	2:1	0.057	0.026	0.076	0.024	0.075
NILI	(0,1)	Large	1,000	1:1	0.054	0.026	0.067	0.024	0.064
NILI	(-0.5,1.3)	Small	300	2:1	0.080	0.035	0.086	0.032	0.077
NILI	(-0.5,1.3)	Small	300	1:1	0.071	0.037	0.082	0.032	0.068
NILI	(-0.5,1.3)	Small	500	2:1	0.064	0.029	0.071	0.026	0.061
NILI	(-0.5,1.3)	Small	500	1:1	0.060	0.030	0.065	0.028	0.056
NILI	(-0.5,1.3)	Small	1,000	2:1	0.044	0.019	0.047	0.018	0.041
NILI	(-0.5,1.3)	Small	1,000	1:1	0.042	0.021	0.047	0.020	0.039
NILI	(-0.5,1.3)	Medium	300	2:1	0.088	0.040	0.108	0.036	0.094
NILI	(-0.5,1.3)	Medium	300	1:1	0.083	0.043	0.103	0.039	0.085
NILI	(-0.5,1.3)	Medium	500	2:1	0.069	0.031	0.083	0.029	0.072
NILI	(-0.5,1.3)	Medium	500	1:1	0.064	0.032	0.077	0.029	0.063
NILI	(-0.5,1.3)	Medium	1,000	2:1	0.048	0.021	0.056	0.020	0.049

NILI	(-0.5,1.3)	Medium	1,000	1:1	0.046	0.023	0.056	0.021	0.045
NILI	(-0.5,1.3)	Large	300	2:1	0.098	0.047	0.136	0.043	0.119
NILI	(-0.5,1.3)	Large	300	1:1	0.094	0.050	0.129	0.044	0.105
NILI	(-0.5,1.3)	Large	500	2:1	0.080	0.038	0.108	0.034	0.095
NILI	(-0.5,1.3)	Large	500	1:1	0.075	0.038	0.098	0.034	0.082
NILI	(-0.5,1.3)	Large	1,000	2:1	0.053	0.025	0.073	0.023	0.063
NILI	(-0.5,1.3)	Large	1,000	1:1	0.051	0.026	0.067	0.024	0.055
NILI	(-0.5,0.7)	Small	300	2:1	0.082	0.040	0.105	0.034	0.089
NILI	(-0.5,0.7)	Small	300	1:1	0.079	0.041	0.097	0.037	0.082
NILI	(-0.5,0.7)	Small	500	2:1	0.064	0.030	0.079	0.027	0.070
NILI	(-0.5,0.7)	Small	500	1:1	0.062	0.032	0.075	0.030	0.065
NILI	(-0.5,0.7)	Small	1,000	2:1	0.048	0.022	0.057	0.020	0.052
NILI	(-0.5,0.7)	Small	1,000	1:1	0.046	0.023	0.054	0.022	0.050
NILI	(-0.5,0.7)	Medium	300	2:1	0.089	0.043	0.120	0.040	0.106
NILI	(-0.5,0.7)	Medium	300	1:1	0.085	0.045	0.113	0.042	0.095
NILI	(-0.5,0.7)	Medium	500	2:1	0.070	0.034	0.093	0.031	0.083
NILI	(-0.5,0.7)	Medium	500	1:1	0.068	0.036	0.089	0.034	0.075
NILI	(-0.5,0.7)	Medium	1,000	2:1	0.049	0.023	0.065	0.022	0.057
NILI	(-0.5,0.7)	Medium	1,000	1:1	0.049	0.026	0.065	0.024	0.055
NILI	(-0.5,0.7)	Large	300	2:1	0.099	0.049	0.140	0.046	0.126
NILI	(-0.5,0.7)	Large	300	1:1	0.092	0.050	0.130	0.047	0.111
NILI	(-0.5,0.7)	Large	500	2:1	0.077	0.038	0.108	0.036	0.097
NILI	(-0.5,0.7)	Large	500	1:1	0.074	0.040	0.102	0.037	0.087
NILI	(-0.5,0.7)	Large	1,000	2:1	0.054	0.027	0.078	0.025	0.070
NILI	(-0.5,0.7)	Large	1,000	1:1	0.052	0.028	0.072	0.027	0.061

Notes. [1]Location of non-invariance (I = invariant indicator, NIL = non-invariant loading, NII = non-invariant intercept, NILI = non-invariant loading and intercept), [2]Latent variable distribution of Group 2 (mean, variance), [3]Magnitude of non-invariance, [4]Total sample size, [5]Balance of sample sizes.

Table 5

Marginal RMSE by Group 2 sample size for d_{MACS}, SDI_2, and UDI_2.

Effect size	Group 2 sample size					
	100	150	167	250	333	500
d_{MACS}	0.099	0.095	0.077	0.074	0.054	0.052
SDI_2	0.102	0.095	0.078	0.074	0.055	0.052
UDI_2	0.100	0.093	0.078	0.073	0.055	0.051

Table 6

Marginal RMSE by Group 1 sample size for WSDI and WUDI.

Effect size	Group 1 sample size					
	150	200	250	333	500	667
WSDI	0.041	0.039	0.032	0.030	0.023	0.021
WUDI	0.040	0.038	0.031	0.029	0.021	0.021

Table 7

Descriptive statistics of each effect size by location of non-invariance for the small magnitude condition.

