STRUCTURAL EQUATION MODELING
2026, VOL. 33, NO. 1, 81–88
ISSN: 1070-5511 | E-ISSN: 1532-8007

OPEN ACCESS

Quantifying Measurement Non-Invariance Beyond Simple Structure: The Closed Formulas of Universal Effect Size Measures for MI

Tizian Maria Benedikt Schuhbeck[a], Philipp Sterner[a,b,c], and David Goretzko[a,d]

[a]Ludwig-Maximilians-Universität München; [b]Ruhr-Universität Bochum; [c]Deutsches Zentrum für Psychische Gesundheit; [d]Utrecht University

ABSTRACT
Investigating measurement invariance is an essential prerequisite for meaningful evaluations of psychological questionnaires. Because the widely-used approach of treating MI as a binary decision does not enable inferences about the practical implications of non-invariance, effect sizes for MI – most notably d^{MACS} – were proposed. So far, such effect sizes have only been defined for models with simple structure, which consequently limits their applicability. In this paper, we extend d^{MACS} and subsequent effect sizes of measurement nonequivalence to general factor analytic models, rendering them universally applicable. We provide corresponding implementations with optimal computational complexity.

KEYWORDS
Differential item functioning; dMACS; effect size; factor analysis; measurement invariance

1. Introduction

Psychological research commonly revolves around inferences of latent variables, such as motives, attitudes, or character traits. Such variables are unobservable and need to be measured indirectly. This typically involves measurement models constructed from data of observable indicators, usually given by the responses to a survey or test item. For any meaningful comparisons of latent variables across groups or time, *measurement invariance* (MI) is often seen as a prerequisite (e.g., Meredith, 1993; Putnick & Bornstein, 2016; Van De Schoot et al., 2012). This follows the idea that values obtained from measures can only be assumed to be readily comparable, if the corresponding measures function identically across all groups of interest.[1] Consequently, several methods of assessing MI to explicitly detect sources of non-invariance and item biases have been developed. The most widely established method – *multi-group confirmatory factor analysis* (MG-CFA) – enables researchers to evaluate levels of MI by differentiating between several types of invariance across groups (e.g., Van De Schoot et al., 2012; Vandenberg & Lance, 2000): *configural invariance* (identical measurement model structure & applicability of the same constraints), *metric invariance* (identical loading parameters), *scalar invariance* (identical intercept parameters), and *strict invariance* (identical residual variances). This method, however, is not without shortcomings: MG-CFA is not equipped to handle continuous covariates and only has limited capacities when dealing with many groups. Most notably, it is restricted to models of confirmatory factor analysis. To address these limitations, numerous alternatives have been developed (e.g., Asparouhov & Muthén, 2023; Bauer, 2017; De Roover, 2021; Goretzko & Sterner, 2024; Kim et al., 2016; Muthén & Asparouhov, 2012; Sterner & Goretzko, 2023). Kim et al. (2017) provides an overview of modern approaches to test for MI in the setting of CFA, while Sterner et al. (2024a) illustrate newly developed methods that are built on *exploratory factor analysis* (EFA).

Although there are some exceptions [such as alignment optimization, in which *measurement non-invariance* (MNE) is considered to be a minimizable quantity, see also Luong and Flake (2023)], the majority of methods treat MI testing as a binary decision – *either* MI holds *or* a scale is determined to be non-invariant. In particular, this is the case for MG-CFA, in which the use of χ^2 significance tests and binary threshold rules, comparing model fit indices across models with varying parameter constraints, are considered good practice (e.g., Putnick & Bornstein, 2016; Van De Schoot et al., 2012; Vandenberg & Lance, 2000). The main issue with this approach is that significance tests as well as model fit indices depend on sample size (Goretzko et al., 2024) – as with other hypothesis tests, larger samples lead to greater statistical power, which will eventually lead to a significant result, in this case indicating the violation of MI.[2] Thus, focusing solely on statistical significance is not sufficient for a meaningful evaluation of MNE. To be able to determine whether a detected violation of MI does in fact constitute a severe bias for subsequent statistical analyses, or the amount of MNE can be seen as negligible, researchers need to be able to quantify MI. For this reason, MI effect size measures have been proposed.

1.1. Effect Size Measures for Measurement Nonequivalence

Most approaches to quantifying the amount of MNE are inspired by common effect size measures that are used to quantify mean differences between two or more groups (e.g., in the context of t-tests or ANOVAs). Arguably the most popular effect size measure in psychology is *Cohen's d* (Cohen, 2013) which standardizes the mean difference between two independent groups by dividing it by the pooled standard deviation:

$$d = \frac{\bar{y}_1 - \bar{y}_2}{\mathrm{SD}_{\mathrm{pooled}}}.$$

This idea was translated to the context of MI by Nye and Drasgow (2011) introducing d^{MACS}, where *MACS* refers to the *mean and covariance structure*. Gunn et al. (2020) introduced different variants of d^{MACS} that (dependent on the specific empirical data set and researcher question) can have some advantages over the original approach. Recently, with the development of f^{MACS} – an analogue to *Cohen's f* – the concept of d^{MACS} has also been extended to the multigroup case (Lai et al., 2025). Although these effect sizes cover different contexts and provide standardized values for group differences in the measurement model, they are only applicable to unidimensional models as well as factor models with perfect simple structure (i.e., models without cross-loadings). This is highly restrictive, especially since many EFA-based tools for investigating MI have been developed recently (Sterner et al., 2024a) and measurement models assuming perfect simple structure usually do not fit the data too well (Goretzko et al., 2024). When applying EFA-based methods such as mixture multigroup factor analysis (De Roover et al., 2022) or EFA trees (Sterner & Goretzko, 2023), or allowing cross-loadings in a multidimensional factor model, d^{MACS} (in its current form and implementation) as well as its alternatives are not applicable.

