PARE

Practical Assessment, Research, and Evaluation

A peer reviewed, open-access electronic journal: ISSN 1531-7714

Assessing the Measurement Equivalence of a set of Items: Item-specific Diagnostics on an Interpretable Metric

Johan Braeken, *University of Oslo*
Saskia van Laar, *Maastricht University*

Abstract: The characteristic of 'measuring the same thing repeatedly in the same way' makes psychological tests with equivalent items an attractive choice for one-off assessments and progress monitoring. A hierarchy of factor analysis measurement models formalizes the global equivalence of the item set. The traditional model comparison approach provides a binary statistical significance decision about the global level of equivalence, but interpretable local diagnostics to assess the degree of equivalence for specific items on a meaningful metric are not yet available. We introduce such item-specific effect-size diagnostics through smart use of the effect-coding identification rule to set the to-be-measured latent variable's scale.

Keywords: Factor analysis, Effect size, Measurement equivalence, Congeneric, Parallel

Introduction

In specific psychological, educational, and health assessment contexts, there is a tradition to develop tests consisting of items that are intended to be as equivalent as possible. Tests consisting of equivalent items are easily interpreted because scores on such items are directly comparable (Jöreskog, 1971) and the testing procedure aligns well with people's natural understanding of measurement, validity, and reliability. Furthermore, the test format with equivalent items also facilitates verifying intra-individual consistency in performance (Levine & Rubin, 1979). The characteristic of 'measuring the same thing repeatedly in the same way' makes testing procedures with equivalent items an attractive choice in practice for both one-off assessments as well as progress monitoring. In early child development for instance, an oral reading fluency test (e.g., Arnesen et al., 2017; Fuchs et al., 2001; van der Velde et al., 2024) would typically contain multiple short texts where each text needs to be read aloud by the child and words correct per minute is the measured outcome. The short texts are constructed such that they are as similar as possible in many textual and reading aspects (e.g., word frequency, sentence length and complexity) to ensure that they are of equal difficulty. Having a set of such equivalent texts that do differ in text type (e.g., narrative/expository) and contents but not in psychometric characteristics, allows for a desired flexibility in assessment practices. For instance, some texts can be used as exercise material whereas others are kept for the actual test moment (test security) or children can be repeatedly tested on the same time point to check their consistency or across time points to track their development. The equivalence allows for easy and clear comparison of the child's performance across texts. All of this without having to worry about potential memory effects and other repetition artifacts that would apply when one repeatedly uses the same short text as stimulus.

While creating equivalent items, the objective is to sample items from a single universe of items. As there are so many item features and aspects to consider, constructing such equivalent items is a challenging endeavor even for expert item writers. Modern approaches, using large language models and automatic item generation templates (e.g., Bezirhan & von Davier, 2023; Götz et al., 2023), can assist in the construction of equivalent items, but we would also hope for the availability of accessible and informative methodological procedures that can provide empirical evidence to assess the extent that the goal of making equivalent items has succeeded.

Hierarchy of Measurement Equivalence Models

In psychometrics, methodological procedures to assess item equivalence are framed in the context of a hierarchy of measurement models (e.g., Jöreskog, 1971; Lord & Novick, 1968). A set of items ($i \in \{1,\ldots,I\}$) are considered *perfect measures* of a latent variable F if all continuous item scores Y_i measure F perfectly without measurement error and by implication scores on all items would be exactly equal (i.e., duplicates) for each individual: $Y_i = F$ and $Y_1 = \ldots Y_i = \ldots = Y_I$. Such items would be fully interchangeable at the individual measurement level and would allow to administer a measurement instrument consisting of a single item from the item set, where the specific item could even vary across the individuals being measured and still lead to comparable measurement outcomes. However, the existence of perfect measures, with the implied set of deterministic equalities and absence of measurement error[1], is hardly considered realistic in practice and at a philosophical level one can debate whether this prescribes a measurement model or a mere tautology.

In contrast, a set of items are considered *parallel measures* if all continuous item scores Y_i conform to the following measurement model

$$\begin{gathered} Y_i = F + \varepsilon_i \\ \forall i\colon \sigma^2_{\varepsilon_i} = \sigma^2_{\varepsilon} \end{gathered} \tag{1}$$

Each item is assumed to measure the same underlying latent factor F in the same units of measurement on the same scale and with the same precision (cf. homoscedasticity assumption on the measurement error variance across items in Equation 1) (Lord & Novick, 1968). Measurement error is assumed to be on average 0 and be independent of the latent variable F and across items. This implies that all items capture the same true score F (cf. Classical Test Theory, Novick, 1966; Spearman, 1904) for an individual, but that individual observed item scores Y_i and Y_j $(i \neq j)$ are not necessarily equivalent due to the presence of item-specific measurement error ε_i and ε_j. However, inferences made based on any of these items are expected to be interchangeable because, if the parallel measurement model holds, item score distributions are implied to be

