Different Kinds of DIF: A Distinction Between Absolute and Relative Forms of Measurement Invariance and Bias

Denny Borsboom, Gideon J. Mellenbergh, and Jaap van Heerden
Department of Psychological Methods, University of Amsterdam

In this article, a distinction is made between absolute and relative measurement. Absolute measurement refers to the measurement of traits on a group-invariant scale, and relative measurement refers to the within-group measurement of traits, where the scale of measurement is expressed in terms of the within-group position on a trait. Relative measurement occurs, for example, if an item induces a within-group comparison in respondents. These distinctions are discussed within the framework of measurement invariance, differentiating between absolute and relative forms of measurement invariance and bias. It is shown that items for relative measurement will produce bias as classically defined if the mean and/or variance of the trait distribution differ between groups. This form of bias, however, does not result from multidimensionality but from the fact that measurement is on a relative scale. A logistic regression procedure for the detection of relative measurement invariance and bias is proposed, as well as a model that allows for the incorporation of items for relative measurement in test analysis. Implications of the distinction between absolute and relative measurement are discussed and prove to be especially relevant for the domain of personality research. *Index terms: construct validity, differential item functioning, item bias, measurement invariance.*

Questions concerning test validity are central to test theory and scientific progress, but also to ethical, legal, and political issues related to test use (Cronbach, 1988; Messick, 1989). Within validity theory, the development of concepts such as measurement invariance and item bias has provided an important conceptual framework for thinking about these issues. However, the relation between construct validity and measurement invariance is not yet entirely clear. This article purports to provide some insight into this relation by presenting a distinction between different kinds of measurement invariance and bias and by evaluating these within a construct validity perspective. Especially, the authors are concerned with the meaning of measurement invariance and bias in the domains of personality and attitude testing.

The ideas of item bias and measurement invariance were first conceived of in item response theory (IRT) by Lord (1980), who proposed that measurement invariance with respect to group membership holds if an item follows the same item characteristic curve (ICC) in all groups. In IRT for dichotomous item responses, this requirement means that the probability of a given response is the same for members of different groups with the same position on the trait measured by the test (Mellenbergh, 1989; Millsap & Everson, 1993). The notion of measurement invariance can be generalized to cover a wider range of models by making the more general requirement that the distribution function of the item response be invariant across groups, conditional on the latent trait (Meredith, 1993). Thus, an item j, answered by Participant i and assumed to measure latent trait T, is measurement invariant with respect to selection on variable V if and only if the following

Applied Psychological Measurement, Vol. 26 No. 4, December 2002, 433–450
DOI: 10.1177/014662102237798

equation holds for the distribution function F of the item response X_{ij}:

$$F\left(X_{ij} = x_{ij} | T = t_i, V = v_i\right) = F\left(X_{ij} = x_{ij} | T = t_i\right) \tag{1}$$

for all $\{x, t, v\}$.

This definition corresponds to unobserved conditional invariance (UCI) as discussed by Millsap and Everson (1993). Whenever the above condition is violated, the item is said to be biased. In the IRT literature, the more neutral term differential item functioning (DIF) is often preferred. In this article, the terms are used interchangeably. Thus, item bias occurs if and only if

$$F\left(X_{ij} = x_{ij} | T = t_i, V = v_i\right) \neq F\left(X_{ij} = x_{ij} | T = t_i\right) \tag{2}$$

for some $\{x, t, v\}$.

As can be seen from Formula 2, item bias amounts to an influence of group membership on the item response in subpopulations with the same position on the trait. Item bias may, for example, occur in an IQ-test if men score better on an item than women, although there is no difference in intelligence. Item bias is to be sharply distinguished from impact, which amounts to differences in test scores that are due to differences in trait distributions (Millsap & Everson, 1993). For the above example, impact would occur if the better scoring of men was due to higher mean intelligence. In this case, the differential performance can be entirely attributed to a difference in location of the latent trait distributions.

Item bias bears directly on construct validity. When the intention is to measure a unidimensional concept, one would intuitively expect that a biased item is necessarily invalid. Indeed, item bias has been equated with multidimensionality (Kok, 1988). Shealy and Stout (1993, p. 198) remarked that "test bias occurs if the test under consideration is measuring a quantity in addition to the one the test was designed to measure, a quantity that both groups do not possess equally." On the item level, item bias is seen as the effect of an unwanted, additional variable on the item response. In this view, bias with respect to group membership is produced by an association of this additional variable with group membership, thus influencing item responses differently in each group. Consequently, if the intention is to measure a single trait, removing items that are biased with respect to group membership from the test seems a plausible strategy to enhance construct validity. One of the objectives of this study, however, is to show that this is not always the case.

The conceptual framework of measurement invariance has been developed from the perspective of cognitive testing, and this is the primary field where DIF-analyses are used. One reason for this is that the concepts of measurement invariance and bias are most salient in individual decisions that are "high-stake," for example, when tests are used for college admissions or personnel selection—domains where cognitive tests are of primary importance. For scientific purposes, however, the importance of questions concerning measurement invariance and bias is not restricted to any specific theoretical domain. Indeed, there have been some recent applications in personality testing (Ellis, Becker, & Kimmel, 1993; Huang, Church, & Katigbak, 1997; Smith & Reise, 1998), and the screening for DIF is equally important in the field of personality psychology as in any other domain of psychological measurement.

Now, the technical aspects of measurement invariance and bias can be applied to domains other than cognitive testing without any specific problems, because they are of a mathematical nature and thus entirely syntactical. However, the meaning of measurement invariance and bias may change with the field of application. The authors will be concerned with one specific shift of meaning that occurs when the concepts of measurement invariance and bias are used in the area of personality and attitude testing. Especially, the meaning of measurement invariance when items invoke a frame of reference, for example, by inducing a within-group comparison, will be looked at. It will be argued that such items will display bias as defined above. However, from a construct validity perspective, such items are not necessarily invalid, but rather there is a misfit between the model that is used and the cognitive processes that are involved in the item response. To deal with this problem, the framework of measurement is extended. A distinction between relative and absolute forms of measurement is introduced, and corresponding forms of measurement invariance and bias are defined. It is shown that items that induce a within-group comparison will lead to absolute, but not to relative, bias. Following this distinction, it will be argued that items showing absolute, but no relative, bias do not necessarily have to be eliminated from a test. Upon proper analysis, these items—although biased according to current standards—can enhance test validity and do not necessarily produce test bias.

