Investigating black-white differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC and a critique of the method of correlated vectors

Conor V Dolan & Ellen L Hamaker

Department of Psychology, University of Amsterdam.
Roetersstraat 15, 1018WB Amsterdam, The Netherlands.
Email: c.v.dolan@uva.nl

This chapter appeared as:

Dolan, C.V. and Hamaker, E.L. (2001). Investigating black-white differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC, and a critique of the method of correlated vectors. In: Frank Columbus, Editor, Advances in Psychological Research vol. VI, pp. 31-60. Huntington, NY: Nova Science Publishers, Inc.

See http://users.fmg.uva.nl/cdolan (paper 3.5, for LISREL INPUT FILES)

Abstract

According to Spearman's hypothesis, observed black-white (b-w) differences in psychometric measures of cognitive abilities are attributable to a fundamental difference in general intelligence. Efforts to investigate this hypothesis have centered on the method of correlated vectors. This method involves calculating the correlation between the factor loadings of cognitive tests on the common factor identified as g and the (standardized) observed differences in means on the tests. This correlation is called a Spearman correlation. A large and positive Spearman correlation is supposed to be indicative of the importance of general intelligence as the source of b-w differences.

Several commentators have argued that the issue of b-w differences in psychometric IQ test should be addressed by means of multi-group confirmatory factor analysis (MGCFA). The software to carry out such analyses is in place, as are the exact requirements within this model for a meaningful comparison of groups. The aims of the present chapter are to provide a detailed account of the problems associated with the method of correlated vectors in the light of MGCFA, and to present suitable multi-group factor models to investigate Spearman's hypothesis and other competing hypotheses. These models are fitted to a published data set. The upshot of our critique of the method of correlated vectors is that the the Spearman correlation, as a test of the importance of g in b-w differences, is lacking in specificity. The results of our MGCFA's suggest that it is very difficult to distinguish between competing hypotheses concerning the latent sources of b-w differences.

Introduction

The present chapter concerns the analysis of black-white (b-w) differences in psychometric measures of cognitive abilities (henceforth simply IQ test scores). Differences between black and whites in IQ test scores in the United States are well established, but, within differential psychology, consensus concerning their meaning or causes is conspicuously absent. This state of affairs may well in part be due to the controversial nature of the subject matter. Jencks and Phillips (1998) suggest that most social scientists have chosen safer topics and hoped that the problem would go away. As is evident from good modern textbooks (e.g., Mackintosh, 1998), recent edited volumes (e.g., Jencks and Phillips, 1998; Devlin, Fienberg, Resnick and Roeder, 1997), and articles (e.g., Nyborg and Jensen, 2000), the problem has not gone away.

In so far as the subject of b-w differences in IQ test scores have been on the scientific agenda, it is due in no small part to the work of Jensen. His efforts in this area spans over 30 years and has culminated in the publication of his book "The g factor. The science of mental ability" (Jensen, 1998). This books addresses many aspects relating to the hypothetical construct of general intelligence, or simply g as it is denoted. Two chapters of this book, which addresses many aspects relating to the hypothetical construct of general intelligence, are concerned with population differences in IQ test scores. Jensen devotes a large part of his chapter 11 to presenting and testing the hypothesis that the relative size of the b-w mean differences on diverse IQ tests is a positive function of each test's g loading (i.e., the regression coefficients in the regression of the test on the common g factor in a factor analysis). As the relationship between g loadings and b-w mean differences was first noted by Spearman (1927), Jensen refers to this hypothesis as Spearman's hypothesis. Jensen distinguishes two versions of this hypothesis, a strong version (g as the sole source of group differences) and a weak version (g as the predominant source of group differences). Jensen attaches great importance to this hypothesis because, if proven true, it would identify g as the source of both within population (i.e., individual) and between population (i.e., mean) differences in IQ test scores. This in turn would imply that a scientific understanding of the nature of b-w differences depends on understanding the nature of g (Jensen, 1998, p. 371).

To test Spearman's hypothesis, Jensen devised the method of correlated vectors. This method essentially involves correlating the standardized b-w mean differences on a set of tests with the tests' loadings on the common factor identified as g. The latter are obtained from some form of common factor analysis, which may include g as a second order common factor. A positive and substantial correlation is viewed as support for Spearman's hypothesis. The rationale of this test is simple: if g is the main source of between and within group differences, then there should be a positive relationship between a test's g-loading and the b-w difference in mean on the test. Jensen prescribes the specific conditions that have to be met in applying the method of correlated vectors. The application of this method to a large number of data sets over the past 20 years has produced a mean correlation of .63 (Jensen, 1998, p. 378).

Spearman's hypothesis and the method of correlated vectors have received considerable attention. They have been the subject of a Brain and Behavioral Science target paper (Jensen, 1985), and of special issues of both Multivariate Behavioral Research (Guttman, 1992) and Cahiers de Psychologie Cognitive (Schönemann, 1997). Invited commentaries in these issues have addressed many aspects of the hypothesis and the method of correlated vectors. Furthermore, this subject is still generating commentaries (or commentaries on commentaries, e.g., Dolan and Lubke, 2001), and the method of correlated vectors is being applied on an ever increasing scale in the United States and beyond (Lynn and Owen, 1994; te Nijenhuis and van der Flier 1997; te Nijenhuis, Evers and Mur, 2000; Rushton, 1999).

The present chapter has two aims. First we identify a number of methodological weaknesses of the method of correlated vectors by considering Spearman's hypothesis in the light of the underlying multi-group confirmatory factor model (henceforth MGCFM). Second we apply multi-group confirmatory factor analysis (MGCFA), which we consider to be a superior method to investigate Spearman's hypothesis (both versions), to a published data set (Naglieri and Jensen, 1987). Our comment on the method of correlated vectors is based on our own work (Dolan, 1997; 2000; Dolan and Lubke, 2001; Lubke, Dolan and Kelderman, 2001), but also draws on a number of the published commentaries (Gustafsson, 1992; Millsap, 1997b; Horn, 1997; Loehlin, 1992). The suggestion of applying MGCFA to investigate b-w differences is not new (see Gustafsson 1992; Horn, 1997; Millsap, 1997b; Dolan, 1997), but has not, to our knowledge, resulted in many applications. Dolan (2000) has discussed this methodology and applied it to a published data set (Jensen and Reynolds, 1982). It would appear that the general usefulness of this methodology is not fully recognized. For instance, the recent book by Jencks and Phillips (1998) contains no references to the relevant literature (e.g., Sörbom,1974; Meredith, 1993).

The present chapter is organized as follows. We first present Jensen's procedure for carrying out the method of correlated vectors. Next we present various MGCFM's that may be applied in studying b-w differences. We show how specific cases of these models represent the weak and strong version of Spearman's hypothesis, as well as other, competing hypotheses. With these models in place, we discuss the methodological weaknesses of the method of correlated vectors. We subsequently apply MGCFA to the data set published in Naglieri and Jensen (1987). The models that we consider are the same as those applied by Dolan (2000). We conclude the chapter with a discussion.

The scope of the present chapter is necessarily limited. We do not address the theoretical aspects of g. We recognize that there are those who do subscribe to the idea of g (Jensen, 1998) and those who do not (Horn and Noll, 1997). Concerning covariance structure analysis, we accept that the linear common factor model is a useful point of departure in analyzing IQ test scores, especially in the light of Meredith's work on measurement invariance within the factor model (Meredith, 1993). We assume that the dichotomy black-white, like the dichotomy male-female, is realistic and defensible. However, we make no claims concerning the biological, or indeed the socio-cultural underpinnings of this dichotomy. The nature of group differences depends on the particular phenotype under investigation. For instance, those who study sickle cell anaemia may justifiably view the black-white dichotomy as biological. Those who study group differences in hairstyle may justifiably view the male-female dichotomy as socio-cultural. Beyond the proposition that MGCFA provides a useful framework to consider the possible contribution of genetic and environmental factors, we make no claim concerning the role of these factors. Our application of MGCFA in the present chapter is aimed primarily at investigating how blacks and whites differ phenotypically, not why they differ.

The method of correlated vectors

As mentioned above, Jensen devised the method of correlated vectors to test Spearman's hypothesis. The crux of this test is the relation between the vector of differences in means and factor loadings. Both Pearson product moment correlation and Spearman's rho have been used to quantify this relation. We shall refer to these correlations as Spearman correlations. In calculating these correlations, Jensen (1985; 1992; 1998) advocates a procedure consisting of the following steps. 1) Exploratory factor analyses of IQ test data are carried out in representative samples of blacks and whites separately to extract factor loadings of the tests on g. The IQ tests are required to be many in number, and diverse in content to ensure that the factor loadings are sufficiently variable. 2) Establishing factorial invariance over the groups by calculating measures of factorial congruence of the factor loadings in the white and the black samples. 3) Standardization of differences in means by dividing by the pooled standard deviations. 4) Correlating the standardized differences in means with the g factor loadings. We refer to the procedure incorporating these steps as Jensen's procedure.

A number of commentators and researchers have adopted the method of correlated vectors. They view the reported Spearman correlations as strongly supportive of the weak version of the hypothesis. Schönemann (1997a, p. 666-667) provides a brief survey of these positive views. Others, however, are critical of this method. Notably Schönemann (1997a, 1997b), in a special issue of Cahiers de Psychologie Cognitive (CPC), claims to have proven that the method of correlated vectors is fundamentally flawed. Schönemann purports to show that the correlations between the factor loadings and the differences in means are necessarily positive. A detailed discussion of Schönemann's CPC criticism is provided in Dolan and Lubke (2001). Below we show that the positive correlation is an implication of strict factorial invariance, and, as such, neither trivial nor necessary. We disagree with Schönemann that the method of correlated vectors as used by Jensen is flawed in any fundamental way. However, in so far as Schönemann's critique may be interpreted as addressing the specificity of this test in establishing the role of g in b-w differences, his critique is well taken. Specificity concerns the question whether a large and positive correlation is necessarily indicative of the central role of g.