Location	d_{MACS}				
	Min[1]	Q1[2]	Mean[3]	Q3[4]	Max[5]
I[6]	0.00	0.05	0.08	0.11	0.41
NIL[7]	0.00	0.09	0.13	0.17	0.49
NII[8]	0.00	0.17	0.21	0.25	0.52
NILI[9]	0.00	0.17	0.21	0.26	0.55
	UDI_2				
	Min	Q1	Mean	Q3	Max
I	0.00	0.04	0.07	0.09	0.38
NIL	0.00	0.08	0.12	0.15	0.48
NII	0.00	0.16	0.21	0.25	0.52
NILI	0.00	0.16	0.20	0.24	0.56
	WUDI				
	Min	Q1	Mean	Q3	Max
I	0.00	0.02	0.03	0.04	0.15
NIL	0.00	0.03	0.05	0.06	0.18
NII	0.00	0.07	0.09	0.11	0.22
NILI	0.01	0.07	0.09	0.11	0.26
	SDI_2				
	Min	Q1	Mean	Q3	Max
I	-0.38	-0.04	0.00	0.04	0.32
NIL	-0.40	-0.08	-0.04	0.01	0.28
NII	-0.14	0.16	0.20	0.25	0.52
NILI	-0.19	0.13	0.18	0.23	0.56
	WSDI				
	Min	Q1	Mean	Q3	Max
I	-0.14	-0.02	0.00	0.02	0.15
NIL	-0.14	-0.03	-0.01	0.01	0.12
NII	-0.04	0.07	0.09	0.11	0.22
NILI	-0.06	0.06	0.09	0.11	0.26

Notes. [1]Minimum value of effect size, [2]First quartile of effect size, [3]Mean value of effect size, [4]Third quartile of effect size, [5]Maximum value of effect size, [6]Indicator with an invariant loading and intercept [1]Indicator with a non-invariant intercept, [2]Magnitude of non-invariance, [3]Indicator with a non-invariant loading, [4]Latent variable distribution for Group 2, [5]Indicator with a non-invariant loading and intercept, [6]Indicator with an invariant loading and intercept, [6]Indicator with an invariant loading and intercept, [7]Indicator with a non-invariant loading, [8]Indicator with a non-invariant intercept, [9]Indicator with a non-invariant loading and intercept.

Table 8

Cohen's d values and average value of d_{MACS} for each meaningful pairwise or simple pairwise comparison.

Subgroup	Comparison Group 1	Comparison Group 2	Comparison Group 1 Mean	Comparison Group 2 Mean	Cohen's *d*
NII[1]	Small mag[2]	Medium mag	0.21	0.41	2.89
	Small mag	Large mag	0.21	0.61	5.70
	Medium mag	Large mag	0.41	0.61	2.79
NIL[3] & Small Mag	LVD2[4](-0.5,1.3)	LVD2(0,1)	0.14	0.13	0.27
	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	0.14	0.13	0.25
NIL & Medium Mag	LVD2(-0.5,1.3)	LVD2(0,1)	0.34	0.28	0.73
	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	0.34	0.28	0.70
NIL & Large Mag	LVD2(-0.5,1.3)	LVD2(0,1)	0.56	0.46	1.22
	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	0.56	0.45	1.27
NILI[5] & Small Mag	LVD2(0,1)	LVD2(-0.5,1.3)	0.24	0.20	0.60
	LVD2(0,1)	LVD2(-0.5,0.7)	0.24	0.20	0.64
NILI & Medium Mag	LVD2(0,1)	LVD2(-0.5,1.3)	0.52	0.43	1.29
	LVD2(0,1)	LVD2(-0.5,0.7)	0.52	0.39	1.75
	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	0.43	0.39	0.49
NILI & Large Mag	LVD2(0,1)	LVD2(-0.5,1.3)	0.82	0.68	1.70
	LVD2(0,1)	LVD2(-0.5,0.7)	0.82	0.61	2.67
	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	0.68	0.61	0.99
I[6]	N = 300	N = 500	0.11	0.08	0.47
	N = 300	N = 1000	0.11	0.06	1.05
	N = 500	N = 1000	0.08	0.06	0.64

Notes. Comparisons appear in the order they are discussed in the narrative. [1]Indicator with a non-invariant intercept, [2]Magnitude of non-invariance, [3]Indicator with a non-invariant loading, [4]Latent variable distribution for Group 2, [5]Indicator with a non-invariant loading and intercept, [6]Indicator with an invariant loading and intercept.

Table 9

Cohen's d values and average value of SDI_2 for each meaningful pairwise or simple pairwise comparison.

Subgroup	Comparison Group 1	Comparison Group 2	Comparison Group 1 Mean	Comparison Group 2 Mean	Cohen's *d*
NIL[1] & Small Magnitude	LVD2[2](0,1)	LVD2(-0.5,1.3)	-0.001	-0.051	0.72
	LVD2(0,1)	LVD2(-0.5,0.7)	-0.001	-0.058	0.76
NIL & Medium Magnitude	LVD2(0,1)	LVD2(-0.5,1.3)	0.000	-0.146	1.80
	LVD2(0,1)	LVD2(-0.5,0.7)	0.000	-0.167	1.93
	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	-0.146	-0.167	0.25
NIL & Large Magnitude	LVD2(0,1)	LVD2(-0.5,1.3)	-0.003	-0.268	2.70
	LVD2(0,1)	LVD2(-0.5,0.7)	-0.003	-0.291	2.86
	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	-0.268	-0.291	0.22
NII[3] & Small Magnitude	LVD2(-0.5,0.7)	LVD2(0,1)	0.225	0.200	0.35
	LVD2(-0.5,0.7)	LVD2(-0.5,1.3)	0.225	0.183	0.60
	LVD2(0,1)	LVD2(-0.5,1.3)	0.200	0.183	0.27
NII & Medium Magnitude	LVD2(-0.5,0.7)	LVD2(0,1)	0.446	0.401	0.61
	LVD2(-0.5,0.7)	LVD2(-0.5,1.3)	0.446	0.368	1.07
	LVD2(0,1)	LVD2(-0.5,1.3)	0.401	0.368	0.49
NII & Large Magnitude	LVD2(-0.5,0.7)	LVD2(0,1)	0.671	0.602	0.90
	LVD2(-0.5,0.7)	LVD2(-0.5,1.3)	0.671	0.553	1.57
	LVD2(0,1)	LVD2(-0.5,1.3)	0.602	0.553	0.70
NILI[4] & Small Magnitude	LVD2(0,1)	LVD2(-0.5,0.7)	0.217	0.179	0.49
	LVD2(0,1)	LVD2(-0.5,1.3)	0.217	0.151	0.95
	LVD2(-0.5,0.7)	LVD2(-0.5,1.3)	0.179	0.151	0.39
NILI & Medium Magnitude	LVD2(0,1)	LVD2(-0.5,0.7)	0.494	0.364	1.44
	LVD2(0,1)	LVD2(-0.5,1.3)	0.494	0.318	2.09
	LVD2(-0.5,0.7)	LVD2(-0.5,1.3)	0.364	0.318	0.53
NILI & Large Magnitude	LVD2(0,1)	LVD2(-0.5,0.7)	0.834	0.583	2.35
	LVD2(0,1)	LVD2(-0.5,1.3)	0.834	0.536	2.83
	LVD2(-0.5,0.7)	LVD2(-0.5,1.3)	0.583	0.536	0.44

Notes. Comparisons appear in the order they are discussed in the narrative. [1]Indicator with a non-invariant loading, [2]Latent variable distribution for Group 2, [3]Indicator with a non-invariant intercept, [4]Indicator with a non-invariant loading and intercept.

