Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Intelligence 31 (2003) 543–566
INTELLIGENCE

On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model

Gitta H. Lubke[a,*], Conor V. Dolan[b], Henk Kelderman[c], Gideon J. Mellenbergh[d]

[a]*Graduate School of Education and Information Studies (GSEIS), University of California, Los Angeles (UCLA), Moore Hall, Box 951521, Los Angeles, CA 90095-1521, USA*
[b]*Developmental Psychology, University of Amsterdam, Amsterdam, The Netherlands*
[c]*Department of Work and Organizational Psychology, Free University Amsterdam, Amsterdam, The Netherlands*
[d]*Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands*

Received 16 January 2001; received in revised form 20 March 2003; accepted 27 March 2003

Abstract

Investigating sources of within- and between-group differences and measurement invariance (MI) across groups is fundamental to any meaningful group comparison based on observed test scores. It is shown that by placing certain restrictions on the multigroup confirmatory factor model, it is possible to investigate the hypothesis that within- and between-group differences are due to the same factors. Moreover, the modeling approach clarifies that absence of measurement bias implies common sources of within- and between-group variation. It is shown how the influence of background variables can be incorporated in the model. The advantages of the modeling approach as compared with other commonly used methods for group comparisons is discussed and illustrated by means of an analysis of empirical data.

Keywords: Common factor model; Within-group differences; Between-group differences; Measurement invariance; Confirmatory factor analysis

* Corresponding author. Tel.: +1-310-206-2053; fax: +1-310-206-6293.
E-mail address: glubke@ucla.edu (G.H. Lubke).

0160-2896/03/$ – see front matter © 2003 Elsevier Science Inc. All rights reserved.
doi:10.1016/S0160-2896(03)00051-5

1. Introduction

Investigating within- and/or between-group differences on test scores is the focus of a large number of research studies. The variance of an item or subscale score *within* a group indicates the individual differences within the group. Individual differences with respect to multiple observed variables may be summarized in a within-group variance–covariance (or correlation) matrix. The structure of this matrix can be investigated using confirmatory (or exploratory) factor analysis. Confirmatory factor analysis (CFA) is concerned with "explaining" the common content of observed variables captured by their covariances with a smaller number of underlying latent variables called factors. As CFA is applied to the covariance matrix within a single group, the common factors can be regarded as the sources of systematic within-group differences. Differences *between* groups, on the other hand, are often tested by comparing the groups with respect to the means of the observed scores or with respect to the means of the factors underlying the observed scores. The latter may be viewed as an analysis of the sources of between-group differences and can be done by carrying out multigroup CFA.

To render the group comparisons meaningful, it is necessary to address the issue of measurement invariance (MI) and demonstrate that a given test measures the same underlying factors across groups. We use the expression "same factor" to indicate that a factor has exactly the same conceptual interpretation across groups. The interpretation of a factor depends on the content of the observed items or subscales that are related to the factor and the strength of those relations. Consequently, for a factor to have an identical interpretation across groups, it is necessary that the relations of the observed variables and the underlying factor are exactly the same across groups.

The present paper focuses on the relation between these three aspects of group comparisons, namely the relation between the sources of within-group differences (i.e., which factors explain individual differences within a group?), the sources of between-group differences (i.e., which factors explain the differences between groups?), and MI (i.e., does the test measure the same factors in all groups?). Although all three aspects have been extensively investigated separately, the relation between the three has not been clearly examined. In addition, we show that hypotheses concerning these three issues can be tested using multigroup confirmatory factor models.

Two areas, in which the relation between within- and between-group differences and MI is important, are ethnic group differences in IQ test scores and the seemingly linear increase over time in mean IQ test scores, termed the "Flynn effect" (Flynn, 1987, 1999). In these areas of research, it has been frequently noted that sources of within-group differences and sources of between-group differences are not necessarily identical (Lewontin, 1970). Differences between ethnic groups on an IQ test may be due to other factors than those which contribute to the individual differences within each of the groups. Although several studies focus explicitly on the issue of within- and between-group differences, both in a more general context (Turkheimer, 1990, 1991) and specifically with respect to the Flynn effect (Flynn, 2000; Rodgers, 1998), it is common practice to investigate the two sources of variance separately. Examples are single group factor analysis and multiple regression, which are

based on within-group differences, and (M)ANOVA, in which groups are compared with respect to their observed means. Another common strategy is to compare the test score means adjusted for the influence of some variable of interest (Phillips, Brooks-Gunn, Duncan, Klebanov, & Crane, 1998). As will be shown, this approach is based on the implicit assumption that sources of within-group variance are the same as sources of between-group variance. This assumption is usually not tested in practice. In order to show that within- and between-group differences are indeed due to the same factors, it is necessary to analyze the means and the covariances of the observed scores simultaneously.

As mentioned above, if comparisons of item or subscale scores are to be valid, the test has to measure the same underlying factors in all groups. The concept of MI provides a theoretical framework, which includes the necessary conditions to establish whether a given test measures the same factors in the groups under consideration. The definition of MI states that, conditional on the factor scores, observed scores do not depend on group membership. This means that members of different groups who have the same score on the factor (e.g., the same level of ability) have on average the same observed scores. The definition of MI implies that groups may differ only with respect to the means and covariances of the factors that are measured by the observed scores. In practice, MI can be investigated by fitting multigroup CFA models to a given data set. To represent MI, certain model parameters are restricted to be equal across groups. Both the restricted model and a less restricted model are fitted to the data. The models may be compared by means of a likelihood ratio test. The test can provide evidence that MI is tenable (for applications, see Dolan, 2000; Dolan & Hamaker, 2001).

The central issue of the present paper concerns the relation between MI on the one hand and within- and between-group differences on the other hand. Specifically, the definition of MI across groups implies that between-group differences cannot be due to factors with a different conceptual interpretation than the factors that account for the within-group differences. Although the importance of MI has been acknowledged (Byrne, Shavelson, & Muthén, 1989; Dolan, 2000; Lubke, Dolan, & Kelderman, 2001; Marsh, 1994; McArdle, 1998; Oort, 1998), this implication is not well recognized. Hence, if in practice the hypothesis of MI is not rejected, one can conclude with some confidence that within- and between-group differences are attributable to the same factors.

Given the importance of conclusions in areas such as ethnic differences and/or the Flynn effect, it is surprising that, at least to our knowledge, few of the recent studies use multigroup CFA. A possible reason for the lack of using more state-of-the-art methods may lie in the rather technical character of publications discussing implications of MI (Bloxom, 1972; Ellis, 1993; Meredith, 1993). Although some technical formulation is unavoidable, it is our aim to explain the relation between MI and common sources of within- and between-group in an accessible way and to discuss the advantages of using multigroup confirmatory factor models rather than other commonly used methods. The approach proposed in this paper is applicable to a wide range of research questions. The approach is adequate if groups are to be compared on tests that consist of a larger number of continuous items or subscale scores, which are assumed to measure a smaller number of underlying factors. This includes group comparisons on multidimensional test batteries (e.g., IQ test batteries) as well group comparisons on personality, mood, and attitude questionnaires or combinations of these.

The paper is organized as follows. First, the multigroup CFA model is presented. We show that observed scores are decomposed into common factor scores and a regression residual, which comprises measurement error and item specific error. This decomposition has the advantage that groups can be compared with respect to the means and covariances of the factors. Second, we explain the concept of MI on a theoretical level and on a more practical level in the context of the multigroup common factor model. The multigroup common factor model corresponding to MI is characterized by a set of invariance restrictions across groups. Third, we show that MI implies that between-group differences are unlikely to be due to other factors than those capturing systematic within-group differences. We discuss how this result can be used in practice. By comparing a model with the invariance restrictions across groups to a less restricted model in a likelihood ratio test, one can examine not only whether MI holds but also whether between-group differences are due to differences in the same factors as the within-group differences. Fourth, we discuss how the multigroup model can be extended to include background variables. The way in which background variables are integrated can be guided by the outcome of tests of MI (Oort, 1992, 1998). Finally, we briefly discuss the advantages of multigroup CFA as compared with other commonly used methods and present, for the purpose of illustration, an analysis of scores of African and Caucasian Americans on an IQ test (Osborne, 1980).