1.2. Aim of This Paper

In this paper, we generalize all existing effect size measures for MNE to multi-factorial models with arbitrary loading patterns (i.e., allowing cross-loadings and factor-correlations) by extending the defining integrals to their multivariate analogue. Since the predominant approach of implementing effect sizes of MNE via numerical approximation (Gunn et al., 2020; Nye & Drasgow, 2011) would result in the extended effect sizes having computational complexity *exponential* in the number of factors – this is especially problematic when several groups are compared and d^{MACS} (or an alternative) has to be calculated multiple times – we provide a general method to obtain closed formulas with optimal computational complexity (i.e., linear in the number of factors). This method can be applied to all generalized effect size measures that rely on the principle of standardized mean differences, most notably d^{MACS}. Corresponding implementations in R have been made publicly available on https://github.com/TiziSchuh/ExpectedDifferenceMeasures.

2. Preliminaries

In order to analyze and quantify MNE, a formal notion of group-specific measurement models is required. We begin by introducing all factor analytic notions necessary for the rigorous formulation of mean predicted response models – the core concept of factor analysis – followed by a review of d^{MACS} – the effect size measure for MNE that lays the foundation of this paper.

2.1. The Common Factor Model

We refer to Mulaik (2009) for a more detailed introduction to the key concepts of factor analysis. Consider a random vector of p observed variables $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$ (*indicators*) representing responses to a given questionnaire consisting of p items. The core assumption of the *common factor model* is that the observations $\boldsymbol{Y}$ can be causally related to a fewer number of q unobserved variables in terms of a linear relationship. In more precise terms, consider a random vector of continuous latent variables $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)^\top$ (the common *factors*) and a linear model H with parameters $\boldsymbol{\tau} \in \mathbb{R}^p$ (*intercepts*) and $\Lambda \in \mathbb{R}^{p\times q}$ (*factor loadings*). Then the common factor model can be formulated as

$$\begin{aligned} \boldsymbol{Y} &= H(\boldsymbol{\eta}) + \boldsymbol{\epsilon} \\ &= \boldsymbol{\tau} + \Lambda\boldsymbol{\eta} + \boldsymbol{\epsilon}, \end{aligned}$$

where $\boldsymbol{\epsilon}$ describes a normally distributed error term with mean vector $\boldsymbol{0} \in \mathbb{R}^p$ and covariance matrix $\Psi = \mathrm{Diag}(\mathrm{Var}(Y_1), \ldots, \mathrm{Var}(Y_p)) \in \mathbb{R}^{p\times p}$. The common factor model forms the basis of *confirmatory factor analysis* (CFA), which aims to confirm the validity of such a model with fixed parameters and *exploratory factor analysis* (EFA), which aims to find the smallest number of factors that allow model parameters sufficiently fitting the observations in the above relation. Given specific model parameters $\boldsymbol{\tau}$ and Λ, a measurement model for item j called the *mean predicted response* is then defined as

$$\begin{aligned} \hat{Y}_j(\boldsymbol{\eta}) &= \mathbb{E}\left[Y_j|\boldsymbol{\eta}\right] \\ &= \boldsymbol{\tau}_j + {\Lambda_j}^\top\boldsymbol{\eta}, \end{aligned}$$

where Λ_j denotes the j-th row of Λ. In more concrete terms, such a model projects an individual with latent scores $\boldsymbol{s} = (s_1, \ldots, s_q) \in \mathbb{R}^q$ to respond to item j on average with the value given by

$$\hat{Y}_j(\boldsymbol{s}) = \boldsymbol{\tau}_j + {\Lambda_j}^\top\boldsymbol{s}.$$

Throughout the literature, it is common to drop the dependency on $\boldsymbol{\eta}$ in the notation and only use $\hat{Y}_j$ to denote the mean expected response.

Multi-group factor analysis considers partitions of the observed data into groups $g \in \mathcal{G}$, such as “male” and “female”, and allocates a distinct mean predicted response model to each specific group. The group-specific measurement model for group g is denoted

$$\hat{Y}_{jg} = \boldsymbol{\tau}_{jg} + {\Lambda_{jg}}^{\top}\boldsymbol{\eta}$$

where $\boldsymbol{\tau}_{jg}$ and Λ_{jg} denote the corresponding group-specific intercepts and factor loadings.