Absolute and Relative Forms of Bias

Consider the following thought experiment. Imagine a world where the development of measurement theory in the social sciences has preceded measurement in the natural sciences. In this world, psychological research on attitudes and self-efficacy is common practice, whereas concepts such as height or weight are still to be invented. A psychologist might then conceive of a person's height as a useful construct for the explanation of certain types of behavior, such as the predisposition of some individuals to participate in basketball and the difficulty others experience when reaching for the upper shelves of a closet. However, because a measurement apparatus for the assessment of height has not yet been invented, he or she can only use social science's measurement methods to assess height. For this reason, the psychologist would probably go about constructing a questionnaire consisting of items such as "I have trouble getting a book from the upper shelves in a library," "Sometimes I have to bend over in order to see my face in a mirror," and "When sitting on somebody else's chair, I cannot usually reach the ground with my feet." Suppose the psychologist would have constructed a questionnaire consisting of the aforementioned three items, and would add a fourth on the basis of his or her intuitions concerning the relation between height and basketball: "I would do well on a basketball team."

Although this item has high face validity, a formal test of DIF points out that the item shows DIF with respect to sex; women have a higher probability of answering yes than do men of the same height. Formally, if one calls the item response X_{ij} (scored dichotomously with yes $= 1$ and no $=$ 0), takes height to represent the latent trait T, lets V denote sex (say, $V = 0$ for men and $V = 1$ for women), and P the probability of an item response, then

$$P\left(X_{ij} = 1|T = t_i, V = 0\right) < P\left(X_{ij} = 1|T = t_i, V = 1\right), \tag{3}$$

for at least some values of T, so the item has DIF. To increase test validity, the psychologist removes the item from the test. But is this a sensible thing to do? The authors think it is not, and this has to do with the nature of the sex difference. A woman, 5.8 feet tall, may imagine a basketball team consisting of women and conclude that she would do well because she is relatively tall—considering her sex. A man of the same height may correctly judge himself to be relatively short—considering his sex—and conclude the opposite. Because of the within-group comparison made by both sexes, the item has absolute bias: men and women of the same height do not have the same probability of an affirmative answer. However, men and women with the same relative height within their own group (for example, a standard deviation above the group mean) do have identical probabilities of an affirmative answer. Thus, although the item is biased with respect to absolute height, it is not biased with respect to relative height.

This notion is now formalized. Denote the relative position on the trait by W, taking on values w_i. Then, for the item under consideration, although it is true that

$$P\left(X_{ij} = 1|T = t_i, V = 0\right) < P\left(X_{ij} = 1|T = t_i, V = 1\right) \tag{4}$$

for some $\{t\}$, it is also true that

$$P\left(X_{ij} = 1|W = w_i, V = 0\right) = P\left(X_{ij} = 1|W = w_i, V = 1\right) \tag{5}$$

for all $\{w\}$. Following this insight, two forms of measurement can be distinguished. *Absolute* measurement refers to a procedure to measure the trait on an absolute scale (e.g., “I have trouble getting a book from the upper shelves in a library”), and *relative* measurement refers to a procedure to measure the trait on a relative scale (e.g., “I would do well on a basketball team”), where the measurement unit is expressed in terms of the relative position within the group to which the participant belongs. The different forms of measurement imply different definitions of measurement invariance and bias. Accordingly, absolute and relative measurement invariance and their corresponding forms of bias are differentiated as follows:

Definition 1. For an item, generating item response X_{ij} and measuring trait T, *absolute measurement invariance* with respect to selection on variable V occurs if and only if

$$F\left(X_{ij} = x_{ij}|T = t_i, V = v_i\right) = F\left(X_{ij} = x_{ij}|T = t_i\right) \tag{6}$$

for all $\{x, t, v\}$. *Absolute bias* with respect to selection on variable V occurs if and only if

$$F\left(X_{ij} = x_{ij}|T = t_i, V = v_i\right) \neq F\left(X_{ij} = x_{ij}|T = t_i\right) \tag{7}$$

for some $\{x, t, v\}$. Note that these are the usual definitions of measurement invariance and bias (Mellenbergh, 1989; Millsap & Everson, 1993).

Definition 2. For an item, generating item response X_{ij} and measuring trait T, *relative measurement invariance* with respect to selection on variable V occurs if and only if, for the item response conditional on W (the relative within-group position on the trait T),

$$F\left(X_{ij} = x_{ij}|W = w_i, V = v_i\right) = F\left(X_{ij} = x_{ij}|W = w_i\right) \tag{8}$$

holds for all $\{x, w, v\}$. *Relative bias* with respect to selection on variable V occurs if and only if

$$F\left(X_{ij} = x_{ij}|W = w_i, V = v_i\right) \neq F\left(X_{ij} = x_{ij}|W = w_i\right) \tag{9}$$

for some $\{x, w, v\}$.

Now the problem occurs how to specify W. This depends primarily on the nature of the cognitive processes involved in answering personality items, which at present is unknown for most tests. However, it is obvious that W should be some transformation of the trait T. The form of this transformation might be different for different tests and items, and it could, in principle, even vary over groups. So, in the general definitions, the exact form of the transformation should not play a role. However, to apply the concepts of relative measurement and bias, some form for the transformation has to be assumed. W will be conceived of as the within-group standardized transformation of T. There are three reasons for this. First, this assumption leads to precise and testable hypotheses. Second, the Z-transformation has many desirable mathematical properties that will become apparent in the next section. Third, even if the actual comparison is not made on a standardized within-group scale (for example, if it is in terms of absolute deviations from the mode), the Z-score will often provide a reasonable approximation. In the remainder of this article, therefore, W will be equated with the within-group standardized transformation of T, which will be denoted as Z, taking on possible values z_i. Note that the first two moments of the distribution of Z are—by definition—the same across groups: It has mean 0 and variance 1 within each of the groups. Furthermore, if the original trait distributions are normal, the resulting Z-distribution is the same in each group. This observation has implications for the theory of multidimensionality, which are discussed below.

By equating W with the within-group standardized transformation of T, Definitions 8 and 9 are altered by substituting Z and z_i for W and w_i, respectively, and the resulting concepts may be coined "standardized relative measurement invariance and bias." To avoid an overload of terminology, however, in text the authors will continue to speak of relative measurement invariance and bias, with the understanding that an assumption concerning the form of the transformation has been made; consequently, Z instead of W will be used in the formulae. Finally, the authors would like to stress that different forms of the transformation could be used and that the appropriateness of the chosen form of the transformation represents a testable hypothesis. Thus, although the exact form of the transformation does not play a role in the general definitions given above, it does play a role in the consequences and assessment of relative measurement invariance and bias. As a consequence, the results derived hereafter do depend on the appropriateness of the Z-transformation.

