METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Beyond Scalar Invariance: Evaluating the Validity of Person-Level Score Comparisons

Gregor Sočan[1]

[1] *Department of Psychology, University of Ljubljana, Ljubljana, Slovenia.*

Methodology, 2026, Vol. 22(2), 109–126, https://doi.org/10.5964/meth.18849

Received: 2025-07-14 • **Accepted:** 2026-02-08 • **Published (VoR):** 2026-06-30

Handling Editor: Eduardo Estrada, Autonomous University of Madrid, Madrid, Spain

Corresponding Author: Gregor Sočan, Filozofska fakulteta, Oddelek za psihologijo, Aškerčeva c. 2, 1000 Ljubljana, Slovenia. E-mail: gregor.socan@ff.uni-lj.si

Supplementary Materials: Code, Materials [see Index of Supplementary Materials]

Abstract

Scalar invariance is widely regarded as essential for comparing test-score means across groups. However, it is less clear when test scores can be meaningfully compared at the individual level — specifically, whether individuals from different groups who share the same observed score have the same expected value of the latent trait. I show that scalar invariance alone is insufficient for meaningful person-level comparisons based on sum scores. In addition to scalar invariance, person comparison invariance requires equality of latent variable means and omega coefficients across groups. Nevertheless, non-invariance effects can be relatively small if the omega coefficients are high and similar in magnitude across groups. I relate person comparison invariance to predictive invariance and provide R code to test person comparison invariance and to visualise the effects of non-invariance.

Keywords

measurement invariance, predictive invariance, factor analysis, individual diagnostics, McDonald's omega

Measurement invariance has become a standard topic in research involving comparisons across cultures (Davidov et al., 2018), organisations (Somaraju et al., 2022), developmental stages (Putnick & Bornstein, 2016) and other subpopulations in which a test may perform differently. The question of whether the test measures the latent trait in the same way in different groups can be investigated either within the framework of factor analysis or item response theory (in the latter case it is referred to as differential item functioning;

Wells, 2021). The factor analytic approach seems to have gained acceptance in psychometric practice, possibly because it can be easily integrated into more comprehensive structural equation models and because a generally accepted standard procedure exists.

In the case of complete measurement invariance, the conditional probability distribution of the observed scores X given a latent variable[1] ξ does not depend on the group membership *G* (Millsap, 2011, p. 52):

$$P(X|\xi, G) = P(X|\xi). \tag{1}$$

However, this definition is both overly restrictive and difficult to evaluate empirically. In practice, more specific aspects of the score distribution are tested. Typically, a series of increasingly restrictive nested models are tested, two of which are of particular interest:

Scalar invariance stipulates the equality of the expected values of the item scores, conditional on the common factor, across all groups:

$$\mathrm{E}(X|\xi, G) = \mathrm{E}(X|\xi). \tag{2}$$

This implies the equality of both the factor loadings and the intercepts between the groups.

Strict invariance additionally requires equality of the residual variances across groups. A generally accepted view seems to be that scalar invariance is generally sufficient in practice (Leitgöb et al., 2023). This is because two persons with the same standing on the measured trait have the same expected test score regardless of the group membership (Millsap, 2011, p. 50). This also enables meaningful comparisons of mean test scores across groups: if the factor means are equal across groups, the mean values of the observed scores are equal as well. Nevertheless, some authors advocated the strict invariance as the model that should be achieved in practice. For instance, Meredith (1993) noted that differences in residual variances affect the likelihood of admission in a selection context (p. 530) and concluded that strict (rather than scalar) invariance is essential when comparing individuals (p. 542).

Apart from scalar and strict invariance, it is possible to test the hypotheses regarding the distribution of the common factor(s). The equality of means, variances and possibly covariances of the factors is classified under the generic term structural invariance (Byrne et al., 1989). Structural invariance is not considered a problem of measurement and therefore structural non-invariance is interpreted as a substantive result and not as an indicator of a deficient measurement.

As mentioned earlier, the literature on measurement invariance has focused on group-level comparisons, typically on the conditions for comparability of means between groups. On the other hand, it could be argued that the scalar (or strict) invariance

1) In accordance with the factor analytic terminology, we will refer to the latent variable as the common factor.

perspective may not be satisfactory for a practitioner conducting individual-level diagnostics. This is because in practical situations a psychologist must base their conclusions on the observed scores and does not know the values of the common factor for particular individuals. Consider a selection or diagnostic situation in which we compare two individuals from different groups who have obtained the same test score. From a practitioner's perspective, it is more relevant and natural to ask whether individuals with the same observed score have the same expected value of the measured trait than to ask whether individuals with the same level of the measured trait have the same expected test score.

This issue has largely been overlooked in the basic psychometric literature on measurement invariance, possibly due to the focus on measurement in the research setting.

In order to formalise the perspective of diagnostics on an individual level, we introduce the concept of "person comparison invariance": Under this condition, the expected value of the common factor, conditional on the observed test score S, does not depend on group membership:

$$\mathrm{E}(\xi|S,\ G) = \mathrm{E}(\xi|S) \tag{3}$$

If the person comparison invariance holds, a practitioner can conclude that the person with the higher test score is more likely to have the higher standing on the measured trait as well, compared to a person with the lower test score. In this paper, I define the test score S as the sum of the item scores, which is the common approach to scoring for tests based on either classical test theory or factor analysis.