As a means of demonstrating the centrality of g in IQ related b-w differences, Spearman correlations are patently crude and circuitous (Dolan, 2000; Lubke, Dolan and Kelderman, 2001). The correlation, which is so central to the test of Spearman's hypothesis, may be difficult to interpret and potentially misleading. As its methodological weaknesses are especially clear when the method of correlated vectors is compared to MGCFA, we first present multi-group confirmatory factor models. As we discuss below, a number of the models that we consider incorporate the strong and weak version of Spearman's hypothesis.

Multi-group confirmatory factor models

MGCFA is a well established technique to investigate group differences in means and covariances within the common factor model (e.g., Lawley and Maxwell, 1971; Jöreskog, 1971; Sörbom, 1974; Rock, Werts, and Flaugher, 1978; Hanna, 1984; Byrne, Shavelson and Muthén, 1989; Marsh and Grayson, 1990, 1994; Horn and McArdle, 1992; Schaie, Willis, Jay and Chipuer, 1998; Millsap and Everson, 1991; Dolan and Molenaar, 1994; Little, 1997). Standard software can be used to carry out MGCFA, such as EQS (Bentler, 1990), Mx (Neale, 1997), or LISREL (Jöreskog and Sörbom, 1999).

Here we consider the same three models as Dolan (2000). We call these the 1st order multi-group confirmatory factor model (MGCFM), the 2nd order MGCFM, and the 1st order alternative MGCFM (aMGCFM). The 2nd order MGCFM is particularly well suited to investigate Spearman's hypothesis (both versions). We include the other models, because it is desirable to compare the goodness of fit of a set of competing models, in evaluating Spearman's hypothesis (Dolan, 2000).

We present the models including various identifying and substantive constraints. As we discuss below, the substantive constraints are imposed to ensure that strict factorial invariance (SFI) holds (Meredith, 1993). SFI is necessary to ensure that the group comparisons are meaningful. The identifying constraints are necessary to estimate variances of common factors and differences between groups in means of these factors. The identifying constraints are well known in the literature (e.g., Bollen, 1989).

Let $\mathbf{y}_{ij}$ denote the observed p-dimensional random column vector of subject j ($j=1\ldots N_i$) in population i. As we are only concerned with blacks or whites, we have i=b or i=w. The following factor model is assumed to hold for the observations $\mathbf{y}_{ij}$ (see Sörbom, 1974):

$$\mathbf{y}_{ij} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_{ij} + \boldsymbol{\varepsilon}_{ij}, \qquad (1)$$

where $\boldsymbol{\eta}_{ij}$ is a q-dimensional random vector of correlated common factor scores ($q<p$), and $\boldsymbol{\varepsilon}_{ij}$ is a p-dimensional vector of residual scores unique to each observed variable. These scores include specific (i.e., to the test) and random measurement error (Sörbom, 1974; Meredith, 1993). The (pxq) matrix $\boldsymbol{\Lambda}$ contains factor loadings and the (px1) vector $\boldsymbol{\nu}$ contains intercepts. We assume that $\boldsymbol{\Lambda}$ is full column rank. In confirmatory factor analysis, one usually has a good idea about the relationship between the observed variables and the common factors so that many elements in $\boldsymbol{\Lambda}$ will be fixed to zero. We further assume that $\boldsymbol{\varepsilon}_{ij}\sim N_p(\mathbf{0},\boldsymbol{\Theta})$ and $\boldsymbol{\eta}_{ij}\sim N_q(\boldsymbol{\alpha}_i,\boldsymbol{\Psi}_i)$, where the (pxp) diagonal matrix $\boldsymbol{\Theta}$ is positive, and the (qxq) covariance matrix $\boldsymbol{\Psi}_i$ is positive definite. As the residual scores include both random measurement error and unique systematic effects, the specification that $\boldsymbol{\varepsilon}_{ij}\sim N_p(\mathbf{0},\boldsymbol{\Theta})$ represents a strong assumption (Meredith, 1993; Sörbom, 1975). The observed variables are distributed $\mathbf{y}_{ij}\sim N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$, where, assuming $E[(\boldsymbol{\eta}_{ij}-\boldsymbol{\alpha}_i)\boldsymbol{\varepsilon}_{ij}^t]=\mathbf{0}$,

$$\boldsymbol{\mu}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha}_i \qquad (1a)$$

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}\boldsymbol{\Psi}_i\boldsymbol{\Lambda}^t + \boldsymbol{\Theta}. \qquad (1b)$$

The variances of the common factors are not identified. The variances are identified by equating them with the variances of a given observed variable in the analysis or by standardization. Here we employ the former method. This involves fixing appropriate elements in $\mathbf{\Lambda}$ to equal 1. Likewise, the latent means ($\boldsymbol{\alpha}_i$) are not identified (Sörbom, 1974). Sörbom suggests the following reparametrization to ensure identifiability of the latent mean differences. We choose one of the groups as the reference group, say group 1. In this group we set the latent means to equal zero. In all other groups, we estimate the latent difference relative to group 1, rather than the latent means themselves. So with the appropriate subscripts and the whites as the reference group, we have:

$$\boldsymbol{\mu}_w = \boldsymbol{\nu} \quad (1c)$$

$$\boldsymbol{\Sigma}_w = \boldsymbol{\Lambda}\ \boldsymbol{\Psi}_w\ \boldsymbol{\Lambda}^t + \boldsymbol{\Theta} \quad (1d)$$

$$\boldsymbol{\mu}_b = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\delta} \quad (1e)$$

$$\boldsymbol{\Sigma}_b = \boldsymbol{\Lambda}\ \boldsymbol{\Psi}_b\ \boldsymbol{\Lambda}^t + \boldsymbol{\Theta}, \quad (1f)$$

where $\boldsymbol{\delta}$ represents the (qx1) vector of differences in common factor means ($\boldsymbol{\alpha}_b$-$\boldsymbol{\alpha}_w$). We call this the 1st order MGCFM. The 1st order aMGCFM, which is a special case of this model, is presented below.

In Jensen's procedure, the common factor g features as a second order common factor (e.g., Jensen and Reynolds, 1982; Naglieri and Jensen, 1987) and is usually extracted using Schmid-Leiman exploratory factor analysis (Schmid and Leiman, 1957). To accommodate this conceptualization of g within the present model, we extend the model as follows (see Jöreskog and Sörbom, 1989). Let $\boldsymbol{\eta}_{ij} = [\boldsymbol{\Gamma}\boldsymbol{\xi}_{ij} + \boldsymbol{\zeta}_{ij}]$, where $\boldsymbol{\Gamma}$ is a (qxr) matrix of loadings of the q first order factor scores, $\boldsymbol{\eta}_{ij}$, on the r second order factor scores, $\boldsymbol{\xi}_{ij}$, and $\boldsymbol{\zeta}_{ij}$ is a q-dimensional vector of random (first order) residual terms. Throughout this paper, we consider only one second order factor (i.e., the g factor), so that we have r=1. The model for the observations is now:

$$\mathbf{y}_{ij} = \boldsymbol{\nu} + \boldsymbol{\Lambda}[\boldsymbol{\Gamma}\boldsymbol{\xi}_{ij} + \boldsymbol{\zeta}_{ij}] + \boldsymbol{\varepsilon}_{ij}. \quad (2)$$

We assume that $\zeta_{ij} \sim N_q(\boldsymbol{\alpha}_i, \boldsymbol{\Psi}^*{}_i)$, where the asterisk indicates that the covariance matrix is diagonal, and $\xi_{ij} \sim N_r(\boldsymbol{\kappa}_i, \boldsymbol{\Phi}_i)$. Furthermore, we assume that $E[(\zeta_{ij}-\boldsymbol{\alpha}_i)\,\boldsymbol{\varepsilon}_{ij}{}^t] = E[(\xi_{ij}-\boldsymbol{\kappa}_i)\,\boldsymbol{\varepsilon}_{ij}{}^t] = E[(\xi_{ij}-\boldsymbol{\kappa}_i)(\zeta_{ij}-\boldsymbol{\alpha}_i)^t] = \mathbf{0}$. So we have $\mathbf{y}_{ij} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where

$$\boldsymbol{\mu}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha}_i + \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\kappa}_i \qquad (2a)$$

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}[\boldsymbol{\Gamma}\boldsymbol{\Phi}_i\boldsymbol{\Gamma}^t + \boldsymbol{\Psi}^*{}_i]\boldsymbol{\Lambda}^t + \boldsymbol{\Theta}. \qquad (2b)$$

Again for reasons of identification, we estimate latent mean differences, instead of latent means. Retaining the whites as the reference group and employing appropriate subscripts we have:

$$\boldsymbol{\mu}_w = \boldsymbol{\nu} \qquad (2c)$$

$$\boldsymbol{\Sigma}_w = \boldsymbol{\Lambda}[\boldsymbol{\Gamma}\boldsymbol{\Phi}_w\boldsymbol{\Gamma}^t + \boldsymbol{\Psi}_w{}^*]\boldsymbol{\Lambda}^t + \boldsymbol{\Theta} \qquad (2d)$$

$$\boldsymbol{\mu}_b = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\delta} + \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\tau} \qquad (2e)$$

$$\boldsymbol{\Sigma}_b = \boldsymbol{\Lambda}[\boldsymbol{\Gamma}\boldsymbol{\Phi}_b\boldsymbol{\Gamma}^t + \boldsymbol{\Psi}_b{}^*]\boldsymbol{\Lambda}^t + \boldsymbol{\Theta}, \qquad (2f)$$

where $\boldsymbol{\tau} = \boldsymbol{\kappa}_b - \boldsymbol{\kappa}_w$ and $\boldsymbol{\delta} = \boldsymbol{\alpha}_b - \boldsymbol{\alpha}_w$. We call this the 2nd order MGCFM. As it stands the model requires further restrictions. Specifically, estimation of the second order mean difference ($\boldsymbol{\tau}$) and all the differences in means due to the first order residuals ($\boldsymbol{\delta}$) is not possible. At least one component of the (qx1) dimensional vector $\boldsymbol{\delta}$ has to be fixed to zero. In addition, we fix appropriate values of both $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ to equal 1 so that we may estimate variances in $\boldsymbol{\Phi}_i$ and $\boldsymbol{\Psi}_i{}^*$, respectively. Note that the covariance matrices of the first order factor residuals are not constrained to be equal over the groups. Throughout this chapter, we consistently allow any latent variable that contributes to the differences in means also to contribute to differences in variances. The Schmid-Leiman hierarchical factor analysis (Schmid and Leiman, 1957) that Jensen uses to extract g as a second order factor (e.g., Naglieri and

Jensen, 1987) can be viewed as the exploratory version of the model in equation 2b.