2. The multigroup model

The basic idea in multigroup CFA as opposed to single group analysis is to fit factor models in several groups simultaneously. The factor model fitted within a group is a linear regression model, which relates observed item or subscale scores to a smaller number of latent variables called factors. Say we have $i = 1, \ldots, I$ observed scores, $\boldsymbol{Y}$, measuring $l = 1, \ldots, L$ factors. Suppose further that the total sample consists of $j = 1, \ldots, J$ subjects each belonging to one of $s = 1, \ldots, S$ groups. If $I = 6$, $L = 2$, $J = 300$, and $S = 2$, we would have a test with six items or subscales measuring two factors and 300 test takers that are divided over two groups, for instance, 180 subjects in one group and 120 in the other. The within-group model for the score of subject j on item i can be expressed as follows:

$$y_{ij} = v_i + \sum_{l=1}^{L} \lambda_{il}\eta_{jl} + \varepsilon_{ij}. \tag{1}$$

As can be seen, the observed score y is modeled as the sum of a regression intercept, v, the scores on the different factors, $\boldsymbol{\eta}$, each multiplied with the corresponding slope parameters, λ, and a residual, ε. In the context of the factor model, the parameters for the slope, λ, are called factor loadings. Note that the intercepts and the factor loadings are the same for all subjects but may differ across items; hence, the intercept v and the factor loading λ have no subscript j but the item subscript i in Eq. (1). The factor score $\boldsymbol{\eta}$ is specific to each subject and has subscript j for the subject and subscript l to indicate which of the L factors we are referring to. A factor score represents, for instance, the individual's math ability level. Consequently, it

does not vary across items. The residual contains error that is specific for a given item and random measurement error of a given person (Bollen, 1989). It has subscripts *ij*, meaning that the regression residual may vary across items and subjects.

The multigroup confirmatory factor model can be fitted to the means and covariances of the observed items or subscales instead of the raw scores (Sörbom, 1974). To obtain the mean of item $\boldsymbol{Y}_i$ within a group, we average across subjects in that group. The mean of the residual is assumed to be zero; hence, the mean of item $\boldsymbol{Y}_i$ is

$$\mathscr{E}(\boldsymbol{Y}_i) = v_i \sum_{l=1}^{L} \lambda_{il} \mathscr{E}(\boldsymbol{\eta}_l), \tag{2}$$

where $\mathscr{E}$ denotes the expected value or mean. Commonly, the mean of an observed item score is denoted as μ and the mean of a factor as α Hence, we get

$$\mu_i = v_i + \sum_{l=1}^{L} \lambda_{il} \alpha_l. \tag{3}$$

For I items or subscales, we have I of these equations

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_I \end{bmatrix} + \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1L} \\ \lambda_{21} & \cdots & \lambda_{2L} \\ \vdots & \ddots & \cdots \\ \lambda_{I1} & \cdots & \lambda_{IL} \end{bmatrix} \times \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_L \end{bmatrix}. \tag{4}$$

which can be summarized in matrix notation as

$$\boldsymbol{\mu}_s = \boldsymbol{\nu}_s + \boldsymbol{\Lambda}_s \boldsymbol{\alpha}_s. \tag{5}$$

The subscript s is added to indicate that this is the equation for the means of group s, $s = 1, \ldots, S$. The dimensions of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are $1 \times I$, $\boldsymbol{\Lambda}$ is $I \times L$, and $\boldsymbol{\alpha}$ is $1 \times L$.

The matrix equation for the variances and covariances can be derived in a similar way. Again, Eq. (1) is the point of departure, and the variances and covariances of observed variables are expressed in terms of underlying factors and residuals. The residuals are assumed not to be correlated with the factors and not intercorrelated. The intercepts are constants and have therefore zero covariance with the factors and the residuals. Hence, the (co)variances of the observed scores, $\boldsymbol{Y}$, equal the sum of the factor (co)variances pre- and postmultiplied with the corresponding factor loadings and the variances of the residual scores. We adopt again commonly used notation and denote the covariance matrix of the items as $\boldsymbol{\Sigma}$,

the covariance matrix of the factor scores as Ψ, and the variances of the residuals as Θ. The resulting equation for the variances and covariances in group s is

$$\Sigma_s = \Lambda_s \Psi_s \Lambda_s^t + \Theta_s. \quad (6)$$

It is important to note that the equation for the means and the equation for the covariances are both derived from the same regression equation presented in Eq. (1). Recall that Eq. (1) described the regression of the observed item scores on underlying factor(s). Hence, Eq. (5) describes the means of the item scores in terms of the means of those factors, whereas Eq. (6) describes the (co)variances of the items in terms of the covariances of the factors. The factor loadings, Λ, are the same in the model for the means and the model for the covariances, that is, for a given model a factor loading has the same value in the mean model (Eq. (5)) and in the covariance model (Eq. (6)).

The multigroup confirmatory factor model comprises the model for the means as shown in Eq. (5) and the model for the covariances as shown in Eq. (6). The parameters of the full model are the regression intercepts, the factor loadings, the factor means, the factor (co)variances, and the residual variances. These parameters can be used to impose a specific structure on the means of the observed scores and their covariances. Parameters can be restricted to take the same value in each group (i.e., to be invariant across groups) and/or to take a specific value. By comparing restricted models with less restricted models, one can test the hypothesis that the restrictions are tenable.

By fitting the full model to the means and (co)variances from several groups simultaneously, one can compare (1) the means across groups (i.e., the between-group differences) *and* (2) the covariances across groups (i.e., the within-group differences). These two comparisons are done simultaneously in a single analysis. Note, however, that between-group comparisons with respect to factor means and covariances are meaningful only if the observed scores are not biased, that is, if the observed scores are measurement invariant across groups. In Section 3, we elaborate on MI (i.e., unbiasedness) and show that it requires some of the model parameters to be invariant across groups. Imposing and testing these restrictions is straightforward in the multigroup model. This is followed by a section in which we show that multigroup model restricted to represent MI implies that between- and within-group differences are due to the same factors.

3. Measurement invariance

MI has been defined in a very general context, independent of the sort of data at hand (e.g., binary items, continuous items or subscales, etc.) or the type of model for the data. Essentially, it is a statement that the distribution of observed variables given the underlying factor scores is the same in all groups. In the context of IQ scores for instance, this means that given a certain level of, say, verbal ability, all test takers have the same probability of answering a verbal item correctly, independent to which group the test takers belong.

As an approach to the definition of MI, we utilize the concept of selection (Bloxom, 1972; Ellis, 1993; Meredith, 1993). Groups can be conceptualized as being derived by applying a function to a selection variable. Suppose the selection variable is "ethnicity," then a suitable selection function may be one, which assigns the value 0 to Hispanics, the value 1 to Caucasians, and the value 2 to African Americans. This selection function results in three ethnic groups. The selection variable is sometimes called grouping variable because it indicates group membership. We refer to it by the letter V.