2.2. Effect Size Measures for Measurement Non-Invariance

Because measurement models in FA are defined on an item level, effect sizes of measurement non-invariance are defined on an item level as well. We note, that using the collection of all effect sizes obtained from the p items, usually referred to as the corresponding effect size index Nye and Drasgow (2011), statements about the questionnaire as a whole can be made. Given an item j and groups to be compared, the general approach to quantifying measurement non-invariance involves calculating expressions of the form

$$\text{effect size} = \frac{\text{model discrepancy}}{\text{normalization term}}.$$

Consequently, constructing such an effect size measure equates to finding a suitable functional $\mathcal{D} : \{\text{models } \hat{Y} : \mathbb{R}^q \to \mathbb{R}\}^2 \to \mathbb{R}$ such that evaluating $\mathcal{D}(\hat{Y}_{jg}, \hat{Y}_{jk})$ reflects the measurement model discrepancy when comparing groups g and k.

2.3. The Effect Size Index d^{MACS}

To the best of our knowledge, the effect sizes of measurement non-invariance introduced for common factor models thus far have only been explicitly designed for models with simple structure, i.e., with items only loading onto a single factor each (Gunn et al., 2020; Lai et al., 2025; Nye & Drasgow, 2011). The groundwork was laid by Nye and Drasgow (2011), who propose quantifying the discrepancy with a weighted *average difference* of predicted responses realized by the integral

$$\mathcal{D}(\hat{Y}_{jr}, \hat{Y}_{jg}) = \sqrt{\int (\hat{Y}_{jr} - \hat{Y}_{jg}|\eta)^2 f(\eta)\mathrm{d}\eta}. \qquad (1)$$

where the group r (*reference group*) is meant to contain the majority of the sample and g is the group of interest (*focal group*) that may possibly violate MI. As will be discussed in more detail in the following section, differences are weighted by the probability density function f of the latent factor, which is assumed to be normally distributed with *mean and covariance structure* (MACS) estimated from the focal group g. Normalizing the above expression using the pooled standard derivation results in the definition

$$d_j^{\text{MACS}} = \frac{1}{\text{SD}_{j\text{pooled}}}\sqrt{\int (\hat{Y}_{jr} - \hat{Y}_{jg}|\eta)^2 f(\eta)\mathrm{d}\eta}.$$

Note that unlike traditional effect sizes, d^{MACS} will always be non-negative. Thus, the authors suggest the complementary use of the following signed variant

$$d_j^{\text{MACS\_Signed}} = \frac{1}{\text{SD}_{j\text{pooled}}} \int (\hat{Y}_{jr} - \hat{Y}_{jg}|\eta) f(\eta)\mathrm{d}\eta,$$

which encodes 1) the *direction* of the effect, making it possible to determine which of the two groups predicts higher values on average, as well as 2) the *amount of cancelation* that occurs globally, enabling statements about the actual impact on the observed mean difference (Nye et al., 2019). Nye and Drasgow (2011) themselves warn that a “more complicated formula” is needed to assess items with multiple factor loadings, giving rise to the central question of this paper:

Is There a Natural Extension of d^{MACS} for Multiple-Factor Models?

This involves addressing two distinct issues. Firstly, does the addition of further factors in 2.3 yield a sensible multivariate analogue – and still so when allowing the factors to be correlated? Secondly, numerical approximation of the integral – the approach to calculating d^{MACS} for one-factor models (Nye & Drasgow, 2011) – will be exponentially complex in the number of factors q, so does this render the use of any possible extension impracticable for applications such as EFA trees (Sterner, 2025) which require a significant number of consecutive evaluations?

2.4. This Work

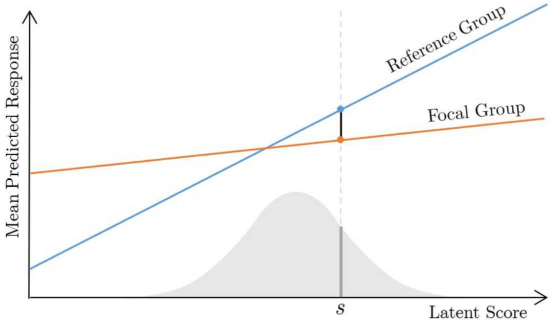
In the following, we will resolve the former issue by restating 2.3 being careful to use unambiguous notation, which will enable a canonical generalization to general common factor models. We present resulting extended versions for all SotA proposals written in terms of expected model differences, which will make respective correspondences to Cohen’s d, Glass’ Δ and Cohen’s f visually apparent, increasing accessibility and interpretability. To address the latter issue of computability in the most general setting, we introduce the class of *expected difference measures* that includes d^{MACS} and all other SotA extensions. We then show that any measure of this class has a closed formula with optimal computational complexity $\mathcal{O}(q)$.

3. Extending d^{MACS} to Multiple-Factor Models

In the following, we explicitly extend the definition of d^{MACS} to general common factor models with an arbitrary number of latent factors, allowing cross-loadings and factor correlations.

3.1. A Change of Notation

We begin by reformulating 2.3 to clarify the role of the latent factor η within the expression, as the notation dη (meant to signal an average over all possible values of η) makes no distinction between *values* of the latent variable and the latent variable *itself*. For this, note that the concept of averaging model predictions over all possible latent scores can be captured more precisely using a distinct variable s for the values (*scores*) of the latent variable η.