The models presented so far are well known in the literature. We now consider an alternative to the 1st order MGCFM. As presented, this model allows one to model observed differences in means as a function of differences in q common factor means. However, one may also test the hypothesis that the groups differ with respect to a subset of the q common factors in the 1st order MGCFM (see Figure 1 and Table 1 below). For instance, given three correlated Verbal, Memory and Spatial factors, it is possible that two groups differ only with respect to the unobserved causes of the Verbal factor scores. Because the factors are correlated, it is reasonable to assume that the causes of the Verbal factor scores overlap partially with the causes of the Memory and Spatial factor scores. Due to this overlap of causes, the differences with respect to the causes of the Verbal factor will also be evident in the other two common factors. Dolan and Molenaar (1994) presented a model that takes into account the secondary effects on other common factors of primary differences with respect to a subset of common factors. Their model is based on the strong assumption that the differences between the groups can be conceived of as the result of selection effects at the level of the (subset of) common factors. One of the populations, chosen arbitrarily, is defined as the reference population, the other is defined as the selected population. Here the white population features as the former, and the black population features as the latter. In white population the first order factor model holds as shown in Eqs. 1c & 1d. The black population is supposed to arise from the white population by selection of the basis of a subset of the common factors. This process of selection is modeled by taking into account the consequences of selection based on the factor scores. Factor scores are not actually calculated, but the weight matrix used to calculate factor scores does feature in the derivation of the model. There are a number of ways to calculate such a weight matrix (Saris, de Pijper and Mulder, 1978; Lawley and Maxwell, 1971). Here we limit our attention to the factor scores regression matrix (pxq), which is calculated as follows, $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}\mathbf{\Lambda}^{t}\mathbf{\Psi}_{W}$. Dolan and Molenaar demonstrate that the exact choice of regression matrix does not affect the results in any serious way. If the groups differ with respect to one or more components of the vector of factor scores, $\boldsymbol{\eta}=\boldsymbol{\Omega}^t\mathbf{y}$, the covariance matrix and mean vector in the black group may be modeled as follows (see Dolan and Molenaar, 1994):

$$\mu_b = \nu + \Sigma_W \, \Omega \, \delta \;= \nu + \Lambda \, \Psi_W \, \pi \qquad (3a)$$

$$\Sigma_b \;= \Sigma_W + \Sigma_W \, [\Omega \, \Delta \, \Omega^t] \Sigma_W \;= \Lambda [\Psi_W \, \Delta \; \Psi_W{}^t \; + \; \Psi_W] \Lambda^t \; + \; \Theta \, , \qquad (3b)$$

where $[\Psi_W \, \pi]$ represents the vector of latent differences in means, and $[\Psi_W \, \Delta \, \Psi_W{}^t + \Psi_W]$, the covariance structure of the common factors in the black group. We call this model the 1st order alternative MGCFM (aMGCFM). The (qx1) vector $\boldsymbol{\pi}$ and the (qxq) symmetric matrix $\boldsymbol{\Delta}$ are known functions of the differences in means and the (co)variances of the selection variable in the white and the black group (Dolan and Molenaar, 1994). Because we are considering differences with respect to a subset of common factors, both $\boldsymbol{\pi}$ and $\boldsymbol{\Delta}$ will contain zero elements. For instance, let us suppose that we specify that the groups differ primarily with respect to the second of three common factors, i.e., the Memory factor. As mentioned, differences in one common factor implies differences in the other common factors, because the correlated common factors display over lap in their causes (hence the factors are correlated). The differences are modeled by parameters in $\boldsymbol{\pi}$ and $\boldsymbol{\Delta}$. For instance, in the present case, we have

$$\boldsymbol{\Delta} \quad = \quad \begin{matrix} 0 & 0 & 0 \\ 0 & \Delta_{22} & 0 \\ 0 & 0 & 0 \end{matrix} \qquad (4a)$$

and

$$\boldsymbol{\pi}^t \;=\; [0 \quad \pi_2 \quad 0]. \qquad (4b)$$

Should we wish to test the hypothesis that the groups differ primarily with respect to the two first order factors (Verbal and Memory), we would specify:

$$\boldsymbol{\Delta} = \begin{matrix} \Delta_{11} & \Delta_{12} & 0 \\ \Delta_{21} & \Delta_{22} & 0 \\ 0 & 0 & 0 \end{matrix} \qquad (5a)$$

and

$$\boldsymbol{\pi}^t = [\pi_1 \quad \pi_2 \quad 0], \qquad (5b)$$

where $\Delta_{21} = \Delta_{12}$. The manner in which groups differ in this and other models is illustrated in Figure 1 and Table 1. The 1st order aMGCFM coincides with the 1st order MGCFM (Eqs. 1c to 1f), if all elements of the symmetric matrix $\boldsymbol{\Delta}$ and the vector $\boldsymbol{\pi}$ are estimated. In this case $[\boldsymbol{\Psi}_w\boldsymbol{\pi}]$ equals $\boldsymbol{\delta}$, in Eq. 1e, and $[\boldsymbol{\Psi}_w\boldsymbol{\Delta}\boldsymbol{\Psi}_w{}^t+\boldsymbol{\Psi}_w]$ equals $\boldsymbol{\Psi}_b$ in Eq. 1f. Note that $\boldsymbol{\Delta}$ and $\boldsymbol{\pi}$ will contain zero elements, but the latent mean vector $[\boldsymbol{\Psi}_w\boldsymbol{\pi}]$ and covariance matrix $[\boldsymbol{\Psi}_w\boldsymbol{\Delta}\boldsymbol{\Psi}_w{}^t + \boldsymbol{\Psi}_w]$ will not, because the covariance matrix $\boldsymbol{\Psi}_w$ will not generally contain zero elements. As mentioned 1st order aMGCFM is derived using an explicit selection mechanism. Analyses of simulated data suggest that this model may provide a good approximation of group differences regardless of the exact manner in which the groups come to differ with respect to a subset of correlated factors.

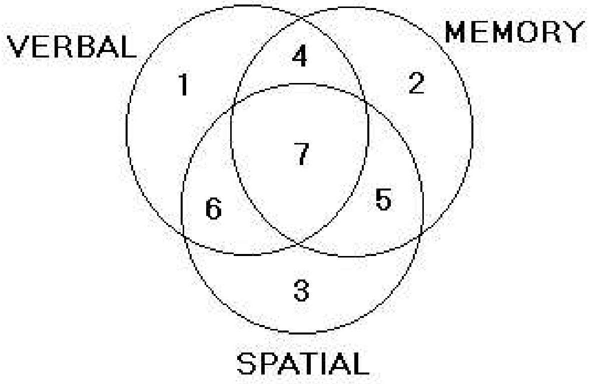

Figure 1: A Venn-diagrammatic representation of the sets of causes of the common factors V,M and S (see Table 1).

Table 1: Sources of group in latent group differences in models A4 to C6. The sources of group differences are numbered. These numbers refer to sets of causes shown in Figure 1.

Models incorporating SFI	source of group differences	empty	Spearman's hypothesis
A4 1st MGCFM	1,2,3,4,5,6,7 (V&M&S)	none	no
B2 2nd MGCFM	7 (g)	4,5,6	strong
B3 2nd MGCFM + residual	7,1 (g&V)	4,5,6	weak
B4 2nd MGCFM + residual	7,2 (g&M)	4,5,6	weak
B5 2nd MGCFM + residual	7,3 (g&S)	4,5,6	weak
C1 1st aMGCFM	1,4,6,7 (V)	none	no
C2 1st aMGCFM	2,4,5,7 (M)	none	no
C3 1st aMGCFM	3,5,6,7 (S)	none	no
C4 1st aMGCFM	1,2,4,5,6,7 (V&M)	none	no
C5 1st aMGCFM	1,3,4,5,6,7 (V&S)	none	no
C6 1st aMGCFM	2,3,4,5,6,7 (S&M)	none	no

Strict factorial invariance

The models presented above share the property that the differences between the groups in observed means and covariance structure are due to differences in means and covariance structure of the common factors. In these models, the factor loadings ($\mathbf{\Lambda}$ and $\mathbf{\Gamma}$), the specific variances ($\mathbf{\Theta}$), and intercepts ($\boldsymbol{\nu}$) are equal over the groups. Together these constraints comprise the hypothesis of strict factorial invariance (SFI; Meredith, 1993). Millsap (1997a) points out that given SFI, "two individuals who have identical common factor scores would be expected to have identical observed scores regardless of their group membership" (p. 250). This implies that we are measuring the same constructs in both groups. Relaxation of any of the constraints mentioned would complicate this identical construct interpretation of the common factors in the model. The finding that factorial invariance is tenable suggests that the tests are unbiased with respect to group (Mellenbergh, 1989; Meredith, 1993; Millsap, 1997a, 1997b; Muthén and Lehman, 1985; Oort, 1996). Although it is an important aspect of SFI, the equality constraint on the covariance matrices of the residual variance terms is sometimes dropped (i.e., $\Theta_w \neq \Theta_b$). One reason for this is that these terms contain both random measurement error and specific terms (Meredith, 1993). As groups may differ with respect to these specific terms, which are substantive latent variables like the common factors, Little (1997; see also Muthén and Lehman, 1985) advocates estimating separate Θ matrices in each group in order to reduce the possible effects of unequal specific variance terms on the other parameters in the model. The hypothesis $\Theta_w = \Theta_b$ can be tested readily.