Table 10

Cohen's d values and average value of UDI$_2$ for each meaningful pairwise or simple pairwise comparison.

Subgroup	Comparison Group 1	Comparison Group 2	Comparison Group 1 Mean	Comparison Group 2 Mean	Cohen's *d*
I[1]	LVD2[2](-0.5,0.7)	LVD2(-0.5,1.3)	0.077	0.066	0.24
NIL[3] & Medium Magnitude	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	0.296	0.274	0.29
	LVD2(-0.5,1.3)	LVD2(0,1)	0.296	0.256	0.54
	LVD2(-0.5,0.7)	LVD2(0,1)	0.274	0.256	0.23
NIL & Large Magnitude	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	0.537	0.463	0.77
	LVD2(-0.5,1.3)	LVD2(0,1)	0.537	0.450	0.95
NII[4] & Small Magnitude	LVD2(-0.5,0.7)	LVD2(0,1)	0.227	0.203	0.37
	LVD2(-0.5,0.7)	LVD2(-0.5,1.3)	0.227	0.186	0.63
	LVD2(0,1)	LVD2(-0.5,1.3)	0.203	0.186	0.27
NII & Medium Magnitude	LVD2(-0.5,0.7)	LVD2(0,1)	0.446	0.401	0.61
	LVD2(-0.5,0.7)	LVD2(-0.5,1.3)	0.446	0.368	1.08
	LVD2(0,1)	LVD2(-0.5,1.3)	0.401	0.368	0.49
NII & Large Magnitude	LVD2(-0.5,0.7)	LVD2(0,1)	0.671	0.602	0.90
	LVD2(-0.5,0.7)	LVD2(-0.5,1.3)	0.671	0.553	1.57
	LVD2(0,1)	LVD2(-0.5,1.3)	0.602	0.553	0.70
NILI[5] & Small Magnitude	LVD2(0,1)	LVD2(-0.5,0.7)	0.229	0.200	0.44
	LVD2(0,1)	LVD2(-0.5,1.3)	0.229	0.178	0.84
	LVD2(-0.5,0.7)	LVD2(-0.5,1.3)	0.200	0.178	0.36
NILI & Medium Magnitude	LVD2(0,1)	LVD2(-0.5,0.7)	0.516	0.402	1.42
	LVD2(0,1)	LVD2(-0.5,1.3)	0.516	0.388	1.71
NILI & Large Magnitude	LVD2(0,1)	LVD2(-0.5,1.3)	0.874	0.670	2.15
	LVD2(0,1)	LVD2(-0.5,0.7)	0.874	0.649	2.31
	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	0.670	0.649	0.23

Notes. Comparisons appear in the order they are discussed in the narrative. [1]Indicator with an invariant loading and intercept, [2]Latent variable distribution for Group 2, [3]Indicator with a non-invariant loading, [4]Indicator with a non-invariant intercept, [5]Indicator with a non-invariant loading and intercept.

Table 11

Cohen's d values and average value of WSDI for each meaningful pairwise or simple pairwise comparison.

Subgroup	Comparison Group 1	Comparison Group 2	Comparison Group 1 Mean	Comparison Group 2 Mean	Cohen's *d*
NIL[1] & Small Magnitude	LVD2[2](0,1)	LVD2(-0.5,1.3)	0.000	-0.012	0.38
	LVD2(0,1)	LVD2(-0.5,0.7)	0.000	-0.015	0.47
NIL & Medium Magnitude	LVD2(0,1)	LVD2(-0.5,1.3)	0.000	-0.032	1.00
	LVD2(0,1)	LVD2(-0.5,0.7)	0.000	-0.042	1.28
	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	-0.032	-0.042	0.33
NIL & Large Magnitude	LVD2(0,1)	LVD2(-0.5,1.3)	-0.001	-0.055	1.54
	LVD2(0,1)	LVD2(-0.5,0.7)	-0.001	-0.071	1.97
	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	-0.055	-0.071	0.47
NII[3]	LVD2(0,1)	LVD2(-0.5,1.3)	0.189	0.171	0.22
	Small mag[4]	Medium mag	0.091	0.182	2.81
	Small mag	Large mag	0.091	0.273	5.30
	Medium mag	Large mag	0.182	0.273	2.56
NILI[5] & Small Magnitude	LVD2(0,1)	LVD2(-0.5,0.7)	0.098	0.081	0.65
	LVD2(0,1)	LVD2(-0.5,1.3)	0.098	0.078	0.54
NILI & Medium Magnitude	LVD2(0,1)	LVD2(-0.5,0.7)	0.211	0.161	1.40
	LVD2(0,1)	LVD2(-0.5,1.3)	0.211	0.160	1.33
NILI & Large Magnitude	LVD2(0,1)	LVD2(-0.5,0.7)	0.338	0.249	2.02
	LVD2(0,1)	LVD2(-0.5,1.3)	0.338	0.257	1.84
NII	Balanced	Unbalanced	0.192	0.172	0.25
NILI	Balanced	Unbalanced	0.195	0.168	0.29

Notes. Comparisons appear in the order they are discussed in the narrative. [1]Indicator with a non-invariant loading, [2]Latent variable distribution for Group 2, [3]Indicator with a non-invariant intercept, [4]Magnitude of non-invariance, [5]Indicator with a non-invariant loading and intercept.