[1] Throughout the paper, we only consider continuous item responses such that the latent variable measurement models will then essentially be factor analysis models. Extensions to categorical item responses exist, but are not considered as categorical responses have no natural metric unit of measurement. We do return to this point in the discussion.

equivalent. Item scores are expected to have equal means $\mu_{Y_i} = \mu_F$ and equal variances $\sigma^2_{Y_i} = \sigma^2_F + \sigma^2_\varepsilon$ across items, and inter-item covariances $\sigma_{Y_iY_j} = \sigma^2_F (\forall i \neq j)$ are all equal. As a consequence, inter-item correlations $\rho_{Y_iY_j} = \sigma^2_F/(\sigma^2_F + \sigma^2_\varepsilon)\ \ (\forall i \neq j)$ are equal, as are correlations ρ_{Y_iX} of items with another variable X. Furthermore, in terms of measurement, each item is an unbiased measure of the true score and has the same measurement reliability $\omega_i = 1 - \sigma^2_\varepsilon/(\sigma^2_F + \sigma^2_\varepsilon)$.

Measurement models implying less strict forms of item equivalence can be obtained by relaxing some of the explicit (and implicit) across-item equality constraints in Equation 1. Items are considered *tau-equivalent* measures if they still measure the same underlying latent variable F in the same units of measurements on the same scale, but now with potentially varying precision. Hence, the homoscedasticity assumption for the measurement error variance is dropped and $\sigma^2_{\varepsilon_i}$ can vary across items. While the mean scores μ_{Y_i} and inter-item covariances $\sigma_{Y_iY_j}$ remain equal across items, the correlations $\rho_{Y_iY_j}$ and ρ_{Y_iX} will differ across items due to items now differing in their total variance $\sigma^2_{Y_i} = \sigma^2_F + \sigma^2_{\varepsilon_i}$. Correlations involving items with larger $\sigma^2_{\varepsilon_i}$ (i.e., lower measurement precision) will be more attenuated towards zero than correlations involving items with lower $\sigma^2_{\varepsilon_i}$ (i.e., higher precision). Yet, when predicting an item score Y_i, the regression coefficient of a given predictor X remains equal regardless which one of the items is chosen to be the outcome, such that for studies of determinants of the latent variable F, tau-equivalent measures are still interchangeable. In terms of measurement, each item remains an unbiased measure of the true score, but has differing measurement reliability $\omega_i = 1 - \sigma^2_{\varepsilon_i}/\sigma^2_{Y_i}$.

Items are considered *essentially tau-equivalent* measures if they still measure the same underlying latent variable F in the same units of measurement, but now on a scale that can shift in position of its average on an item-by-item basis and with potentially varying precision (see Equation 2).

$$Y_i = v_i + \mathrm{F} + \varepsilon_i \tag{2}$$

This implies that all item true scores are still the same up to an additive constant (i.e., v_i). Hence, some items might show a form of measurement bias where they tend to, on average, over- or underestimate the true score (cf., positive/negative v_i). The difference in implications for summary statistics, in contrast to tau-equivalence, is that the mean scores μ_{Y_i} are no longer equal across items. Essentially tau-equivalent measures would only be interchangeable after bias-correction, the other implications remain unchanged though. Note that essential-tau-equivalence is also the measurement model underlying the reliability coefficient alpha (Cronbach, 1951).

Items are considered *congeneric* measures if they still measure the same underlying latent variable F, yet in different units of measurement, on shifted scales, and with varying precision (see Equation 3).

$$Y_i = v_i + \mathrm{F}\lambda_i + \varepsilon_i \tag{3}$$

The parameter λ_i either compresses $(\lambda_i < 1)$ or expands $(\lambda_i > 1)$ the original measurement scale of the latent variable F, transforming one unit on the original scale into a smaller/larger distance on the continuous item score Y_i. In this measurement model, one can easily recognize the familiar factor analysis model where λ_i is known as a factor loading and v_i as an item intercept. This congeneric model is the least restrictive measurement model and merely implies that all item scores are expected to be correlated to some degree. The lack of similar measurement units, as reflected in the item-specific loadings, implies that we cannot really speak of measurement-equivalent items, except in a sense that items are still measures of the same latent variable. Note that the congeneric model is the measurement model underlying the reliability coefficient omega (McDonald, 1970).

