Available online at www.sciencedirect.com
SCIENCE DIRECT®
Intelligence 32 (2004) 155–173
INTELLIGENCE

Two failures of Spearman's hypothesis: The GATB in Holland and the JAT in South Africa

Conor V. Dolan*, Willemijn Roorda, Jelte M. Wicherts

Department of Psychology, University of Amsterdam, Roetersstraat 15, Amsterdam 1018 WB, Netherlands

Received 8 July 2002; received in revised form 4 August 2003; accepted 3 September 2003

Abstract

Spearman's hypothesis states that the differences between Blacks and Whites in psychometric IQ are attributable to a fundamental difference in general intelligence (*g*). To investigate this hypothesis, Jensen devised the method of correlated vectors. This method involves calculating the correlation between the factor loadings of the subtest and the observed differences in means. Although the hypothesis concerns U.S. populations, Jensen's test has also been used to investigate other groups. The aim of the present paper is to test Spearman's hypothesis in a published Dutch and a published South African data set. Both data sets were previously analyzed by Jensen's method, and the results were interpreted in support of Spearman's hypothesis. In this paper, we reanalyzed both data sets by Multigroup Confirmatory Factor Analysis (MGCFA). We find that the hypothesis of factorial invariance, which implies that the same construct is measured in the groups, must be rejected. This greatly complicates any comparison of the group with respect to the test scores and makes it impossible to determine the role, if any, of *g* in explaining the observed differences in psychometric IQ. This conclusion runs counter to the conclusion that Spearman's hypothesis is supported in these data sets.

Keywords: Spearman's hypothesis; GATB; JAT; Holland; South Africa

1. Introduction

Spearman's hypothesis states that the differences in psychometric intelligence between Blacks and Whites in the United States are directly related to the tests' general intelligence (*g*) loadings. This implies that the Blacks–Whites (B-W) differences in test scores are not attributable merely to cultural or

* Corresponding author.
E-mail address: cdolan@fmg.uva.nl (C.V. Dolan).

0160-2896/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2003.09.001

linguistic peculiarities in a certain test but to a general factor that permeates all mental tests. Jensen (1985, 1998) formulated two versions of Spearman's hypothesis, a strong version and a weak version. The strong version states that the magnitudes of the B-W differences in psychometric intelligence tests are *solely* a positive function of variation in the tests' *g* loadings. The weak version of the hypothesis states that the B-W differences in psychometric intelligence are *mainly* a positive function of variation in the tests' *g* loadings. This means that the populations may also differ in other cognitive ability factors besides *g*.

To test Spearman's hypothesis, Jensen devised the method of correlated vectors, which involves the following steps. First, exploratory factor analyses of psychometric data are carried out in both samples, separately, to extract factor loadings on *g*. Second, factorial invariance is established by calculating measures of factorial congruence of the factor loadings in both samples. Third, differences in means are standardized. Fourth and last, the standardized differences are correlated with the *g* factor loadings. Using this procedure, Jensen (1998, p. 378) reported a mean correlation of .63 based on 149 psychometric tests and concluded that these results support the weak form of the hypothesis to such an extent that the hypothesis should be viewed as an empirical fact (p. 379; see also Nyborg & Jensen, 2000, p. 594; Te Nijenhuis & van der Flier, 1997, p. 678).