The Relation Between Absolute and Relative Bias

In this section, the relation between absolute and relative measurement invariance and bias is examined. This paragraph is primarily intended to show the mutual incompatibility of absolute and relative measurement invariance. The terminology of IRT will be used because it allows for a clear and comprehensible expression of the concepts of measurement invariance and bias. Later in this article, the authors return to the more general case and also discuss a structural equation modeling (SEM) approach to modeling relative measurement invariance.

In parametric IRT, the probability of a correct response to an item is expressed as a function of a person characteristic (the position on the latent trait) and a number of item characteristics (e.g., the difficulty of the item and the item's ability to discriminate between participants with different trait values). A common form for this relation between the probability of a correct response, the position on the latent trait, item difficulty, and item discrimination is provided by Birnbaum's (1968) two-parameter logistic model:

$$P(X_{ij} = 1|t_i, a_j, b_j) = \frac{\exp\left[a_j(t_i - b_j)\right]}{1 + \exp\left[a_j(t_i - b_j)\right]}, \tag{10}$$

where b_j indicates the difficulty of item j, a_j is its discrimination parameter, and t_i denotes Participant i's position on the latent trait T. For a single item, model Formula 10 gives the ICC, which results from plotting the response probabilities for this item against the latent trait values. The parameter b_j determines the location of the ICC and the parameter a_j its slope in the point $t_i = b_j$, hence their interpretation as item difficulty and item discrimination. Absolute measurement invariance can be expressed as the requirement that the ICCs for different groups are identical: If ICCs are identical across groups, the probability of a correct response, conditional on the latent trait, is the same for participants with the same latent trait values, regardless of their group membership.

The ICC results from plotting the probability of a correct response against the latent trait, for which absolute trait values are used. Following the authors' distinction between absolute and relative measurement, the "classical" ICC discussed above is referred to as an absolute ICC. However, it is also possible to plot the probability of a correct response against relative trait values. This gives us a relative ICC. The relative ICC relates the probability of a correct response to the within-group standardized latent trait Z. Like the absolute ICC, its form is determined by two parameters indicating the relative (within-group) difficulty and slope. These parameters will be denoted as $b_{j_{\text{rel}}}$ and $a_{j_{\text{rel}}}$, and refer to the original absolute parameters as $b_{j_{\text{abs}}}$ and $a_{j_{\text{abs}}}$. The form of the two-parameter variant of the relative ICC is determined by the following formula:

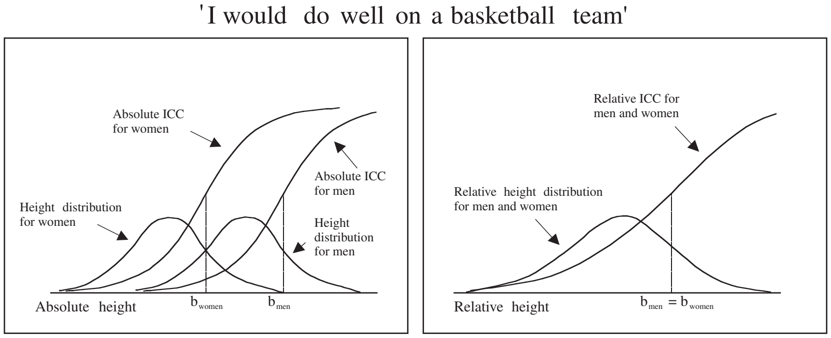
$$P\left(X_{ij} = 1 | z_i, a_{j_{\text{rel}}}, b_{j_{\text{rel}}}\right) = \frac{\exp\left[a_{j_{\text{rel}}}(z_i - b_{j_{\text{rel}}})\right]}{1 + \exp\left[a_{j_{\text{rel}}}(z_i - b_{j_{\text{rel}}})\right]}. \tag{11}$$

As is the case with absolute measurement invariance, the requirement of relative measurement invariance, that the relative ICCs must be equal across groups, can be reformulated as the requirement that the parameters of the relative ICCs, $b_{j_{\text{rel}}}$ and $a_{j_{\text{rel}}}$, are equal across groups.

The question arises how the relative ICC relates to the absolute ICC, or, alternatively, how the relative item difficulty and discrimination parameters relate to the absolute item difficulty and discrimination parameters. In particular, it is interesting to inquire under which conditions absolute and relative measurement invariance may both hold. The relation between absolute and relative measurement will be discussed at an intuitive level before turning to a more precise formulation of the relation between absolute and relative parameters.

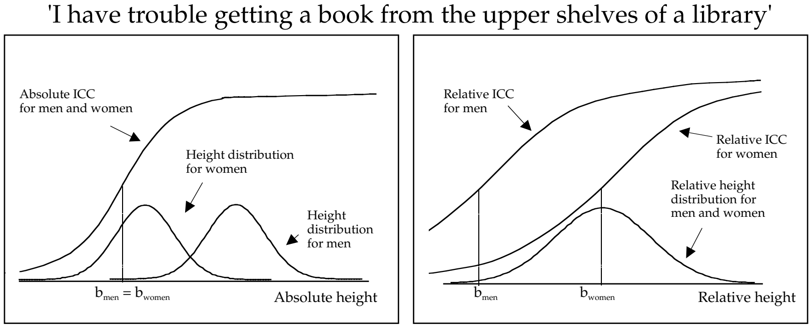
Consider the item for relative measurement in the height test (“I would do well on a basketball team”). The left half of Figure 1 shows, in a single graph, the population distributions of the latent trait and the absolute ICCs for men and women. (The population distributions and the ICCs can be drawn in a single graph because, in IRT, trait parameters and item difficulty parameters are on the same scale.) The ICCs for men and women differ in location (i.e., item difficulty), indicating absolute bias. The right half of Figure 1 shows the relative ICCs, that is, the item response probabilities plotted against relative trait values. Also shown are the population distributions of the relative trait values. These are identical because the trait has been standardized within groups (the distribution has mean 0 and variance 1 in each of the groups). Because the locations of the absolute ICCs relative to the within-group distributions are the same, the relative ICCs are identical for men and women. This indicates that there is no relative bias; the item has relative measurement invariance.