For completeness, a so-called *null* model could be added to the hierarchy of measurement models. This null model would require items to be uncorrelated and essentially has zero equivalence implications given that items do not even measure the same underlying latent variable, with the obvious consequence of having an absolute absence of measurement equivalence across items (see Equation 4).

$$Y_i = v_i + \varepsilon_i \tag{4}$$

Assessing the Degree of Equivalence across Items

The hierarchy of measurement models means that each stricter model is nested in the less restrictive models next in line, and has stronger equivalence implications for the item score distributions. A series of model comparisons through likelihood ratio tests for nested models or other typical structural equation model fit indices (e.g., CFI, RMSEA, or SRMR) would then allow to assess what type of factor analysis measurement model applies to the set of items (e.g., Bentler & Bonett, 1980). The choice of measurement model would then also imply a crude qualitative assessment of the measurement, but such omnibus significance tests remain hard to interpret as outcome (e.g., Marsh, 1998) and the global fit indices do not translate in any obvious effect size measure to help assess the degree of measurement equivalence for specific items. Thus, for practical use, additional diagnostic information and effect size indices that are on a more interpretable metric and more local, than chisquare test statistics and related fit indices, are sought-after.

For this purpose, we propose diagnostics directly related to the measurement units of the individual items by making strategical use of a specific model identification strategy for latent variables. The identification strategy makes use of a type of effect coding of item parameters and essentially centers the latent variable on an "average" metric across items. This strategy has its historical roots in the early days of Rasch modelling and item response theory (IRT) (Lord & Novick, 1968), but has faded out during recent decades, although it has more recently been resurfacing in longitudinal measurement models (Little, 2013; Little et al., 2006).

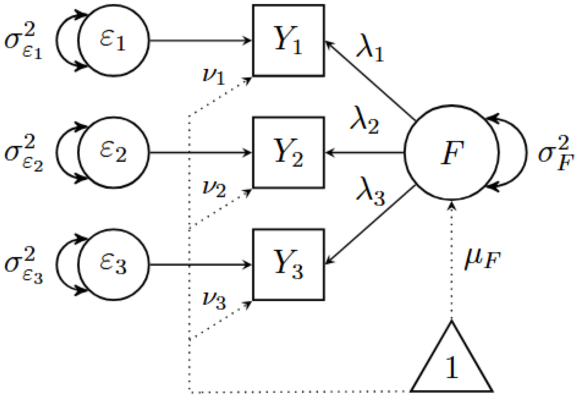
Figure 1 represents a fully-parameterized path diagram of a factor analysis measurement model with a latent variable *F* —with mean μ_F and standard deviation σ_F — underlying observed item responses Y_i $(i = [1, I])$ —with item-specific factor loadings λ_i and intercepts ν_i—, and residuals ε_i —with mean zero and standard deviation σ_{ε_i}. Without any further parameter restrictions, the model in Figure 1 is not identified and cannot be estimated.

A key part of model identification in measurement models is to fix the scale of the latent variable. Whereas one commonly chooses between either standardizing the latent variable $(\mu_F = 0, \sigma_F = 1)$ or fixing a reference item j's loading and intercept (e.g., $\lambda_j = 1, \nu_j = 0$), we consider the following equivalent identification rule:

$$\lambda_I = I - \sum_{i=1}^{I-1} \lambda_i \tag{5}$$

$$v_I = -\sum_{i=1}^{I-1} v_i \tag{6}$$

Figure 1. Fully-Parameterized Factor Analysis Measurement Model with Implied Covariance Matrix and Mean Vector



$$\mathbf{\Sigma_Y} = \begin{bmatrix} \lambda_1\sigma^2_F\lambda_1 & \lambda_1\sigma^2_F\lambda_2 & \lambda_1\sigma^2_F\lambda_3 \\ \lambda_2\sigma^2_F\lambda_1 & \lambda_2\sigma^2_F\lambda_2 & \lambda_2\sigma^2_F\lambda_3 \\ \lambda_3\sigma^2_F\lambda_1 & \lambda_3\sigma^2_F\lambda_2 & \lambda_3\sigma^2_F\lambda_3 \end{bmatrix}$$

$$\mu_{\mathbf{Y}} = \begin{bmatrix} \nu_1 + \lambda_1\mu_F & \nu_2 + \lambda_2\mu_F & \nu_3 + \lambda_3\mu_F \end{bmatrix}$$

Note. Default path diagram conventions hold. Squares correspond to observed variables, circles to latent unobserved variables, and a triangle to a constant. For visual clarity, paths related to the mean structure of the model are drawn as dotted lines.

This identification rule essentially follows an effect coding for the item parameters such that the factor's mean and standard deviation can be freely estimated. The identification rule has some useful properties for our context of assessing the degree of equivalence between the different indicators. First of all, the latent variable is now on a metric reflecting the 'average' metric of the indicators, which makes it more meaningful and easily interpretable in practice. All item parameters are now also defined in direct comparison to this common metric.

A factor loading λ_i would reflect to what extent an item retains the measurement units of the latent variable: A value smaller/larger than 1 contracts/expands the scale (i.e., a multiplicative factor)[2]. A Wald test statistic where the parameter is contrasted against the value of 1 (instead of the default 0) provides a straightforward way of significance testing of the relevant null hypothesis.

