Educational Research and Evaluation: An International Journal on Theory and Practice

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nere20

The formalization of fairness: issues in testing for measurement invariance using subtest scores

Dylan Molenaar [a] & Denny Borsboom [a]

[a] University of Amsterdam , Amsterdam , The Netherlands

Published online: 18 Mar 2013.

To cite this article: Dylan Molenaar & Denny Borsboom (2013) The formalization of fairness: issues in testing for measurement invariance using subtest scores, Educational Research and Evaluation: An International Journal on Theory and Practice, 19:2-3, 223-244, DOI: 10.1080/13803611.2013.767628

To link to this article: http://dx.doi.org/10.1080/13803611.2013.767628

Educational Research and Evaluation, 2013
Vol. 19, Nos. 2–3, 223–244, http://dx.doi.org/10.1080/13803611.2013.767628

The formalization of fairness: issues in testing for measurement invariance using subtest scores

Dylan Molenaar* and Denny Borsboom

University of Amsterdam, Amsterdam, The Netherlands

Measurement invariance is an important prerequisite for the adequate comparison of group differences in test scores. In psychology, measurement invariance is typically investigated by means of linear factor analyses of subtest scores. These subtest scores typically result from summing the item scores. In this paper, we discuss 4 possible problems related to this common practice. Specifically, we discuss (a) nonlinearity of the latent variable to subtest relation; (b) suboptimality of the total score as a proxy for the latent variable measured through the item scores; (c) non-normality of the subtest score; and (d) differences in the nature of the latent variable at the item level as compared to the latent variable at the subtest level. Additionally, we give guidelines to overcome these problems and illustrate the issues by analysing data that pertain to a performal IQ data set.

Keywords: factor analysis; item response theory; measurement invariance; differential item functioning; sum scores; non-normality

Introduction

In psychology and educational assessment, researchers often want to make inferences regarding properties like working memory, mood, extraversion, arithmetic ability, and perceptual organization. As researchers cannot observe these properties directly, they typically focus on observable indicators that are assumed to be determined by the properties of interest. For example, if we want to measure perceptual organization, we may consider the performance of a sample of subjects on the Block Design and Matrix Reasoning subtests of the Wechsler Adult Intelligence Scale III (WAIS-III; Wechsler, 1997), and take performance differences to reflect differences in perceptual organization (Edwards & Bagozzi, 2000).

Formal models designed to represent and test such measurement hypotheses are generically known as latent variable models (Bartholomew, 1987; Borsboom, 2005, 2008). The parameters of reflective measurement models represent measurement characteristics of the test or items, that is, specify how it relates the latent variables of interest to the indicators. Various measurement models are available; these differ in the structure of the latent and observed variables (e.g., categorical or continuous; see Batholomew, 1987; Mellenbergh, 1994a) and in the type of relation that connects them (e.g., monotonic, as in Ellis & Junker, 1997, or linear, as in Jöreskog, 1971). Popular measurement models include the linear factor model (Spearman, 1904, 1927; Thurstone, 1947), which links continuous item data to a continuous latent variable (Mellenbergh, 1994b), and the two-parameter logistic model (2PL; Birnbaum, 1968), which links dichotomous item data to a continuous latent variable.

*Corresponding author. Email: D.Molenaar@uva.nl

As Lord (1980) suggested, meaningful interpretations of observed mean differences, in terms of the latent variables in the model, are facilitated if the same measurement properties hold in the samples under consideration. For a parametric model, this implies that the parameters in the measurement model are the same across groups. This idea was developed by B. O. Muthén (1989) and especially Mellenbergh (1989) and Meredith (1993) to construct the theoretical framework of measurement invariance. Measurement invariance holds if and only if all subjects with the same position on the latent variable have the same observed score distribution, regardless of their group membership. This idea is completely general and applies to all kinds of measurement models, including nonparametric item response theory (IRT) models and latent class models. If group differences in observed scores exist conditional on the latent variable, however, the scores display violations of measurement invariance, also known as differential item functioning (DIF) in the IRT literature. Although formally speaking these concepts designate the same property, in this paper we will generally speak of DIF when discussing individual items, and of measurement invariance with respect to subtests, since this terminology has become common in the literature (see Millsap, 2011). Violations of measurement invariance are designated as bias if they introduce unwanted distortions in the inferences made in the measurement process (this need not be the case in all settings; e.g., see Borsboom, 2006).

In the case of DIF or lack of measurement invariance, observed differences between groups do not necessarily reflect differences on the latent variable. In this case, it is likely that unintended latent variables have contributed to the measurement process, which may implicate validity problems (Messick, 1989). For instance, the "Information" subtest of the WAIS is consistently found to be biased in favour of males (e.g., Dolan et al., 2006; Jensen & Reynolds, 1983; Van der Sluis, Posthuma, et al., 2006). Such bias could possibly be due to the overrepresentation of questions that require solving physics problems. If so, the average male advantage on this subtest is not due to a higher verbal ability – the latent variable that the information subtest is purported to measure – but due to more knowledge about physics (since, on average, males tend to favour courses like physics more often than females do in secondary school). In addition to unintended latent variables that are measured by a test or item, bias could be introduced by differences in research settings that exist across groups. For instance, Wicherts, Dolan, and Hessen (2005) showed that bias can be introduced by altering test instructions in different groups. Specifically, in a sample of female students, an arithmetic test was preluded by stating that females do less well on arithmetic tests as compared to males, while in another sample of female students this statement was omitted. Wicherts et al. found that the different instruction lowered the average score of the former group, indicating that bias was introduced in the former group.

The above definition of test bias has not been uncontroversial in the literature. Where measurement invariance denotes equality of measurement models across samples, an alternative definition given by Cleary (1968) states that a test is biased when the regression of an external criterion (e.g., grade point average) on the test score (e.g., an IQ score) diverges across groups. This definition has been used – or is being used – in influential guides including the Standards for Educational and Psychological Testing (see Borsboom, Romeijn, & Wicherts, 2008). In addition, predictive invariance has been used as operationalization of test bias in a number of (empirical) studies (e.g., Aguinis & Smith, 2007; Evers, Te Nijenhuis, & Van der Flier, 2005; Gamliel & Cahan, 2007; Hunter & Schmidt, 2000; Neisser et al., 1996; Rushton & Jensen, 2005; Sackett, Borneman, & Connely, 2008; Sackett, Schmitt, Ellington, & Kabin, 2001). However, as argued in multiple sources (Borsboom et al., 2008; Millsap, 1997, 1998, 2008; Wicherts & Millsap, 2009),

predictive invariance does not necessarily imply the absence of test bias. Specifically, when predictive invariance holds, severe bias can be present (e.g., Wicherts & Millsap, 2009). Borsboom et al. (2008) concluded that measurement invariance should be preferred over predictive invariance to test for item or test bias because (a) predictive invariance implies a violation of bias as defined by Mellenbergh (1989), which is unacceptable, whereas measurement invariance does not violate this definition; (b) tests on predictive invariance are ambiguous as they depend on the choices concerning the criterion variables; (c) predictive invariance can lead to opposite results when the causal direction between the criterion and the test scores are reversed (i.e., when the test is used as criterion and the criterion is used as test). In this paper, we therefore espouse the definition of bias in terms of lack of measurement invariance or – similarly – the presence of DIF.

Statistical procedures to test for the presence of measurement invariance are well developed (see Millsap, 2011, for an overview). Traditionally, tests on measurement bias originate in factor analysis (Meredith, 1964; Thomson & Lederman, 1939; Thurstone, 1947). The linear factor model is a measurement model that links a continuously distributed latent variable to a set of continuous observed variables. In psychology, the linear factor model is arguably the most widely used method to test for item bias in case of power tests like intelligence and ability tests. For power tests, data are generally dichotomous (scored 1: correct and 0: incorrect). Therefore, the linear factor model is not strictly suitable for these data. Researchers have, however, focused on the analysis of summed item scores, or item parcels, to make the data more suitable for factor analysis. Reasons for using factor analysis lie primarily in the large number of subtests and items that are featured in typical power tests, which make item-level tests more difficult unless samples are very large. Consider for instance the WAIS. The WAIS-III (Wechsler, 1997) contains 14 subtests, each of which consists of 40 to 60 items. An item-level measurement model like the 2PL model would require a second-order structure with 14 first-order latent variables and 4 second-order latent variables. Fitting such a model and comparing its parameters across groups is computationally infeasible. By taking sum scores for each subtest, and using the linear factor model for the analysis, the model simplifies to a linear model with only 4 first-order latent variables, which is numerically less demanding and can be fitted using widely available software. Thus, there may be a reasonable justification for using factor models in such cases, even though one knows that they cannot strictly be correct.