Although Jensen's method has been widely used to test Spearman's hypothesis in a variety of groups (Colom, Juan-Espinosa, Abad, & García, 2000; Colom, Juan-Espinosa, & García, 2001; Flynn, 2000; Helms-Lorenz, van der Vijver, & Poortinga, 2003; Lynn & Owen, 1994; Must, Must, & Raudik, 2003; Naglieri & Jensen, 1987; Nyborg & Jensen, 2000; Rushton, 2001; Te Nijenhuis, Evers, & Mur, 2000; Te Nijenhuis & van der Flier, 1997), it has been the subject of criticism (see special issues of *Multivariate Behavioral Research*, 1992 and *Cahiers de Psychologie Cognitive*, 1997). Not all criticisms are justified (see Dolan & Lubke, 2001), but it has been established empirically that Jensen's method is quite insensitive to detect violations of the various components of Spearman's hypothesis (Lubke, Dolan, & Kelderman, 2001). A comprehensive criticism of Jensen's method is given in Dolan and Hamaker (2001). Much of this criticism originates in the fact that Jensen's method does not allow for explicit and comprehensive model fitting, explicit goodness-of-fit testing, or comparison of competing models. In addition, Jensen's method does not allow for rigorous testing of the necessary conditions for valid group comparisons within the common factor model. Multigroup confirmatory factor analysis (MGCFA), which was already advocated by Gustafsson (1992), Horn (1997), and Millsap (1997) in this context, can be regarded as a superior method as it does offer these possibilities. Lubke, Dolan, Kelderman, and Mellenbergh (2003) discuss the method in conceptual terms. Meredith (1993) presents a rigorous development of multigroup common factor models, including the conditions that have to be met for unbiased group comparisons.

Applications of MGCFA to study B-W differences in psychometric intelligence are given by Dolan (2000), Dolan and Hamaker (2001), Gustafsson (1992), and Lubke et al. (2003). Dolan and Dolan and Hamaker investigated B-W differences in WISC-R and the K-ABC test scores in U.S. samples[1] and concluded that the tests are unbiased with respect to group. This implies that the same constructs are measured in the groups. However, because various competing models fitted the data approximately equally well, it remained difficult to establish the central role of *g* in B-W differences with any confidence.

Although Jensen formulated Spearman's hypothesis in an attempt to gain understanding of B-W differences in the United States, the method has also been applied to study psychometric IQ scores in

[1] These data sets are published in Jensen and Reynolds (1982) and Naglieri and Jensen (1987).

other groups and in other countries. Two of these applications are the focus of this paper. The first concerns a study performed by Te Nijenhuis and van der Flier (1997) in the Netherlands. The data set consists of samples Dutch, Surinamese, Dutch Antilleans, North Africans, and Turks. Te Nijenhuis and van der Flier reported correlations of .42 (Surinamese), .71 (Antilleans), .87 (North Africans), and .87 (Turks) and concluded that *g* is the predominant factor accounting for differences between the majority group (the Dutch group) and the immigrant groups (p. 686). The second data set is published in Lynn and Owen (1994) and concerns a South African data set that consists of samples of Whites, Blacks, and Indians. On the basis of a significant correlation of .62 that was obtained for the B-W differences using the corrected *g* loadings in the Black sample, Lynn and Owen concluded that Spearman's hypothesis was strongly supported (p. 27).

The aim of the present paper is to investigate Spearman's hypothesis by analyzing the data sets published in Lynn and Owen (1994) and Te Nijenhuis and van der Flier (1997). We want to establish whether the conclusion based on Jensen's method of correlated vectors can be corroborated using the more rigorous modeling offered by MGCFA. We first discuss MGCFA in general terms. Next, we present the constraints, which have to be met for valid group comparisons within the common factor model. We relate these constraints to an explicit definition of unbiasedness and discuss the relationship between the models presented and Spearman's hypothesis. We then apply MGFCA to the two data sets and discuss the results. We conclude the paper with a general discussion.

2. Multigroup confirmatory factor analysis

MGCFA is a well-established method to study group differences in means and covariances within the common factor model (e.g., Byrne, Shavelson, & Muthén, 1989; Jöreskog, 1971; Little, 1997; Marsh & Grayson, 1990; Millsap & Everson, 1991; Rock, Werts, & Flaugher, 1978; Sörbom, 1974). Standard software can be used to carry out MGCFA, such as EQS (Bentler, 1990), the freely available Mx (Neale, 1997), or LISREL (Jöreskog & Sörbom, 1999).

We consider the first-order multigroup confirmatory factor model (MGCFM) and the second-order MGCFM (e.g., Dolan, 2000). The latter is particularly well suited to investigate the strong and weak versions of Spearman's hypothesis. The models are presented, including various identifying and substantive constraints. The substantive constraints are imposed to establish factorial invariance, which is a necessary condition for group comparisons to be valid (Lubke et al., 2003; Meredith, 1993). Certain additional constraints are imposed to ensure that parameter estimates are identified (e.g., Bollen, 1989; Long, 1983).