An intercept parameter ν_i now reflects the amount of systematic measurement bias (cf. tau-equivalence= no over/underestimation of the true score by individual items) of the indicator on this very same metric (i.e.,

[2] This interpretation is more easily seen in the following equivalent expression for the parameter constraints in Equation 5, $\frac{\sum_{i=1}^{I}\lambda_i}{I} = 1$.

an additive factor). By definition, this item-specific measurement bias averages out across the item set[3]. The default Wald test of that parameter provides a way to test the null hypothesis that this item-specific bias is zero.

Next to these item-specific diagnostics, we propose to look at the difference in estimated mean and standard deviation of the latent variable across the hierarchy of measurement models. Given the chosen identification rule, these quantities are directly comparable and large changes from model to model would indicate which equivalence restrictions have an observable impact on the estimated metric of the latent variable reflected by the set of assumed-equivalent items. Keeping in line with the identification rule, we suggest a ratio effect size for the latent variable's standard deviation $\phi(\sigma_F) = \sigma_F^{[alternative]}/\sigma_F^{[baseline]}$ and a difference effect size for the latent variable's mean $\Delta(\mu_F) = \mu_F^{[alternative]}/\mu_F^{[baseline]}$, each time with the congeneric model as baseline and one of the more restrictive models in the equivalence hierarchy as alternative.

Empirical Illustration: Oral Reading Fluency

We consider a synthetic dataset of a subset of oral reading fluency items inspired by a study by Arnesen et al. (2017)[4]. The item set consists of three text passages, each designed by experienced content expert item writers to be measurement equivalent. The item responses are continuous variables measured in terms of the number of words that the child read correctly out loud within one minute of time. Figure 2 summarizes the resulting congeneric factor analysis measurement model for this itemset, under both a traditional standardized latent variable identification and the effect-coding identification that is key to the proposed procedure for the assessment of measurement equivalence of a set of items. We included the corresponding sufficient statistics in the form of a covariance matrix and mean vector, and provided, for completeness, conversion formulas for the parameter sets under both identification rules. For three items, the congeneric measurement model is equivalent to the fully saturated model, which allows the reader to easily compare parameter values and trace the path diagram.

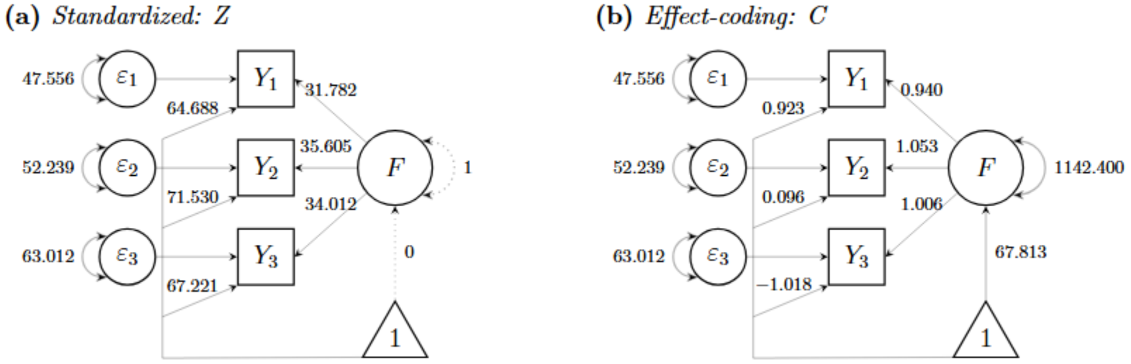
Table 1 summarizes results of a traditional model comparison approach across the hierarchy of factor analysis measurement models for the first set of items. Note that in the case of 3 items, the congeneric measurement model is equivalent to the perfect fitting saturated model. The chisquare likelihood ratio tests rejected all nested models in favor of the congeneric model, leaving us to the binary conclusion that the items are not measurement equivalent. Yet, this does not provide any indication of how badly inequivalent specific items are.

Relying on fit indices (e.g., CFI, RMSEA, or SRMR) to make such an inference does not provide much uniform direction in this respect. Depending on how strict of a rule of-thumb one would adhere to (and which fit indices), we could consider the essential tau equivalent model or even the parallel measurement model as our model of choice, and suddenly reach an entire opposite binary conclusion: The items are measurement equivalent at a specified level in the hierarchy. Relatively speaking, one could conjecture that the main equivalence problem might be in the biased item intercepts as the shift from essential tau-equivalence to tau-equivalence showed the largest difference for most of the fit indices.

[3] This is more easily seen in the following equivalent expression for the parameter constraints in Equation 6, $\sum_{i=1}^{I} v_i = 0$.

[4] A non-parametric bootstrap resample of the empirical data, omitting all identifying information, including the school and class structure of the children.