Although the procedure of taking summed item scores or item parcels is valuable from a practical point of view, several problems arise in applications of the linear factor model to summed item scores when testing for measurement invariance. In the present paper, we analyse these problems and suggest ways to address them. Specifically, we discuss problems associated with (a) nonlinearity of the latent variable to subtest relation, (b) suboptimality of the sum score as a proxy for the latent variable underlying the item scores, (c) non-normality of the subtest score, and (d) differences in the nature of the latent variable as it applies to the item level, as compared to the latent variable that applies to the subtest level. We provide guidelines to evaluate whether factor analysis can be applied safely and whether alternative methods should be considered, such as methods based on item response theory (IRT, e.g., the 2PL model).

An outline of this paper is as follows: We first introduce the linear factor model and discuss how a typical test on measurement invariance is executed. Next, we discuss problems with multi-group factor analyses of total scores. We then illustrate our main points on a dataset that pertains to performal intelligence. We end with some general recommendations with respect to tests for measurement invariance in power tests.

Linear factor analysis and measurement invariance

Linear factor analysis is a statistical method to investigate the hypothesized factor structure for a set of observed variables. In factor analysis, the observed data are linearly regressed on a set of latent variables through a set of factor loadings (i.e., the regression weights), intercepts, and residuals. In doing so, the covariance matrix of the data is modelled in terms of the factor loadings, the residual variances, and the (co)variances of the latent variables. In addition, the mean vector of the observed data can be modelled in terms of the factor loadings, the intercepts and – in multi-group applications – the means of the latent variables.[1] Note that in confirmatory factor analysis – as discussed in this paper – the structure of the factor loadings is typically assumed to be given. That is, it is assumed that we have prior information that can be used to determine which items load on (i.e., are regressed on) which of the common factors. To investigate whether the hypothesized factor model gives an adequate description of the data, various model fit indices can be consulted (see, for an overview, e.g., Schermelleh-Engel, Moosbrugger, & Müller, 2003).

Measurement invariance involves the sequential application of increasingly restrictive factor models to the observed data of multiple groups (Meredith, 1993; see, e.g., Horn & McArdle, 1992; Widaman & Reise, 1997). To enable application of these models, a multivariate normal distribution is assumed for the data (Meredith, 1993), implying a normal distribution for all subtests. The models that subsequently are fitted to the data are described next.

The first step, *configural invariance*, involves the question whether the same factor structure holds within each group under investigation. To this end, the hypothesized factor structure is fitted to the data of all groups and goodness of fit is established. Note that all parameters of the factor model are allowed to vary between groups. The only restriction in place is that the factor loading configuration is equal across groups. The next step, *metric invariance*, involves the question whether the factor loadings are invariant across groups. To this end, factor loadings are constrained to be equal across groups, and the deterioration of model fit is established. When measurement invariance holds, model fit should not deteriorate significantly and goodness-of-fit measures (like the root mean square error of approximation [RMSEA] and Akaike's information criterion [AIC], which take into account model parsimony) should indicate that the model does not fit worse than the preceding model.

Next, *equal residual variances* are specified between the groups (Lubke & Dolan, 2003). Again, when measurement invariance holds, this model should be favoured by the model fit indices. The final step is to test for *strict measurement invariance* (Meredith, 1993). In this model, the intercepts are equated across groups, and the factor means are freed in all groups but an arbitrary reference group, allowing for factor mean differences across groups (Sörbom, 1974). To establish measurement invariance, one needs to justify the conclusion that the strict factorial invariance model is the best fitting model among all models considered, in terms of the model fit criteria. The strict factorial invariance model is the only factor model among those considered above that guarantees measurement invariance (Meredith, 1993). All of the other models considered are consistent with violations of measurement variance and hence do not guarantee accurate inferences from observed differences between groups to latent differences.

In the ideal case, each model fits better than the preceding model in terms of the fit indices like the RMSEA, or the model fit does not deteriorate significantly, in terms of a likelihood ratio test. In such cases, the strict measurement invariance model may be preferred over the less parsimonious models (Dolan, 2000). In the case that a model fits

worse than the preceding model, modification indices could be inspected, which could shed a light on the causes of misfit. If misfit is limited to a few parameters, these could be freed across groups, and valid comparisons may still be made on the basis of the subset of items or subtests that do satisfy measurement invariance. In case of too many parameters that violate invariance, measurement invariance should be rejected. Importantly, this does not disqualify the use of tests for all purposes; for instance, one could still legitimately compare the sign of the correlation between factors across groups as long as the relation between latent and observed variables is monotonic in both groups. However, inferences to latent mean differences from observed mean differences are typically jeopardized in the presence of bias, as in the use of tests for selection purposes that involve members of different groups.

Possible confounds when testing for bias on summed item scores

The most important measurement instruments in psychology are power tests, like intelligence and ability tests, and self-report questionnaires, like mood and personality questionnaires. Items of a power test typically result in dichotomous item responses (correct vs. incorrect). As a result, naturally appropriate measurement models for data gathered with power tests are IRT models like the Rasch model (Rasch, 1960) or the 2PL model (Birnbaum, 1968). In IRT models, however, the relation between the observed item scores and the latent variable is modelled by an S-shaped curve (generally a logistic function or a normal ogive function). This function is called the item characteristic curve (ICC) and, in the case of the Birnbaum (1968) model, is characterized by two parameters: item discrimination (the slope of the curve) and item difficulty (the location of the mid-point of the curve on the latent variable continuum). In case of the Rasch model, all item discriminations are assumed to be equal.

Thus, here we advocate the use of IRT models like the multi-group Rasch model and the multi-group 2PL model to test for measurement invariance. We note, however, that alternative methods are available that are appropriate for discrete item-level data but that do not depend on IRT. These methods include, for instance, the Mantel-Haenszel approach (MH; Mantel & Haenszel, 1959; see, for related approaches, Dorans & Kulick, 1986; Shealy & Stout, 1993) and the logistic regression approach (Swaminathan & Rogers, 1990). A more elaborate overview can be found elsewhere (Millsap, 2011; Osterlind & Everson, 2009; Penfield & Camilli, 2007). Most of the methods, like the MH and the logistic regression approach, test for an association between group membership and item response using the sum score to correct for differences on the underlying construct. As the sum score plays a key role in these procedures, most of these alternative methods are vulnerable to the same criticism as below. On the contrary, no parametric assumptions need to be imposed (e.g., latent variable distributions and local independence).

Thus, as discussed above, in practice, researchers rely heavily on the sum score in testing for measurement invariance. That is, all items of the same subtest are summed and analysed as continuous variables in the linear factor model. In such cases, the investigator purposefully misspecifies the measurement model, which may lead to various problems.