Figure 1
Absolute and Relative Item Characteristic Curves (ICCs) for an Item With Relative Measurement Invariance but Absolute Bias


'I would do well on a basketball team'

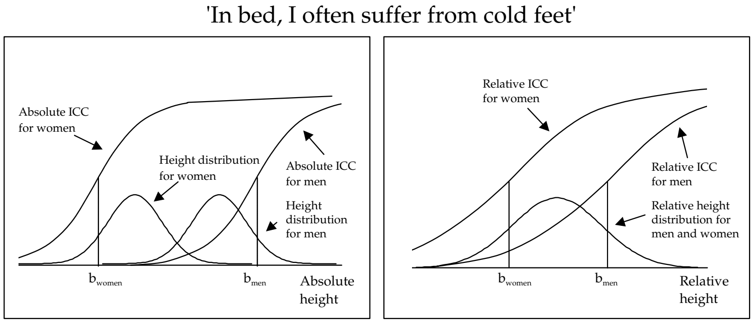
In contrast, Figure 2 shows an item with absolute measurement invariance (“I have trouble getting a book from the upper shelves in a library,” scored yes = 0 and no = 1). The absolute ICCs, shown in the left half of the figure, are identical for men and women, indicating absolute measurement invariance. However, the absolute ICC is located relatively farther away from the mean of the trait distribution for men than it is for women; moderately short women have the same probability of an affirmative response as do extremely short men. As a consequence, the relative ICCs are different, as is shown in the right half of Figure 2, and the item has relative bias.

Figure 2
Absolute and Relative Item Characteristic Curves (ICCs) for an Item With Absolute Measurement Invariance but Relative Bias


'I have trouble getting a book from the upper shelves of a library'
Absolute ICC for men and women
Height distribution for women
Height distribution for men
$b_{men} = b_{women}$
Absolute height
Relative ICC for men
Relative ICC for women
Relative height distribution for men and women
b_{men}
b_{women}
Relative height

Finally, Figure 3 shows an item for which both the absolute and the relative ICCs are different for men and women, for example, "In bed, I often suffer from cold feet." This indicates that the item has both absolute and relative bias.

Figure 3
Absolute and Relative Item Characteristic Curves (ICCs) for an Item With Both Absolute Bias and Relative Bias



The figures suggest that absolute and relative measurement invariance cannot hold simultaneously if the distribution of the latent trait differs across groups. This is due to the fact that the absolute ICCs cannot be simultaneously located at the same position on the absolute trait (absolute measurement invariance) and have the same location relative to the group means (relative measurement invariance). The authors now turn to a more precise formulation of the relation between absolute and relative parameters.

A relative ICC has been defined in model Formula 11. The relative parameters can be expressed as functions of the absolute parameters, because the relative trait values are linear transformations of the absolute trait values; they are defined by the within-group standardization

$$z_i = \frac{t_i - \mu_{T_v}}{\sigma_{T_v}}, \tag{12}$$

where μ_{T_v} and σ_{T_v} represent the mean and standard deviation of the trait distribution in group v, to which Participant i belongs. This standardization is performed separately for each group, which means that a possibly different linear transformation of the trait values is performed in each group. The relation between absolute and relative item parameters can be expressed as the effect of these transformations on the item parameters.

The absolute difficulty parameter $b_{j_{\text{abs}}}$ is defined as the latent trait value for which the probability of a correct response, given the latent trait, is 0.5. In the standardization, all trait values are rescaled through Formula 12. It follows that the relative difficulty parameter is the relative trait value that is associated with the absolute trait value $t_i = b_{j_{\text{abs}}}$ through the linear transformation given in model Formula 12. So,

$$b_{j_{\text{rel}}} = \frac{b_{j_{\text{abs}}} - \mu_{T_v}}{\sigma_{T_v}}. \tag{13}$$

The relative and absolute difficulty parameters are related through a linear transformation that is possibly different for each group. Whether the transformation is different depends on differences in the trait distribution between groups. It follows that, if the mean and/or variance of the trait distribution differ between groups, absolute and relative measurement invariance in the difficulty parameters cannot hold simultaneously.

A similar effect holds for the discrimination parameters. It is intuitively plausible that differences in trait variances have an effect on the slope of the relative ICC. Because the standardization changes the distances between trait values by a factor $1/\sigma_{T_v}$, the slopes of the absolute and relative ICCs can be expected to differ by a factor σ_{T_v}. Formally, this result is derived as follows. Equations 10 and 11 may be set equal within each group, because the standardization of trait values is a linear transformation, and consequently the probability of a response for each value of $T = t$ and the corresponding value of $Z = z$ must be the same within each group. Substituting the right-hand sides of Equations 12 and 13 for z_i and $b_{j_{\text{rel}}}$, respectively, and solving for $a_{j_{\text{rel}}}$, the following is obtained:

$$a_{j_{\text{rel}}} = \sigma_{T_v} a_{j_{\text{abs}}}. \tag{14}$$

From this relation, it follows that, if the variance of the trait distribution differs over groups, either absolute or relative bias in the discrimination parameter will occur.

Thus, absolute and relative measurement invariance cannot hold simultaneously if groups differ in means and/or variances of the latent trait. This relation also holds for other than dichotomous item responses, for example, polytomous or continuous item responses. This statement is not formally proven, because it is thought rather obvious: Any type of item can only be simultaneously measurement invariant with respect to T and with respect to Z if the transformation that leads from T to Z is identical across groups. This transformation can only be identical if the means and variances of the population distributions on the latent variable are the same. Thus, absolute and relative measurement invariance can hold simultaneously, but only if there are no differences in the means and variances of these population distributions. If there are differences in the means and/or variances of these distributions, absolute measurement invariance will lead to relative bias, and relative measurement invariance will lead to absolute bias.

DIF-Detection and Modeling

The question arises how to detect relative DIF in an empirical situation. The formulation of relative measurement invariance as the requirement that relative ICCs are identical across groups opens a range of possibilities. It makes IRT-based techniques for the assessment of absolute DIF available for the detection of relative DIF. Thus, methods based on area measures, such as signed and unsigned area tests, as well as statistics for the equality of item parameters, or Mantel-Haenszel-based procedures, could in principle be used to assess relative DIF (see Holland & Wainer, 1993, or Camilli & Shepard, 1994, for overviews of available techniques for the detection of absolute DIF).

For the dichotomous case, an adaptation of the logistic regression approach (Swaminathan & Rogers, 1990) for the detection of relative measurement invariance and DIF will be presented, because it is simple and instructive. The logistic regression approach is based on the idea that, in a regression of the binary item response on the continuous latent trait, group membership should not contribute significantly to the prediction once the trait has been included as a predictor in the regression equation. An item can be tested for DIF by fitting the full regression

$$P(X = 1) = \frac{\exp(c_0 + c_1 t_i + c_2 v_i + c_3 t_i v_i)}{1 + \exp(c_0 + c_1 t_i + c_2 v_i + c_3 t_i v_i)}, \tag{15}$$

where the c_0 to c_3 are regression parameters, t_i represents the latent trait value of Participant i, and v_i is a dummy variable coding for group membership. In this procedure, one checks whether the parameters c_2 and c_3 differ from zero. Here, c_2 represents the main effect of group membership and c_3 the interaction between group membership and the latent trait. A significant parameter value for c_3 would indicate nonuniform DIF, which occurs when the amount of DIF changes across trait values (Mellenbergh, 1982). If the parameter value for c_3 is not significant, but the parameter value for c_2 is, this indicates uniform DIF, that is, a constant amount of DIF across trait values. Usually, the latent trait values are unknown and are replaced by sum scores. If this is done, and the interaction term is dropped, the logistic regression procedure tests the same hypothesis as the Mantel-Haenszel procedure (Swaminathan & Rogers, 1990).