Figure 2. Equivalence between the Standardized and Effect-coding Identification Rule for the Latent Variable in the Congeneric Factor Analysis Measurement Model



Convert from Effect to Standardized

$$\sigma_F^{[Z]} = \sigma_F^{[C]}/\sigma_F^{[C]}$$
$$\mu_F^{[Z]} = \mu_F^{[C]} - \mu_F^{[C]}$$
$$\lambda_i^{[Z]} = \lambda_i^{[C]}\sigma_F^{[C]}$$
$$\nu_i^{[Z]} = \nu_i^{[C]} + \lambda_i^{[C]}\mu_F^{[C]}$$
$$\sigma_{\varepsilon_i}^{[Z]} = \sigma_{\varepsilon_i}^{[C]}$$

Convert from Standardized to Effect

$$\sigma_F^{[C]} = \overline{\lambda_i^{[Z]}}$$
$$\mu_F^{[C]} = \overline{\nu^{[Z]}}$$
$$\lambda_i^{[C]} = \lambda_i^{[Z]}/\overline{\lambda^{[Z]}}$$
$$\nu_i^{[C]} = \nu_i^{[Z]} - \overline{\nu^{[Z]}}\lambda_i^{[Z]}/\overline{\lambda^{[Z]}}$$
$$\sigma_{\varepsilon_i}^{[C]} = \sigma_{\varepsilon_i}^{[Z]}$$

$$\mathbf{S_Y} = \begin{bmatrix} 1060.570 & & \\ 1134.712 & 1323.579 & \\ 1083.953 & 1214.331 & 1223.198 \end{bmatrix}$$

$$\mu_\mathbf{Y} = \begin{bmatrix} 64.688 & 71.530 & 67.221 \end{bmatrix}$$

$$n = 362$$

Note. Default path diagram conventions hold (for notation, see Figure 1). Squares correspond to observed variables, circles to latent unobserved variables, and a triangle to a constant. Paths related to fixed parameters are drawn as dotted lines. $\boldsymbol{S_Y}$ = model-implied covariance matrix; $\mu_\mathbf{Y}$ = model-implied mean vector. $\bar{\lambda}$ and $\bar{v}$ represent the average loading and average intercept across items, respectively.

If one looks purely at how well the latent variable is measured by the set of items, we can notice that the latent factor's mean and standard deviation remain relatively stable across measurement models (see Table 2). Given the effect-coding identification rule, the factor's mean can be interpreted as the grand mean of item true scores across the item set. In this case, this comes down to about 68 words correctly read within one minute, with a standard deviation of about 34 words. The variation across measurement models is merely in the digits behind the decimal, with latent variable's mean differences $\Delta(\mu_F)$ near 0 and latent variable's standard deviation ratios $\phi(\sigma_F)$ near 1. The same holds for the item set's reliability, which is estimated at about .98. Hence, as a set, the items' measurement properties seem quite excellent and stable.

This estimated latent variable metric also provides good guidance on what to expect with respect to summary statistics for a novel oral reading fluency text that would be generated in similar fashion as the current three texts.

Table 1. Model Comparison across the Hierarchy of Measurement Models

	χ^2	df	p	CFI	RMSEA	SRMR	$\Delta\chi^2$	$\Delta(df)$	p
Congeneric	0.000	0		1.000	0.000	0.000			
EssTau	48.667	2	<.001	0.975	0.254	0.056	48.667	2	<.001
Tau	175.519	4	<.001	0.907	0.344	0.074	126.852	2	<.001
Parallel	187.028	6	<.001	0.902	0.289	0.077	11.509	2	<.001
Null	1854.234	3	<.001	0.000	1.306	0.551			

Table 2. Latent Variable Summary Statistics across the Hierarchy of Measurement Models

	μ_F	$\Delta(\mu_F)$	σ_F	$\phi(\sigma_F)$	reliability
Congeneric	67.813		33.799		0.984
EssTau	67.813	.000	33.586	.994	0.983
Tau	67.399	-.414	33.600	.994	0.980
Parallel	67.813	.000	33.722	.998	0.980

Note. With the congeneric model acting as baseline, $\Delta(\mu_F) = \mu_F^{[alternative]}/\mu_F^{[baseline]}$ and $\phi(\sigma_F) = \sigma_F^{[alternative]}/\sigma_F^{[baseline]}$.

Table 3 provides a more zoomed-in perspective directly on the item-level. Although the estimated reliability of each item as single-item measure of the latent trait is very high (i.e., $\omega_i > .95$ for all three items), there are some minor item-specific equivalence deviations. The factor loadings are not all equal to one and hence items differ in how a unit difference on the true score scale translates into extra words correctly read, with loadings below/above 1 leading to a contracted/expanded scale unit (i.e., one unit counts for less/more than one word on that item). On average the measurement units can be off by 4%, with item 1 being contracted by about 6% and item 2 expanded by about 5% (Δ in Table 3). In practice, individual differences as measured by item 1 will be somewhat deflated, and for item 2 somewhat inflated. Looking further, the intercepts are not all equal to zero and hence, items also differ in the amount of measurement bias, yet only slightly with about a full word over/estimated for item 1 and 3, respectively.