Nonlinearity of the latent variable to subtest relation

An important problem is pointed out by Tucker-Drob (2009) and illustrated in Figure 1. Figure 1 shows the relationship between the latent variable and the total score on 20, 40, and 60 items that follow a Rasch model. Three scenarios are considered: (a) the difficulties

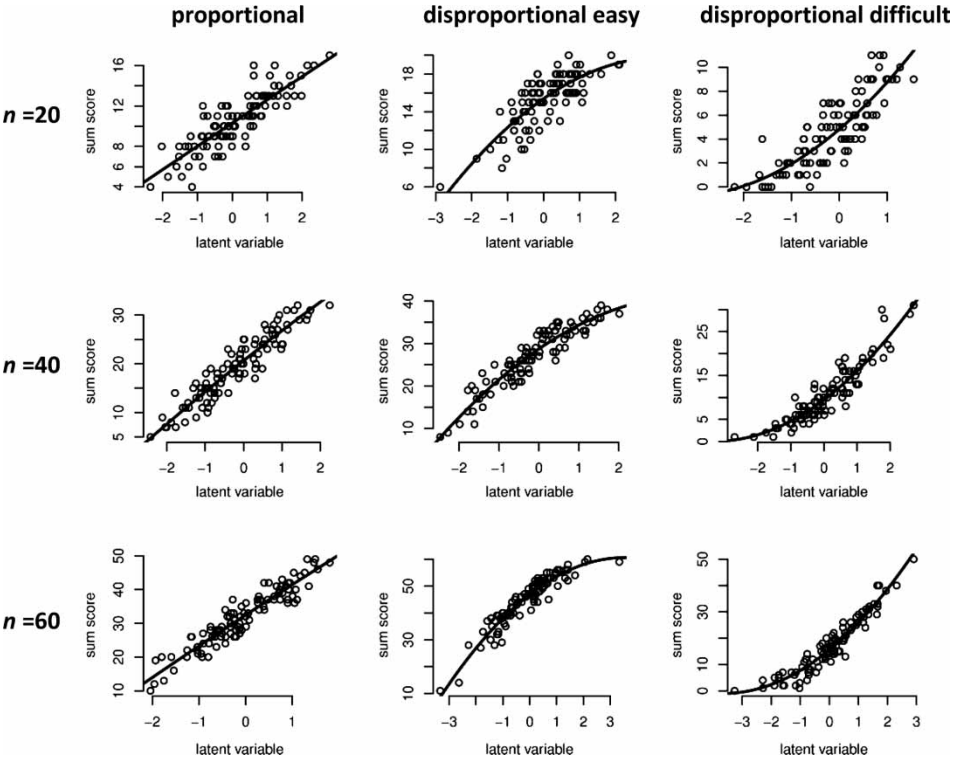

Figure 1. Relation between the summed item scores and the latent variable when the items follow a Rasch model where the item difficulties are (a) proportionally varied, (b) have a disproportional number of difficult items, and (c) have a disproportional number of easy items. Cases are depicted for 20, 40, and 60 items.

of the items are proportionally varied (equal amount of difficult and easy items); (b) there are a disproportional number of difficult items; and (c) there are a disproportional number of easy items. As can be seen, in the first scenario the function approximates a linear function reasonably well (although strict linearity is not attainable due to the fact that the total score is bounded from below and from above). For the second and third scenarios, however, the relation between the sum score and the latent variable becomes strongly nonlinear due to a floor and a ceiling effect, respectively. Such nonlinearity is problematic for the application of factor analysis to testing measurement invariance. For instance, Bauer (2005) showed analytically that, when an invariant nonlinear relation between the observed and latent variables underlies the data, but the linear factor model is applied, tests on MI will be distorted. As a result, the factor loadings and intercepts will diverge across groups; see Bauer (2005) for a detailed explanation of the expected parameter values in the groups under nonlinearity.

An important implication of Figure 1 and the results by Bauer (2005) is that, in the analysis of total scores, spurious nonlinearity can arise due to over- or underrepresentation of difficult and easy items, and that this nonlinearity can result in wrongfully rejecting measurement invariance. In very large samples, measurement invariance would in fact be expected to be generally violated: Because the relation between total score and latent variable cannot be strictly linear as a matter of principle, rejecting measurement invariance is guaranteed as the sample size approaches infinity. Severe violations of linearity also occur when dichotomous items have different discrimination parameters, as can be seen

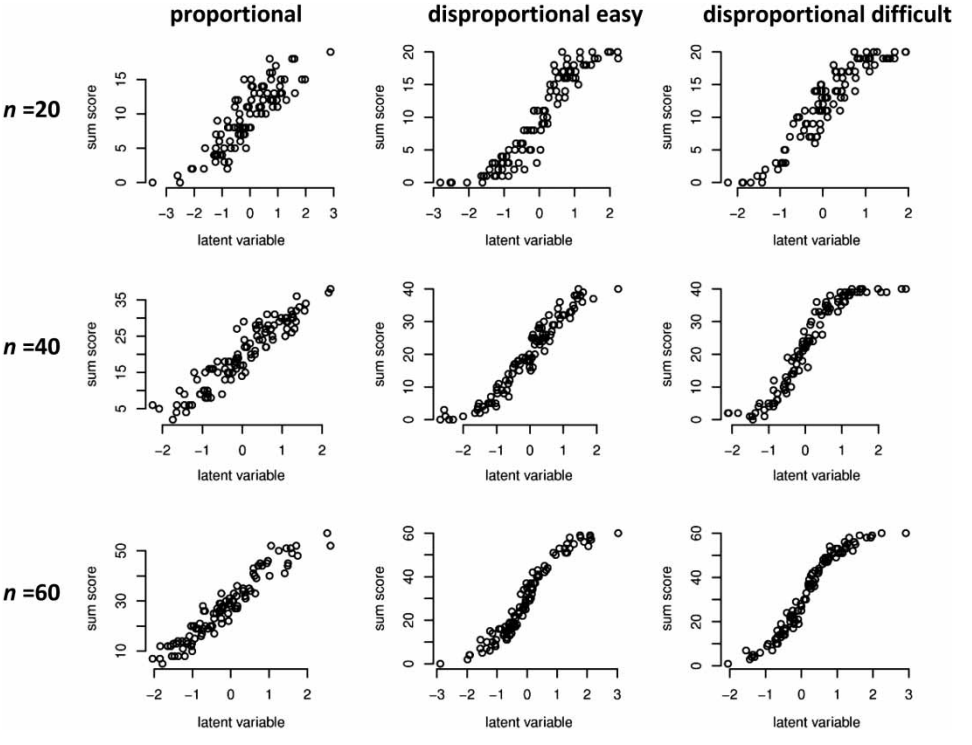

Figure 2. Relation between the summed item scores and the latent variable when the items follow a two-parameter logistic model where the item discriminations (a) do not differ between items, (b) differ mildly, and (c) differ substantially across items. Cases are depicted for 20, 40, and 60 items.

in Figure 2. In that case, the relation between the latent variable and the summed item scores has an S shape. Thus, nonlinearity can arise when a sample is characterized by restriction of range, for example, in case of the application of an intelligence test to a sample of highly educated subjects. In Figure 2, subjects will be mainly positioned on the upper end of the latent variable, which will typically result in stronger nonlinearity.

The total score as a proxy for the latent variable

In executing measurement invariance tests at the total score level, one basically assumes that the relevant total score is a good proxy for a first-order latent variable that is measured directly by the items of a subtest. Thus, when one enters a total score into a factor analysis routine, one typically assumes that the items themselves are unidimensional. To the extent that this is not the case, violations of measurement invariance at the subtest level are very hard to interpret, because they may arise either as a result of item bias or because of erroneous assumptions regarding the dimensionality of the item scores.

Ideally, one would in fact have a Rasch model that holds at the item level, because in this case the unweighted total score is a sufficient statistic for the latent variable. This means that, conditional on the total score, the item scores are independent of the latent variable, so that the total score contains all relevant information with regard to the latent variable distribution (Andersen, 1973). However, when the Rasch model is violated, summing item scores will discard some information about the discriminatory ability of the items. A

weighted sum score, where the item scores are weighted by their discriminatory power, may in this case provide a better proxy, as this weighted score would be a sufficient statistic for the latent variable if the weights were known (see e.g., Verhelst & Glas, 1995). However, discriminatory power of the items, which regulates their weights in computing the weighted score, is a parameter that needs to be estimated. In addition, a weighted total score is still suboptimal, as estimated discrimination is only an approximation to the true discriminatory power of an item.

A practical solution is not to estimate weights at all, but instead rely on the robustness of the total score, which is guaranteed to be monotonically related to the latent variable provided all the individual items are (i.e., has monotone likelihood ratio; see Grayson, 1988). However, although these results guarantee certain regularity properties of comparisons of individuals that belong to the same population (specifically, the expected latent variable score is strictly higher for people with higher total scores), they do not guarantee that the ICCs are equal across groups and hence do not guarantee measurement invariance – especially not when a misspecified factor model is fitted. A related issue in using the total score is that a lack of measurement invariance on the item level, that is, DIF, is not necessarily expressed on subtest level. It is possible that DIF cancels out at the level of the total score (Borsboom, 2005), or that the power to detect DIF on the subtest level is small. Thus, if DIF is present at item level, it need not be present at the subtest level.