Table 12

Cohen's d values and average value of WUDI for each meaningful pairwise or simple pairwise comparison.

Subgroup	Comparison Group 1	Comparison Group 2	Comparison Group 1 Mean	Comparison Group 2 Mean	Cohen's *d*
NIL[1] & Medium Magnitude	LVD2[2](-0.5,1.3)	LVD2(-0.5,0.7)	0.115	0.109	0.23
	LVD2(-0.5,1.3)	LVD2(0,1)	0.115	0.109	0.25
NIL & Large Magnitude	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	0.195	0.178	0.55
	LVD2(-0.5,1.3)	LVD2(0,1)	0.195	0.182	0.44
NII[3]	Small magnitude	Medium magnitude	0.093	0.183	2.83
	Small magnitude	Large magnitude	0.093	0.274	5.35
	Medium magnitude	Large magnitude	0.183	0.274	2.56
	LVD2(0,1)	LVD2(-0.5,1.3)	0.189	0.173	0.21
NILI[4] & Small Magnitude	LVD2(0,1)	LVD2(-0.5,0.7)	0.104	0.090	0.45
	LVD2(0,1)	LVD2(-0.5,1.3)	0.104	0.087	0.58
NILI & Medium Magnitude	LVD2(0,1)	LVD2(-0.5,1.3)	0.220	0.183	1.12
	LVD2(0,1)	LVD2(-0.5,0.7)	0.220	1.820	0.11
NILI & Large Magnitude	LVD2(0,1)	LVD2(-0.5,1.3)	0.355	0.299	1.32
	LVD2(0,1)	LVD2(-0.5,0.7)	0.355	0.288	1.51
	LVD2(-0.5,1.3)	LVD2(-0.5,0.7)	0.299	0.288	0.24
NIL	Small magnitude	Medium magnitude	0.050	0.111	2.37
	Small magnitude	Large magnitude	0.050	0.185	4.79
	Medium magnitude	Large magnitude	0.111	0.185	2.42
NII & Small Magnitude	Balanced	Unbalanced	0.099	0.088	0.36
NII & Medium Magnitude	Balanced	Unbalanced	0.193	0.172	0.63
NII & Large Magnitude	Balanced	Unbalanced	0.289	0.259	0.88
NILI & Small Magnitude	Balanced	Unbalanced	0.100	0.087	0.47
NILI & Medium Magnitude	Balanced	Unbalanced	0.210	0.180	0.85
NILI & Large Magnitude	Balanced	Unbalanced	0.340	0.288	1.12

Notes. Comparisons appear in the order they are discussed in the narrative. [1]Indicator with a non-invariant loading, [2]Latent variable distribution for Group 2, [3]Indicator with a non-invariant intercept, [4]Indicator with a non-invariant loading and intercept.

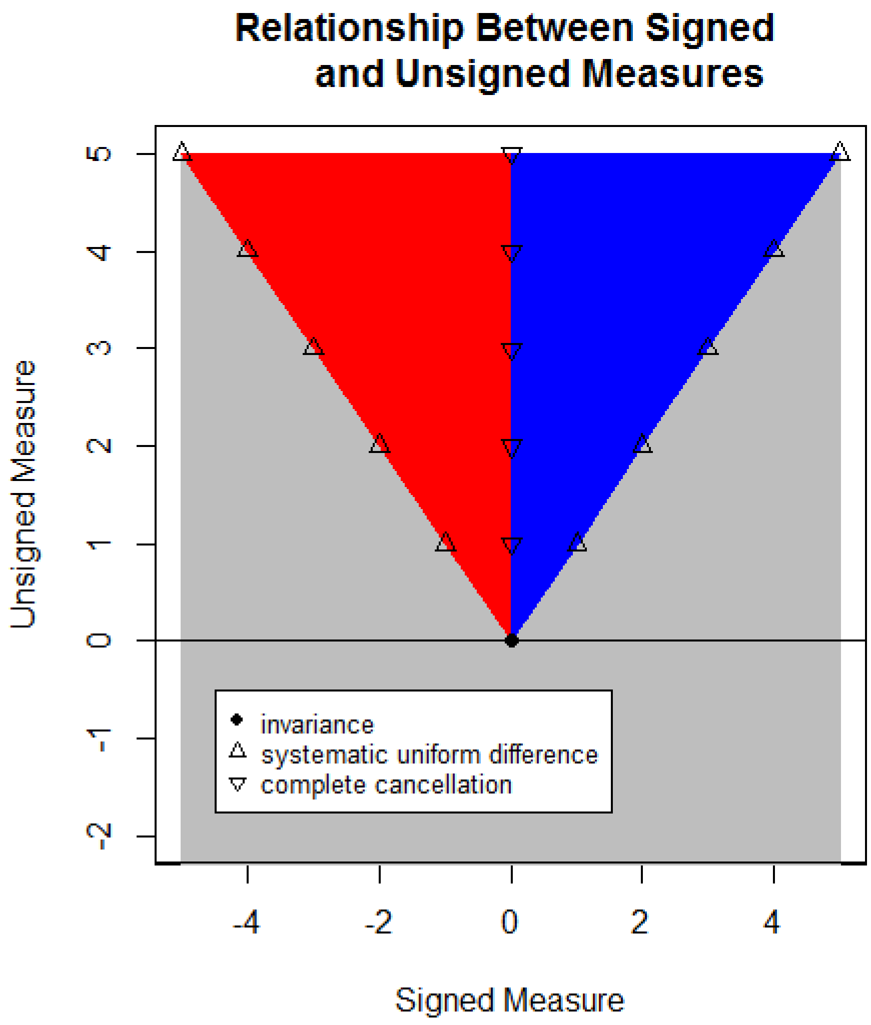

Figure 1. Illustration of possible values for signed and unsigned effect size measures. The grey area represents impossible values. The red area denotes possibilities where Group 2 has higher expected indicator scores, on average. The blue area denotes possibilities where Group 1 has higher expected indicator scores, on average.