Relative measurement invariance can also be tested using logistic regression. Because the concept of relative measurement invariance requires that there be no effect of group membership, given the relative position on the trait, z_i is substituted for t_i in the regression (recall that z_i is the within-group standardized value of t_i, so that one needs a set of absolute items to estimate t_i before this procedure can be carried out). This gives

$$P(X = 1) = \frac{\exp(c_0 + c_1 z_i + c_2 v_i + c_3 z_i v_i)}{1 + \exp(c_0 + c_1 z_i + c_2 v_i + c_3 z_i v_i)}. \tag{16}$$

Again one proceeds by checking the significance of the parameters c_2 and c_3, but now significant parameter values indicate relative DIF instead of absolute DIF. Analogous to the absolute case, a significant value for the parameter c_3 indicates an interaction between group membership and the latent trait, corresponding to nonuniform relative DIF. A significant value for the interaction parameter c_3 without a significant value for c_2 indicates uniform relative DIF.

If an item shows relative measurement invariance but absolute DIF, the item may be used as a relative indicator of the trait in question. This requires modeling relative measurement, which implies that the absolute and relative items be treated differently. For absolute items, item parameters should be equal across groups as usual. For relative items, however, the (absolute) item parameters will differ across groups if the trait distributions differ (see the previous section). Now, under relative measurement invariance, the differences in discrimination and slope are functions of the difference in trait distributions. The relations between the absolute parameters in both groups are simple and can be deduced from Formulae 13 and 14. Setting the right-hand side of Formula 13 equal for two groups and solving for the absolute difficulty in Group 1 gives

$$b_{j1_{\text{abs}}} = \sigma_{T_1} \left[\frac{b_{j2_{\text{abs}}} - \mu_{T_2}}{\sigma_{T_2}} \right] + \mu_{T_1} \tag{17}$$

for the difficulty parameters, where the second subscript on these parameters indicates group. For the discrimination parameters, the following is obtained:

$$a_{j1_{\text{abs}}} = \frac{\sigma_{T_2}}{\sigma_{T_1}} a_{j2_{\text{abs}}}. \tag{18}$$

Modeling relative item responses can be carried out using these relations. A (slightly ad hoc) method for doing this would consist of the following three steps. First, estimate the means and variances of the trait distributions in both groups using only a set of absolute items. This provides estimates for the means and variances of the trait distribution in the different groups. Second, estimate the absolute item parameters for the relative items in one group (use the largest group for better parameter estimation). This provides estimates for the absolute item parameters for the relative items in one group, so that the difficulty and discrimination parameters for each relative item can be inserted into the right-hand side of Formulae 17 and 18. Finally, fix the absolute parameters for the relative items in the second group at the values given by Formulae 17 and 18. This method is somewhat ad hoc but has the advantage of being simple and easy to implement in widely available software. Also, this procedure yields the possibility to assess the fit of the entire model with absolute and relative items, thus testing the fit of the absolute and relative part of the model simultaneously.

Another option that may be taken, which is especially useful in a SEM approach, is to conceptualize the relative within-group dimension as a separate latent variable. SEM programs such as LISREL (Jöreskog & Sörbom, 1993) are flexible enough to specify an absolute latent variable for the absolute items and a relative latent variable for the relative items. The relative latent variable is then restricted in such a way that it becomes a within-group standardized rescaling of the absolute latent variable. This requires that the relative latent variable correlates perfectly with the absolute latent variable within groups and that it has a mean of zero and a variance of one within each of the groups. To provide a within-group correlation of one between the absolute and relative latent variable, the covariance matrix of these latent variables must be subjected to nonlinear restrictions. Furthermore, the mean and variance of the relative latent variable are fixed at zero and one, respectively, and specified to be invariant across groups. The between-group differences in means and variances for the absolute latent variable, however, are freely estimated. Then one subjects the entire model to a test for strict factorial invariance (Meredith, 1993) to test for relative measurement invariance. The formal details of this model are outlined in the appendix. This is an elegant procedure for fitting the relative model and a useful extension of the SEM framework. To the authors' knowledge, widely available IRT software does not allow the required restrictions to be imposed. For dichotomous item responses, this approach can therefore only be taken indirectly through the analysis of tetrachoric correlations with SEM programs.

Illustration 1

Some of the ideas and procedures set forth in this article will be illustrated by analyzing a Dutch version of the Personality Research Form-E (PRF-E), a widely used personality questionnaire due to Jackson (1974). The PRF-E was administered to 157 male and 279 female undergraduate psychology students. Absolute and relative measurement invariance will be assessed with respect to sex.

A nice property of the concept of relative measurement invariance is that it is possible to do a quick scan of a scale to see whether it may contain items for relative measurement—which is difficult with absolute measurement invariance. The reason for this is that the restrictions of relative measurement invariance imply that the p values of relative items are equal across groups. So, if a scale consists of a number of items with unequal p values across groups, but there is also a set of items with equal p values, this may indicate that the items with equal p values are items for relative measurement of the trait in question.

Several scales in the PRF-E showed this pattern, but it is most pronounced in the subscale Impulsivity. Because this analysis is a mere illustration of some of the ideas presented in this article, the present analysis is limited to this scale. The pattern of p values for males and females is shown in Table 1.

Table 1
p Values for the Items in the PRF-E Subscale Impulsivity

Item	*p* (males)	*p* (females)
1. Often I stop in the middle of one activity in order to start something else.	.64	.62
2. I often say the first thing that comes into my head.	.50	.63
3. When I go to a store, I often come home with things I had not intended to buy.	.42	.58
4. Many of my actions seem to be hasty.	.47	.46
5. I have often broken things because of carelessness.	.50	.47
6. Most people feel that I act impulsively.	.41	.45
7. Sometimes I get several projects started at once because I don’t think ahead.	.59	.59
8. I find that thinking things over very carefully often destroys half the fun of doing them.	.44	.56
9. I am careful to consider all sides of an issue before taking action.	.52	.64
10. I am pretty cautious.	.31	.34
11. Rarely, if ever, do I do anything reckless.	.71	.71
12. Emotion seldom causes me to act without thinking.	.41	.66
13. I have a reserved and cautious attitude toward life.	.32	.46
14. My thinking is usually careful and purposeful.	.36	.60
15. I am not one of those people who blurt things out without thinking.	.46	.61
16. I generally rely on careful reasoning in making up my mind.	.31	.45

Note. Negative items (Items 9 to 16) have been recoded so that all p values represent the proportion of indicative responses. *Note*. PRF-E = Personality Research Form-E.