2.1. First-order MGCFM

We employ matrix notation to present the factor model (e.g., Bollen, 1989; Long, 1983). Let $\mathbf{y}_{ij}$ denote the observed *p*-dimensional random column vector of subject *j* in population *i*. The following linear factor model is assumed to hold for the observation $\mathbf{y}_{ij}$:

$$\mathbf{y}_{ij} = \boldsymbol{\nu}_i + \boldsymbol{\Lambda}_i \boldsymbol{\eta}_{ij} + \boldsymbol{\varepsilon}_{ij}, \tag{1}$$

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 random vector of residuals. The residuals in $\boldsymbol{\varepsilon}_{ij}$ contain both random measurement and

systematic error. The systematic error consists of specific scores unique to each observed variable (Meredith, 1993, p. 532). Eq. (1) may be viewed as a regression equation in which the $(p \times q)$ matrix $\boldsymbol{\Lambda}_i$ contains regression coefficients (factor loadings) in the regression of $\mathbf{y}_{ij}$ on $\boldsymbol{\eta}_{ij}$, the $(p \times 1)$ vector $\boldsymbol{\nu}_i$ contains intercepts, and $\boldsymbol{\varepsilon}_{ij}$ are residuals. With respect to the distribution of the random variables, we assume that $\boldsymbol{\varepsilon}_{ij} \sim N_p(0,\boldsymbol{\Theta}_i)$[2] and $\boldsymbol{\eta}_{ij} \sim N_q(\boldsymbol{\alpha}_i,\boldsymbol{\Psi}_i)$. The $(p \times p)$ covariance matrix $\boldsymbol{\Theta}_i$ of the residuals $\boldsymbol{\varepsilon}_{ij}$ is diagonal and positive (i.e., the residuals are uncorrelated). The $(q \times q)$ covariance matrix $\boldsymbol{\Psi}_i$ is positive definite (see Wothke, 1993). Given these assumptions, the observed variables are distributed $\mathbf{y}_{ij} \sim N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$, where, assuming the covariance between $\boldsymbol{\eta}_{ij}$ and $\boldsymbol{\varepsilon}_{ij}$ is 0 ($E[(\boldsymbol{\eta}_{ij} - \boldsymbol{\alpha}_i)\boldsymbol{\varepsilon}_{ij}^t] = 0$):

$$\boldsymbol{\mu}_i = \boldsymbol{\nu}_i + \boldsymbol{\Lambda}_i\boldsymbol{\alpha}_i \tag{2}$$

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}_i\boldsymbol{\Psi}_i\boldsymbol{\Lambda}_i^t + \boldsymbol{\Theta}_i. \tag{3}$$

The superscript t denotes transposition. We call this model the first-order MGCFM (e.g., Bollen, 1989; Long, 1983; Sörbom, 1974).

2.2. Second-order MGCFM

In Jensen's tests of Spearman's hypothesis, the common factor g features as a second-order common factor (e.g., Jensen, 1998; Jensen & Reynolds, 1982; Naglieri & Jensen, 1987), which is assumed to account for the covariance among the first-order factors (i.e., $\boldsymbol{\Psi}_i$ in Eq. (3)). Let $\boldsymbol{\eta}_{ij} = \boldsymbol{\Gamma}_i\boldsymbol{\xi}_{ij} + \boldsymbol{\zeta}_{ij}$, where $\boldsymbol{\Gamma}_i$ $\boldsymbol{\Gamma}_i$ is a $(q \times r)$ matrix of loadings of the q first-order factors $\boldsymbol{\eta}_{ij}$ 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. The model of the observations is now:

$$\mathbf{y}_{ij} = \boldsymbol{\nu}_i + \boldsymbol{\Lambda}_i(\boldsymbol{\Gamma}_i\boldsymbol{\xi}_{ij} + \boldsymbol{\zeta}_{ij}) + \boldsymbol{\varepsilon}_{ij}. \tag{4}$$