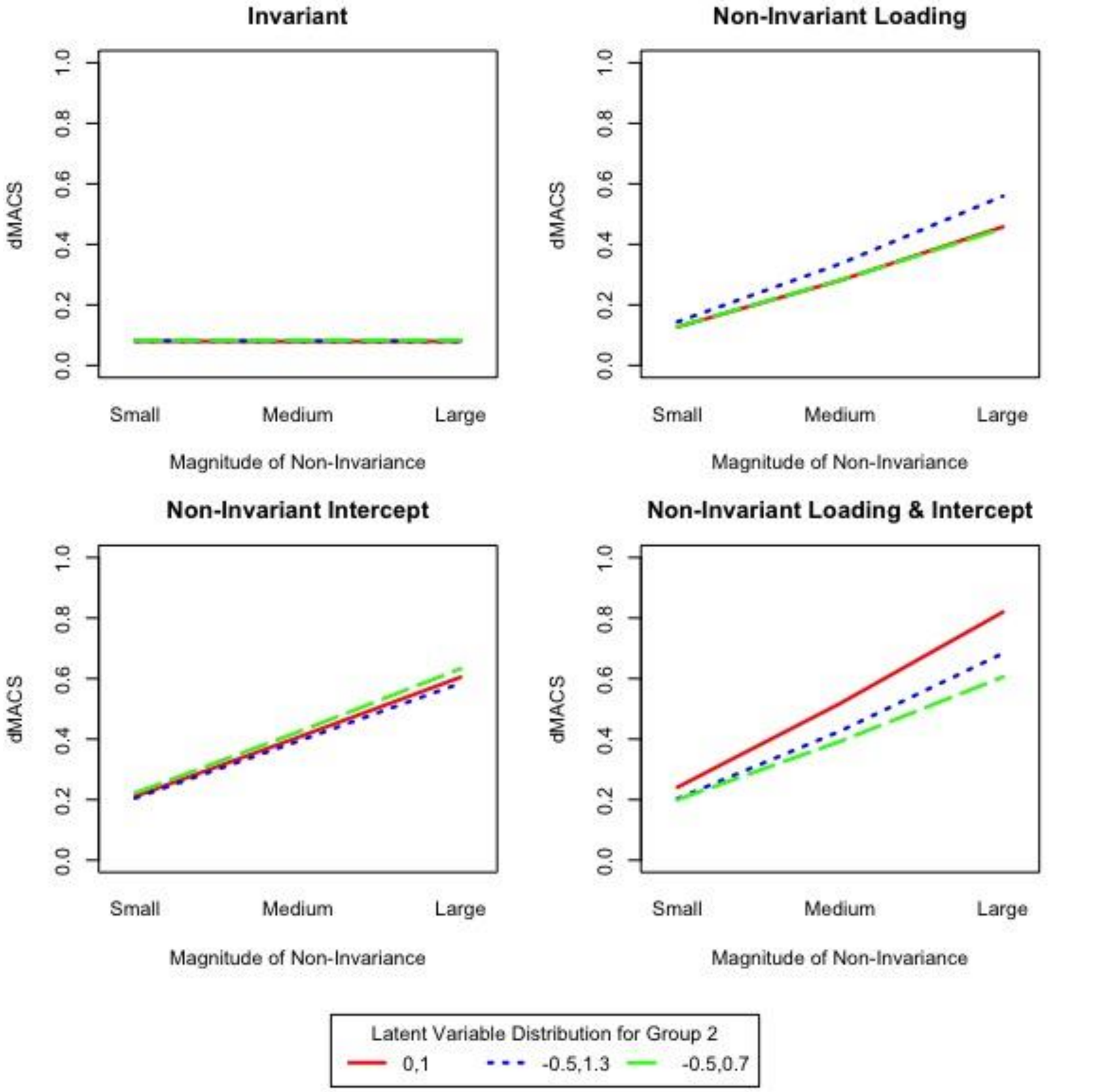

Figure 2. Simple two-way interactions of magnitude × latent variable distribution for Group 2 by location of non-invariance on value of d_{MACS}. The average values of d_{MACS} by condition are plotted.