Fitting the models to observed means and covariance matrices may be done using any of the available programs for multi-group covariance and mean structure modeling (e.g., Neale, 1997; Jöreskog and Sörbom, 1993). Assuming the data are approximately normally distributed, normal theory maximum likelihood (ML) estimation may be applied to obtain estimates, χ^2 goodness of fit indices based on the likelihood ratio, as well as a variety of other fit indices (Jöreskog, 1993).

MGCFM's incorporating Spearman's hypothesis

Millsap (1997b) discussed the relationship between Jensen's test of Spearman's hypothesis and factorial invariance. He showed that the collinearity of differences in expected means and factor loadings, will hold if the groups differ mainly with respect to a single dominant first order common factor in the 1st order MGCFM. This becomes clear when we consider the model for the means. As above we, have $\mu_w = \nu$ and $\mu_b = \nu + \Lambda\delta$, where δ is a scalar representing the difference in mean of the single common factor ($\alpha_b - \alpha_w$). Clearly, in this case the (px1) vector of expected mean differences ($\mu_b - \mu_w = \Lambda\delta$) and the (px1) vector of factor loadings (Λ) are collinear. A problem with the single common factor model in the present context is that it does not fit the covariance structure of tests batteries, like the WISC-R or the K-ABC.

The 2nd order MGCFM is a suitable model for both the weak and the strong versions of Spearman's hypothesis. We first consider the strong version. Here the second order common factor is the sole source of group differences in means and covariance structure. The means are modeled as follows: $\mu_w = \nu$ and $\mu_b = \nu + \Lambda\Gamma\tau$, where the scalar τ equals the differences in the means of g (κ_b-κ_w). Here (px1) vector of expected difference in means, $\mu_b - \mu_w = \Lambda\Gamma\tau$, is collinear with the vector (px1) $\Lambda\Gamma$. The weak version of Spearman's hypothesis may also be modeled using the 2nd order MGCFM. Here we include the contributions of the first order factor residuals to the group differences in means and covariance structure. The expression for the means are $\mu_w = \nu$ and $\mu_b = \nu + \Lambda\delta + \Lambda\Gamma\tau$. As mentioned above, we cannot estimate all components of the (qx1) vector δ, i.e., the mean differences in 1st order factor residuals, for reasons of identification. We can estimate at most q-1 components. In this model the expected vector of differences in means is $\Lambda\delta + \Lambda\Gamma\tau$. The correlation between the (px1) vector $\Lambda\Gamma$ and $\Lambda\delta + \Lambda\Gamma\tau$ depends on the size of the components of δ. The weak version does prescribe that generally the contributions of $\Lambda\Gamma\tau$ should exceed those of $\Lambda\delta$. In conclusion, the relationship between differences in means and factor loadings is one implication of certain MGCFM's incorporating SFI (Millsap, 1997b).

A critique of the method of correlated vectors

In comparing the MGCFM's discussed above with the method of correlated vectors, it is not difficult to see that the latter has a number of methodological weaknesses. These are the following.

1) Spearman's hypothesis comprises a number of components, relating to the covariance structure of the IQ test scores, the equality of factor loadings over groups, and of course the collinearity of differences in means and factor loadings. The aspects are investigated piecemeal in the various steps of Jensen's procedure. MGCFA allows one to cast all aspect of the hypothesis in a single comprehensive model and fit this model directly to the data using ML estimation.

2) Jensen's procedure does not include explicit goodness of fit testing. Lack of fit may accumulate at various points, without this becoming evident. MGCFA, in contrast, can be carried out using well disseminated software that provides a range of goodness of fit indices, relating to overall fit. In addition such software provides information concerning local (i.e., within the model) misfit, such as modification indices and standardized residuals (Jöreskog, 1993).

3) Because its focus is on the Spearman correlation, Jensen's procedure does not lend itself readily to the investigation of alternative hypotheses. Dolan (2000) has argued that support for the central role of g in b-w differences is convincing, only if it can be demonstrated that 'g-consistent' models fit the data well and that they fit better than competing models. Here again the issue of goodness of fit is crucial in comparing models. Because we want to compare models, we have considered a variety of models in the preceding sections. Fitting and comparing competing models is a standard procedure in covariance and mean structure analysis (Jöreskog, 1993).

4) A high value of the congruence coefficient regarding the factor loadings in the black and white samples is a necessary, but not a sufficient condition of factorial invariance. As we have seen, the present MGCFM's incorporate the restrictions relating to SFI. Meredith's (1993) definition of SFI provides the conditions that are required for a meaningful comparison of groups within the MGCFM. As we have seen above, these conditions relate to the equality of intercepts, factor loadings and residual variances. The finding that SFI is tenable supports the hypothesis that measurement bias is absent, or that measurement invariance holds. Equivalently, one can state that given SFI the within-group and between group differences may be considered to be differences on the same latent dimensions (Lubke, Dolan, Kelderman and Mellenbergh, 2001). This is an essential aspect of Spearman's hypothesis, which is not tested adequately by means of measures of factorial congruence. Also the congruence coefficient may be difficult to interpret. How sensitive is this coefficient to violations of the equality of (one or more) factor loadings ? At what value of the congruence measure should one reject the hypothesis that the factor loadings are equal over the groups ?

5) The weak version of Spearman's hypothesis is ill-defined (Dolan, 2000). At what value of the Spearman correlation does one reject the hypothesis ? Even if this question could be answered clearly, it remains uncertain whether the weak version actually holds. The Spearman correlation may assume intermediate values (say, .6), not because first order residuals are contributing to the b-w differences (i.e., consistent with the weak version), but rather because of undetected model violations, which accumulate at the various steps of Jensen's procedure. The correlation produced by the method of correlated vectors may be hard to interpret. Values as low as .5 may be interpreted in support of the hypothesis.

6) Standardization of differences in means is meant to express the mean differences in standard deviation units. Within Jensen's procedure, the pooled standard deviations are used to this end. The aspect of the procedure is problematic because it ignores the systematic b-w differences in variances. It is unlikely that groups that differ with respect to common factors, will only display differences in factor means. Indeed b-w differences in IQ test score variances are well documented (Jensen, 1998). Jensen's procedure does not utilize the expected differences in variances. Rather by pooling this information is lost. With respect to differences in covariance structure, it should be noted that the models presented above specify group differences in observed means and covariance matrices. In comparing blacks and whites, the hypothesis is sometimes tested that the covariance or correlation matrices are equal (Jensen and Reynolds, 1982; Naglieri and Jensen, 1987). Note that the rejection of this hypothesis is hard to interpret: it may be due to differences that are or are not compatible with SFI. The rejection of this hypothesis cannot therefore be taken to mean that covariance matrices of blacks and whites are qualitatively different. MGFCM's subject to SFI predict differences in expected mean vectors and covariance matrices.

Spearman correlations lack specificity

The identification of methodological weaknesses of a given test procedure leaves unanswered the question how serious the weaknesses actually are. We have tried to address this issue in two ways. Lubke, Dolan and Kelderman (2001) investigated the specificity of

Jensen's procedure to detect violations of the various components of Spearman's hypothesis. To compare Jensen’s approach of analyzing b-w differences to MGCFA, Lubke, Dolan and Kelderman (2001) constructed population covariance matrices, which incorporated violations in various degrees. Specifically, the violations concerned either (1) the assumption underlying Jensen’s procedure that g exists, or (2) that differences in g (mainly) account for the differences in observed scores, or (3) both. Parameters that were not manipulated for the study were chosen to be similar to the data of Jensen and Reynolds (1982). Based on Jensen’s procedure only a combination of severe violations was detected: the Spearman correlation dropped to 0.27 when the g-contribution to b-w differences was clearly less than 50% and there were additional second order factors. Conversely, the power to (correctly) reject Spearman’s hypothesis using MGCFA was larger than 0.8 even in cases of milder violations, which were not detected with the Spearman correlation (e.g., correlations exceeding 0.6).

Dolan (2000) fitted a variety of first and second order confirmatory factor models to a published data set (Jensen and Reynolds, 1982) comprising WISC-R data observed in 1868 whites and 305 blacks. Dolan found that strict factorial invariance was tenable and that certain models would be rejected quite readily (including the model incorporating the strong version of Spearman's hypothesis). However, it proved very difficult to distinguish between other competing models. These included models compatible with the weak version of Spearman's hypothesis and models which did not include the g-factor. This failure to distinguish between models may be due in part to the fact that the WISC-R comprises only 13 tests. To detect the often subtle differences between the models, one may require a larger number of observed tests. Dolan (2000) emphasized that the inability to distinguish between competing models using MGCFA should engender some reticence in inferring the importance of g on the basis of Spearman correlation. For the Jensen and Reynolds data set, the reported Spearman correlations are .75 (whites) and .64 (blacks). It seems premature to interpret these in support of Spearman's hypothesis, when we know that these values and the data are compatible with competing ('non g-related') hypotheses.

The main problem with the Spearman correlation is that its behavior given violations of the underlying hypothesis (e.g., the model expressed in Eqs. 2c to 2f) is poorly understood. The results of Lubke et al. (2001) and Dolan (2000) suggest that these correlations are lacking in specificity. Because we are convinced that MGCFA provides a superior means to investigate b-w differences in IQ test scores, we shall not pursue this issue here. Rather we shall devote the rest of this chapter to the application of the MGCFA to a published data set (Naglieri and Jensen, 1987). This data set comprises scores of 86 black and 86 white children on 11 subtests of the Wechsler Intelligence Scale for Children-Revised (WISC-R) and 13 subtests of the Kaufman Assessment Battery for Children (K-ABC). The large number of tests may make it easier to distinguish between competing models.