The exact relation between measurement invariance at the item level and at the subtest level is therefore not as simple as one may think. Even if a unidimensional IRT model holds across groups, this does not guarantee that the total score will have an invariant relation to the latent variable, particularly if the groups differ in the location or variance of the latent distribution. Further research into this issue would be useful; however, it would seem safe to say that a minimum requirement for testing measurement invariance using subtest scores is evidence that the items of the relevant subtests are themselves unidimensional. If this is not the case, then all bets are off with respect to measurement invariance tests through applications of the linear factor model.

Non-normality of the subtest scores

In essence, summed dichotomous item scores are count data. Count data are seldom characterized by a normal distribution, because the data are bounded from above and from below. More appropriate distributions for count data include the Poisson distribution or the lognormal distribution. In single-group applications, it is known that non-normality in the data causes biased parameter estimates and goodness-of-fit measures of the factor models (Curran, West, & Finch, 1996).

In case of testing for measurement invariance, this could be problematic, especially when departures from normality differ between the groups under investigation. For instance, an invariant measurement model may underlie the data of two groups, but due to a mean difference on the latent variable, the total score distribution could be less normal in the high-scoring group as compared to the low-scoring group. This is illustrated in Figure 3, where total scores are depicted for a reference group (Group 1) and an advantaged group (Group 2) which features a higher location of the latent variable distribution. Note again, that in this case, measurement invariance holds across the groups.

As can be seen, the shape of the distributions differs across the groups, which is due to the fact that the sum score is bounded between zero and the number of items. Therefore, the advantage of Group 2 over Group 1 does not simply result in a shift of the distribution across the x-axis, but instead results in a distribution of a different shape. This may

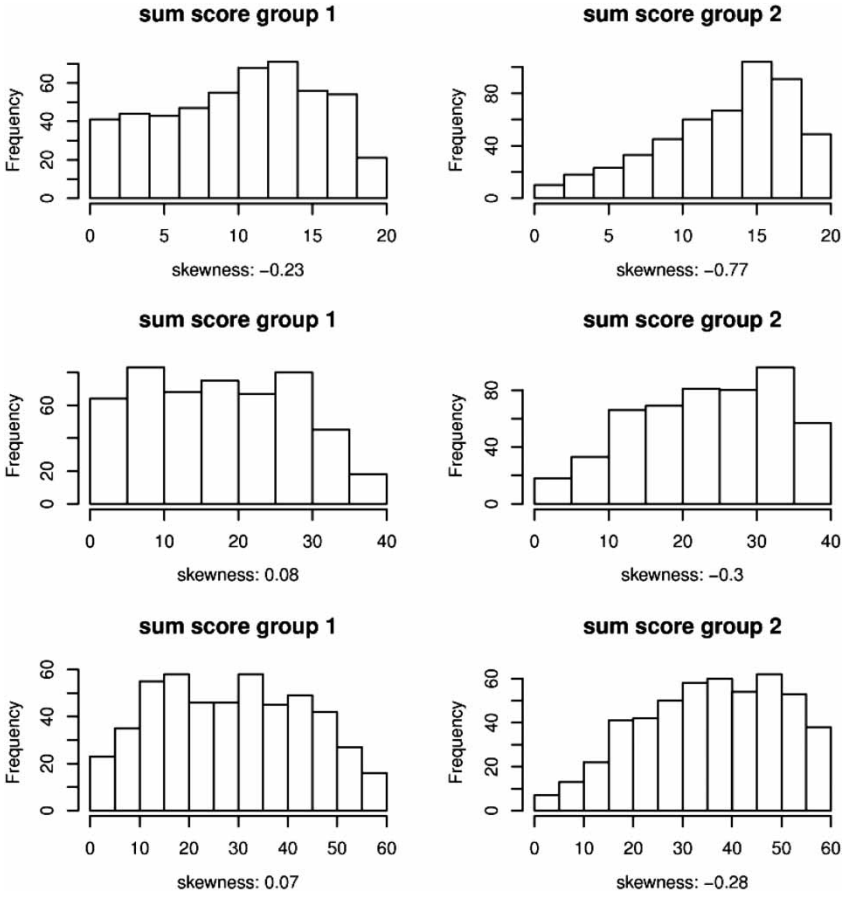

Figure 3. Summed item score distribution in two groups for 20, 40, and 60 items, where Group 2 has mean advantage on the latent variable as compared to Group 1.

distort the fit of the linear factor model (as pointed out by Curran et al., 1996) to a different extent in both groups, resulting in biased results.

Different nature of the latent variables

When testing for measurement invariance in the linear factor model, multiple subtest scores are needed, as a single subtest cannot be tested for bias. As a consequence, multiple subtests are used to measure the same (more general) second-order construct. For instance, if we have an IQ test battery, and we are interested in testing for measurement invariance for the Information subtest of the WAIS, we also need to consider other subtests of the same domain, for example, Vocabulary and Comprehension. As a result, we test the subtest Information for bias with respect to a general latent variable like verbal intelligence. If we had conducted the analysis on item level, that is, if we had submitted the item scores of the Information subtest to a DIF analysis, we would have tested this subtest for bias with respect to the latent variable "general knowledge".

Many researchers would say that these tests answer somewhat different questions, as "general knowledge" is part of, but certainly not equal to, "verbal intelligence". This is

also illustrated in Figure 4. The circles represent subscales that consist of items. Overlap between circles denotes that the respective subtest scores share variance. Now, consider the subscale that is represented by the upper left circle. If we test this subtest for bias at the item level, we investigate whether we can fairly compare subjects on the basis of the striped area, that is, the part that represents the variance of only one of the subtests. If we test for bias at the subtest level, however, we need to take into account the other three subtests, and we are actually investigating whether we can compare subjects on the basis of the black area, that is, the variance that is common to all subtests.

Thus, unless the items are unidimensional across all subtests (i.e., a unidimensional model would fit the responses to all items analysed jointly), measurement invariance with respect to the item level is not the same as measurement invariance at the subtest level: The latent variable that is conditioned on in the definition of measurement invariance (Meredith, 1993) is not identical in both cases. This is consistent with the fact that measurement invariance is a property of test scores rather than of tests; however, in many cases researchers use the invariance properties of subtest scores to say something about the fairness of the test in general. Since any test is essentially a collection of items, it is not entirely clear to what extent such generalizations should be considered to be supported.

Illustration

To illustrate the above points, we analysed both item-level data and subtest-level data of four performal subtests of the Dutch version of the *Intelligenz Struktur Test* (IST; Amthauer,

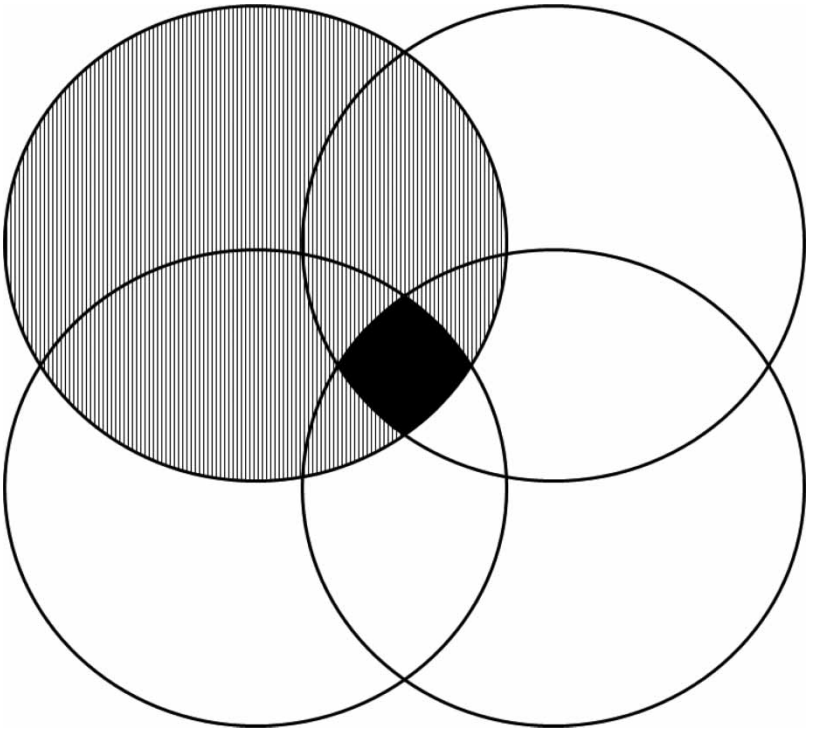

Figure 4. Schematic representation of the nature of the latent variables in an item analysis and in a subtest analysis. Circles represent variance in subtests; overlap represents a correlation between subtests.