The definition of MI (or absence of bias) with respect to the selection variable V has been given by Mellenbergh (1989) as

$$f(\boldsymbol{Y} \mid \boldsymbol{\eta}, s) = f(\boldsymbol{Y} \mid \boldsymbol{\eta}), \tag{7}$$

where, as before, $\boldsymbol{Y}$ and $\boldsymbol{\eta}$ are observed scores and factor scores, respectively. A distribution function is denoted as $f(\cdot)$ and $V=s$ is the selection variable, which determines group membership s. This definition states that, given a subject's factor score(s) $\boldsymbol{\eta}$, the subject's observed scores $\boldsymbol{Y}$ do not depend on group membership. The definition of MI focuses on the distribution of the observed scores and is therefore not confined to a specific model. It is applicable regardless of the model for the observed scores.

Meredith (1993) has introduced a weak form of MI (WMI), in which invariance is limited to the means (denoted as $\mathscr{E}(\cdot)$) and covariances (denoted as $\boldsymbol{\Sigma}(\cdot)$) of the distribution of $\boldsymbol{Y}$: $\mathscr{E}(\boldsymbol{Y}|\boldsymbol{\eta},s)=\mathscr{E}(\boldsymbol{Y}|\boldsymbol{\eta})$ and $\boldsymbol{\Sigma}(\boldsymbol{Y}|\boldsymbol{\eta},s)=\boldsymbol{\Sigma}(\boldsymbol{Y}|\boldsymbol{\eta})$. Since the multigroup model discussed here is based on the assumption of multivariate normality of the factor scores and normality of the residuals, WMI and MI coincide.

A number of researchers have provided methods to investigate MI in the context of the common factor model (see, for instance, Marsh, 1994; McArdle, 1998; Mellenbergh, 1994a, 1994b; Oort, 1998). The present article is based on the work by Meredith (1993). Meredith has elaborated the relation between MI (and WMI) and the multigroup factor model. More specifically, he has shown how the parameters of the multigroup model have to be restricted such that the model represents MI. There are three sets of restrictions implied by MI. First, the regression intercepts, $\boldsymbol{\nu}$ in Eq. (5), have to be invariant across groups. Second, the factor loadings, $\boldsymbol{\Lambda}$ in Eqs. (5) and (6), have to be invariant across groups. The third restriction concerns the regression residuals and consists of three parts: the residual variances, $\boldsymbol{\Theta}$ in Eq. (6), have to be invariant across groups, all residual covariances have to be zero (i.e., no correlated errors), and the residual means have to be zero. As mentioned above, zero residual means and zero residual covariances are part of the model described above. Together, these restrictions are called "strict factorial invariance" (SFI). Meredith has provided proofs that SFI almost certainly ensures MI.

On a more conceptual level, the necessity of the three restrictions can be understood as follows. First, suppose a regression intercept of one of the observed items differs across two groups, say in Group 1 the intercept equals 2 and in Group 2 it equals 1. All other parameters being invariant across groups, this means that the first group scores consistently higher on that item than the other group for each value of the factor. The

regression line corresponding to the regression of that item on the factor in Group 1 is parallel to that of Group 2 but is located higher. As a result, the means of the observed scores depend on group membership, i.e., $\mathscr{E}(\boldsymbol{Y}|\boldsymbol{\eta},s)$ does not equal $\mathscr{E}(\boldsymbol{Y}|\boldsymbol{\eta})$. Hence, MI does not hold.

Secondly, differences in factor loadings (i.e., differences in the slopes of the regression of observed variables on the factor) across groups can be interpreted in terms of an interaction between group membership and factor scores. Again, take the two groups as an example, but now suppose that they have different slopes of the regression of a given item on a factor. Consequently, the regression lines of the two groups cross at some level of the factor score. Only at that level of the factor score do the groups have the same average observed score. For lower factor scores, one group scores lower than the other, and for higher factor scores, it is the other way around. Such an interaction is a form of bias because, again, for a given factor score (excluding the value where the regression lines cross), the observed scores depend on group membership. Thus, factor loadings that differ for the groups are also inconsistent with the definition of MI.

Third, regarding the residual variances, suppose that a test is used for an admission decision and a certain level of ability (as measured by the factor) is required. If the decision is based on the observed scores, then the number of incorrect admissions and incorrect rejections is higher in the group with the larger residual variance (see also Meredith, 1993). More formally, given equal factor loadings, $\boldsymbol{\Sigma}(\boldsymbol{Y}|\boldsymbol{\eta},s)$ does not equal $\boldsymbol{\Sigma}(\boldsymbol{Y}|\boldsymbol{\eta})$ and MI does not hold. The argument with respect to the necessity of zero residual means is the same as with respect to the necessity of equal intercepts. If in a data set the residual means are nonzero, this would be manifested in terms of the intercepts. Group differences in residual means show as differences in intercepts because it is part of the multigroup model that the residual means are restricted to zero. As mentioned above, unequal intercepts implies that one group scores consistently higher than another group and MI is absent. Finally, residual covariances have to be zero. A nonzero residual covariance between two observed items implies that these items have something in common in addition to the factors. Hence, a residual covariance can be interpreted in terms of an additional factor. One may think that restricting the residual covariances (i.e., the correlated errors) to be equal across groups would be sufficient for MI, but in fact it is not. Regard the residual covariance between the two items in terms of an additional factor. For the same reason as discussed above, the factor loadings of both items on the additional factor have to be equal across groups to achieve MI, it is not sufficient to fix their covariance or correlation to be equal.[1]

The requirement of group-invariant intercepts, group-invariant factor loadings, and residuals with group-invariant variances, zero means and zero covariances can be easily specified in the multigroup model. Intercepts, loadings, and residual variances are specified to

[1] If the residual covariance is represented in terms of an additional factor, then, conditional on the other factors, the residual covariance equals the product of the factor loadings and the variance of the additional factor. To achieve MI, not this product but both factor loadings have to be equal. The variance of the additional factor may in fact differ across groups.

be invariant across groups, and the residual covariances are fixed to zero.[2] The restricted measurement invariant model can be represented as follows:

$$\boldsymbol{\mu}_1 = \boldsymbol{\nu}, \tag{8}$$

$$\boldsymbol{\mu}_s = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha}_s, \qquad s = 2, \ldots, S \tag{9}$$

$$\boldsymbol{\Sigma}_s = \boldsymbol{\Lambda}\boldsymbol{\Psi}_s\boldsymbol{\Lambda}^t + \boldsymbol{\Theta}. \tag{10}$$

One of the requirements necessary for the estimation of the model is to fix the factor means, α, to zero in one of the groups; here, we have chosen Group 1 (Sörbom, 1974). The requirement is due to the fact that the scale of the factors is arbitrary and has to be fixed. As a result, the parameters α estimated in the remaining groups represent factor mean differences with respect to the first group.

Comparing Eqs. (5) and (6) with Eqs. (8)–(10), one can see that in the restricted model, only the factor means and the factor covariances may differ across groups in the MI model. Neither of the other model parameters has a group subscript. This is exactly what the definition of MI requires.

4. MI implies that between-group differences cannot be due to other factors than those accounting for within-group differences

The statement that between-group differences are attributable to the same sources as within-group differences (or a subset thereof) is another way of saying that mean differences between groups cannot be due to other factors than the individual differences within each group. To confirm this statement, we have to show that two propositions are tenable by the usual statistical criteria: (1) that the same factors are measured in the model for the means as in the model for the covariances and (2) that the same factors are measured across groups.

The first part follows directly from the way the multigroup model has been derived. We have shown that the two parts of the multigroup model, the model for the means and the model for the covariances, have been deduced from the same regression equation (Eq. (1)). Eq. (1) specifies the relation between observed scores and underlying factors. To derive the multigroup model, we have taken the mean of Eq. (1) (as shown in Eq. (5)) and the variances and covariances (see Eq. (6)). Taking means and (co)variances does not change the relation between observed scores and their underlying factors as specified in Eq. (1). The factors in the model for the means are the same as in the model for the covariances because both submodels are derived from the same regression equation of observed variables on the factors.