Researchers often appear to treat scalar invariance (Equation 2) and person comparison invariance (Equation 3) as equivalent conditions. For example, Olino (2020) concluded that "when wanting to meaningfully compare levels of functioning across clients of different demographic characteristics or identities or compare across time within client to monitor changes in functioning, measurement invariance is critical evidence to seek" (p. 728).

Another example is the European Federation of Psychologists' Association's (2013) test review model, which largely focuses on diagnostic tests and explicitly states that with scalar invariance "raw scores have the same meanings and can be compared across groups" (p. 54).

This paper aims to scrutinise such conjectures. In particular, it aims to:

1. Show that standard models of measurement invariance, especially scalar and strict invariance, are inadequate in the context of person-level comparisons of sum scores.
2. Describe the conditions under which such comparisons are valid.
3. Illustrate the extent of non-equivalence in person-level comparisons that can be expected in practice.

Notation and Preliminaries

The following notation will be used: x_{ij} is the item score for person i on item j, S_i is the sum score for person i, ξ_i is the common factor value for person i, τ_j is the factor intercept for item j, λ_j is the common factor loading for item j, ${}_s\lambda_j$ is the completely standardised common factor loading for item j, δ_{ij} is the residual (or the unique factor value, respectively) of person i on item j, θ_j is the variance of these residuals, κ is the common factor mean, ϕ is the common factor variance, $\mathbf{\Lambda}$ and $\boldsymbol{\tau}$ are $p\times 1$ vectors of the common factor loadings and factor intercepts, respectively, and $\mathbf{\Theta}$ is a $p\times p$ diagonal matrix of the residual variances. The subscripts R and F denote the reference and focal group (G), respectively. Although only the case with two groups is explicitly treated, the results can be generalised to any number of groups. The parameter values are assumed to be realistic, that is, $\mathbf{1}`\mathbf{\Lambda} > 0$, $\mathbf{1}`\mathbf{\Theta 1} > 0$, and $\phi > 0$.

Consider the standard linear one-factor model (e.g., Kaplan, 2000, p. 68):

$$x_{ij} = \tau_j + \lambda_j \xi_i + \delta_{ij} \tag{4}$$

Let us assume that the linear one-factor model holds and that the item scores can be treated as continuous variables (for a discussion on this topic, see e.g., Rhemtulla et al., 2012). Let us further assume that scalar invariance holds, i.e. $\mathbf{\Lambda}_R = \mathbf{\Lambda}_F$ and $\boldsymbol{\tau}_R = \boldsymbol{\tau}_F$. Following standard practice, we set the mean value of the common factor in the reference group to zero, $\kappa_R = 0$.

The sum score on the test for person i can be decomposed as follows:

$$S_i = \sum_{j=1}^{p} x_{ij} = \sum_{j=1}^{p} \tau_j + \xi_i \sum_{j=1}^{p} \lambda_j + \sum_{j=1}^{p} \delta_{ij} \tag{5}$$

and its expected value can be written as

$$\mathrm{E}(S_i) = \sum_{j=1}^{p} \tau_j + \kappa \sum_{j=1}^{p} \lambda_j \tag{6}$$

If the factor mean is set to 0, the mean test score is simply equal to the sum of intercepts. In matrix form, the one factor model can be written as

$$\mathbf{\Sigma} = \mathbf{\Lambda}\phi\mathbf{\Lambda}' + \mathbf{\Theta} \tag{7}$$

For further developments, it is useful to note that, in the case of a perfectly fitting one-factor model, the variance of the sum score can be decomposed into the explained and the unexplained part:

$$\sigma_S^2 = \mathbf{1}'(\mathbf{\Lambda}\phi\mathbf{\Lambda}' + \mathbf{\Theta})\mathbf{1} = \mathbf{1}'\mathbf{\Lambda}\phi\mathbf{\Lambda}'\mathbf{1} + \mathbf{1}'\mathbf{\Theta 1} = \phi\Big(\sum_{j=1}^{p} \lambda_j\Big)^2 + \sum_{j=1}^{p} \theta_j \tag{8}$$

The covariance between the latent trait and the *j*-th item score is

$$\sigma_{\xi x_j} = {}_s\lambda_j s_{x_j}\phi^{1/2} = \left(\lambda_j \phi^{1/2} s_{x_j}^{-1}\right) s_{x_j}\phi^{1/2} = \lambda_j \phi \tag{9}$$

and the covariance between the sum score and the latent trait is

$$\sigma_{\xi S} = \sum_{j=1}^{p} \sigma_{\xi x_j} = \phi \sum_{j=1}^{p} \lambda_j \tag{10}$$

Recall that the squared correlation between the common factor and the sum score is equivalent to McDonalds's omega[2]:

$$r_{\xi S}^2 \equiv \omega = \frac{\phi\left(\sum_{j=1}^{p}\lambda_j\right)^2}{\sigma_S^2} \tag{11}$$