We assume that $\boldsymbol{\zeta}_{ij} \sim N_q(\boldsymbol{\alpha}_i,\boldsymbol{\Psi}_i^*)$ and $\boldsymbol{\xi}_{ij} \sim N_r(\boldsymbol{\kappa}_i,\boldsymbol{\Phi}_i)$. The asterisk in $\boldsymbol{\Psi}_i^*$ is employed to indicate that this matrix is diagonal (in contrast to $\boldsymbol{\Psi}_i$ in Eq. (3)). Furthermore, we assume that $\boldsymbol{\zeta}_{ij}$ and $\boldsymbol{\varepsilon}_{ij}$ are uncorrelated ($E[(\boldsymbol{\zeta}_{ij} - \boldsymbol{\alpha}_i)\boldsymbol{\varepsilon}_{ij}^t] = 0$) and that $\boldsymbol{\xi}_{ij}$ and $\boldsymbol{\zeta}_{ij}$ are uncorrelated ($E[(\boldsymbol{\xi}_{ij} - \boldsymbol{\kappa}_i)(\boldsymbol{\zeta}_{ij} - \boldsymbol{\alpha}_i)^t] = 0$). This implies that $\mathbf{y}_{ij} \sim N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$, where

$$\boldsymbol{\mu}_i = \boldsymbol{\nu}_i + \boldsymbol{\Lambda}_i\boldsymbol{\alpha}_i + \boldsymbol{\Lambda}_i\boldsymbol{\Gamma}_i\boldsymbol{\kappa}_i \tag{5}$$

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}_i(\boldsymbol{\Gamma}_i\boldsymbol{\Phi}_i\boldsymbol{\Gamma}_i^t + \boldsymbol{\Psi}_i^*)\boldsymbol{\Lambda}_i^t + \boldsymbol{\Theta}_i. \tag{6}$$

We call this the second-order MGCFM. Given the hypothesis of a single second-order common factor (g), we would have $r = 1$ in this model.

So far, we have simply presented multigroup version of the well-known first-order and second-order common factor models. We now consider the constraints, which we require to obtain a model that properly represents the hypothesis that g is the main source of B-W differences in psychometric IQ test scores.

[2] The notation $\boldsymbol{\varepsilon}_{ij} \sim N_p(0,\boldsymbol{\Theta}_i)$ means that $\boldsymbol{\varepsilon}_{ij}$ is p-variate normally distributed with mean 0 and covariance matrix $\boldsymbol{\Theta}_i$.

2.3. Testing Spearman's hypothesis using MGCFA

For group comparisons to be meaningful, we have to establish that we are measuring the same construct(s) in each group, or stated otherwise, we must establish that our measurements are unbiased with respect to group (Lubke et al., 2003; Meredith, 1993). To establish this, we adopt the formal definition of unbiasedness as presented by Mellenbergh (1989). According to this definition, bias is absent if $f(\mathbf{Y}|\boldsymbol{\eta})=f(\mathbf{Y}|\boldsymbol{\eta},s)$. This means that the distribution of the observed variable $\mathbf{Y}$ (IQ subtests) conditioned on the common factors, $f(\mathbf{Y}|\boldsymbol{\eta})$, equals the distribution of the observed variable $\mathbf{Y}$ (IQ subtests) conditioned on the common factors and group (s), $f(\mathbf{Y}|\boldsymbol{\eta},s)$. If $f(\mathbf{Y}|\boldsymbol{\eta})$ does not equal $f(\mathbf{Y}|\boldsymbol{\eta},s)$, we state that one or more of the tests (components of $\mathbf{Y}$) are biased with respect to group. In his discussion of racial bias in testing, Jencks (1998) refers to this type of bias as content bias. First, we consider the constraints that are required for this definition to hold in the factor model. Subsequently, we discuss this definition of measurement invariance in relationship to the factor model.