[2] In most software programs for confirmatory factor analysis, it is not necessary to explicitly fix the residual means to zero.

The second part is implied by the concept of MI. The concept of MI has been developed by Meredith (1993) to provide the necessary and sufficient conditions to determine whether a set of observed items actually measures the same underlying factor(s) in several groups. MI states that the only difference between groups concerns the factor means and the factor covariances *but not the relation of observed scores to their underlying factors*. Only if the relation of an observed variable to an underlying factor differs across groups, one can argue that a "different factor" is measured in those groups. If Eq. (1) holds across groups with identical parameter values, with the understanding that the mean and the covariances of the factors, $\boldsymbol{\eta}$ in Eq. (1), may differ, then one can conclude that the proposition that same factors are measured across groups is tenable.

To illustrate our argument, we discuss two scenarios that show why differences in the sources of within- and between-group differences are inconsistent with MI. First, we discuss the case that all factors underlying between-group differences are different from the factors underlying within-group differences. Second, we consider a situation in which the within-group factors are a subset of the between-group factors, that is, the two types of factors coincide but there are additional between-group factors that do not play a role in explaining the within-group differences. In addition, we show that the case, where between-factors are a subset of the within-factors, is consistent with MI and that the modeling approach provides the means to test which of within-group factors does not contribute to the between-group differences.

Suppose observed mean differences between groups are due to entirely different factors than those that account for the individual differences within a group. The notion of "different factors" as opposed to "same factors" implies that the relation of observed variables and underlying factors is different in the model for the means as compared with the model for the covariances, that is, the pattern of factor loadings is different for the two parts of the model. If the loadings were the same, the factors would have the same interpretation. In terms of the multigroup model, different loadings imply that the matrix $\boldsymbol{\Lambda}$ in Eq. (9) differs from the matrix $\boldsymbol{\Lambda}$ in Eq. (10) (or Eqs. (5) and (6)). However, this is not the case in the MI model. Mean differences are modeled with the same loadings as the covariances. Hence, this model is inconsistent with a situation in which between-group differences are due to entirely different factors than within-group differences. In practice, the MI model would not be expected to fit because the observed mean differences cannot be reproduced by the product of $\boldsymbol{\alpha}$ and the matrix of loadings, which are used to model the observed covariances. Consider a variation of the widely cited thought experiment provided by Lewontin (1974), in which between-group differences are in fact due to entirely different factors than individual differences within a group. The experiment is set up as follows. Seeds that vary with respect to the genetic make-up responsible for plant growth are randomly divided into two parts. Hence, there are no mean differences with respect to the genetic quality between the two parts, but there are individual differences within each part. One part is then sown in soil of high quality, whereas the other seeds are grown under poor conditions. Differences in growth are measured with variables such as height, weight, etc. Differences between groups in these variables are due to soil quality, while within-group differences are due to differences in genes. If an MI model were fitted to data from such an experiment, it would be very likely

rejected for the following reason. Consider between-group differences first. The outcome variables (e.g., height and weight of the plants, etc.) are related in a specific way to the soil quality, which causes the mean differences between the two parts. Say that soil quality is especially important for the height of the plant. In the model, this would correspond to a high factor loading. Now consider the within-group differences. The relation of the same outcome variables to an underlying genetic factor are very likely to be different. For instance, the genetic variation within each of the two parts may be especially pronounced with respect to weight-related genes, causing weight to be the observed variable that is most strongly related to the underlying factor. The point is that a soil quality factor would have different factor loadings than a genetic factor, which means that Eqs. (9) and (10) cannot hold simultaneously. The MI model would be rejected.

In the second scenario, the within-factors are a subset of the between-factors. For instance, a verbal test is taken in two groups from neighborhoods that differ with respect to SES. Suppose further that the observed mean differences are partially due to differences in SES. Within groups, SES does not play a role since each of the groups is homogeneous with respect to SES. Hence, in the model for the covariances, we have only a single factor, which is interpreted in terms of verbal ability. To explain the between-group differences, we would need two factors, verbal ability and SES. This is inconsistent with the MI model because, again, in that model the matrix of factor loadings has to be the same for the mean and the covariance model. This excludes a situation in which loadings are zero in the covariance model and nonzero in the mean model.

As a last example, consider the opposite case where the between-factors are a subset of the within-factors. For instance, an IQ test measuring three factors is administered in two groups and the groups differ only with respect to two of the factors. As mentioned above, this case is consistent with the MI model. The covariances within each group result in a three-factor model. As a consequence of fitting a three-factor model, the vector with factor means, α in Eq. (9), contains three elements. However, only two of the element corresponding to the factors with mean group differences are nonzero. The remaining element is zero. In practice, the hypothesis that an element of α is zero can be investigated by inspecting the associated standard error or by a likelihood ratio test (see below).

In summary, the MI model is a suitable tool to investigate whether within- and between-group differences are due to the same factors. The model is likely to be rejected if the two types of differences are due to entirely different factors or if there are additional factors affecting between-group differences. Testing the hypothesis that only some of the within factors explain all between differences is straightforward. Tenability of the MI model provides evidence that measurement bias is absent and that, consequently, within- and between-group differences are due to factors with the same conceptual interpretation.

5. Testing the measurement invariant model

The multigroup model can be fitted using standard software such as M*plus* (Muthén & Muthén, 2002), Lisrel (Jöreskog & Sörbom, 1999), EQS (Bentler, 1993), or Mx (Neale, M.C.,

Boker, S.M., Xie, G., & Maes, H.H., 2002). Tenability of the MI model may be evaluated on the basis of measures of goodness-of-fit and/or likelihood ratio tests (see, for instance, Bollen, 1989; Bollen & Long, 1993). Since MI is a composite hypothesis consisting of three restrictions, rejection of MI can have several causes. In order to distinguish between the causes of misfit, it may be useful to conduct the analysis in two steps (Mandys, Dolan, & Molenaar, 1994). In both steps, the mean model and the covariance model are fitted simultaneously. First, only the factor loadings and the residual variances are restricted to be equal across groups. The residual covariances are fixed to zero. The factor means are fixed equal to zero in all groups and the restriction of equal intercepts over groups is relaxed. Eqs. (8)–(10) change to

$$\boldsymbol{\mu}_s = \boldsymbol{\nu}_s \tag{11}$$

$$\boldsymbol{\Sigma}_s = \boldsymbol{\Lambda}\boldsymbol{\Psi}_s\boldsymbol{\Lambda}^t + \boldsymbol{\Theta}. \tag{12}$$

Here, the mean model is saturated. By equating observed means to the regression intercepts, for each observed mean a separate parameter is estimated, which, in addition, may vary across groups. Fitting this model serves to investigate whether the restricted model for the covariances holds across groups (i.e., within-group differences due to the same factors across groups). Some researchers allow group-specific residual variances (Little, 1997). As mentioned before, group-specific residual variances may be problematic (see also Lubke & Dolan, 2003; Meredith, 1993). If equal factor loadings and error variances across groups result in the rejection of the covariance model, modification indices may help to identify the variables that cause the misfit (i.e., identify the biased variables).[3] In case one has a hypothesis that a given background variable is causing the bias, one may proceed by extending the model. Incorporating such a variable in the model may eliminate the misfit (see Section 6). If the model fits adequately, in a second step, covariances and the *restricted* mean model (i.e., Eqs. (8)–(10)) are fitted simultaneously. In this step, it is tested whether observed mean differences can be accounted for by mean differences in the same factors as within-group differences. Appreciable decrease of model fit with respect to the first step of the analysis will occur if observed mean differences are not solely due to mean differences in the factor underlying within-group differences. The decrease can be interpreted in terms of absence of MI because intercept differences between groups are due to other factors in addition to the factors underlying the within-group differences.