Table 1. Correlations for the four subtests for males and females (males below diagonal).

	FC	DR	AR	MR
FC	1.00	0.33	0.06	0.39
DR	0.47	1.00	0.08	0.36
AR	0.18	0.13	1.00	0.10
MR	0.40	0.39	0.14	1.00

Brocke, Liepmann, & Beauducel, 2001). These subtests concern: Figure Completion (FC), Dice Rotation (DR), Arithmetic Reasoning (AR), and Matrix Reasoning (MR), consisting of 20 items each (scored correct: 1, incorrect: 0), and were completed by 1,473 psychology freshman for course credit (428 males; 1,045 females). For the subtest AR, we omitted the first two items, as none of the subjects answered these items incorrectly. See Table 1 for the correlations among the subtest scores and Table 2 for the means and standard deviations.

On these data, we first conducted a nonlinear factor analyses to assess whether we could assume a linear measurement model for the sum scores. Next, we tested measurement invariance with respect to gender in the linear factor model, using the summed item scores for each subtests. We also tested for predictive invariance. Next, we tested the items of each subtest on measurement invariance in a two-parameter logistic item response model. To asses model fit, we relied on the RMSEA, comparative fit index (CFI), and AIC. In addition, we used a likelihood ratio test (LRT) to test restrictions across models on significance. For all tests, we will use a nominal level of significance of 0.01.

Analyses at the subtest level: testing measurement invariance

To first investigate nonlinearity, we tested whether the factor-to-subtest relationship was nonlinear. To this end, we fitted a nonlinear single-factor model to the subtest data. Results are represented in Table 3. As can be judged from the standard errors of the parameter estimates, in the male group, subtest FC is characterized by nonlinear factor loadings ($p = .002$). However, in the female group, all factor loadings are linear.

Next, we started the actual measurement analyses. Results are represented in Table 4. First, we fitted a single-factor model to the data in both groups. This baseline model, denoted configural invariance (Model 1), in which the parameters are free to vary across groups, fitted well, $\chi^2(4) = 2.09$, $p = 0.72$, RMSEA = 0.00. Next, we introduced equality constraints for the factor loadings, denoted metric invariance (Model 2). This did not affect the model fit significantly, as judged by the likelihood ratio test (LRT; $p = 0.09$). In addition, the RMSEA and CFI showed good model fit. AIC showed a minor deterioration, but the difference was so small that we considered it neglectable. We conclude that metric invariance is tenable.

Table 2. Means and standard deviations of the four subtest scores for males and females.

		FC	DR	AR	MR
Males	**Mean**	11.91	11.47	9.70	11.12
	SD	3.63	4.64	6.72	3.23
Females	**Mean**	11.43	10.08	7.34	11.72
	SD	3.42	4.22	5.91	2.97

Table 3. Results of nonlinear factor analysis of subtest-level data.

Parameter(s)	subtest	Males	Females
linear loadings	FC	1.000	1.000
	DR	1.136 (0.181)	1.171 (0.105)
	AR	0.566 (0.163)	0.390 (0.138)
	MR	0.701 (0.115)	1.002 (0.142)
nonlinear loadings	FC	–0.052 (0.018)	–0.040 (0.022)
	DR	–0.009 (0.030)	0.038 (0.021)
	AR	–0.060 (0.056)	–0.049 (0.068)
	MR	–0.049 (0.023)	–0.047 (0.024)
intercepts	FC	12.265 (0.244)	11.586 (0.146)
	DR	11.537 (0.353)	9.931 (0.170)
	AR	10.111 (0.524)	7.527 (0.314)
	MR	11.460 (0.225)	11.906 (0.124)
residual variance	FC	6.069 (1.154)	7.725 (0.635)
	DR	12.553 (1.507)	12.416 (0.708)
	AR	42.642 (2.006)	34.260 (0.844)
	MR	6.874 (0.669)	4.788 (0.616)
Factor variance		6.907 (1.327)	3.920 (0.661)

Note: Standard errors are in brackets.

We proceeded by restricting the residual variance to be equal across groups (Model 3a). The likelihood ratio test indicated that the restrictions were tenable ($p = 0.02$), however, the RMSEA showed a drop in model fit as it nearly doubled. We therefore consulted the modification indices. These indicated that for subtest AR, the restriction was possibly too stringent. We therefore freed this parameter (Model 3b). This resulted in an improvement in model fit: The RMSEA decreased to a level comparable with that of Model 2, and the likelihood ratio showed no significant difference as compared to Model 2 ($p = 0.37$). In addition, AIC indicated that Model 3b was the best fitting as compared to the previous models. Inspection of the results indicated that, in the male sample, subtest AR was associated with a residual variance of 43.18, and in the female sample this variance equalled 34.21. We therefore concluded that equal residuals were tenable for all subtests but AR, which had a significantly larger residual in the male sample.

In the next step, we tested for strict factorial invariance (Model 4a). This model showed a significant deterioration in model fit, as indicated by the likelihood ratio ($p < .001$), the CFI, AIC, and RMSEA. We consulted the modification indices to diagnose the source of the misfit. Subtest MR was associated with the largest modification index. However,

Table 4. Testing for measurement invariance: model fit statistics for the subtest-level analysis.

		χ^2	*df*	LRT *models*	$\chi^2(df)$	RMSEA	CFI	AIC
1	Configural Invariance	2.09	4	–	–	0.000	1.00	32779.44
2	Metric Invariance	8.56	7	2 vs 1	6.47 (3)	0.017	1.00	32779.91
3a	Equal residual variances	20.02	11	3a vs 2	11.46 (4)	0.033	0.99	32783.77
3b	*– variance AR free*	11.70	10	3b vs 2	3.14 (3)	0.015	1.00	32777.05
4a	Strict MI	111.18	13	4a vs 3b	99.48 (3)	0.101	0.835	32870.53
4b	*– intercept MR freed*	47.48	12	4b vs 3b	35.78 (2)	0.063	0.94	32808.83
4c	*– intercept AR freed*	20.42	11	4c vs 3b	8.72 (1)	0.034	0.98	32783.77

freeing the relevant intercept (Model 4b) did not improve model fit sufficiently, that is, likelihood ratio was still significant as compared to Model 3b ($p < 0.001$), and the AIC, RMSEA, and CFI were too high as compared to the previous models. The intercept of MR was estimated to be 11.12 in the male sample and 12.53 in the female sample. After consulting the modification indices, we freed the intercept of subtest AR (Model 4c). Again, model fit did not improve sufficiently, that is, the likelihood ratio was still significant as compared to Model 3b ($p = 0.003$), and AIC, RMSEA, and CFI were still worse as compared to Models 1, 2, and 3b. The intercept of AR was estimated to be 9.70 in the male sample and 7.71 in the female sample. Despite the fact that model fit was still insufficient, we could not free more intercepts (as with four subtests the minimal number of intercepts required for strict factorial invariance equals two. Freeing an additional intercept causes the mean model to be saturated, that is, in this case the number of parameters equals the number of observed means).

Therefore, the general conclusion of these analyses is that strict factorial invariance is not tenable. This means that we cannot assume measurement invariance for these subtests. At least subtest MR and AR are associated with intercept differences, with a higher intercept for the males for subtest AR, and a larger intercept for females on the MR subtest. In addition, subtest AR is associated with a residual variance that differs across males and females, with the larger variance in the male sample. Finally, it is likely that the intercepts of FC and/or DR also differed between the samples, but we were not able to test for this due to the presence of already two intercept differences.