Hence the model discrepancy as described in 2.3 can be disambiguated to read

$$\int_{\mathbb{R}} \hat{Y}_{jr}(s) - \hat{Y}_{jg}(s)|^2 f_{\eta g}(s)\mathrm{d}s,$$

where $f_{\eta g}$ – in line with the convention of writing $\hat{Y}_{jg}$ – denotes the probability distribution function of η with parameters (i.e. mean and variance) estimated from group g and consequently

$$d_j^{\text{MACS}} = \frac{\sqrt{\mathbb{E}_{\eta g}\left[\left|\hat{Y}_{jr} - \hat{Y}_{jg}\right|^2\right]}}{\text{SD}_{j\text{pooled}}}, \tag{2}$$

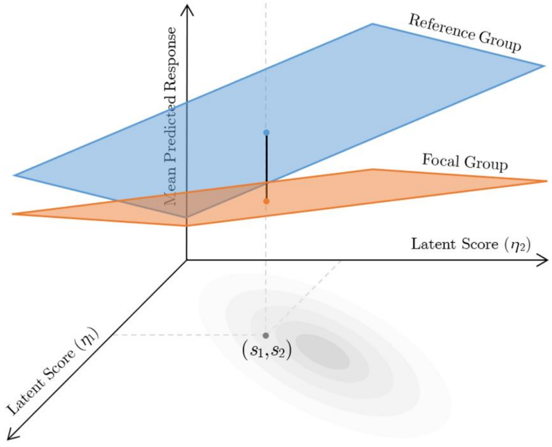
where $\mathbb{E}_{\eta g}$ is meant to denote expectation w.r.t the normally distributed factor η with mean and variance estimated from group g.

3.2. Extension via Multivariate Integration

The above reformulation of d^{MACS} now allows for straightforward transitioning to multiple latent factors, by letting the average range over all possible *score combinations* $\boldsymbol{s} = (s_1, \ldots, s_q) \in \mathbb{R}^q$

$$\left(\int_{\mathbb{R}^q} \left|\hat{Y}_{jr}(s) - \hat{Y}_{jg}(s)\right|^2 f_{\eta g}(s)\mathrm{d}s\right)^{1/2}$$

weighted according to the distribution function $f_{\boldsymbol{\eta} g}$ of the latent vector $\boldsymbol{\eta}$ with mean vector and covariance matrix estimated from group g, illustrated below for $q = 2$ factors.



Consequently, 3.1 can naturally be generalized as

$$d_j^{\text{MACS}} = \frac{\sqrt{\mathbb{E}_{\boldsymbol{\eta} g}\left[\left|\hat{Y}_{jr} - \hat{Y}_{jg}\right|^2\right]}}{\text{SD}_{j\text{pooled}}},$$

describing the desired extension of d^{MACS} to arbitrary common models with possible cross-loadings. Note, that factor correlations will be reflected in the distribution function $f_{\boldsymbol{\eta} g}$ and pose no further problems, so this definition is in fact well-defined for *all* possible common factor models rendering it theoretically suitable for any setting of EFA and CFA.

4. Expected Difference Measures

The above extension can be applied to all effect size measures of measurement non-invariance analogously. We provide an overview of the resulting extensions, again using the notation

$$\mathbb{E}_{\boldsymbol{\eta} g}[X] = \int_{\mathbb{R}^q} X(\boldsymbol{s}) f_{\boldsymbol{\eta} g}(\boldsymbol{s})\mathrm{d}\boldsymbol{s},$$

where $f_{\boldsymbol{\eta} g}$ denotes the probability density function of the normally distributed latent factors $\boldsymbol{\eta}$ with mean $\boldsymbol{\kappa}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$ estimated from group g. We will present these extensions grouped according to their corresponding "standard effect size" analogue, which will become visually more apparent. We proceed by introducing a notion of the general class of such effect size measures – the class of *expected difference measures*.

4.1. Analogue to Cohen's d

As the name implies, the pair of effect size measures d^{MACS} & $d^{\text{MACS\_Signed}}$ proposed by Nye and Drasgow (2011) with extended forms

$$d_j^{\text{MACS}} = \frac{\sqrt{\mathbb{E}_{\boldsymbol{\eta} g}\left[\left|\hat{Y}_{jr} - \hat{Y}_{jg}\right|^2\right]}}{\text{SD}_{j\text{pooled}}} \quad \text{and respectively} \quad d_j^{\text{MACS\_Signed}} = \frac{\mathbb{E}_{\boldsymbol{\eta} g}\left[\hat{Y}_{jr} - \hat{Y}_{jg}\right]}{\text{SD}_{j\text{pooled}}}$$

is based on Cohen's d_s (also known as Hedges g) which is defined as

$$d_s = \frac{\bar{y}_r - \bar{y}_g}{\text{SD}_{\text{pooled}}}.$$

4.2. Analogue to Glass' Δ

Nye and Drasgow (2011) note that the standard deviation of the reference group may be used instead. We refer to the corresponding effect size measures as d^{MACS} & $d^{\text{MACS\_Signed}}$, defined

$$\text{UDI}_j = \frac{\mathbb{E}_{\boldsymbol{\eta} g}\left[\left|\hat{Y}_{jr} - \hat{Y}_{jg}\right|^2\right]}{\text{SD}_{jr}} \quad \text{and respectively} \quad \text{SDI}_j = \frac{\mathbb{E}_{\boldsymbol{\eta} g}\left[\hat{Y}_{jr} - \hat{Y}_{jg}\right]}{\text{SD}_{jr}},$$

as they encode the closely related concept of Glass' Δ

$$\Delta = \frac{\bar{y}_r - \bar{y}_g}{\mathrm{SD}_r}.$$

The so-called *unsigned* & *signed differences in expected indicator score* (UDI & SDI) proposed by Gunn et al. (2020) are defined similarly, using the standard deviation from the focal group for normalization

$$\mathrm{UDI}_j = \frac{\mathbb{E}_{\boldsymbol{\eta}g}\left[\left|\hat{Y}_{jr} - \hat{Y}_{jg}\right|\right]}{\mathrm{SD}_{jg}} \text{ and respectively } \mathrm{SDI}_j = \frac{\mathbb{E}_{\boldsymbol{\eta}g}\left[\hat{Y}_{jr} - \hat{Y}_{jg}\right]}{\mathrm{SD}_{jg}}.$$