Using Equations 8 and 10, $r_{\xi S}^2$ can also be written as the standardised covariance:

$$\begin{aligned} r_{\xi S}^2 \equiv \omega &= \frac{\sigma_{\xi S}^2}{\phi\sigma_S^2} = \frac{\phi^2\left(\sum_{j=1}^{p}\lambda_j\right)^2}{\phi\left(\phi\left(\sum_{j=1}^{p}\lambda_j\right)^2 + \sum_{j=1}^{p}\theta_j\right)} = \frac{\phi\left(\sum_{j=1}^{p}\lambda_j\right)^2}{\phi\left(\sum_{j=1}^{p}\lambda_j\right)^2 + \sum_{j=1}^{p}\theta_j} \\ &= \frac{\left(\sum_{j=1}^{p}\lambda_j\right)^2}{\left(\sum_{j=1}^{p}\lambda_j\right)^2 + \phi^{-1}\sum_{j=1}^{p}\theta_j} \end{aligned} \tag{12}$$

Conditional Expectations and Person Comparison Invariance

The expected sum score of an individual, conditional on the common factor [3], is obtained by taking the expectation of Equation 5:

$$\mathrm{E}(S_i|\xi) = \sum\tau + \xi_i\sum\lambda + \mathrm{E}\left(\sum\delta\right) = \sum\tau + \xi_i\sum\lambda \tag{13}$$

Alternatively, under the standard assumption of a linear relationship between the common factor and its indicators, we can express this expectation by the basic linear regression equation:

2) The formulas for the coefficient omega (e.g., McDonald, 1999, p. 89) typically do not take into account the common factor variance, as they assume the use of loadings that are standardised for the latent variable (for instance, obtained by the option std.lv in lavaan (Rosseel, 2012)).

3) To avoid notational clutter, I omit the item indices in the remainder of the paper.

$$\mathrm{E}(S_i|\xi) = \alpha + \beta\xi_i \tag{14}$$

The regression slope can be expressed as the product of the correlation and the ratio of the standard deviations. Expressing the correlation as in Equation 12, and the sum score standard deviation as in Equation 8, we obtain:

$$\begin{aligned}\beta &= r_{\xi S}\frac{\sigma_S}{\phi^{1/2}} = \sqrt{r_{\xi S}^2\frac{\sigma_S^2}{\phi}} = \sqrt{\frac{(\sum\lambda)^2}{(\sum\lambda)^2+\phi^{-1}\sum\theta}\bullet\frac{\phi(\sum\lambda)^2+\sum\theta}{\phi}} \\ &= \sqrt{\frac{(\sum\lambda)^2}{(\sum\lambda)^2+\phi^{-1}\sum\theta}\bullet\frac{(\sum\lambda)^2+\phi^{-1}\sum\theta}{1}} = \sum\lambda\end{aligned} \tag{15}$$

The regression constant can then be expressed using the means and the slope. After expressing $E(S)$ as in Equation 13 and the slope as in Equation 15, it becomes:

$$\alpha = \mathrm{E}(S) - \kappa\beta = \sum\tau + \kappa\sum\lambda - \kappa\sum\lambda = \sum\tau \tag{16}$$

This illustrates the well-known implication of scalar invariance: if the loadings and intercepts are invariant across groups, individuals with the same value of the common factor will have the same expected test score regardless of the group membership. In our case, however, we are interested in the expectation of the common factor given the sum score. Therefore, we need to find the parameters of the equation

$$\mathrm{E}(\xi_i|S) = a + bS_i \tag{17}$$

We can first express the slope in the same way as before:

$$\begin{aligned}b &= r_{\xi S}\frac{\phi^{\frac{1}{2}}}{\sigma_S} = \sqrt{r_{\xi S}^2\frac{\phi}{\sigma_S^2}} = \sqrt{\frac{(\sum\lambda)^2}{(\sum\lambda)^2+\phi^{-1}\sum\theta}\bullet\frac{\phi}{\phi(\sum\lambda)^2+\sum\theta}} \\ &= \sqrt{\frac{(\sum\lambda)^2}{(\sum\lambda)^2+\phi^{-1}\sum\theta}\bullet\frac{1}{(\sum\lambda)^2+\phi^{-1}\sum\theta}} = \sqrt{\frac{(\sum\lambda)^2}{\left((\sum\lambda)^2+\phi^{-1}\sum\theta\right)^2}} = \frac{\sum\lambda}{(\sum\lambda)^2+\phi^{-1}\sum\theta}\end{aligned} \tag{18}$$

Since the denominator is the same as in Equation 12, we can express b as

$$b = \frac{\omega}{\sum\lambda} \tag{19}$$

Accordingly, ω is equal to $b\sum\lambda$. Using these results and Equation 6, and noting that $\mathrm{E}(\xi) \equiv \kappa$, we express the constant as:

$$a = \mathrm{E}(\xi) - b\mathrm{E}(S) = \kappa - \frac{\omega}{\sum\lambda}\left(\sum\tau + \kappa\sum\lambda\right) = \kappa - \frac{\omega}{\sum\lambda}\sum\tau - \omega\kappa = \kappa(1-\omega) - \frac{\omega}{\sum\lambda}\sum\tau = \kappa(1 - \omega) - \frac{\sum\lambda}{(\sum\lambda)^2 + \phi^{-1}\sum\theta}\sum\tau \tag{20}$$

Note that, when using the conventional approach to identification, the part $\kappa(1 - \omega)$ equals 0 in the reference group.