MGCFA of the Naglieri and Jensen (1987) data set[1]

Naglieri and Jensen (1987) provide the covariance matrices, mean vectors and standard deviations of the WISC-R and K-ABC test scores. The WISC-R comprises the following tests: Information (I), Similarities (S), Arithmetic (A), Vocabulary (V), Comprehension (C), Digit Span (DS), Picture completion (PC), Picture arrangement (PA), Block design (BD), Object assembly (OA) and Coding (CO). These test are described at length in the relevant literature. The K-ABC includes the following tests: Hand movement (HM), Gestalt closure (GC), Number recall (NR), Triangles (T), Word order (WO), Matrix analogies (MA), Spatial memories (SM), Photos series (PS), Faces and places (FP), Arithmetic (AR), Riddles (R), Reading/Decoding (RD), Reading/Understanding (RU). Naglieri and Jensen (1987) provide a brief description of the K-ABC subtests. Jensen and Reynolds (1982) investigated the factor structure of the WISC-R. Keith and Dunbar (1984) and Jensen (1984) investigated the factor structure of the K-ABC. Keith (1997) presents confirmatory factor analyses of the K-ABC and the WISC-R.

The black and white children were matched, using a matched-pair procedure, on the following variables: sex, school, age (± 3 months), and socio-economic status. This resulted in 86 black-white pairs, of which 40 were males and 46 were females. The mean age for whites was 10.7 years (sd = 8.2 months; range = 9.3 to 12.4 years). The mean age for blacks was 10.8 years (sd = 8.0 months; range = 9.4 to 12.4 years). SES was measured on a 5-point scale (similar to the one presented in the WISC-R manual) based mainly on the parent's level of occupation. The mean SES for whites was 3.4 (sd = 1.2) and for blacks 3.4 (sd = 1.2). The WISC-R and the K-ABC were administered individually in two sessions a week apart. Within each black-white pair the two tests were given in the same order.

Both test batteries admit a 3 common factor model with factors relating verbal abilities (V), spatial abilities (S), and memory (M). Naglieri and Jensen (1987) subjected the complete set of 24 variables to a Schmid-Leiman factor analysis and derived a single 2nd order common factor (g) and three 1st order common factors (M,V,S). Browne and Cudeck (1993) using newly developed goodness of fit indices suggest that given the small sample size three factors is reasonable in the white sample.

Exploratory factor analyses and first order MGCFM's

We used LISREL 8.3 to carry out all analyses (Jöreskog and Sörbom, 1999). Assuming the data are approximately multivariate normal, we apply normal theory maximum likelihood estimation throughout (e.g., Sörbom, 1974). To assess and compare goodness of fit we report a variety of fit indices, as recommended by Bollen and Long (1993). Although there is some redundancy among these indices, we consider the non-normed fit index (NNFI), Akaike's information criterion (AIC), the related CAIC, the expected cross validation index (ECVI), the root mean square error of approximation (RMSEA), the χ^2, and χ^2/df (df stands for degrees of freedom). The χ^2 is treated as a measure of (badness of) fit, rather than as a formal test-statistic (Jöreskog, 1993). The index χ^2/df is a simple measure of fit which takes into account the degrees of freedom of the model. The RMSEA (Steiger, 1990; Browne and Cudeck, 1993) is a measure of the error of approximation of the specified model covariance and mean structures to the covariance and mean structures in the population(s). As a rule

[1] All LISREL input files used in this chapter are available on request to the first author (same applies to LISREL output files, if required).

of thumb, Browne and Cudeck (1993) suggest that an RMSEA of .05, or less, is indicative of a good approximation. The ECVI provides an indication of the discrepancy between the fitted covariance matrices in the analyzed samples and the expected covariance matrices that would be obtained in a second sample of the same size. The models with low values of ECVI are preferable to models with large values. This rule also applies to AIC and CAIC: lower values of AIC and CAIC indicate better fitting models. Compared to AIC, CAIC favors more parsimonious models. To judge the adequacy of models, we also report the range and median of the standardized residuals. The standardized residuals are a function of the difference between the observed data and the expected data under the specified model. If the model provides a good description of the data (strictly speaking, if the specified model is true, and the data are multivariate normally distributed), the standardized residuals follow a standard normal distribution. Finally, to assess local misfit (i.e., within a given model), we report selected modification indices (MI's; Sörbom, 1989). LISREL 8.3 calculates a MI for each fixed or constrained parameter in the model. Each modification index measures how much the χ^2 of the model is expected to decrease if the associated parameter is estimated freely.

We subjected correlation matrices of the whites and the blacks to separate exploratory factor analyses. To arrive at an oblique structure that satisfied simple structure, we subjected the factor loadings to varimax rotation with Kaiser normalisation, followed by promax rotation (Lawley and Maxwell, 1971)[2]. In Table 2 we report the rotated factor loadings obtained in the white and black samples. These are quite similar to those reported by Naglieri and Jensen (1987). The correlations among the factors ranged from .36 to .49. Based on these results, we derived the pattern of factor loadings as shown in Table 2 (last three columns). This is the pattern of the matrix Λ that features in the equations above. The goodness of fit indices for the 3 factor solutions are shown in Table 3.

Table 2: Promax rotated factor loadings and the derived pattern of factor loadings.

	whites			blacks			pattern of loadings		
	V	M	S	V	M	S	V	M	S
I	0.84	0.00	-.05	0.85	-.11	0.07	λ	0	0
S	0.72	0.03	0.08	0.68	-.07	0.23	λ	0	0
A	0.18	0.52	0.10	0.42	0.19	0.07	λ	λ	0
V	0.87	-.07	0.10	0.89	0.03	-.08	1*	0	0
C	0.70	0.04	-.07	0.74	0.08	0.07	λ	0	0
DS	-.09	0.85	-.01	-.06	0.74	-.06	0	1*	0
PC	0.17	-.07	0.47	0.10	-.09	0.68	0	0	λ
PA	0.05	-.04	0.45	0.19	-.07	0.52	0	0	λ
BD	0.18	0.02	0.69	0.18	0.08	0.70	0	0	1*
OA	0.15	-.05	0.63	-.12	-.10	0.77	0	0	λ
CO	-.23	0.05	0.46	-.09	0.28	0.22	0	0	λ
HM	0.15	0.47	0.08	-.06	0.45	0.14	0	λ	0
GC	0.33	-.24	0.26	0.11	-.07	0.52	0	0	λ
NR	-.11	0.73	0.01	0.03	0.66	-.13	0	λ	0
T	-.04	0.10	0.77	-.04	0.00	0.78	0	0	λ
WO	-.16	0.73	0.11	-.02	0.48	0.04	0	λ	0
MA	0.15	0.20	0.41	0.23	0.29	0.35	0	λ	λ
MS	-.24	0.08	0.60	-.09	0.25	0.52	0	0	λ
PS	0.27	-.05	0.29	-.10	0.11	0.54	0	0	λ
FP	0.91	-.14	-.05	0.81	-.14	0.05	λ	0	0
AR	0.44	0.46	0.02	0.45	0.17	0.08	λ	λ	0
R	0.78	-.06	0.13	0.73	-.02	0.21	λ	0	0
RD	0.61	0.40	-.09	0.70	0.23	-.21		dropped	
RU	0.60	0.23	-.01	0.77	.07	-.03	λ	0	0

[2] Promax rotation depends on a parameter which Lawley and Maxwell (1971) denote m. In LISREL, this parameter m is set to equal 3.

Table 3: Goodness of fit indices of the fitted models.

	df	$\chi2$	$\chi2$/df	RMSEA	ECVI	AIC	CAIC	NNFI
EFA 24	414	511.1	1.23	-	-	-	-	-
EFA 23	374	429.7	1.15	-	-	-	-	-
A1	448	594.3	1.32	.053	5.02	853.9	1476.0	.91
A2	471	625.4	1.33	.053	4.91	835.5	1362.2	.91
A3	494	685.3	1.38	.058	4.96	843.4	1274.4	.89
A3r	493	662.5	1.34	.056	4.91	834.1	1269.6	.90
A4	513	692.8	1.35	.055	4.80	816.9	1169.4	.90
B1	498	669.6	1.34	.057	4.90	833.6	1248.3	.90
B2	520	701.3	1.35	.056	4.78	812.6	1136.1	.90
B3(v)	518	700.3	1.35	.056	4.80	815.2	1147.0	.90
B4(m)	518	694.4	1.34	.055	4.76	809.6	1141.0	.90
B5(s)	518	701.1	1.35	.056	4.80	816.1	1147.0	.90
C1(v)	520	705.2	1.36	.057	4.83	821.8	1145.4	.90
C2(m)	520	715.4	1.37	.058	4.85	825.1	1148.6	.89
C3(s)	520	702.2	1.35	.055	4.77	811.1	1134.0	.90
C4(v&m)	517	699.8	1.35	.056	4.81	817.6	1153.6	.90
C5(v&s)	517	699.2	1.35	.056	4.82	819.3	1155.2	.90
C6(m&s)	517	695.8	1.34	.055	4.77	811.1	1147.1	.90

note: goodness of fit indices are discussed in the text.

Having obtained a plausible pattern of the matrix Λ, we fit 4 oblique factor models, A1 to A4. In going from model A1 to A4, we introduce the equality constrains associated with SFI in a number of steps. In models A1 to A3, we do not include the means. First we constrain the factor loadings to be equal over the groups ($\Lambda_w=\Lambda_b$; A2), then we constrain the residual variances to be equal ($\Theta_w=\Theta_b$; A3) in addition to the factor loadings, and finally we constrain the expected differences in means to be a function of the differences in common factor means (A4; see Eqs. 1c to 1f). The sequence of models is shown in Table 4. Table 5 displays the nesting among the models.

Table 4: Fitted MGCFM's.