Analyses at the subtest level: testing predictive invariance

To illustrate that predictive invariance does not necessarily imply the absence of test bias, which is not always fully appreciated in the literature on fairness, we tested the four subtests for predictive invariance. To do so, we considered the scores of the subjects on a memory subtest of the IST. This subtest (ME) consisted of 33 items measuring the ability to memorize pictures, numbers, and words. We took this subtest and used it as a criterion in the test on predictive invariance. For each subtest, an ordinary least squares regression model was fitted with the memory test, ME, as dependent variable, and the corresponding subtest as independent variable. Also included in each regression model was the effect of gender and the interaction effect between gender and the subtest score.

For predictive invariance to hold, the same regression model should hold in both groups, that is, the effect of gender should be insignificant (implying equal intercepts), and the interaction effect should be insignificant (implying equal factor loadings). All subtest variables were standardized in the total sample to prevent spurious interactions. Results are represented in Table 5. As can be seen, predictive invariance held for FC and MR, as both the interaction effect and the gender effect were non-significant. For DR

Table 5. Tests on Predictive Invariance for each subtest.

	Intercept	Main effect	Gender	Interaction
FC	–0.10 (0.05)	**0.33 (0.04)**	0.14 (0.06)	–0.07 (0.05)
DR	**–0.14 (0.05)**	**0.30 (0.04)**	**0.18 (0.06)**	–0.12 (0.05)
AR	–0.12 (0.05)	**0.18 (0.04)**	**0.16 (0.06)**	–0.03 (0.05)
MR	–0.02 (0.05)	**0.38 (0.04)**	0.03 (0.05)	–0.06 (0.05)

Note: Effects that significant at $\alpha = 0.01$ are in bold face.

and AR, predictive invariance did not hold, as intercepts differed significantly between males and females, with the higher intercept in the female group for both subtests.

Analysis at the item level

For the analysis at the item level, we used the R-package "difR" (Magis, Béland, Tuerlinckx, & De Boeck, 2010). Specifically, within this package, we used the Raju method (Raju, 1988, 1990), which involves a test on the signed area between the item characteristic functions in two groups. If this area departs significantly from 0, measurement invariance does not hold for that item, that is, the item is said to show DIF.

Results are represented in Table 6. As evident from the table, subtest FC contains 6 items with a p value smaller than 0.01, and subtest DR contains 1 significant item. For AR, none of the p values exceeds the nominal level of significance, and for subtest MR 10 items are significant according to the Raju method. As we did multiple testing, conclusions should be drawn with care as the number of false positives increases when a large number of statistical tests are conducted. However, a clear pattern is visible in Table 6, that is, the items from subtest DR and AR are largely unbiased, while some items of FC are associated with DIF. In addition, subtest MR shows a considerable extent of DIF.

Comparison of the item-level and subtest-level analyses

In Table 7, we provide an overview concerning the subtest analyses on measurement invariance, the item-level analysis on DIF, and the analysis on predictive invariance.

Table 6. Testing for measurement invariance: model fit statistics for the item-level analysis.

item no.	FC		DR		AR		MR	
	χ^2_I	p	χ^2_I	p	χ^2_I	p	χ^2_I	p
1	9.12	**0.003**	0.31	0.577	–	–	103.84	**<0.001**
2	2.43	0.119	1.08	0.300	–	–	0.08	0.773
3	9.99	**0.002**	1.00	0.320	0.10	0.753	1.61	0.206
4	2.56	0.110	1.82	0.178	3.20	0.074	14.36	**<0.001**
5	7.51	**0.006**	0.17	0.683	5.02	0.025	0.81	0.367
6	6.25	0.012	1.23	0.266	5.06	0.024	9.67	**0.002**
7	5.76	0.017	3.03	0.082	1.04	0.308	1.04	0.309
8	3.13	0.077	1.99	0.159	0.16	0.691	0.88	0.349
9	73.27	**<0.001**	1.10	0.296	0.83	0.363	3.39	0.065
10	6.76	**0.009**	0.09	0.767	3.80	0.052	0.26	0.607
11	1.32	0.252	6.40	0.011	0.11	0.741	9.00	**0.003**
12	4.20	0.040	0.13	0.716	0.77	0.382	35.52	**<0.000**
13	6.81	**0.009**	0.13	0.718	1.85	0.173	57.30	**<0.000**
14	0.25	0.619	1.39	0.240	2.96	0.085	22.66	**<0.000**
15	0.17	0.680	9.73	**0.002**	2.31	0.129	75.52	**<0.000**
16	0.08	0.779	2.96	0.085	3.61	0.057	42.25	**<0.000**
17	0.72	0.396	0.62	0.431	3.20	0.073	21.62	**<0.000**
18	0.30	0.582	2.40	0.120	0.01	0.915	51.70	**<0.000**
19	1.32	0.250	2.53	0.113	0.16	0.686	30.47	**<0.000**
20	0.25	0.618	0.32	0.572	1.93	0.165	0.10	0.757

Note: χ^2_I is the Raju's method test statistic, p is the corresponding p value. All p values smaller than 0.01 are in bold face.

Table 7. Comparison of the results concerning bias using the different methods.

Subtest	Measurement Invariance	Predictive Invariance	DIF
FC	bias	no bias	bias
DR	no bias	bias	no bias
AR	bias	bias	no bias
MR	bias	no bias	bias

We first note that the results concerning predictive invariance yield a pattern that is exactly opposite to that shown by the measurement invariance analyses. That is, when a subtest is flagged as biased by the item analysis, the predictive invariance analysis labels this item as fair, and vice versa. The predictive invariance analysis and the measurement invariance analysis only agree on the AR subtest. Both methods indicated this subtest as biased. These results illustrate the fact that one cannot assume predictive invariance and measurement invariance to be exchangeable or even mutually supportive (Borsboom et al., 2008).

Next, we compare the results concerning DIF at the item level and measurement invariance at the subtest level. For the MR subtest, the results agree. That is, both analyses indicate that this subtest is biased. Figure 5 represents the item characteristic curves for the male and female samples for the biased items as flagged in Table 6. For almost all items, the females have a slight advantage, as items are mostly easier for this sample (Item 4 is the sole exception). In the subtest analysis, we further found an intercept difference on subtest MR, where the higher intercept was associated with the female sample. This is consistent with the observation that all items are easier in the female sample.

In the DIF and measurement invariance analysis results of subtest AR, an inconsistency occurs. This subtest was flagged by the measurement invariance analyses as being biased (due to a larger residual variance and intercept in the male sample). However, the item analyses revealed no bias in this subtest. Figure 6 suggests an explanation for this inconsistency. In the figure, the subtest scores are plotted for the male and female samples. As can be seen, the distribution of the AR test in the male sample is severely non-normal with a clear ceiling effect, possibly causing a higher residual variance and intercept in this sample.

For the FC and DR subtests, results are difficult to compare. It is unclear from the subtest analysis which of the two is associated with an intercept difference. However, given that the subtest analysis correctly picked up the bias in subtest MR, it is likely that the intercept difference is associated with FC, as the item analysis showed some bias in this subtest.

Thus, comparing the subtest and item analysis revealed some consistency in case of subtest MR, and some large inconsistency in case of the subtest AR. Inconsistencies in the results concerning AR are likely due to violations of normality, clearly illustrating why one should be cautious in analysing total scores when testing for measurement invariance. Finally, results of the predictive invariance analyses were largely inconsistent with results of the other methods. This clearly bolsters earlier critiques concerning the use of predictive invariance to investigate item or test bias.