Table 3. Item-level Effect Size Measures for Measurement Equivalence

	Est	95%CI	Δ	SE	Wald	p
λ_1	0.940	[0.924, 0.956]	-0.060	.009	-6.457	<.001
λ_2	1.053	[1.033, 1.066]	0.053	.010	5.605	<.001
λ_3	1.006	[0.993, 1.027]	0.006	.010	0.634	0.526
v_1	0.923	[-0.444, 2.289]		.697	1.323	.186
v_2	0.096	[-1.312, 1.503]		.718	0.133	.894
v_3	-1.018	[-2.485, 0.449]		.748	-1.361	.174

Note. The null hypothesis for the Wald test assumes $\lambda_i = 1$ and $v_i = 0$, for loadings and intercepts, respectively. Item-specific reliability (communality) ω_i with 95%CI: $\omega_1 = .955[.943,.967]$, $\omega_2 = .960[.949,.972]$, and $\omega_3 = .948[.935,.961]$.

Table 4 translates these item-specific results into expected score differences between item pairs for persons of different latent proficiency level F. For low-ability persons located 2 standard deviations below the factor mean, the expected differences in score between any pair of the items does not exceed 2 words. For average-ability persons, score differences vary between 2.54 and 6.84 words depending on the item pair that is compared. With the standard deviation of item-specific residuals σ_{ε_i} being estimated at approximately 7 words for each item, these expected score differences are still within the anticipated margins of error. Only for high-ability persons, score differences between item pairs were on average expected to be about 7 words up to 14.47 for the difference between item 2 and 1. When considered from this item-level perspective, the equivalence issue appeared more a matter of differences in measurement units (i.e., loadings) than of measurement bias (i.e., intercepts), as reflected in the larger expected differences in words read correctly for more able compared to less able persons. This in contrast to the prior conjecture made on the comparison of the size of the fit indices differences across the model hierarchy.

Fit indices are in general hard to interpret (e.g., van Laar & Braeken, 2022), whereas differences in the actual metric of the items, number of words read correctly (within a minute) seem the more natural effect size quantity to inspect here. Note that the impact of the minor equivalence deviations in our specific application context of oral reading fluency, with pairwise score differences being more outspoken at high ability levels, is not illogical as these are also the children that would read further through each text, which creates more opportunities for small idiosyncrasies to occur while reading.

Table 4. Pairwise Differences in Expected Observed Item Scores

Person's F	$E(Y_1) - E(Y_2)$	$E(Y_1) - E(Y_3)$	$E(Y_2) - E(Y_3)$
-2SD	.80	1.93	1.12
Average	-6.84	-2.54	4.30
+2SD	-14.47	-6.99	7.48

Note. Compare the differences relatively to the estimated measurement error standard deviation σ_{ε_i} of approximately 7 words.

Although the parallel measurement model is not the best-fitting model, and indeed statistically significantly deviates from the congeneric model, the evidence provided here would support a claim of approximate measurement equivalence for the items in this set. Granted, it does depend on a value judgement that given the shared metric of 68 words on average with a standard deviation of 34, differences of only a few words between items do not matter in the larger scheme. The major practical implication of agreeing with this inference, would be that we would conclude to be sufficiently comfortable using any of these items as interchangeable subsets of indicators of oral reading fluency (except perhaps for the high proficient participants, where differences can grow beyond measurement error bounds, see Table 4). A cautious disclaimer is warranted; the choice of items for measurement is not interchangeable if exact scores for single individuals are crucial as for instance in high-stakes clinical diagnosis contexts, because then minor equivalence deviations across items could potentially still have large consequences for decisions that are taken about a specific to-be-measured individual. In our specific example, where expected score differences due to minor equivalence deviations are more outspoken for high-ability persons, this would be less of a concern, as generally speaking practitioners would be more concerned when it would affect low-ability participants (e.g., higher stakes involved in a special needs education context).

Transparency and Openness

The proposed diagnostic procedure can be implemented in any statistical software supporting structural equation modelling. We have written custom functions in the free software environment for statistical computing and graphics, R (version 4.4.2, R Core Team, 2024) that make use of the popular `lavaan` package (Rosseel, 2012) to facilitate the dissemination of the procedure. A function automatically fits the hierarchy of measurement equivalence models for a specified set of items, summarizes their model fit (cf. Table 1), and provides the proposed diagnostics at latent variable level (cf. Table 2) and at item level (cf. Table 3). The custom functions and code behind the analyses in this manuscript have been made publicly available under a CC BY-NC-SA license at the Open Science framework and can be accessed at https://osf.io/a9n3k/?view_only=679bdbed9c5f483c937c1d31950ee44d.