Therefore, the expected common factor value for a person with a sum score S_i is:

$$\mathrm{E}(\xi_i|S) = \kappa(1-\omega) - \frac{\omega}{\sum\lambda}\sum\tau + \frac{\omega}{\sum\lambda}S_i = \kappa(1-\omega) + \frac{\omega}{\sum\lambda}\left(S_i - \sum\tau\right) \tag{21}$$

Or, alternatively, expressing ω as in Equation 12:

$$\mathrm{E}(\xi_i|S) = \kappa(1 - \frac{\sum\lambda}{(\sum\lambda)^2 + \phi^{-1}\sum\theta}) + \frac{\sum\lambda}{(\sum\lambda)^2 + \phi^{-1}\sum\theta}\left(S_i - \sum\tau\right) \tag{22}$$

Under scalar invariance, the value of the scaled test score $(S_i - \Sigma\tau)/\Sigma\lambda = C$ in Equation 21 does not depend on the group membership. Setting the value of κ_R to zero, the difference between the expected common factor values for two persons who belong to different groups and have the same test score can be expressed as:

$$\begin{gathered}\Delta\mathrm{E}(\xi_i|S) = \mathrm{E}(\xi_i|S, G = \mathrm{F}) - \mathrm{E}(\xi_i|S, G = \mathrm{R}) = \\ \kappa_F(1-\omega_F) + \omega_F C - \left[0(1-\omega_R) + \omega_R C\right] = \kappa_F(1-\omega_F) + (\omega_F - \omega_R)C\end{gathered} \tag{23}$$

Equation 23 is the key result of this paper. It implies that scalar invariance is not sufficient for person comparison invariance: if either the common factor means or the coefficients omega differ across the groups, individuals with the same sum score but belonging to different groups will have different expected common factor values. The value of $\Delta\mathrm{E}(\xi_i|S)$ can be used as a measure of the non-invariance effect size.[4] A positive value indicates that an individual from the focal group has a higher expected value of the factor than an individual from the reference group with the same test score. The value is a sum of a constant part, which is the difference between factor means, weighted by the "unreliability" $(1 - \omega_F)$, and a part depending on the observed score value. In particular, C, that is, the scaled deviation of the test score from the mean of the reference population (cf. Equation 6), is weighted by the difference between the coefficients omega in both groups. The effect of non-invariance is therefore more pronounced for more extreme test

4) It might seem that the value obtained by Equation 23 depends on the choice of the reference group, since in general $\kappa_F(1-\omega_F) \neq \kappa_R(1-\omega_R)$. Appendix A of the Supplementary Materials (see Sočan, 2026) shows that the choice of the reference group is in fact irrelevant.

scores (with respect to the mean of the reference group). On the other hand, it follows from Equation 23 that, for $\omega_R \neq \omega_F$, the effect is zero when

$$S = -\frac{\kappa_F(1-\omega_F)\sum\lambda}{\omega_F - \omega_R} + \sum\tau \tag{24}$$

If the mean factor value is equal in both groups (and therefore $\kappa_F = \kappa_R = 0$), this point corresponds to the mean test score. Otherwise, it can be either above or below the mean test score, depending on the value and the sign of the difference between the factor means and between the coefficients omega.

Special Cases

Let us now consider a special case of structural invariance, where the common factor mean and variance are equal across groups, $\kappa_R = \kappa_F$ and $\phi_R = \phi_F$. In this case, we are free to set their values to the standardised metric ($\kappa = 0$ and $\phi = 1$), and the expectation of the common factor value (see Equation 22) then simplifies to:

$$\mathrm{E}(\xi_i|S) = \frac{\sum\lambda}{(\sum\lambda)^2 + \sum\theta}(S_i - \sum\tau) = \frac{\sum\lambda}{(\sum\lambda)^2 + \sum\theta}(S_i - \mathrm{E}(S_i)) = W(S_i - \mathrm{E}(S_i)) \tag{25}$$

The expected common factor value equals the weighted deviation of the sum score from the mean sum score (note that the mean score is now equal in both groups). The weight W is invariant across groups ($W_R = W_F$) if and only if, in addition to scalar invariance, the coefficient omega is invariant. In this particular case, it also follows that the sums of residual variances are equal across groups, because both the factor loadings and factor variances are invariant. We shall denote this condition as "omega invariance". With invariant factor means and variances, the strict invariance is sufficient but not necessary for the omega invariance, because it is only the sum of the residual variances that matters.

It is also evident from Equation 25 that the value of the weight W is inversely related to the sum of the residual variances. Therefore, among two individuals with the same test score, the person from the population with the lower sum of residual variances (or, equivalently, higher value of coefficient omega) has the larger absolute expected value of the common factor. In other words, the interpretation of the test scores that does not take the group membership into account will be biased against the group in which the measurement is more reliable.

Let us now consider the case where the common factor means are equal across groups ($\kappa_F = \kappa_R$), but the common factor variances differ ($\phi_R \neq \phi_F$). After simplifying Equation 22, the expected value of the common factor can now be expressed as:

$$\begin{aligned} \mathrm{E}(\xi_i|S) &= 0(1 - \frac{\sum\lambda}{(\sum\lambda)^2 + \phi^{-1}\sum\theta}) + \frac{\sum\lambda}{(\sum\lambda)^2 + \phi^{-1}\sum\theta}(S_i - \sum\tau) \\ &= \frac{\sum\lambda}{(\sum\lambda)^2 + \phi^{-1}\sum\theta}(S_i - \sum\tau) = W(S_i - \sum\tau) \end{aligned} \tag{26}$$

In this case, the person comparison invariance holds if and only if the ratio $\frac{\sum\theta}{\phi}$ is invariant across groups, in addition to scalar invariance. Note that also in this case the weight W is invariant across groups if $\omega_R = \omega_F$. The omega invariance is therefore the necessary and sufficient condition for the person comparison invariance in this case as well. In case of non-invariant omegas, a person belonging to the group with the smaller value of $\frac{\sum\theta}{\phi}$ will have a larger absolute expected value of the common factor, compared to a person with the same test score belonging to the group with the larger value of the residual-to-factor variance ratio.