The decrease in model fit can be tested statistically with a likelihood ratio test because the models are nested: a model with a restricted mean structure is compared with the same model with an unrestricted mean structure. Given the assumption of multivariate normality of the observed scores and a sufficiently large sample, the decrease in model fit can be evaluated if the restricted model is correct. Under these conditions, the difference in χ^2 resulting from fitting the two models is χ^2 distributed. The degrees of freedom of the likelihood ratio test

[3] Modification indices are provided by the output of standard software for structural equation modeling and indicate the decrease in χ^2 if a restricted parameter is freed and the model is reestimated (Sörbom, 1989).

equals the difference in free parameters between the two models. In case the first step results in an adequate fit of the covariances and the likelihood ratio test is not statistically significant, one can be reasonably sure that MI is tenable and that between-group differences are not due to other factors than within-group differences. Other criteria may also be employed to determine tenability of the model. Measures of goodness-of-fit such as RMSEA, BIC, CAIC, and others are discussed by Bollen and Long (1993). For an illustration, see Dolan and Hamaker (2001).

Testing MI in the multigroup model can serve to investigate specific hypotheses. One can compare groups at one point in time or at different time points.[4] For instance, consider the Flynn effect. The Flynn effect concerns a seemingly linear increase over time in observed IQ test scores, which has been consistently observed way in several countries. The effect is most notable on items requiring problem solving ability. The gains happen too fast to be of genetic origin (Flynn, 1987, 1999). Researchers doubt that it is in fact an increase in intelligence factors but attribute the effect to various other influences such as increased environmental complexity, improved nutrition (Schooler, 1998; Sigman & Whaley, 1998), or, as Flynn (2000) has proposed, increased time investment in abstract problem solving. To reject the hypothesis that there is an increase in mean intelligence factors and to conclude that the gains in observed scores are due to other variables than the intelligence factors, one can start with a model that does not include any of the potential explanatory variables. The MI model is fitted to the means and covariances of two or more generation groups and compared with the fit of a less restricted model. If the higher observed scores of later generations are not due to the intelligence factors measured by the test, the MI model should fit appreciably worse than the less restricted model, because the generation mean difference in observed scores cannot be explained by mean differences in the intelligence factors. Such an analysis would also reveal which of the observed items or subscales show gains that are not explained by gains in the factors. By inspecting the content of these items, one can develop hypotheses concerning other variables that may explain the gains and include these variables as background variables in the model (see Section 6).

The MI model can also serve to compare different groups at one time point. Examples of analyses of MI using data previously analyzed by Jensen and Reynolds (1982) and Naglieri and Jensen (1987) are described by Dolan (2000), Dolan and Hamaker (2001), and Gustafsson (1992).

6. Model extension with background variables

Frequently, researchers have data concerning the subjects in addition to the test scores they want to analyze. There are two ways of integrating background variables in the multigroup model. First, one can specify the hypothesis that background variables influence only the

[4] Investigating the same group across time is called latent growth modeling and is thoroughly described elsewhere (Muthén, 2001).

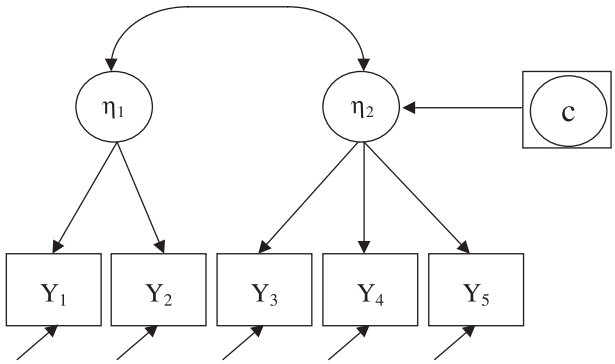

Fig. 1. Option 1: Factor scores η_2 are regressed on a background variable, c which can be latent or observed.

factor(s). We will call this Option 1. Option 1 serves to investigate structural relations, for instance, the hypothesis that nutrition has an impact on IQ factors. Importantly, the influence of the background variable on the observed scores is indirect since it is conveyed through the factors (see Fig. 1). Second, one can specify that a background variable has a direct influence on observed scores in addition to the influence on the factors, i.e., Option 2. This option is suitable if one wants to eliminate the influence of a variable, for instance, age, and estimate the influence of the factors on the observed scores while controlling for age. The two options are depicted in Figs. 1 and 2.

The background variable c may be an observed variable or a factor and is therefore depicted in a circle as well as in a rectangle. The observed means of Items 1 and 2 in Fig. 1 consist of the sum of an regression intercept and a direct effect of the factor η_1, whereas the observed means of Items 3–5 consist of the sum of an intercept, the direct effect of η_2, and the indirect effect of c, which is conveyed through η_2.

Option 2 (see Fig. 2) can be accommodated by simply adding direct effects of the background variable on the observed item(s).

The choice between the two options can be guided by testing a measurement invariant multigroup model.[5] First, a model is fitted without the background variables. The mean model is unrestricted, that is, Eqs. (11) and (12) specify the model. Rejection in the first step of the analysis indicates that the observed within-group covariances cannot be modeled with the same factor structure across groups. MI does not hold without even including the means. There are variables that influence the observed scores differentially across groups. To investigate whether one of the background variables is causing the deviation from MI, Option 2 can be specified to include direct effects of the background variables on the observed items. If, on the other hand, the first step leads to an acceptable fit, the full MI model with restricted means is fitted and compared with the less restricted model. This is done again without the background variables. If the decrease in model fit of the MI model as compared with the less restricted model is significant, one may inspect the modification

[5] See Oort (1992, 1998) for a single group approach of using background variables to detect bias.

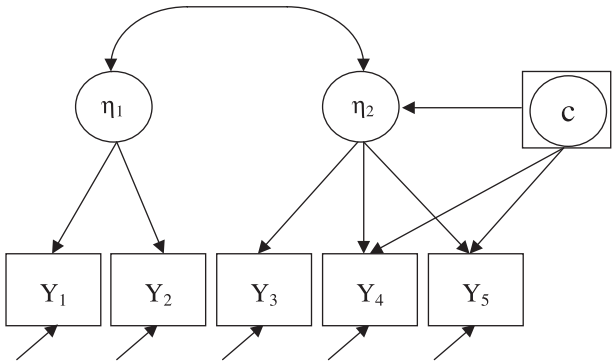

Fig. 2. Option 2: The latent or observed background variable, *c*, has additional direct effects on observed variables $\boldsymbol{Y}_4$ and $\boldsymbol{Y}_5$.

indices pertaining to the intercepts to identify items that have mean differences that are not explained well by the factors. As before, Option 2 can be specified to include direct effects of the background variable(s) on the items that are not measurement invariant. Suppose the model decrease is not significant, one can conclude that the within-group variation is due to variation in the factors and that observed mean differences between groups can be explained by mean differences in the factors. MI is tenable. In that case, it is not necessary to specify direct effects of the background variables on the observed variables because the only variation that remains to be explained is the within- and between-group variation in the factors.[6] Hence, Option 1 can be specified to estimate the regression of the factors on the background variables.