One subtle difference is the use of absolute differences for the unsigned version. which is – as Gunn et al. (2020) argue – a better approach compared to using squared differences, as it results in both versions operating on the same scale.

4.3. WUDI & WSDI

Based on UDI & SDI, Gunn et al. (2020) further propose an effect size for measurement invariance for two groups (1 and 2) that has the advantage of circumventing the notion of a focal and reference group entirely, by first calculating the respective discrepancies relative to the corresponding single group model $\hat{Y}_j$ and then adding the results weighted by the sample size proportions $p_1 = \frac{n_1}{n_1+n_2}$ and $p_2 = \frac{n_2}{n_1+n_2}$, yielding the effect size measure

$$\mathrm{WUDI}_j = p_1 \frac{\mathbb{E}_{\boldsymbol{\eta}1}\left[\left|\hat{Y}_{j1} - \hat{Y}_j\right|\right]}{\mathrm{SD}_{j1}} + p_2 \frac{\mathbb{E}_{\boldsymbol{\eta}2}\left[\left|\hat{Y}_j - \hat{Y}_{j2}\right|\right]}{\mathrm{SD}_{j2}}$$

and respectively

$$\mathrm{WSDI}_j = p_1 \frac{\mathbb{E}_{\boldsymbol{\eta}1}\left[\hat{Y}_{j1} - \hat{Y}_j\right]}{\mathrm{SD}_{j1}} + p_2 \frac{\mathbb{E}_{\boldsymbol{\eta}2}\left[\hat{Y}_j - \hat{Y}_{j2}\right]}{\mathrm{SD}_{j2}}.$$

4.4. Analogue to Cohen's f

Recently Lai et al. (2025) proposed f^{MACS} which has the extended form

$$f_j^{\mathrm{MACS}} = \frac{\sqrt{\frac{1}{G}\sum_{g=1}^{G} p_g \mathbb{E}_{\boldsymbol{\eta}g}\left[\left|\hat{Y}_{jg} - \overline{\hat{Y}}_j\right|^2\right]}}{\mathrm{SD}_j},$$

comparing multiple groups to the *grand mean* $\overline{\hat{Y}}_j = \left(\sum_g p_g \boldsymbol{\tau}_{jg}\right) + \left(\sum_g p_g \boldsymbol{\Lambda}_{jg}\right)\boldsymbol{\eta}$ with group sample proportions $p_g = \frac{n_g}{n_1+\cdots+n_G}$. As the name implies, this mirrors Cohen's f

$$f = \frac{\sqrt{\frac{1}{G}\sum_{g=1}^{G} p_g \left(\bar{y}_g - \bar{y}\right)^2}}{\mathrm{SD}}.$$

Analogously, their general version allowing groups to be weighted has the extended form

$$f_j^{\mathrm{MACS}} = \frac{\sqrt{\frac{1}{G}\sum_{g=1}^{G} w_g \mathbb{E}_{\boldsymbol{\eta}g}\left[\left|\hat{Y}_{jg} - \overline{\hat{Y}}_j\right|^2\right]}}{\mathrm{SD}_j},$$

for non-negative weights $w_g \geq 0$ with $w_1 + \ldots + w_G = 1$. If d^{MACS} is calculated using the item's (rather than the pooled) standard deviation, the relation $2f_j^{\mathrm{MACS}} = d_j^{\mathrm{MACS}}$ (Lai et al., 2025) for two groups r and g still holds in the generalized case and under the same prerequisites, most notably for balanced group sizes $n_r = n_g$ with identical latent distributions $f_{\boldsymbol{\eta}r} = f_{\boldsymbol{\eta}g}$.

4.5. The Class of Expected Difference Measures

The common theme of the above effect size measures are expected measurement model differences. As we will show in the next section, the expected value of common factor model differences can be calculated analytically. To ensure maximal generality of our results, we introduce the class of *expected difference measures* (EDMs) consisting of all effect size measures of measurement non-invariance that evaluate model discrepancies using algebraic expressions of the following terms

$$\mathbb{E}_{\boldsymbol{\eta}h}\left[\hat{Y}_{jg} - \hat{Y}_{jk}\right] \tag{3a}$$

$$\mathbb{E}_{\boldsymbol{\eta}h}\left[\left|\hat{Y}_{jg} - \hat{Y}_{jk}\right|\right] \tag{3b}$$

$$\mathbb{E}_{\boldsymbol{\eta}h}\left[\left|\hat{Y}_{jg} - \hat{Y}_{jk}\right|^2\right], \tag{3c}$$

with choices of groups g, k and h allowed to be different for each term. Consistent with the literature thus far (Gunn et al., 2020; Nye et al., 2019), we refer to EDMs constructed using only terms of the form Equation (3a) as *signed* and respectively *unsigned* for EDMs constructed using only terms of the form (3b) and (3c). All previously discussed effect size measures of measurement non-invariance fall under this category. In the following section, we will show that every EDM has a closed formula with computational complexity $\mathcal{O}(q)$ under certain assumptions on the latent distribution that are always given in practice.