Meredith (1993) provides a detailed discussion of necessary conditions for measurement invariance, or unbiasedness, in the common factor model. In this context, Meredith distinguishes between strict and strong factorial invariance. Strict factorial invariance comprises the substantive constraints that the factor loadings ($\boldsymbol{\Lambda}$, and in the second-order factor model, $\boldsymbol{\Gamma}$), the specific variances ($\boldsymbol{\Theta}$), and the intercepts ($\boldsymbol{\nu}$) are equal over groups. If the tests are unbiased, as defined above, these equality constraints should hold to reasonable approximation (Meredith, 1993; Millsap, 1997; Muthén & Lehman, 1985; Oort, 1992, 1996). Meredith further discusses strong factorial invariance, which comprises the same substantive constraints as strict factorial invariance but relaxes the constraint on the covariance matrices of the residual variance terms. In the model representing strong factorial invariance, the covariance matrix of the residuals ($\boldsymbol{\Theta}$) is allowed to vary over groups.

Besides the substantive constraints that are associated with the hypotheses of strict and strong factorial invariance, some identifying constraints are required, which are well established in confirmatory factor analysis (Bollen, 1989; Long, 1983). We introduce sufficient fixed zeroes in $\boldsymbol{\Lambda}$ to avoid rotational indeterminacy, given correlated common factors. Furthermore, in $\boldsymbol{\Lambda}$, we fix certain elements to equal 1 to determine the variances of the common factors, and in the second-order factor model, we fix an element of $\boldsymbol{\Gamma}$ to equal 1, so that we can estimate the variance of $\boldsymbol{\xi}$. Also for reasons of identification, we are bound to model differences in latent means instead of latent means (Sörbom, 1974). In the reference (f) group, which is chosen arbitrarily, the observed means are equated with the vector of intercepts $\boldsymbol{\nu}$. In the nonreference (nf) groups, the differences in latent means $\boldsymbol{\alpha}_{\mathrm{nf}}-\boldsymbol{\alpha}_{\mathrm{f}}$ are estimated. Given this constraint and the constraints that are associated with strict factorial invariance, the first-order MGCFM is:

$$\boldsymbol{\mu}_{\mathrm{f}}=\boldsymbol{\nu} \tag{7}$$

$$\boldsymbol{\Sigma}_{\mathrm{f}}=\boldsymbol{\Lambda}\boldsymbol{\Psi}_{\mathrm{f}}\boldsymbol{\Lambda}^{t}+\boldsymbol{\Theta} \tag{8}$$

$$\boldsymbol{\mu}_{\mathrm{nf}}=\boldsymbol{\nu}+\boldsymbol{\Lambda}\boldsymbol{\delta} \tag{9}$$

$$\boldsymbol{\Sigma}_{\mathrm{nf}}=\boldsymbol{\Lambda}\boldsymbol{\Psi}_{\mathrm{nf}}\boldsymbol{\Lambda}^{t}+\boldsymbol{\Theta}, \tag{10}$$

where the vector of differences in latent means of the common factors, $\boldsymbol{\delta}$, equals $\boldsymbol{\alpha}_{nf} - \boldsymbol{\alpha}_{f}$. The second-order MGCFM, given strict factorial invariance, is defined as:

$$\boldsymbol{\mu}_{f} = \boldsymbol{\nu} \tag{11}$$

$$\boldsymbol{\Sigma}_{f} = \boldsymbol{\Lambda}\left(\boldsymbol{\Gamma}\boldsymbol{\Phi}_{f}\boldsymbol{\Gamma}^{t} + \boldsymbol{\Psi}^{*}_{f}\right)\boldsymbol{\Lambda}^{t} + \boldsymbol{\Theta} \tag{12}$$

$$\boldsymbol{\mu}_{nf} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\delta} + \boldsymbol{\Lambda}\,\boldsymbol{\Gamma}\boldsymbol{\tau} \tag{13}$$

$$\boldsymbol{\Sigma}_{nf} = \boldsymbol{\Lambda}\left(\boldsymbol{\Gamma}\boldsymbol{\Phi}_{nf}\boldsymbol{\Gamma}^{t} + \boldsymbol{\Psi}^{*}_{nf}\right)\boldsymbol{\Lambda}^{t} + \boldsymbol{\Theta}, \tag{14}$$

where $\boldsymbol{\tau}$ equals $(\boldsymbol{\kappa}_{nf} - \boldsymbol{\kappa}_{f})$ and represents the vector of differences in latent means of the second factor.