First order MGCFM's

Model	mean vector	covariance matrix
A1	$\mu_w = \nu_w$	$\Sigma_w = \Lambda_w \Psi_w \Lambda_w^t + \Theta_w$
	$\mu_b = \nu_b$	$\Sigma_b = \Lambda_b \Psi_b \Lambda_b^t + \Theta_b$
A2	$\mu_w = \nu_w$	$\Sigma_w = \Lambda \Psi_w \Lambda^t + \Theta_w$
	$\mu_b = \nu_b$	$\Sigma_b = \Lambda \Psi_b \Lambda^t + \Theta_b$
A3r	$\mu_w = \nu_w$	$\Sigma_w = \Lambda \Psi_w \Lambda^t + \Theta$
	$\mu_b = \nu_b$	$\Sigma_b = \Lambda \Psi_b \Lambda^t + \Theta$
A4	$\mu_w = \nu$	$\Sigma_w = \Lambda \Psi_w \Lambda^t + \Theta$
	$\mu_b = \nu + \Lambda\delta$	$\Sigma_b = \Lambda \Psi_b \Lambda^t + \Theta$
Second order MGCFM's		
B1	$\mu_w = \nu_w$	$\Sigma_w = \Lambda[\Gamma \Phi_w \Gamma^t + \Psi^*]\Lambda^t + \Theta$
	$\mu_b = \nu_b$	$\Sigma_b = \Lambda[\Gamma \Phi_b \Gamma^t + \Psi^*]\Lambda^t + \Theta$
B2	$\mu_w = \nu$	$\Sigma_w = \Lambda[\Gamma \Phi_w \Gamma^t + \Psi^*]\Lambda^t + \Theta$
	$\mu_b = \nu + \Lambda\Gamma\tau$	$\Sigma_b = \Lambda[\Gamma \Phi_b \Gamma^t + \Psi^*]\Lambda^t + \Theta$
B3,4,5	$\mu_w = \nu$	$\Sigma_w = \Lambda[\Gamma \Phi_w \Gamma^t + \Psi_w^*]\Lambda^t + \Theta$
	$\mu_b = \nu + \Lambda(\Gamma\tau+\delta)$	$\Sigma_b = \Lambda[\Gamma \Phi_b \Gamma^t + \Psi_b^*]\Lambda^t + \Theta$
First order aMGCFM's		
C1 to C6	$\mu_w = \nu$	$\Sigma_w = \Lambda \Psi_w \Lambda^t + \Theta$
	$\mu_b = \nu + \Lambda \Psi_w \pi$	$\Sigma_b = \Lambda[\Psi_w \Delta \Psi_w^t + \Psi_w]\Lambda^t + \Theta$

Note: In model B3, B4, and B5 the differences in mean of a first order residual is estimated (i.e., a single component of the vector δ). In these models, the diagonal covariance matrices of the first order residuals have group subscripts, because we consistently let any latent variable that contributes to differences in mean, also contribute to differences in variance. So in these models one of the residual variances is allowed to vary over the groups. The differences between the models C1 to C6 depend on the specification of Δ and π (see Eqs. 4a,b and 5a,b, for examples). In models A3r to C6, the residual variance of the test HM was free to vary over the groups.

Table 5: Nesting of model.

model	df	nested under model(df)
A4 (V&S&M)	513	A3r(493)
A3r	493	A2(471),A1(448),EFA23(374)
A2	471	A1(448),EFA23(374)
A1	448	EFA23(374)
B2(g)	520	B3(518),B4(518),B5(518)
B2(g)	520	B1(498)
B1	498	A3r(493)
C1(V)	520	A4(313),C4(517),C5(517)
C2(M)	520	A4(313),C6(517),C4(517)
C3(S)	520	A4(513),C6(517),C5(517)
C4(V&M)	517	A4(513)
C5(V&S)	517	A4(513)
C6(M&S)	517	A4(513)

Results obtained by fitting model A1 showed that the covariance between the variables RD and RU is too large to be accounted for by the three factor model in the black sample. The modification index of the covariance between the residuals of these variables (i.e., off diagonal element 23,24 in Θ_b) equals 20.2 and the standardized residual of the associated with the covariance between RD and RU was relatively large (4.50) and clearly outlying. Rather than introduce additional parameters to accommodate these findings, we decided to remove the variable RD. All subsequent results are therefore based on the analysis of 23 variables. Having removed RD, the χ^2 goodness of fit index of the model equals 594.3 with 448 df (with RD, $\chi^2(490)$ =694.2). The RMSEA of .053 and the NNFI of .91 suggest that this model is an acceptable point of departure. In the black (white) sample the standardized residuals were between -2.96 (-3.71) and 3.67 (3.54) with a median of 0.0 (0.0). In model A2 the factor loadings are constrained to be equal over the groups, i.e., $\Lambda_w=\Lambda_b$. The ratio of χ^2 to df (1.33) and the RMSEA (.053) remain about the same. ECVI, AIC and CAIC are all lower, so that we conclude that the equality of factor loadings is an acceptable constraint. Of the 69 MI's associated with the factor loadings in the white (black) sample only 3 (8) are larger than 4.0 and none are larger than 6.1. In the black (white) sample the standardized residuals were between -3.62 (-3.09) and 3.80 (3.15) with a median of .21 (-.13). Model A3 adds the constraint of equal residual variances, i.e., $\Theta_w=\Theta_b$. The χ^2 equals 685.3 with 494 df. The RMSEA and NNFI indicate a slight deterioration in fit. Both ECVI and AIC suggest that model A2 provides a better fit of the data. Inspection of the results revealed that the equality over the groups of residual variance of the variables HM was causing the largest MI (about 19.0). Also the largest and clearly outlying standardized residuals are associated with the variance of HM (4.94). As this source of misfit is local and clearly identified, we introduce an extra parameter to accommodate the b-w difference in residual variance in HM. We call this revised model A3r. In subsequent model fitting we retain this additional parameter. The introduction of the extra parameter results in a χ^2 of 662.5 on 493 df. The difference in χ^2 between model A3 and A3r equals 22.8 with one df. The various fit indices of model A3r are comparable to those of model A2. In the black (white) sample the standardized residuals were between -3.41 (-3.32) and 3.88 (3.55) with a median of .21 (-.17). The final model in this sequence is A4 (Eqs. 1c-1f). Except for the unconstrained residual variance of the variable HM, this model incorporates all constraints associated with SFI. The RMSEA (.055) and the NNFI (.90) indicate that the model fits adequately. In the black (white) sample, the standardized residuals were between -3.65 (-3.11) and 3.79 (3.65) with a median of .26 (-.12). The ECVI, AIC and CAIC assume the smallest values in this sequence of model, so that we conclude that the three factor model including the SFI constraints is acceptable (compared to A1, A2 & A3r). We accept the hypothesis that the tests are unbiased with respect to race. However, as this hypothesis pertains to all variables, it is quite possible that a subset of these 23 are biased. More detailed information is obtained by inspecting the standardized residuals of the means and the MI's of the vector of intercepts $\boldsymbol{\nu}$. We plot the standardized residuals of the means in model A4 in Figure 2. Of the 23 MI's (associated wit the vector $\mathbf{\nu}$), 21 are smaller than 3.84 (17 smaller than 2). The largest MI's are 7.55 (FP) and 5.04 (AR).

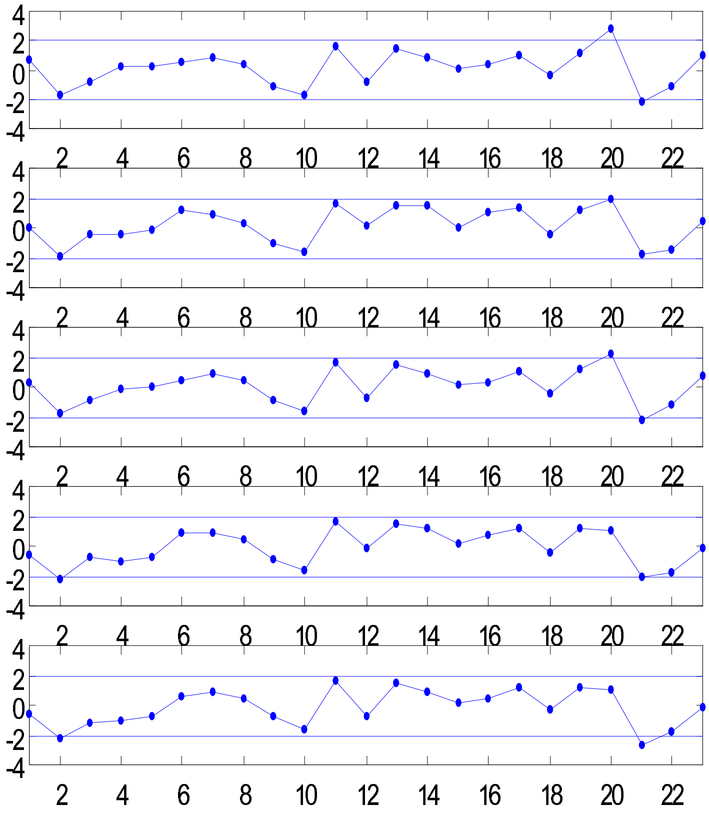

Figure 2: The plot of the standardized residuals of the means. But for sign, there are identical in the black and the white samples. From top to bottom the models are A4, B1, B4, C3, and C6. For the order of the variables, see Table 2.

Second order MFCFM's

We denote the 2nd MFCFM's B1 to B5. These models are shown in Table 4. Table 5 shows the nesting among these models. Model B2 represents the strong version of Spearman's hypothesis, B3, B4 and B5 represent the weak version. B1, which excludes the mean structure, is included to check that the second order factor structure is acceptable. In B1, the loadings of the first order factor ($\mathbf{\Gamma}$) and the first order residuals ($\mathbf{\Psi}$*) are constrained to be equal over the groups. As shown in Figure 1 and Table 1, in this model, g is the only source of group differences in covariance structure (along with the residual of HM).