These results suggest the existence of relative and absolute items in the scale. The Items 1, 4, 5, 6, 7, 10, and 11 show almost identical p values for males and females, which may indicate relative measurement. On the other hand, Items 2, 3, 8, 9, 12, 13, 14, 15, and 16 show higher p values for females, which may indicate a sex difference in latent trait distributions—females being more impulsive. This can be checked by applying the logistic regression procedure. The items hypothesized to be items for absolute measurement are combined in a subscale, generating an absolute total score. This sum score is then standardized within each group to generate a relative score. Subsequently, the amount of DIF for each item is evaluated with respect to the absolute score (to detect absolute DIF), and the relative score (to detect relative DIF) by assessing the effect of sex on the item response. The results yielded by this procedure are reported in Table 2. Only the results concerning uniform DIF are reported, because none of the items showed nonuniform absolute or relative DIF.

Table 2
Standardized Parameter Estimates for the Effect of Sex in the Logistic Regression Procedure

Item	Absolute Bias: Effect of Sex (standardized parameter estimate)	Relative Bias: Effect of Sex (standardized parameter estimate)
1	−1.33	−0.29
2	−0.59	3.04
3	0.82	3.48
4	−3.15	0.43
5	−1.95	0.58
6	−2.12	0.63
7	−2.83	−0.13
8	−0.21	2.45
9	0.65	2.85
10	−1.85	0.63
11	−2.13	−0.05
12	10.03	5.65
13	−0.55	2.85
14	1.37	5.56
15	−0.65	3.18
16	−0.79	3.13

Note. A positive parameter estimate indicates that females have a higher probability of an indicative answer, conditional on their absolute/relative score.

The results are in line with the initial hypothesis. The items hypothesized to be items for relative measurement conform to the idea that they measure relative to the other items in the scale, consistently showing absolute but no relative DIF. An exception is item 1, showing neither absolute nor relative DIF. The theoretical impossibility of such a result, given the difference in absolute score distributions, implies that this is due to a lack of power. The absolute items also behave as expected, consistently showing relative DIF but no absolute DIF, except for Item 12. This item shows both absolute and relative DIF—presumably caused by the explicit use of the word *emotion*—and should probably be removed.

Of course, these results should be interpreted with some caution; although the items do behave as relative items, inspection of the content of the items does not yield obvious reasons why this should be so. Further research should give more insight into the item features that trigger relative measurement. A research strategy that could give some insight in the response processes involved would be to present the items with and without explicit instructions for comparison. So, items could be administered with the instruction to compare oneself to a fixed reference group (e.g., males), with the instruction to compare oneself to a variable reference group (e.g., the group to which one belongs), and without any instruction at all. Comparing ICCs across these situations should provide information on the relevant response processes, which would in turn strengthen the validity of this and other personality scales.

Illustration 2

As mentioned, the concepts of absolute and relative measurement invariance generalize to other latent variable models such as the congeneric model often used in SEM. To illustrate the approach for the SEM model, a subset of data collected by Rodriguez Mosquera, Manstead, and Fischer (in press) is used. They constructed scales to measure several types of honor concerns. A total of 61 male and 61 female Dutch undergraduate psychology students completed the scales. A subset of items of a scale called Feminine Honor Concerns is analyzed, and measurement invariance with respect to sex is evaluated. The items request the participant to rate, on a 7-point scale, how bad he or she would feel if the descriptions given in the items applied to him or her. The content of the items is given in Table 3 along with the means and standard deviations for both sexes.

Table 3
Means and Standard Deviations for Males ($n = 61$) and Females ($n = 61$) on Items in the Scale for Feminine Honor Concerns

Item Content: "How bad would you feel if the following description applied to you?"	Mean (SD) for Males	Mean (SD) for Females
1. Wearing provocative clothes	2.10 (1.14)	2.16 (1.39)
2. Sleeping with someone without starting a serious relationship with that person	2.48 (1.63)	2.77 (1.42)
3. Changing partner often	2.82 (1.38)	3.44 (1.32)
4. Being known as having different sexual contacts	2.77 (1.57)	3.90 (1.38)

A unidimensional model with strict factorial invariance constraints across groups (Meredith, 1993) was fitted to test for measurement invariance. Although the model cannot be rejected, χ^2 (14) = 21.77, $p = .08$, overall fit is not satisfactory (RMSEA = .08), and inspection of modification indices suggests the presence of DIF. Likelihood ratio tests, conducted by individually freeing intercept parameters, reveal uniform bias for item 2, χ^2 (1) = 6.85, $p < .05$, and for item 4, χ^2 (1) = 7.00, $p < .05$.

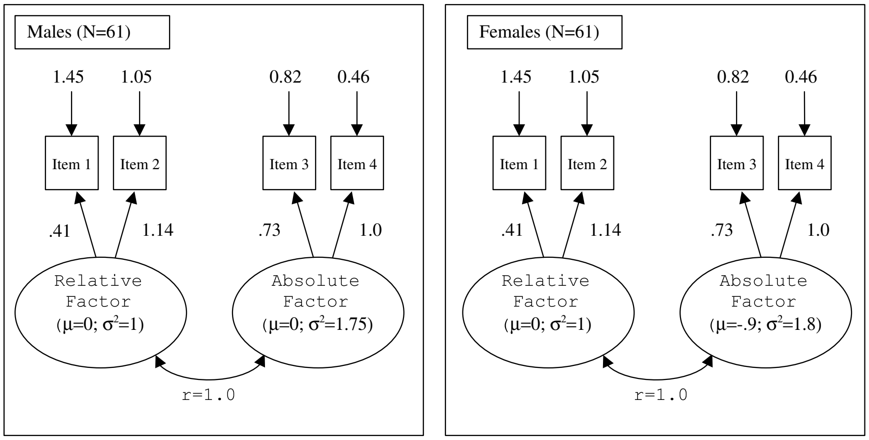
As can be seen from Table 3, however, the observed means of Items 1 and 2 are almost equal across groups. The content of the items "wearing provocative clothes" and "sleeping with someone without starting a serious relationship with that person" suggest that these items may be interpreted differently by men and women. It is not implausible that participants interpret the content of the items conditional on their sex. If this is the case, it implies that these items may be treated as relative within-group indicators. A model specifying these items as relative indicators was fitted by using the SEM procedure described in the previous section (see the appendix for the technical details). A graphical representation of the model is given in Figure 4.