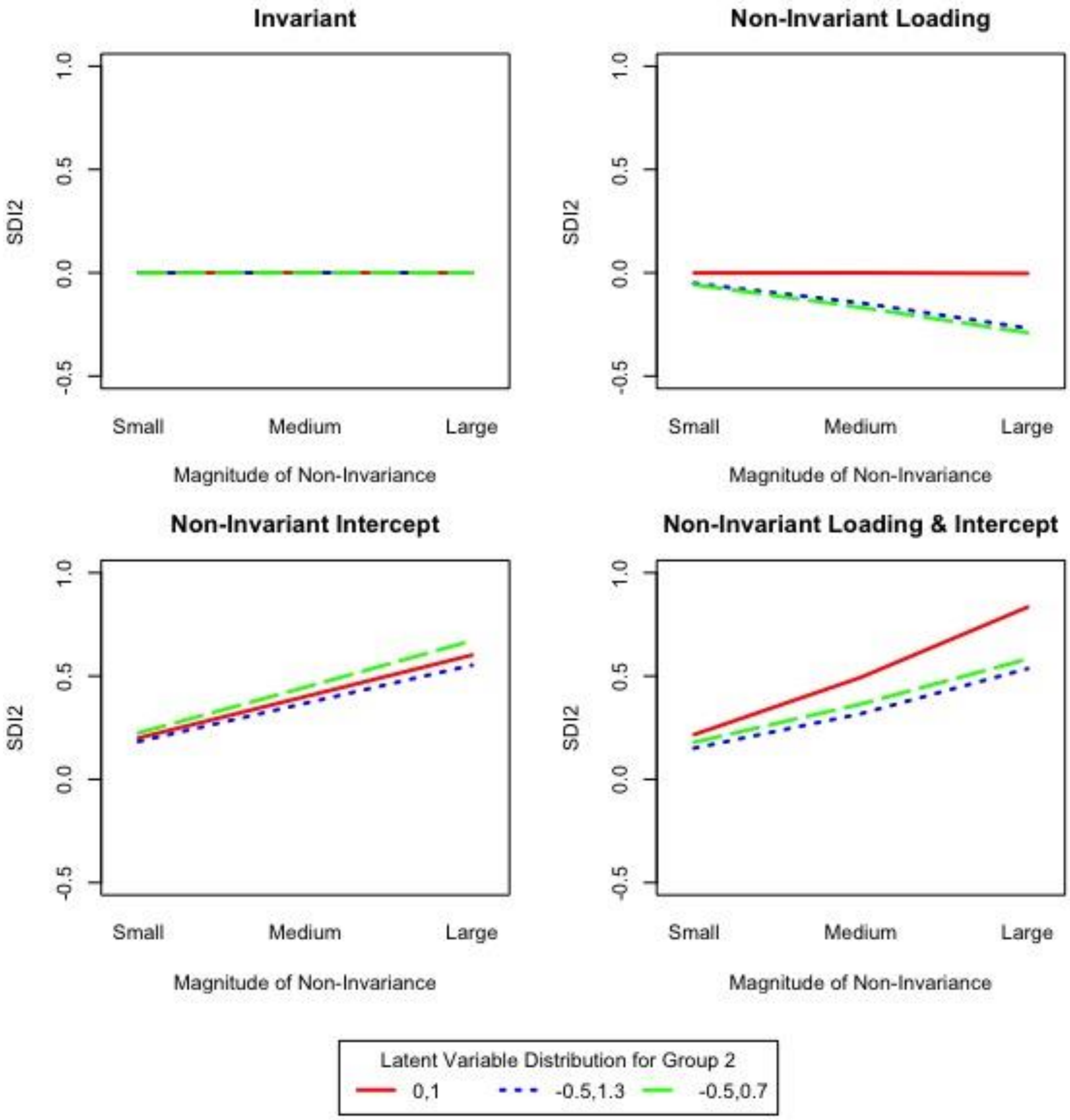

Figure 3. Simple two-way interactions of magnitude × latent variable distribution for Group 2 by location of non-invariance on value of SDI_2. The average values of SDI_2 by condition are plotted.

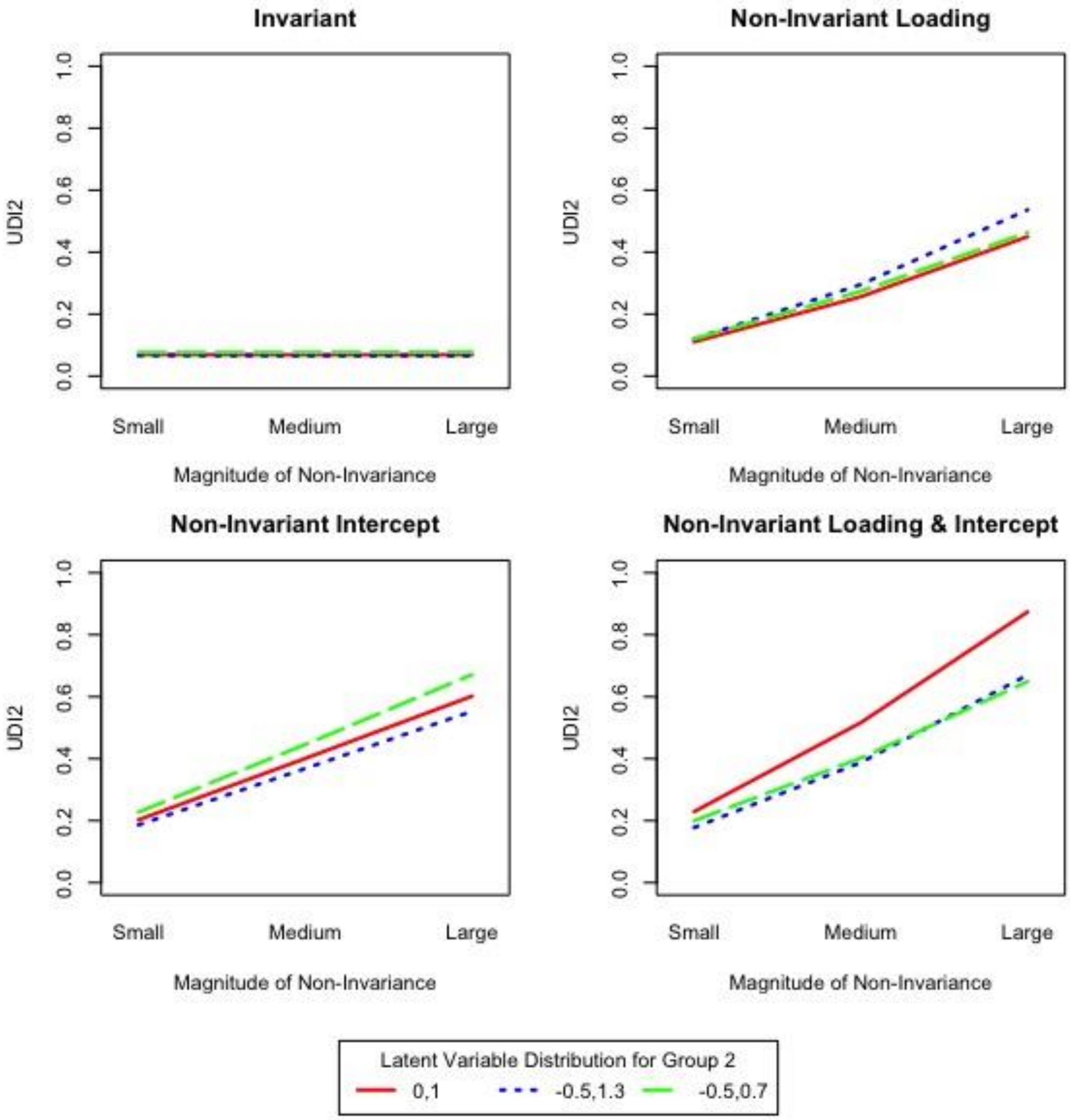

Figure 4. Simple two-way interactions of magnitude × latent variable distribution for Group 2 by location of non-invariance on value of UDI_2. The average values of UDI_2 by condition are plotted.