Finally, the across-group comparisons of individuals' scores are not invariant if the means differ across groups: even if omega invariance holds, the expected values of the common factor for individuals with the same sum score will still differ for a constant value of $\kappa_F(1 - \omega_F)$, see Equation 23. In practice it is important that the effect size of the non-invariance is small if either the mean difference $\kappa_F - \kappa_R$ (which in fact equals κ_F) is small, or the measurement reliability as reflected in coefficient omega is high, because in both cases the value of $\kappa_F(1 - \omega_F)$ will be close to zero.

Relations to Predictive Invariance and Classical Test Theory

Person comparison invariance can also be regarded as a special case of predictive invariance (Millsap, 1995, 1997, 2007), in which the common factor is treated as the dependent variable predicted by the test score. In Appendix B of the Supplementary Material (see Sočan, 2026) it is shown that Equation 18 and Equation 20 can also be derived from the more general results presented by Millsap (1997, 2007). Technically speaking, person comparison invariance is therefore a special case of predictive invariance. Millsap's (1995) duality theorem states that either measurement or predictive invariance can hold empirically, but they hold simultaneously only under constrained conditions. This implies that person comparison invariance can hold even if scalar invariance does not. However, unlike the general case of predictive invariance, in which an external variable is predicted, it is much less likely for person comparison invariance to hold without scalar invariance, because the regression parameters in Equation 17 are functions of the factor parameters. As shown in Appendix B of Sočan (2026), person comparison invariance without scalar invariance requires invariant factor intercepts, and a specific dependence pattern among factor-model parameters. In particular, the ratios

of error-to-factor variance ($\phi^{-1}\sum\theta$) in both groups need to be related through a linear transformation with coefficients depending on the factor loadings in both groups:

$$\phi_F^{-1}\sum\theta_F = \sum\lambda_F\left(\sum\lambda_R - \sum\lambda_F\right) + \frac{\sum\lambda_F}{\sum\lambda_R}\left(\phi_R^{-1}\sum\theta_R\right) \tag{27}$$

It is unlikely that this relationship would hold in practice, therefore the combination of perfect person comparison invariance and violated scalar invariance is unlikely to be observed in real data. Note also that under scalar invariance, where $\sum\lambda_R = \sum\lambda_F$, person comparison invariance requires the equality of error-to-factor variance ratios across groups, as also follows from Equation 18 (see also Equation B7 in Appendix B of Sočan, 2026).

From the classical test theory viewpoint, the common factor is, under certain conditions, a linear transformation of the true score (Jöreskog, 1971). The difference between the expected true scores (T) for two individuals belonging to different groups and having the same sum score S_i can be expressed using the well-known Kelley's formula (Lord & Novick, 1968, p. 65):

$$\begin{aligned}\Delta E(T_i|S) &= E(T_i|S, G = F) - E(T_i|S, G = R) = \mu_F + \rho_F(S_i - \mu_F) - [\mu_R + \rho_R(S_i - \mu_R)] \\ &= \mu_F(1 - \rho_F) - \mu_R(1 - \rho_R) + (\rho_F - \rho_R)S_i\end{aligned} \tag{28}$$

The similarity with Equation 23 should not be surprising due to the syntactic similarity between factor analysis and classical test theory. However, the generality of the classical test theory approach is limited because the true score is defined with the respect to the observed scores on a particular test (namely, as its expected value for an individual), so its range and scale are determined by the number of items and the response scale; on the other hand, the existence of a common factor is, at least in principle, independent of its indicators.

Testing the Hypothesis of Person Comparison Invariance

After scalar invariance has been established, person comparison invariance can be tested by separately testing the equality of both factor means and coefficients omega across groups. The first hypothesis can be tested using the Wald test that is routinely reported by SEM software, or by constraining the factor means across groups and then testing the difference in the model fit. Omega invariance, on the other hand, can be tested by defining the omegas as new model parameters in the scalar invariance model and then either constrain them to equality or (in case of two groups) compute a bootstrap confidence interval for the difference between the omegas. The advantage of the first

approach is the possibility of combining both hypotheses: since the usual statistical practice is to constrain the slopes first and the intercepts second, I propose to test the model with equal omegas against the scalar invariance model in the first step, and to test the model with equal latent variable means against the omega invariance model in the second step. In addition to difference tests, I recommend using the plots as described below to assess the size of the bias; (differences between) fit indices do not seem to be very useful in this respect. R scripts for testing the person comparison invariance and visualization of the effects of non-invariance are available as Supplementary Materials (files PCI_LRtest.R, PCI_boot_omega.R, and PCI_graphs.R, see Sočan, 2026).