Discussion

In this paper, we discussed the effects of testing for measurement invariance on summed item scores. We identified four possible problems that may jeopardize inferences concerning measurement invariance: nonlinearity, information loss in the sum score, non-normality,

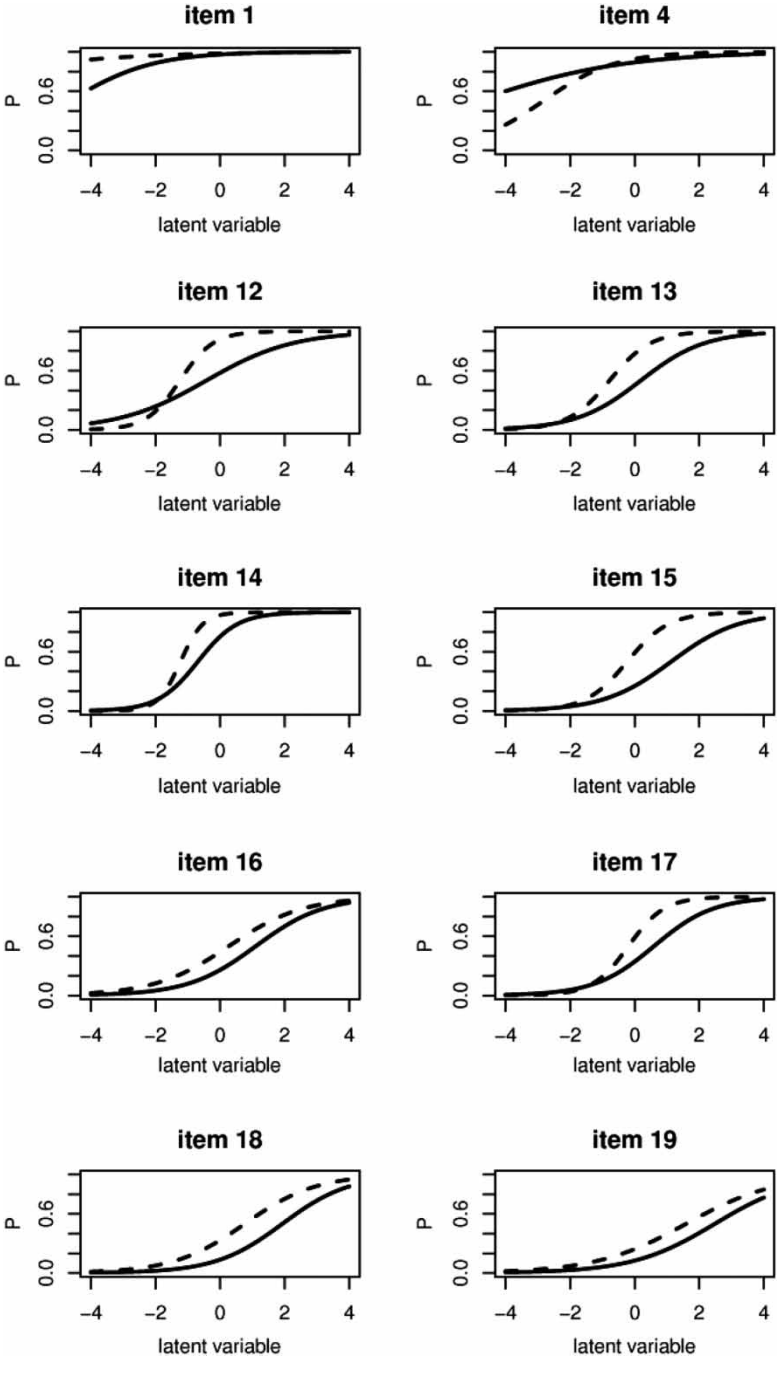

Figure 5. Item characteristic curves for the items flagged as biased in subtest MR for males (solid line) and females (striped line).

and the different nature of the latent variable at item and subtest levels. In a real data set, we showed that inconsistencies in the results for measurement invariance tests were clearly present, and at least some of these inconsistencies could be traced to the problems we

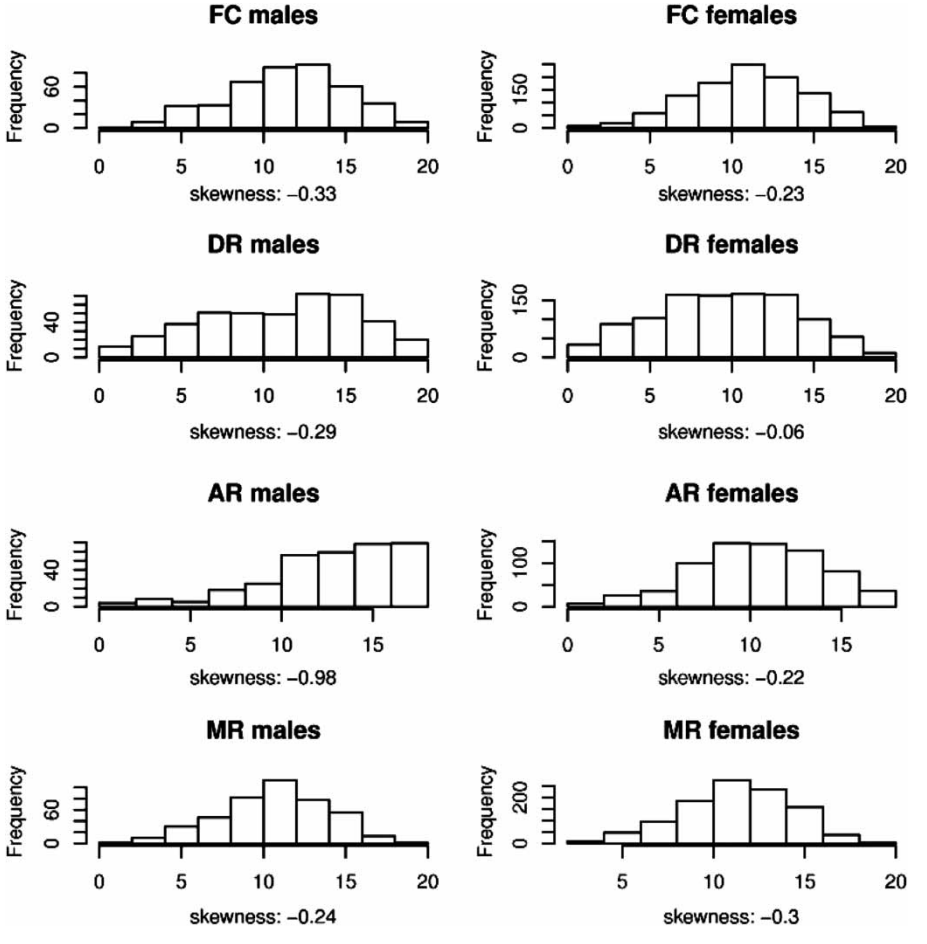

Figure 6. Total score distributions of the four subtests in the male and female sample.

discussed. This shows that caution is warranted when analysing measurement invariance using linear factor models that are applied to total scores.

While we focused on testing measurement invariance on sum scores, results do generally also hold for Likert scales data. That is, analogous to the sum score (Figures 1 and 2), the problem of nonlinearity can also arise in Likert scales when the answer options of the scale are disproportionately used by the subjects. This problem is also referred to as "poor scaling" of the measurement (Molenaar & Dolan, 2012; Van der Sluis, Dolan, Neale, Boomsma, & Posthuma, 2006). A related problem with Likert scales is "censoring" (Tobin, 1958) in which the majority of a sample obtains the highest or lowest possible score. Censoring will result in non-normality by definition, which in turn may distort tests on measurement invariance as discussed in this paper with respect to the sum score. See also Lubke and Muthén (2004), who demonstrated this in simulated Likert scale data. Finally, as with the sum score, treating Likert scales as continuous variables is also associated with the problem of information loss. When Likert scales are analysed with an appropriate IRT model, for example, the graded response model (Samejima, 1969), so-called category thresholds are taken into account. These parameters model the distances between the answer categories on the latent continuum (which are not necessarily the same

in case of ordinal data). By treating the Likert scale data as continuous, however, distances between adjacent categories are assumed to be equal. Again, this can influence tests on measurement invariance (see Lubke & Muthén, 2004).

Importantly, we do not discourage the use of factor analysis in tests on measurement invariance. As we argued in the introduction, in some circumstances such analyses are necessitated by the circumstances. Also, the hypotheses tested in item-level and subtest-level studies of measurement invariance may not precisely align. This means that there may be situations in which one is primarily interested in the invariance properties of subtests, and does not care much for the invariance properties of items. On the other hand, we have rarely come across a serious substantive of methodological motivation that would apply in this kind of situation. It rather seems that many researchers take measurement invariance of items and subtests to be alternative formulations of the same hypothesis (i.e., that the *test* is fair – a hypothesis which, ironically, is tested in neither of the approaches).