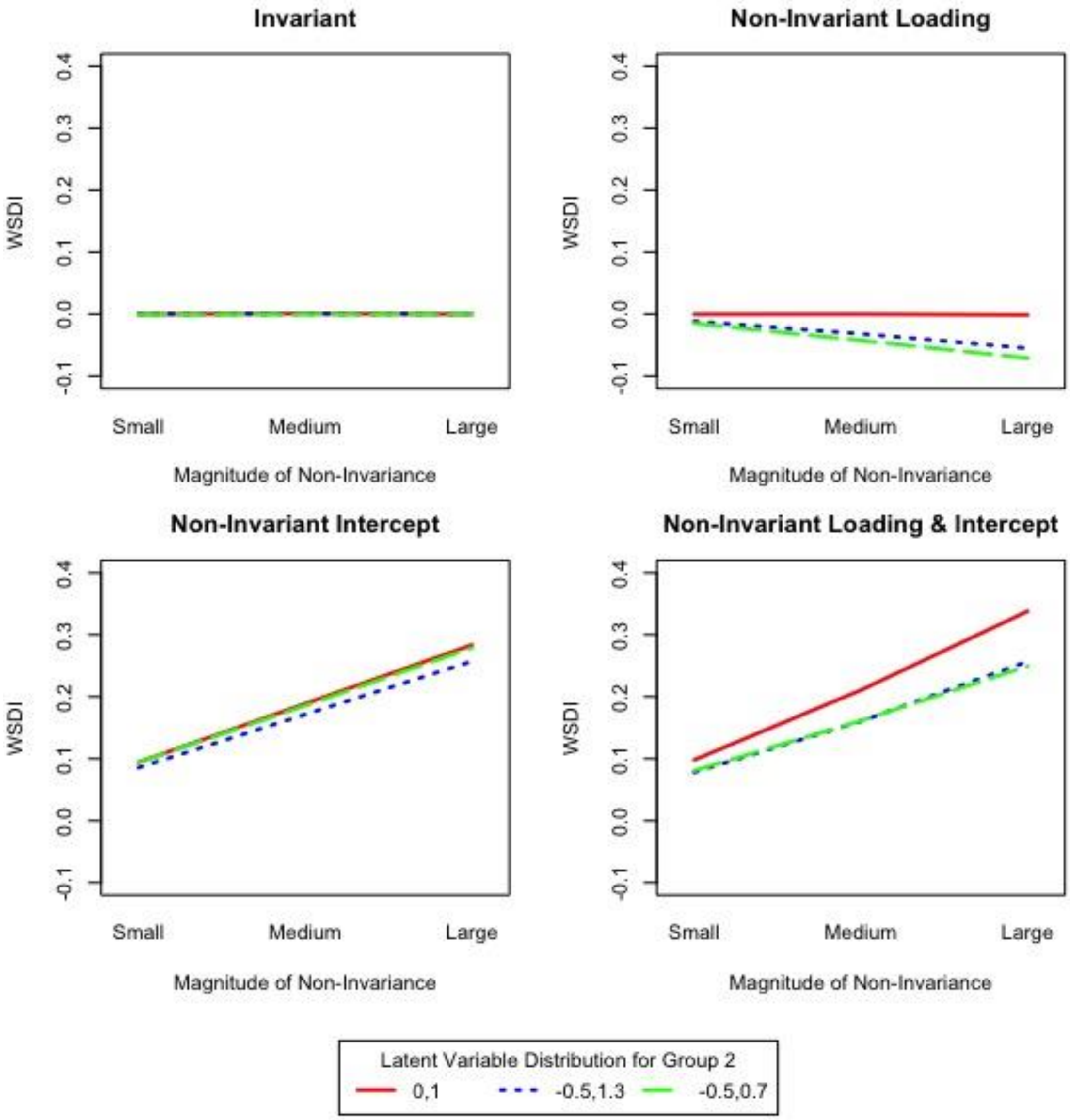

Figure 5. Simple two-way interactions of magnitude × latent variable distribution for Group 2 by location of non-invariance on value of *WSDI*. The average values of *WSDI* by condition are plotted.