7. Disadvantages of other commonly used methods for group comparisons

In what follows we briefly discuss some of the drawbacks of alternative methods for group comparisons. Methods for group comparisons such as those used in recent studies concerning the Flynn effect and ethnic differences hinge on, at times implicit, assumptions about the relation between sources of within-group variance and between-group variance. In addition, the issue of MI is not always addressed adequately. Although the MI model is certainly not restricted to the analysis of IQ test results, we think that this area of research may benefit greatly by adopting a comprehensive model, which integrates MI and within- and between-group sources of variance.

One of the alternative methods is multiple regression. Multiple regression is a method for observed variables. It has been used, among others, to evaluate the effect of variables such as, say, welfare status, on ethnic differences in IQ test scores (Herrnstein & Murray, 1994).

[6] An exception is a background variable, which influences only the within variation but not the between differences of a single item. This is discussed by Lubke and Dolan (2003).

Usually, the item scores are summed and the resulting test score is predicted by several observed variables. Hence, one of the problems of this approach arises from the fact that predictors in multiple regression are assumed to be measured without error, which is unlikely the case with measurements in the social sciences. The consequences of possibly different error variances across groups are neglected. It is well known that neglecting measurement errors can lead to biased estimates (Rock, Werts, & Flaugher, 1978). Moreover, Millsap (1995, 1997) has discussed in detail the paradoxical relation between measurement and prediction bias. He has shown that it is very likely that the prediction of an observed variable from observed predictors will be biased across groups if the predictors are in fact measurement invariant. The problem can be solved by modeling the structural relations between the factors underlying the observed predictors and the test scores instead of using the observed variables.

Another common strategy to examine the effect of a particular variable on mean score differences is to compare the test score means adjusted for the influence of that variable (Phillips et al., 1998). Again, this type of analysis is usually carried out with observed variables (e.g., test sum scores), which makes investigation of MI impossible. In addition, this approach is based on the implicit assumption that within-group variance and between-group variance have common sources. Within-group variation is used as a proxy for between-group variation. Suppose a given factor, say *E*, accounts for 10% of the variance in test scores within both of two groups. It is possible that the same factor varies to a much larger degree between groups. Comparison of mean differences corrected for the within-group influence of *E* leads to meaningful results only under the assumption that the within-group impact of *E* is comparable with the between-group impact. However, this assumption is rarely tested in practice.

The third method that has been repeatedly used to investigate ethnic group differences in IQ test scores has originally been proposed by Jensen (1985) and is called "method of correlated vectors." The method, which has been extensively discussed elsewhere (see special issues in *Multivariate Behavioral Research*, 1992 and *Cahiers de Psychologie Cognitive*, 1997), is designed to test the hypothesis that black–white differences in IQ test scores are fundamentally due to differences in general intelligence (or "g"). For a comparison of this specific method and the multigroup model, the reader is referred to Lubke et al. (2001). Using simulated data, that study shows that Jensen's method lacks specificity. Group differences were ascribed to mean differences in a general intelligence factor when simulated population data were not in accordance with such conclusions. Roorda, Dolan, and Wicherts (2003) drew the same conclusion on the basis of reanalyses of published data set. For a detailed criticism of the method of correlated vectors, see Dolan and Hamaker (2001).

The last issue relates to the search for explanatory factors (i.e., suitable background variables in the MI model). Suppose that the MI model is tenable for IQ test data from two generation groups. Hence, factor mean differences can account for the observed mean differences. The mean difference across generations is usually larger in a problem-solving factor than in a, say, verbal factor, because the gains over time are much more pronounced in nonverbal items involving abstract reasoning. Consequently, when deriving hypotheses concerning the sources of the factor mean difference, one should think of background

variables that have a lesser impact on the verbal factor. This issue has already been stressed by Flynn (1998). The same argument holds for ethnic differences: apparently, Caucasian Americans outperform African Americans on all subtests; however, the disparity seems to be less in memory-related items (Dolan, 2000; Jensen, 1985). Therefore, a background variable accounting for ethnic differences should have less impact on a memory factor. However, in recent studies in this research area, the effects of background variables is only investigated in relation to total test scores, not with respect to certain factors or items. Differential effects of background variables on factors can be easily evaluated using the MI model.

8. Illustration using Osborne's twin data

For the empirical example, we use data published in Osborne (1980). The subjects are African and Caucasian American twins drawn from public and private schools in Kentucky, Georgia, and Indiana. Note that this is clearly not a representative sample for the population of African and Caucasian Americans in the United States; conceptual interpretations of the analysis below are therefore not generalizable. The analysis is included for illustrative purposes.

The data are scores on four subscales of the Primal Mental Ability Test developed by Thurstone and Thurstone (1938). The first scale is verbal meaning, which is designed to measure the ability to understand ideas expressed in words ("verbal"). Secondly, we have number facility, which refers to the ability to work with numbers ("number"). Next, there is reasoning, that is, the ability to solve logical problems ("reason"), and finally, we include spatial relations, which refers to the ability to visualize objects and figures rotated in space and the relations between them ("spatial"). In addition, we have age, sex, and SES status as assessed by the method of Warner, Meeker, and Eells (1949). Our illustration consists of a two-group analysis using the data pertaining to one twin of each pair. Ethnicity is the grouping variable. Age, sex, and SES are used as background variables. Data from the second twin of each pair are used for validation. In Table 1, summary statistics for blacks and whites

Table 1
Summary statistics of the Osborne data

	Sex	Age	SES	Verbal	Numerical	Reason	Spatial
Twin 1							
Blacks	69/36	14.7 (1.5)	16.1 (4.3)	88.4 (12.1)	86.0 (14.6)	86.0 (14.2)	89.1 (13.0)
Whites	38/47	15.4 (1.4)	13.0 (3.7)	102.2 (15.1)	101.3 (18.2)	103.9 (15.9)	102.1 (16.3)
Twin 2							
Blacks	68/37	14.7 (1.5)	16.1 (4.3)	90.6 (12.8)	86.2 (13.7)	86.7 (15.1)	90.8 (13.3)
Whites	50/35	15.4 (1.4)	13.0 (3.7)	105.1 (14.8)	101.2 (16.8)	106.6 (14.3)	101.6 (13.9)

Sex is presented as number of females/males. All other values are averages with standard deviations given between brackets.

Table 2
Observed variable R^2 of the covariance model

	Verbal	Numerical	Reason	Spatial
Blacks	.465	.493	.589	.347
Whites	.638	.664	.744	.518

are given for the twins separately. All analyses are carried out using M*plus* 2.0 (Muthén & Muthén, 2002).

Based on the results of a preliminary exploratory factor analysis, we choose a single-factor model with the four subscales as observed variables to model the data within each group. Following the two-step procedure described above, we first fit the model for the covariances with equal factor loadings and residual variances across groups and leave the model for the means unrestricted. The latter means that in both groups we specify the factor means to be zero and estimate the intercepts. The fit of this model is very good (see fit measures for "covariance model" in Table 3). The largest modification index is smaller than 5. The observed variable R^2, that is, the variance of the subscales that is due to the single factor, is higher in the white group. Note that although groups are restricted to have equal residual variances and factor loadings, they may differ in the factor variance; hence, differences in R^2 as observed here are possible. The observed variable R^2 are shown in Table 2.

To investigate whether these differences are substantial, we fit a covariance model with unrestricted means in which the factor variances are also fixed to be equal across groups. The χ^2 is 21.202 with 12 *df*. The difference in χ^2 with the first covariance model is therefore 8.2. With 1 *df* difference, this is a significant deterioration in model fit. Hence, groups apparently differ in factor variance. In what follows, we leave the factor variance unrestricted across groups.