The model cannot be rejected, χ^2 (14) = 15.84, $p = .32$, and fits the data very well (RMSEA < .01). In accordance with these results, inspection of modification indices does not reveal substantial misfit anywhere in the model. Given the fact that the number of parameters in the model is equal to the number of parameters in the absolute model with strict factorial invariance, the better fit of the relative model suggests that this model should be preferred. This may indicate that Items 1 and 2 do indeed measure relative to Items 3 and 4, which may teach us more about the structure of honor concerns in male and female populations. This, in turn, may provide valuable information for theory development in this area.

Discussion

The theory and research presented in this article provide some insight into the complicated relation between measurement invariance and construct validity. It has been argued that not all items that show DIF in the classical sense are invalid. Rather, a failure to distinguish between absolute and relative forms of measurement will lead to apparent bias of items for relative measurement.

Figure 4
A Structural Equation Model for Relative Measurement
for an Item With Both Absolute Bias and Relative Bias



Note. Items 1 and 2 load on the relative factor, and Items 3 and 4 on the absolute factor. The relative factor is obtained by standardization within groups and correlates unity with the absolute factor.

Items for relative measurement can be valid indicators of a trait within groups, but because of their relative nature, these items are bound to produce bias as defined in the classical sense. If the relative nature of the items is recognized, they do not have to be eliminated from a test. Instead, they can be used as relative indicators of the trait in question.

The distinction between absolute and relative measurement has some implications for the theory of measurement invariance and bias. If the latent trait distribution differs across groups, an item will show either absolute bias, relative bias, or both: Absolute measurement invariance and relative measurement invariance cannot simultaneously hold, unless the trait distributions are identical. If the trait distributions differ, relative measurement invariance of a given item will cause that item to show absolute bias. Bias in the classical sense can therefore result from relative measurement invariance. This is an intriguing result because it contradicts the view that all bias results from multidimensionality.

The relation between bias and multidimensionality should be constructed as follows. Bias is a group difference in the distribution of item responses conditional on the latent trait. Multidimensionality is a possible explanation for the presence of bias. Now, it is sometimes suggested that bias is multidimensionality because a biased item "measures" group membership in addition to the variable of interest. So, in a very general sense, group membership is then conceived of as the second dimension. This line of reasoning may be maintained, but in this case multidimensionality is no longer an explanation of item bias: Such an explanation would be circular because the group difference is exactly the phenomenon that requires an explanation. Thus, in this line of reasoning, all bias is multidimensionality, all multidimensionality is bias, and there does not seem to be a good reason for entertaining two words for the same concept. As a consequence, either of the terms should be dropped from the psychometric vocabulary. We do not endorse such a point of view and take the relation between multidimensionality and bias to be of an explanatory nature. This implies that the second variable that the item measures in addition to the intended trait must be a variable that is distinct from group membership, although it must in some way be related to group membership (otherwise the variable could not influence the item responses differentially). The most sophisticated theory of the relation between this second variable and bias is the theory presented in Shealy and Stout (1993). Shealy and Stout showed that a second variable could produce bias, if the groups differ in the distribution on this variable. In their theory of multidimensionality, group differences in the distribution of the second trait are therefore a necessary condition for bias to occur (Shealy & Stout, 1993, p. 209 ff.). In other words, there has to be some association between this second trait and group membership. However, this is obviously not the case in relative measurement, because the distribution on a relative latent variable will often not be associated with group membership—for example, if the absolute trait distributions are normal. In view of this problem, there are two ways to proceed. Either Shealy and Stout's theory has to be revised to accommodate for the relative position on the measured variable as a second dimension producing bias or it must be concluded that relative measurement does not imply multidimensionality. The first of these options requires that, for example, absolute height and relative height be two different traits. In our view, this would render the concept of multidimensionality rather trivial. We therefore take the second option and submit that relative measurement invariance does not imply multidimensionality but unidimensional measurement of the intended trait within groups. We conclude that not all bias results from multidimensionality.

A failure to recognize the fact that items provide relative measurement may produce distortions in the interpretation of data. For instance, in personality research, researchers obviously assume that the items in a personality scale are items for absolute measurement. This assumption is, however, not self-evident. If the assumption is not fulfilled, this may lead to incorrect conclusions regarding personality differences between groups. This is a direct result from the fact that absence of impact cannot be distinguished from relative measurement invariance without a substantial number of absolute items or a separate criterion. Consider, for example, an assertiveness scale in which most or all items are actually items for relative measurement (i.e., the item responses result from an explicit or implicit comparison of participants with other members of a relevant group). A psychologist obtains responses from American and Japanese participants. Suppose that the American participants are in fact more assertive than the Japanese. What would happen if he started looking for an effect of nationality on assertiveness? He would never find any, because both groups answer the items by comparing themselves to their own reference group, which automatically results in comparable mean scores on the test. This is not an academic point, because virtually nothing is known about the cognitive processes involved in responses to personality items. Whether this kind of distortion occurs, and if so, how grave its consequences are, is of course a question for empirical research. Nevertheless, research in this area may profit from taking the relative nature of items in personality scales into account. An interesting line of research would consist in assessing absolute and relative measurement invariance of personality items with respect to a behaviorally inspired matching criterion. Such research, of course, requires the evaluation of tests at the item level. In this respect, the advantages of the generalized item response theory models (Mellenbergh, 1994) over classical test theory cannot be overemphasized.

The concept of relative measurement invariance could further be applied in a range of other situations. One could, for example, think of cross-cultural research into subjective well-being or happiness: It is not unlikely that people, in responding to items used in these scales, compare themselves with other people in their environment. I may consider myself depressed compared with the people around me, but if I get really depressed and I am admitted for hospitalization, I may consider myself rather happy compared with the people surrounding me there. Concepts such as satisfaction and happiness do seem to have an inherently relative component and are therefore susceptible to relative measurement.