The problem of differences in the nature of the latent variable in the item analyses and in the subtest analyses presents an interesting conceptual puzzle. Even a multidimensional IRT model will not necessarily lead to results comparable to the linear factor model, as the factor model strictly focuses on higher order latent variables which differ from the lower order latent variables analysed in IRT. With respect to this problem, it might be concluded that the linear factor model and the IRT model indeed address different hypotheses concerning measurement invariance. Reasoning further along these lines, one could argue that the linear factor model (as applied to summed item scores) indirectly tests a structural hypothesis; namely, that the relation between the lower order latent variable in the IRT analysis and the higher order latent variable in the factor model is the same across groups. So viewed, the hypothesis tested in the linear factor model approach is not a measurement hypothesis at all. On the other hand, there is no a priori reason to limit the definition of an indicator to item responses; total scores are, after all, observable just as well. Reasoning in this direction, therefore, one would suggest that that the linear factor model does in fact test the invariance of a measurement relation, be it a different one from the IRT modelling approach. This is the position we favour.

The problem of how to sum item scores to obtain the best subtest score for the subsequent factor modelling endeavour is open to debate. If we accept the previous conclusion that IRT and factor analysis address different questions, we may in fact take many routes to construct total scores. For instance, we may take factor scores from a Rasch model or a 2PL model and submit these to factor analysis. To profitably do so, however, we would typically want to assume that DIF is absent at the item level to enable the adequate estimation of factor scores in both groups. Then, however, the question arises whether these estimated factor scores are necessarily measurement invariant if they are based on invariant IRT models. This is likely to differ from model to model, and presents an interesting avenue for further research.

As a practical guideline, we have some suggestions following from the present undertaking. First, when analysing summed item scores, tests on nonlinearity should be routinely considered. These tests are currently straightforward in the Mplus program (L. K. Muthén & Muthén, 2007). Alternatively, as suggested by Bauer (2005), plots between all pairs of item scores can be considered. These should all be approximately linear for the linear factor model to be tenable. As we illustrated in our data analyses, when no apparent nonlinearity is found, one can be more confident in using the linear factor model to test for measurement invariance. When significant nonlinearity is found, it could be safer to use either item-level analyses or reside on the nonlinear factor model. A second suggestion

is that classical test theory item characteristics could be considered to roughly gauge whether the item difficulties are distributed proportionally across the latent variable scale; if this is not the case, linearity is unlikely. To judge whether the distribution of the item difficulties is proportional, uniform QQ-plots of the proportion correct could be considered or a Kolmorogov-Smirnov test for uniform distributions could be conducted (both can be done in SPSS). In addition, the item-total correlations could be considered to get an idea about the differences in discrimination across items. If these correlations are roughly equal (e.g., as judged by their confidence intervals), this would indicate that item discrimination does not differ importantly across item, and the sum score can be more confidently used in a linear factor analysis (given normality and proportionally distributed item difficulties). However, when item-total correlations differ across items, it is better to reside on item-level analysis. A final recommendation is that the total score is ideally based on a large number of items, which will typically make the assumption of a normal distribution more plausible due to the central limit theorem.

Acknowledgements

The research by Dylan Molenaar and Denny Borsboom was made possible by a Top Talent grant and a VIDI grant from the Netherlands Organization for Scientific Research (NWO). We thank Harry Vorst for the data used in the application.

Note

1. Note that the common factor means are only identified in multi-group models given a sufficient number of invariant intercepts (Sörbom, 1974).

References

Aguinis, H., & Smith, M. E. (2007). Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. *Personnel Psychology*, *60*, 165–199.

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *I-S-T 2000R. Intelligenz-Struktur-Test 2000R.* Göttingen, Germany: Hogrefe.

Andersen, E. B. (1973). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B*, *32*, 283–301.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. London, UK: Griffin.

Bauer, D. J. (2005). The role of nonlinear factor-to-indicator relationships in tests of measurement equivalence. *Psychological Methods*, *10*, 305–316.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In E. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison Wesley.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, UK: Cambridge University Press.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*, S176–S181.

Borsboom, D. (2008). Latent variable theory. *Measurement*, *6*, 25–53.

Borsboom, D., Romeijn, J. W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, *13*, 75–98.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, *5*, 115–124.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research*, *35*, 21–50.

Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J. M., Hessen, D. J., & Van der Sluis, S. (2006). Multi-group covariance and mean structure modeling of the relationship between WAIS-III common factors and gender and educational attainment in Spain. *Intelligence*, *34*, 193–210.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.
Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*, 155–174.
Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, *62*, 495–523.
Evers, A., Te Nijenhuis, J., & Van der Flier, H. (2005). Ethnic bias and fairness in personnel selection: Evidence and consequences. In A. Evers, N. Anderson, & O. F. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 306–328). Oxford, UK: Blackwell.
Gamliel, E., & Cahan, S. (2007). Mind the gap: Between-group differences and fair test use. *International Journal of Selection and Assessment*, *15*, 273–282.
Grayson, D. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*, 383–392.
Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144.
Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests. *Psychology, Public Policy, and Law*, *6*, 151–158.
Jensen, A. R., & Reynolds, C. R. (1983). Sex differences on the WISC-R. *Personality and Individual Differences*, *4*, 223–226.
Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109–133.
Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variances across subpopulations mask differences in residual means in the common factor model? *Structural Equation Modeling*, *10*, 175–192.
Lubke, G. H., Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*, 514–534.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Magis, D., Beland, S., Tuerlinckx, F. & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847–862.
Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.
Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.
Mellenbergh, G. J. (1994a). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–307.
Mellenbergh, G. J. (1994b). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*, 223–236.
Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*, 177–185.
Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.
Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*, 248–260.
Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research*, *33*, 403–424.
Millsap, R. E. (2008). Invariance in measurement and prediction revisited. *Psychometrika*, *72*, 461–473.
Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.
Molenaar, D., & Dolan, C. V. (2012). Substantively motivated extensions of the traditional latent trait model. *Netherlands Journal of Psychology*, 67, 48–57.
Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.
Muthén, L. K., & Muthén, B. O. (2007). *Mplus User's Guide* (5th ed.). Los Angeles, CA: Authors.

Neisser, U., Boodoo, G., Bourchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., … Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77–101.
Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks, CA: Sage.
Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125–167). Amsterdam, The Netherlands: Elsevier.
Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502.
Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*, 197–207.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut.
Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, *11*, 235–294.
Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, *63*, 215–227.
Sackett, P. R., Schmitt, N., Ellington, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist*, *56*, 302–318.
Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph, No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf
Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of fit measures. *Methods of Psychological Research*, *8*, 23–74.
Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika*, *58*, 159–194.
Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239.
Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, *15*, 201–293.
Spearman, C. E. (1927). *The abilities of man: Their nature and measurement*. New York, NY: Macmillan.
Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.
Thomson, G. H., & Lederman, W. (1939). The influence of multivariate selection on the factorial analysis of ability. *British journal of psychology*, *29*, 288–305.
Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: university of Chicago Press.
Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, *26*, 24–36.
Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the life span. *Developmental Psychology*, *45*, 1097–1118.
Van der Sluis, S., Dolan, C. V., Neale, M. C., Boomsma, D. I., & Posthuma, D. (2006). Detecting genotype-environment interaction in monozygotic twin data: Comparing the Jinks & Fulker test and a new test based on marginal maximum likelihood estimation. *Twin research and Human Genetics*, *9*, 377–392.
Van der Sluis, S., Posthuma, D., Dolan, C. V., De Geus, E. J. C., Colom, R., & Boomsma, D. I. (2006). Gender differences on the Dutch WAIS-III. *Intelligence*, *34*, 273–289.
Verhelst, N. D., & Glas, C. A. W. (1995). The one-parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds), *Rasch models: Foundations, recent developments, and applications.* (pp. 215–237). New York, NY: Springer.
Wechsler, D. (1997). *Wechsler Adult Intelligence Scale- III (WAISIII).* San Antonio, TX: Psychological Corporation.
Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, *89*, 696–716.

Wicherts, J. M., & Millsap, R. E. (2009). The absence of underprediction does not imply the absence of measurement bias. *American Psychologist*, *64*, 281–283.

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