Illustration on Real Data

I present a re-analysis of the data collected by Kavčič et al. (2023). Among other things, they investigated the measurement invariance of the Connor-Davidson Resilience Scale (CD-RISC) across educational groups (higher vs. lower education). The scale consists of 10 items rated on a 5-point scale; the range of sum scores is from 0 to 40. The confirmatory factor analysis with the MLMV estimator showed a good fit to the scalar invariance model (robust RMSEA = .046, SRMR = .056, robust CFI = .962). The higher education group had a higher mean resilience (standardised factor mean of .342) and a lower coefficient omega (ω_{HI} = .788, ω_{LO} = .840) than the lower education group. Figure 1 shows the effect of these differences on the expected standardised factor values. The solid and dashed lines represent the expected common factor values for individuals with a given sum score belonging to different education groups. The shaded area shows the difference between these expected values, i.e., the bias in person comparisons. The lines intersect at a sum score of 32.8 (this can also be obtained from Equation 24), about one standard deviation above the mean test score in the whole sample. At this point, person comparisons would be invariant. In the range of sum scores below this point, comparisons based on sum scores are biased against individuals with higher education, and vice versa.

In Figure 2, the bias values (i.e., the differences between the expected values) are plotted. The shaded area shows where the absolute bias values are smaller than 0.2 *SD*. By analogy with the commonly accepted interpretation of Cohen's *d* (see also Nye et al., 2019), this is tentatively considered a negligible bias. Using this criterion, we conclude that comparisons of sum scores across groups are notably biased only among participants with the lowest scores (12 or lower). The positive bias in this case means that an individual from the higher education group has a higher expected factor value than an individual from the lower education group with the same sum score.

Figure 1

Expected Common Factor Values for Two Educational Level Groups

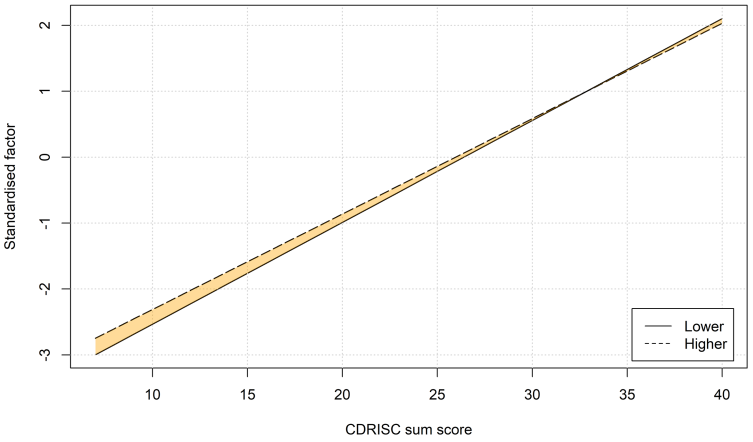
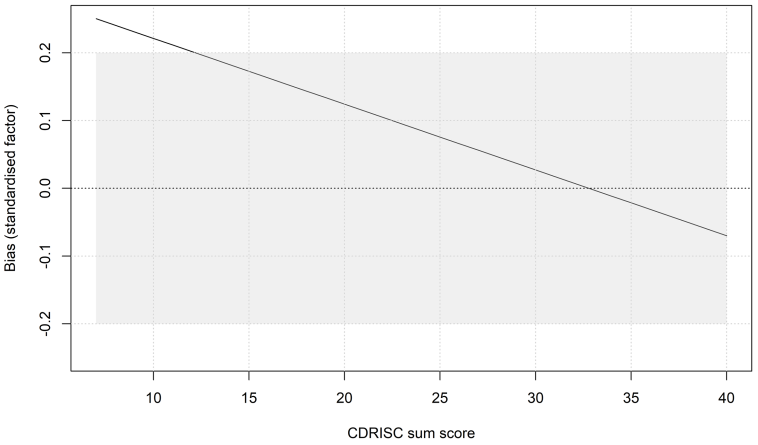


Figure 2

Bias in Person Comparisons Across the Two Educational Level Groups



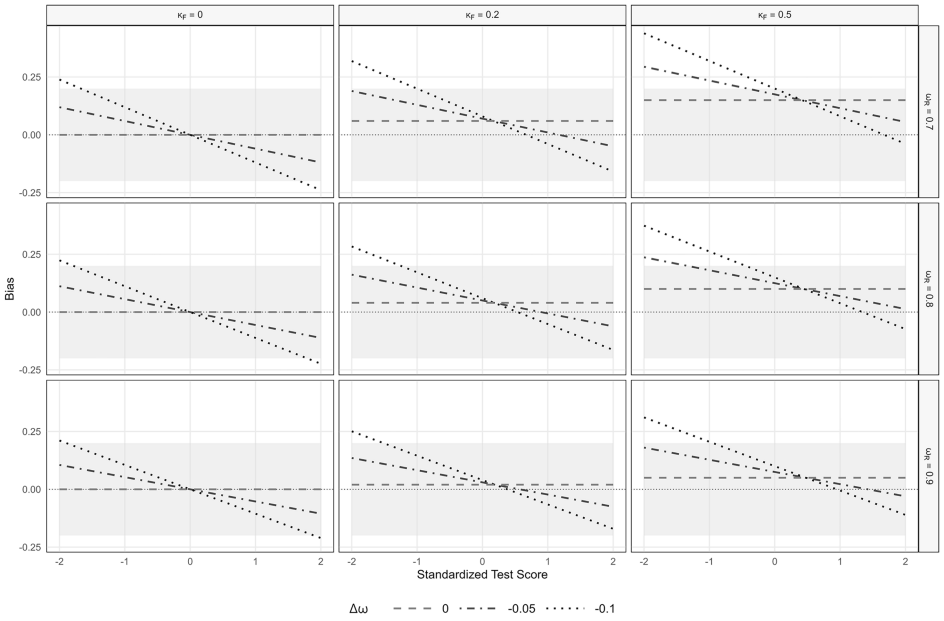
What Size of Bias Should We Expect in Practice?