5. A Closed Formula Derivation Method for EDMs

For any given EDM, deriving a closed formula requires finding explicit solutions for all its expected difference terms. Since solutions for the terms in their most general form (3a), (3b) and (3c) will be applicable to every specific instance, the general problem reduces to finding solutions for these three terms, as substitution into the underlying algebraic expression will then describe a general method of obtaining a closed formula for any given EDM. Exact solutions of the general terms can be obtained using the following two observations:

5.1. Observation 1. Linear Common Factor Model Differences Are Normally Distributed

For any two groups g and k, the random variable of the model difference simplifies to

$$\hat{Y}_{jg} - \hat{Y}_{jk} = (\tau_{jg} - \tau_{jk}) + (\Lambda_{jg} - \Lambda_{jk})^{\top}\boldsymbol{\eta},$$

and is consequently an affine transformation of $\boldsymbol{\eta}$. Recall that normal distributions are stable under affine transformations, so if we assume the factors to be jointly normally distributed $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\kappa}, \Sigma)$, the random vector of the model difference will also be normally distributed with mean

$$\mu = (\tau_{jg} - \tau_{jk}) + (\Lambda_{jg} - \Lambda_{jk})^{\top}\boldsymbol{\kappa}$$

and variance

$$\sigma^2 = (\Lambda_{jg} - \Lambda_{jk})\Sigma(\Lambda_{jg} - \Lambda_{jk})^{\top}.$$

5.2. Observation 2. Basic Integrals of Normally Distributed Variables Have Exact Solutions

Consider a normally distributed random variable X with mean μ and variance σ^2. Then the following equalities hold

$$\begin{aligned}
&\mathbb{E}[X] = \mu \\
&\mathbb{E}[|X|] = \sigma\sqrt{\frac{2}{\pi}}\exp\left(\frac{-\mu^2}{2\sigma^2}\right) + \mu\operatorname{erf}\left(\frac{\mu}{\sqrt{2\sigma^2}}\right) \\
&\mathbb{E}[X^2] = \mu^2 + \sigma^2,
\end{aligned}$$

using the fact that $|X|$ has a folded normal distribution with parameters μ and σ in the second equality and rearranging the definition of the variance to read $\mathbb{E}[X^2] = \mathbb{E}[X]^2 + \mathrm{Var}(X)$ to obtain the third equality.

5.3. Analytic Solutions for the EDM Defining Model Discrepancy Terms

So, assuming joint normal distribution of the factors, we can apply Observation 2 to the difference $Y_{jg} - \hat{Y}_{jk}$, which is thus also normally distributed with mean and variance according to Observation 1, yielding exact solutions for the defining terms of EDMs: The signed difference term (3a) simplifies to

$$\mathbb{E}_{\boldsymbol{\eta}h}\left[\hat{Y}_{jg} - \hat{Y}_{jk}\right] = (\tau_{jg} - \tau_{jk}) + (\Lambda_{jg} - \Lambda_{jk})^{\top}\boldsymbol{\kappa}_h, \tag{4}$$

while the absolute difference term (3b) has the solution

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\eta}h}\left[\left|\hat{Y}_{jg} - \hat{Y}_{jk}\right|\right] = {} & \sqrt{\frac{2}{\pi}(\Lambda_{jg} - \Lambda_{jk})\Sigma_h(\Lambda_{jg} - \Lambda_{jk})^{\top}} \\
& \exp\left(\frac{-((\tau_{jg} - \tau_{jk}) + (\Lambda_{jg} - \Lambda_{jk})^{\top}\boldsymbol{\kappa}_h)^2}{2(\Lambda_{jg} - \Lambda_{jk})\Sigma_h(\Lambda_{jg} - \Lambda_{jk})^{\top}}\right) \\
& + ((\tau_{jg} - \tau_{jk}) + (\Lambda_{jg} - \Lambda_{jk})^{\top}\boldsymbol{\kappa}_h) \\
& \operatorname{erf}\left(\frac{(\tau_{jg} - \tau_{jk}) + (\Lambda_{jg} - \Lambda_{jk})^{\top}\boldsymbol{\kappa}_h}{\sqrt{2(\Lambda_{jg} - \Lambda_{jk})\Sigma_h(\Lambda_{jg} - \Lambda_{jk})^{\top}}}\right),
\end{aligned} \tag{5}$$

and the squared difference term (3c) can be written as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\eta}h}\left[\left|\hat{Y}_{jg} - \hat{Y}_{jk}\right|^2\right] = {} & ((\tau_{jg} - \tau_{jk}) + (\Lambda_{jg} - \Lambda_{jk})^{\top}\boldsymbol{\kappa}_h)^2 \\
& + (\Lambda_{jg} - \Lambda_{jk})\Sigma_h(\Lambda_{jg} - \Lambda_{jk})^{\top}.
\end{aligned} \tag{6}$$