Discussion

With respect to the different identification rules for latent variable measurement models, we consider the use of a reference item to make conceptual sense when that item is the gold standard measure. Yet in the social and behavioural sciences, the lack of gold standard measures is exactly the most pervasive and longstanding measurement issue. The almost automatic reflex to standardize can be seen as a consequence of the same problem: Lacking a good understanding of the original measurement units, one defers to a relative metric in terms of a standard deviation unit of those measurements (i.e., the extent of individual differences in the sample). We hope that the proposed procedure in this short note can help to put an absolute interpretable metric back into measurement models as this, in our opinion, facilitates practical applications. This does not only concern the assessment of the equivalence of an item set, but also applies to the evaluation of other parameter constraints in more complex measurement model settings in longitudinal studies and multi-group studies.

As with many measurement concepts, assessing equivalence is not an absolute yes/no exercise but more about the relative degree of equivalence. In practice, it might even be unrealistic to expect that any items will be exactly equivalent, but we should be able to assess their degree of equivalence and consider whether this is sufficient for our measurement purposes and context. Evaluating the relevance of cumulative theoretically motivated model restrictions (see also Bentler & Bonett, 1980; van Laar & Braeken, 2022) along the lines of the measurement-equivalence model hierarchy is then a logical way to go.

However, a popular alternative approach with users of structural equation models is a so-called modification model-building strategy. In the current case, this strategy would start from the ideal hypothesized parallel measurement model and then use series of statistical significance tests to decide which parameter constraints need to be relaxed to improve and reach acceptable model fit. SEM-software typically provides so-called modification indices for the purpose of searching for such local model misspecification. These modification indices are Lagrange multiplier tests (Chou & Bentler, 1990), testing for each respective individual parameter whether freeing it from its imposed model constraints leads to a statistically significant difference in the score-gradient of the starting model's maximum likelihood. The latter implies an improvement in model fit when the parameter would no longer be constrained. Despite its popularity, the challenges with such model-building approach are well-documented and include among others multiple testing problems, sensitivity to order of modification in the series of models being tested, and inconsistent results in significance outcomes for individual parameters across different steps (for a more extensive discussion, see e.g., MacCallum, 1986; MacCallum et al., 1992). Furthermore, Saris et al. (1987) advocated for the use of an estimate of the expected parameter change related to relaxing an individual parameter from its model constraints, supplementing or even replacing the significance tests, this to reduce the inherent risk of overinterpreting statistically significant but practically non-significant trivial parameter value differences in a modification indices approach. The latter expected parameter change statistics are closer in principle to the item-specific effect-size diagnostics we advocate for in this paper, but are in practice still somewhat challenging as to interpret values of parameters the metric of the latent variable would need to be clear.

Thus, it is clear that significance tests should be supplemented by informative diagnostics and effect sizes (Kirk, 1996). The effect sizes in terms of the natural metric of the item scores and the impact on the key latent variable being measured, as proposed in the current manuscript, should prove informative for this purpose. The caveat remains that this does assess quantitative measurement equivalence, but does not make any inferences on other properties that might be relevant in this context, such as perhaps content coverage and cross-cultural invariance. The effect coding identification rule does bring back the original measurement units into the picture, but does make less sense conceptually when all the individual items are not measured on the same response scale to begin with.

We caution against proceeding to assess measurement equivalence in a fully exploratory fashion across items that have not been designed with this in mind from the start. The metric of the latent factor is determined by the set of items that are simultaneously considered. By definition, its true metric is unknown, it is a latent variable after all; yet the derived shared metric across the item set should serve as a good approximation when items are at least a priori designed to be equivalent. A shared metric can also be derived for a random set of items, but that would obviously be less meaningful.

Furthermore, lacking a core of approximate equivalent items, makes it difficult conceptually to grasp what equivalence would entail but also makes it difficult inferentially as measurement model parameters and related indices might jump into different directions depending on whether specific items are being included or excluded. As soon as the assumed-to-be-equivalent item set consists of more than 3 items, a leave-one-out crossvalidation procedure would be possible that can help diagnose the influence that a single item has on the shared latent factor metric. With more items, these types of cross-validation diagnostics can be further expanded by leaving out complete subsets of items that are suspected to be inequivalent. However, note that such specification searches, trying to iteratively eliminate non-equivalent and retain equivalent items, are not necessarily guaranteed to converge to a unique and optimal solution. Hence, specification searches for reducing a heterogeneous item set to a smaller 'purified' set of equivalent items should be done cautiously, similar to when using modification indices in SEM for model building (e.g., MacCallum, 1986; MacCallum et al., 1992).