Figure 3 illustrates the magnitude of bias (y-axis) across standardised test scores on the reference group metric (-2 ≤ z ≤ 2; the x-axis) under scalar invariance and different parameter combinations:

- Coefficient omega in the reference group (ω_R = .7, .8, or .9).
- Factor mean in the focal group (κ_F = 0, 0.2, or 0.5, following common guidelines for Cohen's *d*); the reference-group factor mean was set to zero.
- The difference between coefficients omega in the focal and reference group, respectively ($\Delta\omega$: = $\omega_F - \omega_R$ = 0, -.05, or -.10).

Figure 3

Bias in Relation to the Sum Score for Different Values of κ_F, ω_R, and $\omega_F - \omega_R$.



The panels correspond to the $\omega_R \times \kappa_F$ combinations, and, within each panel, regression lines correspond to different values of $\Delta\omega$. The shaded region again marks absolute bias values smaller than 0.2 *SD*. In all situations shown, $\kappa_F \geq \kappa_R = 0$, and $\omega_F \leq \omega_R$. If the values were reversed, the basic shape would remain the same, with the lines mirrored. Figures C1–C3, showing the remaining combinations, appear in the Supplementary Materials (Appendix C); see Sočan (2026). The R script to reproduce and modify the plots is also provided there (file Figure_3.R).

In accordance with Equation 23, a positive value means that a member of the focal group has a higher expected factor value than a member of the reference group who has an equal observed test score. A positive value therefore indicates bias against the focal group.

When omega invariance holds ($\Delta\omega = 0$), the bias lines are horizontal because the bias equals $\kappa_F(1 - \omega)$ (see Equation 23), being constant across all score levels. In our range of conditions, the bias never exceeds the threshold, but it would do so with sufficiently low omega values, or sufficiently large latent-mean differences.

When omega coefficients differ across groups, bias also depends on the test score. When $|\Delta\omega| \leq .05$ and $\kappa_F \leq 0.2$, the bias is consistently negligible. When $\kappa_F = 0.5$, bias is notable in the lower score range, especially when ω is lower or $\Delta\omega$ is larger. The bias remains negligible for standardised scores that are either close to κ_F, or somewhat higher.

Two general rules can be inferred from the figures:

1. When $\kappa_F > 0$ (as in Figures 3 and C1) the bias tends to be positive (i.e., against the focal group), and vice versa. This is also evident from Equation 23.
2. When $\kappa_F(\omega_F - \omega_R) < 0$ (as in Figures 3 and C3) the bias is larger in the lower score range, and when $\kappa_F(\omega_F - \omega_R) > 0$ (as in Figures C1 and C2) the bias is larger in the higher score range. The difference in factor means pushes all lines upwards (when $\kappa_F > 0$) or downwards (when $\kappa_F < 0$) and thus amplifies the bias due to the difference in omegas in either the upper or the lower score range. Note that, strictly speaking, in a sufficiently extreme range, both high and low scores are biased whenever $\omega_F \neq \omega_R$.

Discussion

The aim of this paper is to offer a new perspective on measurement invariance, focusing on comparisons between individuals rather than groups. The main conclusion is that the standard models of measurement invariance are not an optimal framework for assessing the comparability of individual scores: scalar invariance is neither necessary nor sufficient for person comparison invariance. Assuming a weakened form of scalar invariance (i.e., the *sums* of factor loadings and the *sums* of intercepts being invariant across groups), individuals with equal sum scores have equal expected factor values when the factor means are equal and omega coefficients are invariant across groups. In the case of invariant omega coefficients and non-invariant factor means, the effect of person comparison non-invariance is constant across the range of scores and depends on both the magnitude of the mean difference and the size of the omega coefficient. If the omega coefficients are not invariant, differences between their values also affect the size of the non-invariance effect. In this case, the non-invariance effect varies across the score range: it is generally higher at the extremes, and there is a point where the effect is zero (i.e., where the individual scores are comparable).

These results accord with several previous findings. Shealy and Stout (1993) noted that conditioning on observed scores distorts the estimation of bias in a studied subtest when true score means differ across groups. It is also well known that latent-trait estimates may be biased, particularly at the extremes (Feuerstahler, 2018). Finally, Meredith's (1993) warning concerning the effect of non-invariant residual variances on person comparability was confirmed.

As the effect-size measure of bias, I propose the difference between the expected values of the measured trait (possibly standardised) for two individuals who share the same sum score but belong to different groups (see Equation 23). The calculation is easy to perform, because it requires only parameters routinely reported by SEM software (factor means, loadings, and intercepts) and the omega coefficient, which has become a standard psychometric quality indicator for tests based on factor analysis.