5.4. Closed Formulas for EDMs

For any given EDM, substituting (3a), (3b) and (3c) according to (4), (5) and (6) respectively within the underlying algebraic expression will yield the desired closed formula. We note that in the setting of models with simple structure, applying this method to f^{MACS} gives rise to the formula given by Lai et al. (2025). Most notably, we can deduce that

$$d_j^{\mathrm{MACS}} = \frac{\sqrt{\Delta^2((\tau_{jr} - \tau_{jg}) + (\Lambda_{jr} - \Lambda_{jg})^{\top}\boldsymbol{\kappa}_g)^2 + (\Lambda_{jr} - \Lambda_{jg})\Sigma_g(\Lambda_{jr} - \Lambda_{jg})^{\top}}}{\mathrm{SD}_{j\mathrm{pooled}}}$$

and respectively

$$d_j^{\mathrm{MACS\_Signed}} = \frac{(\tau_{jr} - \tau_{jg}) + (\Lambda_{jr} - \Lambda_{jg})^{\top}\boldsymbol{\kappa}_g}{\mathrm{SD}_{j\mathrm{pooled}}}.$$

Formulas for all other SotA measures can be derived as described above. We have provided a publicly available R package on https://github.com/TiziSchuh/ExpectedDifferenceMeasures that contains `lavaan` compatible functions implementing the formulas of the extended EDM pairs d^{MACS}&$d^{\mathrm{MACS\_Signed}}$, UDI&SDI, WUDI&WSDI and of the extended EDM f^{MACS}.

6. Discussion

Until now, the effect size measure for measurement nonequivalence d^{MACS} has only been applicable to models without cross-loadings, rendering it unfit for many new developments in MI methodology. In this paper, we extended d^{MACS} and all subsequent effect sizes of MNE to multi-factor models with arbitrary loading patterns (i.e., cross-loadings and factor correlations) and provided closed form expressions with optimal computational complexity. With the growing availability of EFA-based methods to investigate MI (see e.g., Sterner et al., 2024a), extending the established effect size measure d^{MACS} to multi-factor models broadens the capabilities that researchers have at their disposal. This is of great importance for many (sub-)disciplines in psychology and other social sciences as many constructs are conceptualized as multi-factorial such as personality traits (e.g., John et al., 1999), motives (e.g., McClelland, 1987), or intelligence (Schneider & Newman, 2015). To properly assess MI in these contexts, potential dependencies among sub-scales in form of correlated factors or cross-loadings need to considered. This holds also true for respective effect size measures, which is why a multi-dimensional extension of d^{MACS} and its alternatives is highly important.

6.1. Future Research and Limitations

The link to EFA-based assessments of MI also raises the – in our opinion – most important caveat of multi-factor d^{MACS}: rotational dependence. It is well known that estimated EFA solutions are rotationally indeterminate. That is, infinitely many solutions of the loading matrix exist that all have the same fit to the data. What changes with each rotation, however, is the interpretation of results (for an overview of different rotations, see Browne, 2001). Especially in the multi-group context, this is crucial to consider because different rotations (in each group) might change the verdict about whether MI holds for the whole scale or for which items MI is violated. De Roover and Vermunt (2019) introduced a new rotation approach, *multi-group factor rotation* (MGFR), for multi-group models that combines rotation to simple structure *per group* with agreement of solutions *across groups*. This allows for a comparison of loading matrices between groups as well as a disentanglement of differences in loadings from differences in factor correlations.

The values given by d^{MACS} are calculated using specific model parameters and estimates of the factor distribution, and are further directly influenced by cross-loading. Thus, starting from a different rotation (and consequently different model parameters and cross-loadings) may alter the values of d^{MACS} and as a result, the verdict about MI. If different rotations (e.g., a permutation of factors) were to be used for the models of the two groups for which d^{MACS} is calculated, the resulting value would not be interpretable in a meaningful way and even potentially misleading. Future studies should investigate the influence of different rotations on the values of d^{MACS} and possible adaptations of d^{MACS} for direct comparisons of models obtained from distinct rotations. An interesting avenue might also be to assess if MGFR can help to yield statistically and conceptually more sound and interpretable solutions than simply pursuing simple structure in each group. For this, applications to empirical data might be helpful as well, to assess the added benefit of effect size measures in EFA-based MI investigations.

Additionally, although EDMs are well-defined for any conceivable latent distribution, their corresponding formulas as derived above only hold true if the factors are assumed to be jointly normally distributed. If this is not the case, the above formulas will provide wrong results and numerical approximation is most likely the only feasible alternative. Arguably, this does not add new challenges in practice however, because a joint normal distribution is typically already assumed in SEM and, for instance, even a requirement for Maximum Likelihood estimation.

Another possibility for future research stems from the fact, that the definition of EDMs can be directly applied to non-linear models, most notably models of item-response-theory. Because the method of obtaining formulas (as described in this paper) cannot be translated to non-linear models, a different approach to implementing such effect sizes is necessary in this case. This may go hand in hand with an explicit construction of EDMs for ordinal indicators, which so far have been implemented for d^{MACS} using normal graded response models (https://github.com/ddueber/dmacs).