The goodness of fit indices in Table 3 indicate that these constraints are acceptable. Certainly model B1 fits as well as model A3r. ECVI, AIC and CAIC are slightly lower and the difference in χ^2 equals 7.1 with 5 df. We subsequently fit model B2, which incorporates the strong version of Spearman's hypothesis. Figure 1 and Table 1 illustrate the sources of group differences for model B2 to B5. The goodness of fit indices of model B2 suggest that the model fits at least as well as model A4. In the black (white) sample the standardized residuals were between −3.77 (−3.12) and 3.54 (3.59) with a median of .08 (-.08). We plot the standardized residuals of the means in Figure 2. In both the white (black) sample, 22 (21) of the 23 MI's are smaller than 3.84 (17 smaller than 2 in both samples). In the black sample the largest MI's are 6.94 (FP) and 4.50 (AR). In the white sample, the largest MI equals 4.85 (FP). Model B3 represents the weak version of Spearman's hypothesis. In addition to g, the residual of the 1st order factor V contributed to the difference in mean and covariance structure. Model B5 again represents the weak version, but mean and variance of the 1st order residual of S contributing to the group differences. Neither of these models represents any real improvement over the more parsimonious model B2. Model B4 represents the weak version with the mean and variance of the 1st order residual of M contributing to the group differences. This model does fit slightly better than model B2. With the exception of CAIC, the goodness of fit indices are either as good or better than those of model B2. In the black (white) sample the standardized residuals were between −3.09 (−3.64) and 3.59 (3.63) with a median of -.01 (.14). The standardized residuals of the means in model B4 are plotted in Figure 2. We do not consider models with two residual latent mean differences, because these are hardly more parsimonious than model A4.

First order aMFCFM's

We now relinquish the idea of a second order general intelligence factor, and return to the oblique common factor model. In model A4 all 3 common factor contributed to the groups differences. Here we test the hypotheses that the groups differ with respect to a subset of the three common factors. To this end we fit the 1st order aMGCFM. Figure 1 and Table 1 illustrate how the groups are supposed to differ in these models. We first fit models C1, C2, and C3. In each of these models a single common factor is the source of group differences. Model C1 (V as source of group differences) and model C2 (M) provide no improvement over model B2, which has the same number of df. In contrast to models C1 and C2, we find that model C3 (S) does fit the data almost as well as model B2. The standardized residuals of the means in model C3 are shown in Figure 2. In the black (white) sample the standardized residuals are between -3.77 (-3.03) and 3.25 (3.65) with a median of -.08 (.12). In both samples, 22 of the 23 MI's are smaller than 3.84 (16 smaller than 2 in both samples). The largest MI equals 4.49 (variable FP). In the black sample, 21 of the 23 MI's are smaller than 3.84 and 17 are smaller than 2. The largest MI's are 6.33 (FP) and 4.69 (AR).

We next fitted models in which two common factor contributed to group differences: C4 (V&M), C5(V&S), and C6 (S&M). Model C4 and C5 provide no improvement over model B4 or model C3, both of which fit better. Model C6 (M&S) fits slightly better than model C3 and about as well as model B4. The AIC's and ECVI's are equal and CAIC, as expected, favors the more parsimonious C3. In the black (white) sample the standardized residuals are between -3.46 (-3.03) and 3.25 (3.66) with a median of .09 (-.09). The standardized residuals of the means are shown in Figure 2. The MI's were almost identical to those observed in model C3.

Some detailed results of model B4 and C3

It is difficult to identify the best fitting model on the basis of the present results. It appears that model B1, B4, C3 and C6 all fit about equally well. Model C3 should perhaps be preferred to model C6 as this model is more parsimonious and fits as well. Model B4 is consistent with Jensen's finding that the weak version of Spearman's hypothesis holds, with additional differences favoring blacks in the factor M (Jensen, 1985; 1998). The similarity of the models is also evident in the plot of the standardized residuals. These are almost identical for model A4, B1, B4, C3 and C6.

We now consider the results of model C3 and B4 in more detail. Before doing so, we refit the models without the parameters accounting for mean differences, lest there be any doubt concerning their significance within these models. The goodness of fit indices for these models are χ^2(520)=717.3 (B4) and χ^2(498)=722.8 (C3). The decrease in χ^2 is about 23.9 (B4, df=2) and 21.8 (C3, df=1).

In model B4, the mean reliability of the IQ tests in the white (black) sample equals .47 (.43), the median equals .50 (.46), and the standard deviation equals .18 (.19). The lowest and the highest reliabilities equal .07 and .77 (.07 & .76). These are associated with the WISC-R tests CO and V, respectively. These findings were almost identical in model C3. In model B4, g accounts for 52%, 16% and 75% of the variance in the factors V, M, and S, respectively, in the white sample. In the black sample these percentages are 52%, 29% and 75% (remember that this residual variance is free to vary over the groups). The results concerning the factor means and variances are shown in Table 6 for both models. As is to be expected in view of the comparable goodness of fit of the two models, these are quite similar. In model B4, the correlation between first order factors V and M and the second order g is about .72 and .87. The correlation between g and M is lower .40 (whites) and .54 (blacks). The high correlation between g and S suggest that they are closely related.

Table 6: Latent means and covariance matrices (correlation and standard deviations in parentheses) in models B4 and C3.

Model B4

Blacks: covariance matrix of V,M,S, and g.

V	21.15 (4.60)			
M	2.94 (.39)	2.72 (1.65)		
S	24.94 (.62)	6.73 (.47)	76.06 (8.72)	
g	10.89 (.72)	2.94 (.54)	24.94 (.87)	10.89 (3.30)

First order factor residual variance (in regression of factor on g).

V	M	S
10.25 (.52%)	1.93 (29%)	18.97 (75%)

Whites: Covariance matrix of V,M,S, and g.

V	21.29 (4.61)			
M	2.98 (.29)	4.98 (2.23)		
S	25.26 (.62)	6.82 (.35)	76.80 (8.76)	
g	11.03 (.72)	2.98 (.40)	25.26 (.87)	11.03 (3.22)

First order factor residual variance (in regression of factor on g).

V	M	S
10.25 (52%)	4.18(16%)	18.97(75%)

Differences in latent means δ and τ

V	M	S	g
0	0.418	0]	-2.915

Differences in means of 1st order factors ($\Gamma\delta+\tau$)

V	M	S
-2.92	-0.37	-6.67

Table 6 cont.: Latent means and covariance matrices (correlation and standard deviations in parentheses) in models B4 and C3.

Model C3

Blacks: Covariance matrix of V,M, and S.

V	22.25 (4.72)		
M	2.87 (.31)	3.96 (1.99)	
S	27.04 (.64)	7.01 (.39)	78.55 (8.86)

Whites: Covariance matrix of V,M, and S.

V	21.74 (4.66)		
M	2.74 (.29)	3.93 (1.98)	
S	25.55 (.64)	6.63 (.39)	74.24 (8.62)

differences in latent means of factors V,M,S ($\boldsymbol{\Psi}_W \boldsymbol{\pi}$)

V	M	S
-2.30	-0.59	-6.68

Discussion

The aims of the present chapter are to present a comprehensive critique of the method of correlated vectors in the light of MGCFM's and to fit MGCFM's to the Naglieri and Jensen (1987) data set. On the basis of the present, as well as other results (Dolan, 2000), we are convinced that the Spearman correlation cannot be used to demonstrate the importance of g in b-w differences with any confidence. Lubke, Dolan and Kelderman (2001) already showed that the Spearman correlation seems to be lacking in specificity. The results of our MGCFA's underline this problem of the Spearman correlation. Naglieri and Jensen (1987) report a Spearman correlation of .75 for the data set that we have analyzed in this chapter. They conclude that Spearman's hypothesis seems to be strongly substantiated. In terms of the model presented above, Naglieri and Jensen are concluding that the correlation of .75 supports model B2 or, say, model B4. In the light of the present results, however, we do not see how one can draw this conclusion. Many of the models we investigated fitted about equally well. One can of course note the small sample sizes, and the attendant lack of power to distinguish between models. However, this does not salvage the interpretation of the Spearman correlation. The power argument obviously applies equally to Jensen's procedure and to MGCFA. In addition, Dolan (2000) in analyzing WISC-R data obtained in much larger samples (1868 whites and 305 blacks; Jensen and Reynolds, 1982) encountered exactly the same problem. He found it difficult to distinguish between models that are consistent with Spearman's hypothesis and alternative models[3]. The fact that the Spearman correlation is found to be consistently large and positive (Jensen, 1998) does not help. In view of this consistency, Nyborg and Jensen (2000) draw the following conclusion: "The present findings, in addition to those of previous studies, which they strongly replicate, support the conclusion that Spearman's original conjecture that the b-w differences in cognitive tests is predominantly a differences in the g factor should no longer be regarded as just an hypothesis but as an empirically established fact (p. 599)". The repeated demonstration of a positive and large Spearman correlation is a necessary, but not a sufficient condition for inferring the correctness of Spearman's hypothesis. Suppose that it emerges consistently that one cannot distinguish well between competing models using MGCFA. It is possible that the analysis of all available data sets (perhaps using an appropriate meta-analytic procedure) will demonstrate that a model incorporating the weak version of Spearman's hypothesis provides the best description of the data. However, until this work is undertaken, we cannot accept Spearman's hypothesis as an "empirically established fact". We also maintain that researchers, who think that the Spearman correlation can demonstrate the role of g as the major source of b-w differences in IQ tests scores, should provide results to show that 1) the six methodological weaknesses mentioned above can safely be discounted, and that 2) specificity is not a problem.