A last point of discussion to bring up is how this identification rule fairs in the case of categorical item response data where a natural continuous metric unit is unavailable. To start, it should be noted that pure technically speaking, nothing stands in the way of applying this effect-coding identification rule to latent variable models for categorical data (e.g., IRT, Lord & Novick, 1968), it is simply one of many possible equivalent ways of identifying the scale of a latent variable. The absence of a natural metric for the latent scale does imply that the effect-coding identification rule that we advocate for, does lose some of its interpretational charm. Yet, the effect-coding as implied by the constraints in Equation 6 and Equation 5 still contrasts individual items to a grand mean metric scale, only the exact meaning of units is less clear than in the continuous scale as there is no physical unit to correspond to as in the continuous case (cf., number of words in the oral reading fluency example). For instance, while a binary correct/wrong item response has no actual continuous physical metric at the item response level, the item difficulty in a 2-parameter logistic item response model would now become a relative item difficulty stating whether the item is more/less difficult than a hypothetical average item in the item set; Similarly, the item discrimination would be less/more discriminating than a hypothetical average item. Such an interpretation of parameters does relate to the philosophical stance that the magnitude of an object can only be defined in relative comparison (Royall, 1997), and not absolutely. For multi-categorical item responses, such as Likert items rated for instance on a scale from "strongly disagree" over "disagree" to "agree" and "strongly agree", there is also no natural physical metric at the item response level and the additional complexity of psychological distance between response categories surfaces (Wakita et al., 2012). Does an individual regard the distance between the strongly disagree and disagree categories as equal to the distance between the other pairs of consecutive categories? In practice, people tend to make abstraction of the implicit assumptions of equal psychological category distances and might treat the Likert items as continuous. Whether one is comfortable with such an approach leads to an ongoing and likely never-ending debate in both data-analysis as well philosophy of science in the measurement field.

Declarations

Disclosure statement

We have no potential conflict of interest to disclose.

Acknowledgements of support

This study was supported by a research grant [FINNUT342925] of the Norwegian Research Council.

Received: 3/27/2025. **Accepted:** 2/7/2026. **Published:** 2/16/2026.

Citation: Braeken, J., & van Laar, S. (2026). Assessing the measurement equivalence of a set of items: Item-specific diagnostics on an interpretable metric. *Practical Assessment, Research, & Evaluation*, *31*(1)(6). Available online: https://doi.org/10.7275/pare.3048

Corresponding Author: Johan Braeken, Centre for Educational Measurement (CEMO), University of Oslo. Email: johan.braeken@cemo.uio.no

References

Arnesen, A., Braeken, J., Baker, S., Meek-Hansen, W., Ogden, T., & Melby-Lervåg, M. (2017). Growth in oral reading fluency in a semitransparent orthography: Concurrent and predictive relations with reading proficiency in Norwegian, grades 2–5. Reading Research Quarterly, 52, 177–201. https://doi.org/10.1002/rrq.159

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88(3), 588–606.

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. Computers and Education: Artificial Intelligence, 5, 100161. https://doi.org/10.1016/j.caeai.2023.100161

Chou, C.-P., & Bentler, P. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. Multivariate Behavioral Research, 25(1), 115–136. https://doi.org/10.1207/s15327906mbr2501_13

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Educational and Psychological Measurement, 16, 297–334. https://doi.org/10.1007/BF02310555

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. Scientific Studies of Reading, 5(3), 239–256. https://doi.org/10.1207/ S1532799XSSR0503_3

Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. Psychological Methods, 29(3), 494–518. https://doi.org/10.1037/met0000540

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

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56(5), 746–759. https://doi.org/10.1177/ 0013164496056005002

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. Journal of Educational Statistics, 4, 269–290.

Little, T. D. (2013). Longitudinal structural equation modeling. Guilford Press.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. Structural Equation Modeling: A Multidisciplinary Journal, 13(1), 59–72. https://doi.org/10.1207/s15328007sem1301_3

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

MacCallum, R. (1986). Specification searches in covariance structure modeling. Psychological Bulletin, (1), 107–120. https://doi.org/10.1037/0033-2909.100.1.107

MacCallum, R., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. Psychological Bulletin, (3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Marsh, H. W. (1998). The equal correlation baseline model: Comment and constructive alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 5(1), 78– 86. https://doi.org/10.1080/10705519809540090

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. British Journal of Mathematical and Statistical Psychology, 23(1), 1–21. https://doi.org/10.1111/j.2044-8317.1970. tb00432.x

Novick, M. R. (1966). The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

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

Royall, R. (1997). Statistical evidence: A likelihood paradigm. Chapman & Hall.

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. Sociological Methodology, 17, 105– 129.

Spearman, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology, 15(1), 72–101.

van der Velde, M., Molenaar, B., Veldkamp, B. P., Feskens, R., & Keuning, J. (2024). What do they say? assessment of oral reading fluency in early primary school children: A scoping review. International Journal of Educational Research, 128, 102444. https://doi.org/10.1016/j.ijer.2024.102444

van Laar, S., & Braeken, J. (2022). Caught off base: A note on the interpretation of incremental fit indices. Structural Equation Modeling: A Multidisciplinary Journal, 29(6), 935–943. https://doi.org/10.1080/10705511.2022.2050730

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert scale: Comparing different numbers of options. Educational and Psychological Measurement, 72(4), 533–546. https://doi.org/10.1177/0013164411431162