Lastly, although existing benchmark values for d^{MACS} provided by Nye et al. (2019) (which aim to enable easier judgements about what constitutes small, medium and large values of non-invariance) may be applied to the generalized version, the same caveats apply: Lai et al. (2025) urge researchers to be highly cautious in relying on these cutoff values, as they may not be appropriate depending on the context. Their call for fine-grained and adaptive guidelines to guarantee a more sound interpretation is still subject to further research and an important next step in allowing researchers to adequately use and interpret EDMs, including all generalized effect sizes presented here.

6.2. Conclusion

A vast number of methods exist that investigate *whether or not* MI holds. Effect size measures enable a non-binary *quantification* of how much non-invariance is present. We extended the established effect size measure d^{MACS} and some newer alternatives (Gunn et al., 2020; Lai et al., 2025) to multi-factor models with cross-loadings. In this, we enable researchers to combine them with newly developed EFA-based methods as well as CFA applications without overly strict independent clusters assumptions (see also Goretzko et al., 2024). Thus, d^{MACS} can now be applied more appropriately when assessing MI in the process of questionnaire development or when investigating the factor structure of a scale in new contexts. Needless to say, quantifying the effects of non-invariance is no panacea. The interpretation of whether a specific amount of non-invariance is large or possibly "too much" is still a content-related question. But, as with traditional effect sizes measures like Cohen's *d*, effect sizes for measurement nonequivalence, such as d^{MACS}, allow for a quantification on a standardized scale.[3] This increases the comparability of psychometric scales with respect to their amount of non-invariance, hopefully aiding further improvements of their measurement quality.

References

Asparouhov, T., & Muthén, B. O. (2023). Multiple group alignment for exploratory and structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*, 169–191. https://doi.org/10.1080/10705511.2022.2127100

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*, 507–526. https://doi.org/10.1037/met0000077

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150. https://doi.org/10.1207/S15327906MBR3601_05

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

De Roover, K. (2021). Finding clusters of groups with measurement invariance: Unraveling intercept non-invariance with mixture multi-group factor analysis. *Structural Equation Modeling: A*

[3] If the unit of an observed variable is interpretable or decision rules based on raw scores are applied (e.g., in clinical psychology), researchers may also want to calculate the unstandardized expected difference for each item.

Multidisciplinary Journal, *28*, 663–683. https://doi.org/10.1080/10705511.2020.1866577

De Roover, K., & Vermunt, J. K. (2019). On the Exploratory Road to Unraveling Factor Loading Non-invariance: A New Multigroup Rotation Approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 905–923. https://doi.org/10.1080/10705511.2019.1590778

De Roover, K., Vermunt, J. K., & Ceulemans, E. (2022). Mixture multigroup factor analysis for unraveling factor loading noninvariance across many groups. *Psychological Methods*, *27*, 281–306. https://doi.org/10.1037/met0000355

Goretzko, D., Siemund, K., & Sterner, P. (2024). Evaluating model fit of measurement models in confirmatory factor analysis. *Educational and Psychological Measurement*, *84*, 123–144. https://doi.org/10.1177/00131644231163813

Goretzko, D., & Sterner, P. (2024). *Exploratory graph analysis trees-a network-based approach to investigate measurement invariance with numerous covariates*. PsyArxiv.

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

John, O. P., & Srivastava, S., et al. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Kim, E., Cao, C., Wang, Y., & Nguyen, D. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 524–544. https://doi.org/10.1080/10705511.2017.1304822

Kim, E. S., Joo, S.-H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement invariance testing across between-level latent classes using multilevel factor mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 870–887. https://doi.org/10.1080/10705511.2016.1196108

Lai, M. H. C., Zhang, Y., Ozcan, M., Tse, W. W. Y., & Miles, A. (2025). fMACS: Generalizing dMACS effect size for measurement noninvariance with multiple groups and multiple grouping variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *32*, 638–646. https://doi.org/10.1080/10705511.2025.2484812

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

McClelland, D. C. (1987). *Human motivation*. Cup Archive.

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

Mulaik, S. A. (2009). *Foundations of factor analysis*. CRC press.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335. https://doi.org/10.1037/a0026802

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

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

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

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Factorial invariance in multiple populations: A multiple testing procedure. *Educational and Psychological Measurement*, *73*, 713–727. https://doi.org/10.1177/0013164412451978

Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*, 859–870. https://doi.org/10.1080/10705511.2023.2191292

Schneider, W. J., & Newman, D. A. (2015). Intelligence is multidimensional: Theoretical review and implications of specific cognitive abilities. *Human Resource Management Review*, *25*, 12–27. https://doi.org/10.1016/j.hrmr.2014.09.004

Sterner, P. (2025). *On the investigation of measurement invariance* [PhD thesis, lmu].

Sterner, P., De Roover, K., & Goretzko, D. (2024a). New developments in measurement invariance testing An overview and comparison of EFA-based approaches (pp. 1–19).

Sterner, P., & Goretzko, D. (2023). Exploratory factor analysis trees: Evaluating measurement invariance between multiple covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*, 871–886. https://doi.org/10.1080/10705511.2023.2188573

Sterner, P., Pargent, F., Deffner, D., & Goretzko, D. (2024b). A causal framework for the comparability of latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*, 747–758. https://doi.org/10.1080/10705511.2024.2339396

Van De Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486–492.

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. https://doi.org/10.1177/109442810031002