We have not addressed the interpretation or meaning of g in the present chapter. Following Jensen, we simply called the second order factor in the 2nd order MGFCM g. Horn and Noll (1997) provide a critical discussion of the factor analytic evidence purporting to demonstrate the existence of general intelligence. One problem is that it may be very difficult to distinguish between first order factors and second order g when model fitting is used (Gustafsson, 1984). We can illustrate this problem by fixing the residual

[3] Dolan (2000), fitting the same models to the WISC-R data published in Jensen and Reynolds (1982), found that models B5, C4 and A4 fit about equally well. In the present analyses Memory factor is much better represented than in Dolan (2000). We suspect that this has contributed to

variance of the first order factor S equal zero in model B4. In so doing, we equate the first order S and the second order g. The χ^2 for this model is 696.8 (519 df), a decrease of 2.4, with one df, compared to the original B4 (see Table 2). This is not significant by the usual standards, but the present sample sizes do not afford much power to reject the constraint of zero residual variance. Relinquishing g presents no obstacle to the investigation of group differences by means of MGCFA. Seven of the eleven models we considered do not include the second order factor.

We conclude that strict factorial invariance is tenable in comparisons of IQ test scores of blacks and whites. We base this conclusion on the finding that model A4, i.e., the least restrictive model incorporating SFI, fits reasonably well (see also Dolan, 2000). This is an important conclusion, because it implies that measurement bias, as defined by Mellenbergh (1989), is absent. Measurement bias, or content bias as Jencks and Phillips (1998) call it, is generally assumed to be absent (Jencks, 1998). It is nice to find support for this using the appropriate methodology. This finding also sheds light on the issue of the relationship between sources of within-group and between-group differences. It is often stated that these may be distinct. SFI implies that within and between-group differences are differences on the same latent dimensions (Lubke, Dolan, Kelderman and Mellenbergh, 2001). It should be emphasized that the between group differences may be due to a subset of the common factors contributing to within group differences. Locating the source of between group differences within a set of identified sources of within group differences is exactly the point of our comparison of competing models.

the difference in sets preferred models. This demonstrates how tentative conclusions concerning group differences depend on the test being analyzed.

Acknowledgement

The work of CVD has been made possible by a fellowship of the Royal Dutch Academy of Arts and Sciences. We thank Jan te Nijenhuis for regularly bringing to our attention applications of the method of correlated vector (fearlessly including his own). We thank Peter Molenaar for providing ELH with the opportunity to carry out this research, and his general encouragement.

References

Bentler, P. (1990). EQS structural equations program manual. Los Angeles: BMDP Scientific Software.

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

Bollen, K.A. and Long, J.S. (1993). Introduction. In: K.A.Bollen and J.S. Long (Eds.). Testing structural equation models. Newbury Park: Sage Publications.

Browne, M.W. and Cudeck, R. (1993). Alternative ways of assessing model fit. In: K.A. Bollen and J. Scott Long (Eds.). Testing Structural Equation Models. Newbury Park: Sage Publications.

Devlin, B., Fienberg, S.E., Resnick, D.P. and Roeder, K. (1997). Intelligence, Genes, and Success. New York: Springer Verlag.

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. and Lubke, G.H. (2001). Viewing Spearman's hypothesis from the perspective of multi-group PCA: A comment on Schönemann's criticism. Intelligence. In press.

Dolan, C.V. and Molenaar, P.C.M. (1994). Testing specific hypotheses concerning latent group differences in multi-group covariance structure analysis with structured means. Multivariate Behavioral Research, 29, 203-222.

Dolan, C.V. (1997). A note on Schönemann's refutation of Spearman's hypothesis. Multivariate Behavioral Research, 32, 319-325.

Gustafsson, J-E. (1992). The relevance of factor analysis for the study of group differences. Multivariate Behavioral Research, 27, 239-247.

Gustafsson, J-E. (1988). Hierarchical models of individual differences in cognitive abilities. In: R.J.Sternberg (Ed.), Advances in the psychology of human intelligence (35-71). Hillsdale, NJ: Erlbaum.

Gustafsson, J-E. (1984). A unifying model for the structure of intellectual abilities. Intelligence, 8, 179-203.

Guttman, L. (1992). The irrelevance of factor analysis for the study of group differences. Multivariate Behavioral Research, 27, 175-204.

Hanna, G. and Lei, H. (1985). A longitudinal analysis using the LISREL model with structured means. Journal of Educational Statistics, 10, 161-169.

Horn, J.L. and McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. Experimental Aging Research, 18, 117-144.

Horn, J.L. (1997). On the mathematical relationship between factor or component coefficients and differences in means. Cahiers de Psychologie Cognitive, 16, 721-728.

Horn, J.L. and Noll, J. (1997). Human Cognitive Capabilities: Gf-Gc Theory. In: D.P.Flanagan, J.L. Genshaft and P.L. Harrison (Eds). Contemporary Intellectual Assessment. Theories, Tests, and Issues. New York: The Guildford Press.

Jencks, C. and Phillips, M. (1998). The black-white test score gap. Washington: Brookings Institution Press.

Jencks, C. (1998). Racial Bias in Testing. In: C. Jencks and M. Phillips, M. (Eds). The black-white test score gap. Washington: Brookings Institution Press.

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

Jensen, A.R. (1984). The nature of black-white differences on the K-ABC: Implications for future tests. Journal of Special Education, 18, 377-408.

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. (1992). Spearman's hypothesis: Methodology and evidence. Multivariate Behavioral Research, 27(2), 225-233.

Jensen, A.R. (1998). The g factor. The science of mental ability. Westport: Praeger.

Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. Psychometrika, 36, 409-426.

Jöreskog K.G. and Sörbom D. (1999). LISREL 8 [Computer program]. Chicago: Scientific Software International.

Jöreskog, K.G. (1993). Testing Structural Equation Models. In: K.A. Bollen and J. Scott Long (Eds.). Testing Structural Equation Models. Newbury Park: Sage Publications.

Jöreskog, K.G and Sörbom, D. (1989). LISREL VII. A guide to the program and applications. Chicago: SPSS inc.

Keith, T.Z. (1997). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In: D.P.Flanagan, J.L. Genshaft and P.L. Harrison (Eds). Contemporary Intellectual Assessment. Theories, Tests, and Issues. New York: The Guildford Press.

Keith, T.Z. and Dunbar, S.B. (1984). Hierarchical factor analysis of the K-ABC: Testing Alternate Models. Journal of Special Education, 18, 367-375.

Lawley, D.N. and Maxwell, A.E. (1971). Factor analysis as a statistical method. London: Butterworth.

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

Loehlin, J.C. (1992). On Schönemann on Guttman on Jensen, via Lewontin. Multivariate Behavioral Research, 27, 261-263.

Lubke, G.H., Dolan, C.V. and Kelderman, H. (2001). Investigating group differences on cognitive tests using Spearman's hypothesis: An evaluation of Jensen's method. Multivariate Behavioral Research. In press.

Lubke, G.H., Dolan, C.V., Kelderman, H. & Mellenbergh, G.J. (2001). The Flynn effect and group differences in ability and achievement tests: explicit models of within- and between-group differences. submitted.

Lynn, R. and Owen, K. (1994). Spearman’s hypothesis and test scores differences between White, Indian, and Blacks in South Africa. The Journal of General Psychology, 121, 27-36.

Mackintosh, N.J. (1998). IQ and Human Intelligence. Oxford: Oxford University Press.

Marsh, H.W. and Grayson, D. (1990). Public/Catholic differences in the high school and beyond data: A multi-group structural equation modelling approach to testing mean differences. Journal of Educational Statistics, 5, 199-235.

Marsh, H.W. and Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. Structural Equation Modeling, 1, 317-359.

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

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

Millsap, R.E. and Everson, H. (1991). Confirmatory measurement model comparisons using latent means. Multivariate Behavioral Research, 26, 479-497.

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

Millsap, R.E. (1997b). The investigation of Spearman's hypothesis and the failure to understand factor analysis. Cahiers de Psychologie Cognitive, 16, 750-757.

Muthén, B. and Lehman, J. (1985). Multiple group IRT modeling application to item bias analysis. Journal of Educational Statistics, 10, 133-142.

Neale, M. (1997). Mx: Statistical modeling. Richmond: Medical College of Virginia.

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

Nyborg, H. and Jensen, A.R. (2000). Black-white differences on various psychometric tests: Spearman's hypothesis tested on American armed services veterans. Personality and Individual Differences, 28, 593-599.

Oort, F.J. (1996). Using restricted factor analysis in test construction. Ph.D. Thesis, Psychology Faculty, University of Amsterdam.

Rushton, J.P. (1999). Secular gains is IQ not related to the g factor and inbreeding depression — unlike Black-White differences: a reply to Flynn. Personality and Individual Differences, 26, 381-389.

Saris, W.E., de Pijper and Mulder, J. (1978). Optimal procedures for estimating factor scores. Sociological Methods and Research, 7, 85-106.

Schaie, K.W., Willis, S.L. Jay and Chipuer, H.(1998). Structural invariance of cognitive abilities across the adult life span: A cross-sectional study. Journal of Personality and Social Psychology, 25, 652-662.

Schmid, J. and Leiman, J.M. (1957). The development of hierarchical factor solutions. Psychometrika, 22, 53-61.

Schönemann P.H. (1997). Famous artefacts: Spearman's hypothesis. Cahiers de Psychologie Cognitive (Current Psychology of Cognition), 16, 665-694.

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

Sörbom, D. (1975). Detection of correlated errors in longitudinal data. British Journal of Mathematical and Statistical Psychology, 27, 229-239.

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

SPSS inc. (1990). SPSS reference guide. Chicago: SPSS, inc.

Spearman, C. (1927) The abilities of man: Their nature and measurement. New York: Macmillan.

Steiger, J.H. (1990). Structural model evaluation and modification: and interval estimation approach. Multivariate Behavioral Research, 25, 173-180.

te Nijenhuis, J. and van der Flier, H. (1997). Comparability of the GATB scores for immigrants and majority group members: some Dutch findings. Journal of Applied Psychology, 82, 675-687.

te Nijenhuis, J., Evers, A. and Mur, J.P. (2000). Validity of differential aptitude test for the assessment of immigrant children. Educational Psychology, 20, 99-115.