The next model we fit is the full MI model. The intercepts are set equal across groups and the factor mean is fixed to zero in one group and estimated in the other. The choice of groups is arbitrary; here, we estimate the factor mean in the white group. The fit of the model is good (see Table 3, full MI model). The factor mean difference between the two groups is 1.164 standard deviations favoring the whites. The standardization is done using the factor variance of the white group. The fit of the model for the covariances is compared with the fit of the full MI model, still without covariates. The difference in χ^2 equals 2.2 with 3 *df* difference, which is not significant. Hence, both the actual fit of the full MI model

Table 3
Measures of goodness-of-fit

Model	χ^2	*df*	RMSEA	SRMR	CFI
Covariance model	12.998	11	0.044	0.059	0.993
Full MI model	14.236	14	0.013	0.063	0.999
MI model with SES	19.109	21	0.0	0.084	1.0

The *P*-values of all fitted models are >.05.

and the fit compared with the covariance model provide evidence that the Primal Mental Ability Test is measurement invariant on a subscale level across the two ethnic groups. Since the residual variances and factor loadings are equal, differences in the variance of the observed subscales are due to differences in factor variance. Differences in the means of the subscales are due to a difference in factor mean. Since the within-group differences (i.e., the covariances) and the between-group differences (i.e., the means) are modeled simultaneously using the same regression equation (i.e., the same factor loadings, intercepts, and residual variances), we have evidence that within- and between-group variance are due to the same factor.

The last step in our analysis concerns the incorporation of background variables. Since the MI model holds, we choose Option 1 as explained above and regress the factor on the covariates. The aim is to investigate whether factor mean differences are at least partially due to mean differences in the covariates. Hence, we conduct preliminary t tests to compare blacks and whites with respect to the covariates. These preliminary mean comparisons have 1 *df* and show that groups differ significantly with respect to SES ($t_{SES} = 7.4568$). The t values for age and sex are smaller ($t_{age} = -4.5954$ and $t_{sex} = -2.6753$). All differences are statistically significant. Hence, in the model with covariates, we regress the single factor of our model on age, sex, and SES. The results reveal that the regression on age and on sex is not significant in either group. For SES, the results are more pronounced. The regression of the factor on SES is significant in both groups. Holding the regression of the factor on SES measurement invariant across groups, we find that the factor mean difference is slightly decreased and equals 0.977 standard deviations as compared with 1.164 standard deviations in the MI model without SES. Both standardized mean differences are derived using the factor variance in the white group. To avoid the dependence on the factor variance in the white group, one can compute the percentage decrease in factor mean difference, which is due to SES. We find that SES "explains" $\sim$16% of the factor mean difference. Measures of goodness-of-fit of the covariance model, the MI model, and the MI model with SES are shown in Table 3.

The important difference with regression models without latent variables is that here we investigate the relation of a background variable such as SES with the factor underlying the observed IQ test scales *while knowing that these scales actually measure the same factor in both groups*. The relation of SES and the underlying factor is examined in the absence of measurement error, which accounts for a considerable proportion of the observed variance. If the impact of SES were investigated without explicitly taking the proportion of measurement error in each of the subscales into account, this might lead to distorted results (Millsap, 1995).

Since the data published by Osborne (1980) are twin data, we can validate our results using the second twin of each pair as a validation sample. The general line of results is very similar. Although the model fit of all models is slightly inferior, the covariance model with equal factor loadings and residual variances is tenable. The largest modification index is smaller than 7. The decrease in model fit when adding the restrictions in the model for the means (i.e., fitting the full MI model) is not statistically significant (χ^2 difference is <2 with *df* difference = 3; see Table 4). Hence, also in the validation sample, the test is measurement invariant with respect to the ethnic groups. The factor mean difference is somewhat higher in

Table 4
Measures of goodness-of-fit in the validation sample

Model	χ^2	*df*	RMSEA	SRMR	CFI
Covariance model	18.640	11	0.086	0.087	0.969
Full MI model	20.121	14	0.068	0.090	0.975
MI model with SES	25.790	21	0.05	0.084	0.982

The *P*-values of all fitted models are >.05.

the validation sample and equals 1.478 standard deviations. When incorporating the covariates, we find that, as before, age and sex have no significant effect on the factor, but SES does. Regressing the factor on SES results in a decrease of the factor mean difference to 1.181 standard deviations. This corresponds to 20% of the factor mean difference being due to SES. The validation analysis therefore largely confirms the previous results.

9. Discussion

If groups are to be compared on observed test scores, it is necessary to investigate whether the test is measurement invariant across groups. MI in the factor model, that is, absence of bias, implies that within- and between-group variations are due to the same factors. Consequently, establishing MI and investigating between- and within-group differences coincide in the context of the multigroup confirmatory factor model. A model restricted according to MI can be compared with a less restricted model in a likelihood ratio test, which makes MI a testable hypothesis. If the MI model is rejected, it is very likely that between-group differences are due to different factors than within-group differences *and* that the test at hand is biased across the groups under investigation.

The multigroup model, incorporating the restrictions implied by MI, provides a useful tool to investigate the Flynn effect and ethnic group differences. The possibility to extend the model is a further advantage. By adding background variables, one can investigate whether observed mean differences or mean differences in intelligence factors are (partially) due to other factors. Multigroup models with background variables allow a more thorough investigation of many of the proposed explanations of the Flynn effect and ethnic group differences.

There are, of course, limitations of the multigroup approach. Although rather far fetched, there are three possible exceptions to our argument that MI implies common sources of between- and within-group differences. Two exceptions are due to the fact that the regression residual of the confirmatory factor model consists of the sum of random error and a contribution specific to the observed item. First, mean differences in the residual are absorbed by the intercept because residual means are restricted to be zero. Hence, the regression intercept of an observed score can be rewritten as $\nu^* = \nu + \mathscr{E}(\varepsilon)$, where $\mathscr{E}(\varepsilon)$ is the nonzero mean (or expected value) of the residual. Equating the intercept ν^* across groups as part of the set of MI restrictions does not exclude the possibility that groups differ in ν to the same

extend as they differ in $\mathscr{E}(\varepsilon)$ and that the two differences cancel each other out. In that situation, MI would hold although there is a twofold bias. As shown by Meredith (1993), the same indeterminacy exists with respect to the residual variances: if groups differ in specific variance to the same amount as in random error variance, and these group differences compensate each other, then the residual variance is invariant across groups. Since this is not a likely scenario, we think that consideration of these two exceptions should be based on evidence indicating that there might indeed exist two or more sources of bias with effects that compensate each other. If there is a hypothesis at hand concerning potentially biasing factors, these can be measured and included as background variables. The third exception concerns the factor loadings. As mentioned above, it is unlikely that a set of items is related to two conceptually different factors in exactly the same way. However, let us consider this possibility. In terms of the factor model, it would imply that the factor loadings on the conceptually different factors are equal or strictly proportional. As a consequence, $\boldsymbol{\Lambda}\boldsymbol{\eta}_1 + \boldsymbol{\Lambda}\boldsymbol{\eta}_2 = \boldsymbol{\Lambda}(\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2) = \boldsymbol{\Lambda}\boldsymbol{\eta}^*$, meaning that the factors cannot be distinguished. Therefore, equating $\boldsymbol{\Lambda}$ across groups cannot guarantee that the same factors are measured across groups, and similarly, equating $\boldsymbol{\Lambda}$ across the mean and the covariance model cannot guarantee that within- and between-group differences are due to the same factors. However, it is hard to imagine why the interpretation of factors with equal factor loadings should be different across groups or, similarly, why the interpretation of the factor should be different for the mean and the covariance model. The interpretation of factors with respect to their content depends on inspection of the content of the observed variables. When analyzing data with a modeling approach, the derived factors exist merely by grace of their factor loadings: they are what observed variables have in common. If the common content of the variables can be interpreted in terms of conceptually different factors, the researcher can add observed variables to the test in order to arrive at a unique interpretation of the factors. Hence, we conclude that these three exceptions do not represent a serious threat to the interpretation of the model.