With these models in place, we return to the definition of unbiasedness, which is central to the formulation of the models. Lubke et al. (2003) provide a conceptual discussion of this definition of unbiasedness and the associated strict factorial invariance model. In conceptual terms, strict factorial invariance implies that the between-group differences (i.e., differences in means) and within-group differences (i.e., systematic individual differences) are differences on the same common factor(s). This means that there should be no difference in the distribution of the test scores of subjects (regardless of their group membership) if one conditions on the common factor scores. Clearly, if one conditions on the only source of systematic between-group and within-group differences [viz. the common factor(s)], there should be no difference in the resulting conditional distributions. This explains the idea behind the definition of unbiasedness mentioned above [i.e., $f(\mathbf{Y}|\boldsymbol{\eta}) = f(\mathbf{Y}|\boldsymbol{\eta},s)$]. Now, consider the situation in which $\boldsymbol{\eta}$ is not the only source of group differences. In this case, conditioning on $\boldsymbol{\eta}$ cannot remove all group differences so that the requirement for unbiasedness, as defined above, can no longer hold [i.e., $f(\mathbf{Y}|\boldsymbol{\eta}) \neq f(\mathbf{Y}|\boldsymbol{\eta},s)$].

The present definition of unbiasedness gives rise to the MGCFMs presented above (Eqs. (7–(10) and Eqs. (11)–(14); see Meredith, 1993). As noted by Millsap (1997), specific instances of these models incorporate the crux of Jensen's test of Spearman's hypothesis (i.e., the relationship between factor loadings and the group differences in subtest means). To demonstrate this, we limit our discussion to the second-order MGCFM because this model provides the best representation of Jensen's factor analytic conceptualization of *g* (i.e., as a higher-order common factor; see Jensen, 1998). The strong version of Spearman's hypothesis states that observed B-W differences in means are *solely* a positive function of variation in the tests' *g* loadings. This implies that there are no differences in latent means of the first common factor (e.g., $\boldsymbol{\delta} = 0$ in Eq. (13)). In this model, the loadings on the second-order factor (*g*) are, given strict or strong factorial invariance, represented by $\boldsymbol{\Lambda\Gamma}$ and the mean differences by $\boldsymbol{\Lambda\Gamma\tau}$. Because we have only one second-order factor (viz. *g*), $\boldsymbol{\tau}$ is a scalar, so $\boldsymbol{\Lambda\Gamma\tau}$ is collinear with $\boldsymbol{\Lambda\Gamma}$. The correlation between these vectors is thus unity. Millsap discusses this collinearity in the single common factor case.

Note that this collinearity, which is the main focus of Jensen's method, is but *a single aspect* of the full MGCFM considered here. Note also that the crux of Spearman's hypothesis (i.e., that *g* is the main source of both within-group and between-group differences) is based on a formal definition of unbiasedness (Mellenbergh, 1989; Meredith, 1993) and is formalized in an explicit and testable manner in the MGCFM. If $\boldsymbol{\delta} = 0$ in Eq. (13) (i.e., the strong version holds) and $r = 1$, the second-order factor $\boldsymbol{\xi}$ (i.e., *g* in Spearman's hypothesis) accounts for both systematic individual differences on *all* the subtests

(Eqs. (12) and (14)) and the differences in means of the subtests between the groups (Eqs. (11) and (13); $\boldsymbol{\mu}_{nf} - \boldsymbol{\mu}_{f} = \boldsymbol{\Lambda\Gamma\tau}$).

The weak version states that observed B-W differences in means are *mainly* a positive function of variation in the tests' *g* loadings. In this case, both differences in latent means of the first ($\boldsymbol{\delta}$) and differences in latent means the second factor ($\boldsymbol{\tau}$) can account for observed B-W differences in means. We note that if $\boldsymbol{\tau}$ and $\boldsymbol{\delta}$ are both estimated, then, for reasons of identification, one element in $\boldsymbol{\delta}$ has to be fixed to 0.