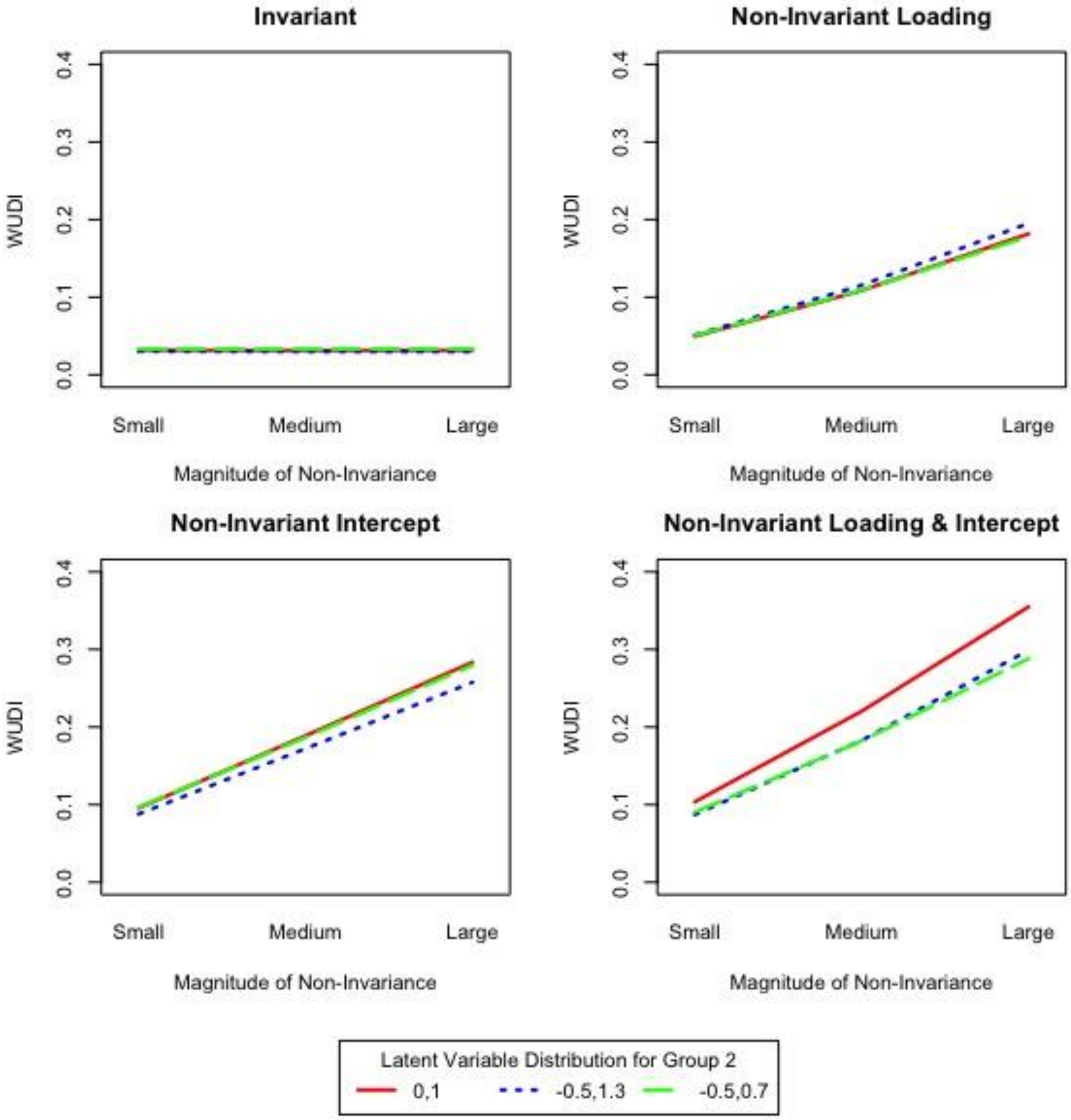

Figure 6. Simple two-way interactions of magnitude × latent variable distribution for Group 2 by location of non-invariance on value of *WUDI*. The average values of *WUDI* by condition are plotted.

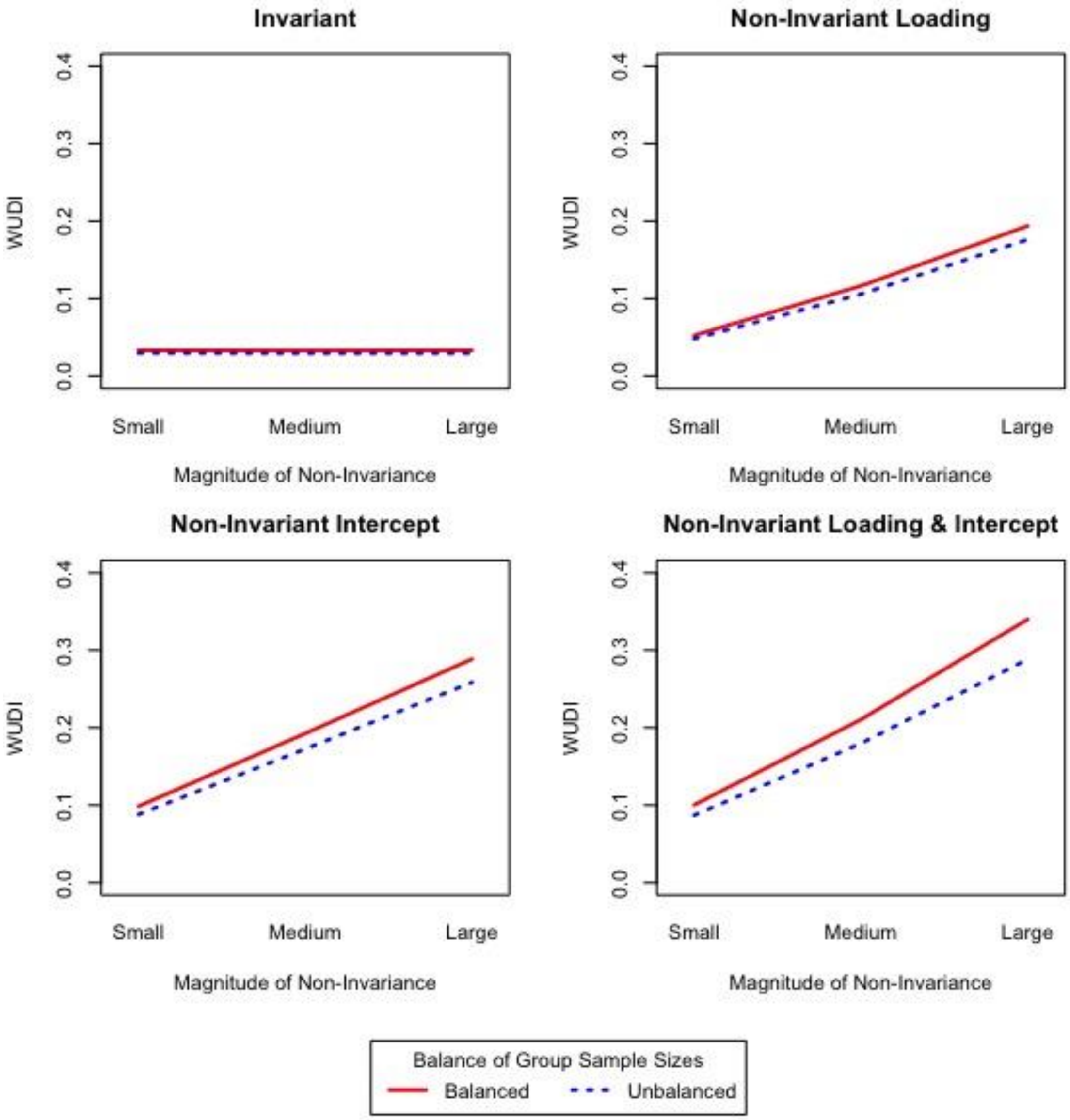

Figure 7. Simple two-way interactions of magnitude × balance of group sample sizes by location of non-invariance on value of *WUDI*. The average values of *WUDI* by condition are plotted.