The more serious problem may concern the power to reject the MI model. For instance, how sensitive is the model with latent means in detecting the influence of a biasing variable on the observed mean difference of a test score? Although an extensive power study is lacking, Lubke, Dolan, Kelderman, and Mellenbergh (2003) have shown that, given a correlation of .3 between the biasing factor and the selection variable and/or the factors underlying the test, a factor loading of size 0.3 on the biasing factor is sufficiently large to reject the MI model even if the reliability of the observed variables was low (e.g., 0.4). In that study, sample sizes were equal across groups ($N_1 = N_2 = 1000$). Large differences in sample size across groups might be problematic. Kaplan and George (1995) have shown that unequal sample sizes across groups have a negative effect on the power to reject a MI-related hypothesis.

Aside from these limitations, simultaneous investigation of sources of within- and between-group variance based on the MI model is a useful approach. The approach described in the present paper is adequate in case the observed variables are continuous and measure one or several underlying factors that are also continuous. Within-group variation is explicitly specified to be due to the same factors as between-group variation. Representation in terms of

these models may facilitate the conceptualization of a given hypothesis regardless of any attempt to actually fit the models to data. Evaluation of a hypotheses can be based on measures of goodness-of-fit. If the hypothesis, that within- and between-group differences are due to the same factors, is rejected, a researcher may attempt to detect the items with mean differences that cannot be explained by the factors than those underlying the test. This may help to identify suitable background variables and include them in the model. Incorporation of background variables as described in the previous sections can serve to substantially strengthen (or refute) many of the currently proposed explanations regarding the Flynn effect and ethnic differences. Although in the present paper we have focused on group comparisons on cognitive tests, the range of possible applications of the multigroup common factor model restricted to investigate MI extends beyond this field of research.

Acknowledgements

The research by the first author was supported through a subcontract to grant 5 R01 HD30995-07 by NICHD. The research of Conor Dolan was made possible by a fellowship of the Royal Netherlands Academy of the Arts and Sciences.

References

Bentler, P. (1993). *EQS structural equations program manual*. Los Angeles: BMDP.

Bloxom, B. (1972). Alternative approaches to factorial invariance. *Psychometrika*, *37*, 425–440.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A., & Long, J. S. (1993). *Testing structural equations models*. Newbury Park: Sage Publications.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466.

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., & Hamaker, E. (2001). Investigating black–white differences in psychometric IQ: Multi-group confirmatory factor analysis and a critique of the method of correlated vectors. In F. Columbus (Ed.), *Advances in Psychological Research*, *vol. 6* (pp. 31–60). Huntington: Nova Science.

Ellis, J. L. (1993). Subpopulation invariance of patterns in covariance matrices. *British Journal of Mathematical & Statistical Psychology*, *46*, 231–254.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ really measures. *Psychological Bulletin*, *101*, 171–191.

Flynn, J. R. (1998). IQ-gains over time: Toward finding the causes. In U. Neisser (Ed.), *The Rising Curve: Long-term gains in IQ and related measures* (pp. 25–66). Washington, DC: American Psychological Association.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist*, *54*, 5–20.

Flynn, J. R. (2000). IQ-gains, WISC subtests and fluid g: g theory and the relevance of Spearman's hypothesis to race. In N. Foundation (Ed.), *The nature of intelligence* (pp. 202–227). Chichester: Wiley.

Gustafsson, J. E. (1992). The relevance of factor analysis for the study of group differences. *Multivariate Behavioral Research*, *27*, 319–325.

Herrnstein, R. J., & Murray, C. (Eds.). (1994). *The bell curve*. New York: The Free Press.

Jensen, A. R. (1985). The nature of the black–white difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences*, *8*, 193–263.

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences*, *3*, 423–438.

Jöreskog, K. G., & Sörbom, D. (1999). Lisrel 8.3. [Computer program]. Chicago: Scientific Software International.

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling*, *2*, 101–118.

Lewontin, R. C. (1970). Race and intelligence. *Bulletin of the Atomic Scientists*, *26*, 2–8.

Lewontin, R. C. (1974). The analysis of variance and the analysis of causes. *American Journal of Human Genetics*, *26*, 400–411.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53–76.

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., Dolan, C. V., & Kelderman, H. (2001). Investigating group differences using Spearman's hypothesis: An evaluation of Jensen's method. *Multivariate Behavioral Research*, *36*, 299–324.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). Absence of measurement bias with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical & Statistical Psychology* (in press).

Mandys, F., Dolan, C. V., & Molenaar, P. C. (1994). Two aspects of the simplex model: Goodness of fit to linear growth curve structures and the analysis of mean trends. *Journal of Educational and Behavioral Statistics*, *19*, 201–215.

Marsh, H. W. (1994). Confirmatory factor models of factorial invariance: A multi-faceted approach. *Structural Equation Modeling*, *1*, 5–34.

McArdle, J. J. (1998). Contemporary statistical models of test bias. In J. J. McArdle, & R. W. Woodcock (Eds.), *Human abilities in theory and practice* (pp. 157–195). Mahwah, NJ: Erlbaum.

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: Quantitative Methods in Psychology*, *115*, 300–307.

Mellenbergh, G. J. (1994b). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*, 223–236.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, *30*, 577–605.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*, 248–260.

Muthén, B. O. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In L. Collins, & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 291–322). Washington, DC: APA.

Muthén, L. K., & Muthén, B. O. (2002). M*plus* 2.1. [Computer program]. Los Angeles, CA: Muthén and Muthén.

Naglieri, J. A., & Jensen, A. R. (1987). Comparison of black–white differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence*, *11*, 21–43.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2002). *MX: Statistical modeling*. Richmond, VA: VCU Department of Psychiatry.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika*, *6*, 150–166.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, *5*, 107–124.

Osborne, R. T. (1980). *Twins black and white*. Athens, GA: Foundation for Human Understanding.

Phillips, M., Brooks-Gunn, J., Duncan, G. J., Klebanov, P., & Crane, J. (1998). Family background, parenting practices, and the black–white test score gap. In C. Jencks, & M. Phillips (Eds.), *The black–white test score gap* (pp. 103–145). Washington, DC: Brookings.

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research*, *13*, 403–418.

Rodgers, J. L. (1998). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, *26*, 337–356.

Roorda, W., Dolan, C. V., & Wicherts, J. M. (2003). Two failures of Spearman's hypothesis: The GATB in Holland and the JAT in South Africa (submitted for publication).

Schooler, C. (1998). Environmental complexity and the Flynn effect. In U. Neisser (Ed.), *The Rising Curve: Long-term gains in IQ and related measures* (pp. 67–80). Washington, DC: American Psychological Association.

Sigman, M., & Whaley, S. E. (1998). The role of nutrition in the development of intelligence. In U. Neisser (Ed.), *The Rising Curve: Long-term gains in IQ and related measures* (pp. 155–182). Washington, DC: American Psychological Association.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical & Statistical Psychology*, *27*, 229–239.

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*, 371–384.

Thurstone, L. L., & Thurstone, T. G. (1938). *Primary mental abilities*. Chicago: Chicago University Press.

Turkheimer, E. (1990). On the alleged independence of variance components and group differences. *European Bulletin of Cognitive Psychology*, *10*, 686–690.

Turkheimer, E. (1991). Individual and group differences in adoption studies of IQ. *Psychological Bulletin*, *110*, 392–405.

Warner, W. L., Meeker, M., & Eells, K. (1949). *Social class in America: A manual of procedure for the measurement of social class*. Chicago: Science Research.