In sum, items with relative measurement invariance but absolute bias are not multidimensional and may be valid within-group indicators of the construct to be measured. Also, the fact that such items occur may lead to theory formation on item response processes outside the cognitive realm. The question then becomes what the practical implications of these findings are, and how they could be of help in practical situations. Should we drastically change the way we make personality tests? Should we be telling participants not to make within-group comparisons? In our opinion, no definitive answers to these questions can, at present, be given. How often the response processes outlined here occur and which item features and person characteristics trigger these processes are questions open to empirical research. Obviously, however, the relation between item response models and item response processes is not clear in domains outside cognitive testing. Within the field of cognitive testing, there is at least a raw image of the response processes that lead to item responses, and to a certain extent these processes have been successfully modeled (see Embretson, 1994, for a good example). Retaining items with relative measurement invariance in cognitive tests does not seem to be a very good idea, for there is little theoretical foundation for such practice. In fact, selecting items with relative measurement would technically be comparable to the item selection rules specified in the Golden Rule Settlement (McAllister, 1993), where the Educational Testing Service agreed to construct tests by giving priority to items showing the least differences between groups. Most psychometricians would agree that this was not a psychometrically sound basis for item selection, because it was based on the presumption that all group differences in the performance on these tests reflect bias. The main reason why retaining items with absolute bias in cognitive tests is not a very good idea, however, is precisely because relative measurement invariance conflicts with the construct definitions. Indeed, if one approaches such items from the perspective of cognitive processes in problem solving, the nature of these processes suggests, or even prescribes, that absolute measurement invariance should hold. This is in sharp contrast with construct definitions and response processes outside the realm of cognitive testing. In fact, it seems somewhat disturbing that the demands of measurement invariance are often generalized to the measurement of personality traits and attitudes, whereas this article clearly shows how a rather simple, and not implausible, response process would destroy measurement invariance in the classical sense. Coupled with the fact that in many research areas, there is very little theory on what happens between item administration and item response, relative measurement invariance may be an important concept, although we cannot, at present, determine its scope or usefulness in practical situations. However, we can safely conclude that the relation between construct validity and measurement invariance is rather intricate, because items without measurement invariance may very well be valid indicators of the construct in question. Therefore, the relation between measurement invariance and construct validity needs to be reconsidered, and theory formation on this subject is called for. Especially, the need to extend the work of Embretson (1994), on the relation between cognitive theories on response processes and latent trait models, to fields other than cognitive testing, seems pertinent.

Appendix

A Relative Modification of the Structural Equation Model

Evaluating measurement invariance for continuous item responses requires testing the strict factorial invariance model of Meredith (1993). This involves the modeling of mean structures through multigroup analysis (Sörbom, 1974). Strict factorial invariance with respect to a selection variable V (here V is taken to indicate group) holds if

$$\mathbf{y}_v = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\alpha}_v \tag{A1}$$

and

$$\boldsymbol{\Sigma}_v = \boldsymbol{\Lambda}\boldsymbol{\Phi}_v\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{A2}$$

where the vector $\mathbf{y}_v$ is a vector of means on observed variables in group v, $\boldsymbol{\tau}$ is a vector of intercepts, $\boldsymbol{\Lambda}$ is the matrix of factor loadings, $\boldsymbol{\alpha}_v$ is the vector of factor means for group v, $\boldsymbol{\Sigma}_v$ is the covariance matrix of the observed variables in group v, $\boldsymbol{\Phi}_v$ is the covariance matrix of the factors in group v, and $\boldsymbol{\Theta}$ is a diagonal matrix containing the variances of the residuals. The strict factorial invariance model specifies that only the factor means and variances may differ between groups.

A simple unidimensional model with one latent variable is taken as the point of departure. This renders $\boldsymbol{\alpha}_v$ and $\boldsymbol{\Phi}_v$ scalars. The model is identified by setting one of the elements in $\boldsymbol{\Lambda}$ to one and the factor mean $\boldsymbol{\alpha}_v$ to zero in one of the groups; $\boldsymbol{\alpha}_v$ is free to vary in the other groups. This is the unidimensional model fitted to the data in Illustration 2.

To cope with relative items, the model is modified as follows. Partition the observed variables into a set of absolute items and a set of relative items. For the absolute items, the strict factorial invariance model is maintained as above. The original single factor is termed the *absolute factor*. For the relative items, a new factor is now invoked, so that $\boldsymbol{\alpha}_v$ is now a 1 × 2 vector and $\boldsymbol{\Phi}_v$ a 2 × 2 symmetric matrix. The relative items are allowed to load only on this second factor. This factor is termed the *relative factor*. To ensure that the relative factor is the within-group standardized variant of the absolute factor, the following restrictions are added to the model. First, the relative factor is required to have the standard normal distribution in each group. Second, the relative factor is required to correlate unity with the absolute factor within each of the groups. This gives the restrictions

$$\boldsymbol{\alpha} = [\alpha_v \;\; 0] \tag{A3}$$

and

$$\boldsymbol{\Phi}_v = \begin{bmatrix} \phi_v & \\ \sqrt{\phi_v} & 1 \end{bmatrix}. \tag{A4}$$

Equation A4 is a nonlinear restriction that can be readily implemented in SEM programs such as LISREL (Jöreskog & Sörbom, 1993). However, admissibility checks should be turned off because $\boldsymbol{\Phi}$ is not positive definite. Equation A4 ensures that the correlation r_{12} between the absolute and the relative factor is equal to $r_{12} = \phi_{12}/\sqrt{\phi_1 \times \phi_2} = \sqrt{\phi_v}/\sqrt{\phi_v \times 1} = 1$, as required. The model with Restrictions 3 and 4 is the relative model fitted to the data in Illustration 2.

References

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

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory. *Journal of Cross Cultural Psychology*, *24*, 133-148.

Embretson, S. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107-135). New York: Plenum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Huang, D. C., Church, T. A., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, *28*, 192-218.

Jackson, D. N. (1974). *Personality Research Form-E*. London: Research Psychologist Press.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8 user's reference guide*. Chicago: Scientific Software International.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models*. New York: Plenum.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

McAllister, P. H. (1993). Testing, DIF, and public policy. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Mellenbergh, G. J. (1982). Contingency table methods for assessing item bias. *Journal of Educational Statistics*, *7*, 105-118.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127-143.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300-307.

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., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing bias. *Applied Psychological Measurement*, *17*, 297-334.

Rodriguez Mosquera, P. M., Manstead, A. S. R., & Fischer, A. H. (in press). The role of honor concerns in emotional reactions to offenses. *Cognition and Emotion*.

Shealy, R., & Stout, W. (1993). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum.

Smith, L. J., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *Journal of Personality and Social Psychology*, *75*, 1350-1362.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *Psychometrika*, *55*, 229-239.

Swaminathan, H., & Rogers, J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

Acknowledgments

The authors wish to thank Harrie C. M. Vorst and Patricia M. Rodriguez Mosquera for making their data available. We also wish to thank Conor Dolan and two anonymous reviewers for some valuable comments. We credit Frits Zegers of Groningen University for being the first to conceive of a height questionnaire.

Author's Address

Send requests for reprints or further information to Denny Borsboom, Department of Psychological Methods, Faculty of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, the Netherlands. E-mail: ml_borsboom.d@macmail.psy.uva.nl.