Two limitations of the proposed approach should be noted. First, it relies on the standard factor analysis model with homoscedastic errors, assuming linear relationships between the common factor and the item scores, and treating the item scores as numeric variables. While this reflects a common practice in applied psychometrics, it can be somewhat inaccurate, especially for items with a smaller number of rating categories. Second, I assume that the residuals in the regression equations have a mean value of zero. Although the expected value of random measurement error is zero by definition, this may not be true for the so-called specific factors (see Millsap, 2011, p. 77). However, this limitation also applies to the standard testing of scalar invariance.

Mathematically speaking, person comparison invariance is a special case of Millsap's (1995, 1997, 2007) predictive invariance. Therefore, it could hold even without scalar invariance, but this would require restrictions unlikely to hold in practice. The added value of this paper relative to the existing literature on predictive invariance is the treatment of the special case where the underlying latent factor is "predicted", and the derivation of an explicit expression for bias and elucidation of its determinants and their interplay in this special case. I also provide estimated bias values for various likely combinations of parameters.

From the practical viewpoint, person comparison invariance can be conceptualised as an extension of the standard measurement invariance testing in situations where specific persons are to be compared. Even for such tests, scalar invariance should be examined as an initial step in order to identify items that function differently across subpopulations. Person comparison invariance is more difficult to achieve in practice than scalar invariance because a larger set of parameters is constrained and because it partially depends on factors beyond the researcher's control, such as the means and variances of the common factor. At first glance, this may appear discouraging to test users. Fortunately, as the presented illustrations show, the size of the bias may be small or even negligible in practice, especially if the measurement reliability is high. Indeed, the bias would vanish if coefficients omega equalled 1, regardless of other parameter

values. In contrast to some liberal recommendations on reliability that can be found in the literature, these findings underscore the importance of striving for a high omega coefficient in test construction. On the other hand, they point to inherent limitations of the ubiquitously used sum scores as indicators of the measured latent variables.

Funding: The author acknowledges the financial support from the Slovenian Research and Innovation Agency (Research Core Funding No. P5-0062 and Grants No. J5-4590 and J7-4599).

Acknowledgments: I wish to thank Tina Kavčič for providing the raw data on CD-RISC.

Competing Interests: The authors have declared that no competing interests exist.

Data Availability: All codes needed to perform the analyses and to reproduce the figures are available in the Supplementary Material (see Sočan, 2026). The manuscript includes an analysis of secondary data; while the raw data are not available, the model parameters needed to reproduce the reported results are available as part of the Supplementary Material (see Sočan, 2026).

Supplementary Materials

Type of supplementary material	Availability/Access
Data	
No study data available	—
Code	
R code files	Sočan (2026)
Material	
Model parameters to reproduce the reported results	Sočan (2026)
Study/Analysis preregistration	
Study was not preregistered	—
Other	
Appendices further illustrating and explaining study issues	Sočan (2026)

References

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

Davidov, E., Schmidt, P., Billiet, J., & Meuleman, B. (Eds.). (2018). *Cross-cultural analysis: Methods and applications* (2nd ed.). Routledge. https://doi.org/10.4324/9781315537078

European Federation of Psychologists' Associations. (2013). *EFPA review model for the description and evaluation of psychological and educational tests* (Version 4.2.6). https://www.efpa.eu/sites/default/files/2023-06/110c_EFPA_BOA_TEST_REVIEW_MODEL_version426.pdf

Feuerstahler, L. M. (2018). Sources of error in IRT trait estimation. *Applied Psychological Measurement, 42*(5), 359–375. https://doi.org/10.1177/0146621617733955

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109–133. https://doi.org/10.1007/BF02291393

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. SAGE Publications.

Kavčič, T., Zager Kocjan, G., & Dolenc, P. (2023). Measurement invariance of the CD-RISC-10 across gender, age, and education: A study with Slovenian adults. *Current Psychology, 42*(3), 1727–1737. https://doi.org/10.1007/s12144-021-01564-3

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30*(4), 577–605. https://doi.org/10.1207/s15327906mbr3004_6

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*(3), 248–260. https://doi.org/10.1037/1082-989X.2.3.248

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*(4), 461–473. https://doi.org/10.1007/s11336-007-9039-7

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

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*(3), 678–709. https://doi.org/10.1177/1094428118761122

Olino, T. M. (2020). Clinical applications of measurement invariance. *Journal of Personality Assessment, 102*(5), 727–729. https://doi.org/10.1080/00223891.2020.1793766

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, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Shealy, R., & Stout, W. A. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159–194. https://doi.org/10.1007/BF02294572

Sočan, G. (2026). *Supplementary Materials to* "Beyond scalar invariance: Evaluating the validity of person-level score comparisons" [R code files to test person comparison invariance and visualise the effects of non-invariance, Appendices further illustrating and explaining study issues]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.22227

Somaraju, A. V., Nye, C. D., & Olenick, J. (2022). A review of measurement equivalence in organizational research: What's old, what's new, what's next? *Organizational Research Methods, 25*(4), 741–785. https://doi.org/10.1177/10944281211056524

Wells, C. S. (2021). *Assessing measurement invariance for applied research.* Cambridge University Press. https://doi.org/10.1017/9781108750561

EUROPEAN ASSOCIATION OF METHODOLOGY

Methodology (METH) is the official journal of the European Association of Methodology (EAM).

zpid Leibniz-Institut für Psychologie

PsychOpen GOLD is a publishing service provided by the Leibniz Institute for Psychology (ZPID), Germany.