3. Applications

In the subsequent sections, we use MGCFA to test Spearman's hypothesis in two previously published data sets (Lynn & Owen, 1994; Te Nijenhuis & van der Flier, 1997). In each data set, we first performed exploratory factor analyses in each group separately to establish the number of first-order common factors and the pattern of the matrix of factor loadings $\boldsymbol{\Lambda}$. Next, MGCFMs were fitted to test hypotheses concerning factorial invariance and the presence of *g*. All analyses were carried out using LISREL 8.30 (Jöreskog & Sörbom, 1999). Assuming that the data are approximately multivariate normal, we applied normal theory maximum likelihood (ML) estimation (e.g., Bollen, 1989; Long, 1983; Sörbom, 1974).[3]

To assess goodness of fit, we considered a variety of fit indices as recommended by Bollen and Long (1993). We considered the nonnormed fit index (NNFI), Akaike's information criterion (AIC), the related consistent/correct AIC (CAIC), the expected cross-validation index (ECVI), the root mean square error of approximation (RMSEA), the χ^2, the χ^2/df, and the standardized residuals. A NNFI of about 0.92 or higher is viewed as indicative of a well-fitting model. The lower the values of the AIC and the CAIC, the better the fit of the model. Compared with the AIC, the CAIC favors more parsimonious models. 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. As with the AIC and the CAIC, models with low values of the ECVI are preferable to models with large values. The RMSEA 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 of thumb, Browne and Cudeck (1993) suggest that a RMSEA of 0.05 or less is indicative of a good approximation. 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 indicator of fit that takes into account the *df* of the model.

The sequence of models that we fitted and the nesting among these models are shown in Table 1. Two models are nested if one model (say, A1 in Table 1) can be derived from the other, more constrained, model (say A2) by relaxing constraints. We first examined first-order oblique factor models. To establish the tenability of strict factorial invariance, we first fitted the model with unstructured means and without any equality constraints over the groups (model A1). Then, in model A2, we tested whether $\boldsymbol{\Lambda}$ is invariant over groups, and in model A3, we tested the invariance of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$. Note that model A3 is nested under A2 (by relaxing the equality constraint on $\boldsymbol{\Theta}$) and A1 (by relaxing the equality constraint on $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$). Finally, in model A4, we extended the model to include the means and tested whether $\boldsymbol{\nu}$ could

[3] The LISREL input files for the analyses carried out here can be found at http://users.fmg.uva.nl/cdolan/. The LISREL input files used in Dolan (2000) and Dolan and Hamaker (2001) can also be obtained at this address.

Table 1
Summary of models

Model	Specification	Nesting[a]
First-order MGCFMs		
A1: pattern invariance	$\boldsymbol{\Lambda}_{nf} \neq \boldsymbol{\Lambda}_{f}$, $\boldsymbol{\Psi}_{nf} \neq \boldsymbol{\Psi}_{f}$, $\boldsymbol{\Theta}_{nf} \neq \boldsymbol{\Theta}_{f}$, $\boldsymbol{\nu}_{nf} \neq \boldsymbol{\nu}_{f}$, $\boldsymbol{\delta} = 0$	
A2: invariance $\boldsymbol{\Lambda}$	A1, but $\boldsymbol{\Lambda}_{nf} = \boldsymbol{\Lambda}_{f}$	A1
A3: invariance $\boldsymbol{\Lambda}$, $\boldsymbol{\Theta}$	A1, but $\boldsymbol{\Lambda}_{nf} = \boldsymbol{\Lambda}_{f}$ and $\boldsymbol{\Theta}_{nf} = \boldsymbol{\Theta}_{f}$	A2
A4: invariance $\boldsymbol{\nu}$	A1, but $\boldsymbol{\Lambda}_{nf} = \boldsymbol{\Lambda}_{f}$ and $\boldsymbol{\Theta}_{nf} = \boldsymbol{\Theta}_{f}$, $\boldsymbol{\nu}_{nf} = \boldsymbol{\nu}_{f}$, $\boldsymbol{\delta} \neq 0$	A3
Second-order MGCFMs		
B1: invariance $\boldsymbol{\Psi}^*$ and $\boldsymbol{\Gamma}$	$\boldsymbol{\Lambda}_{nf} = \boldsymbol{\Lambda}_{f}$, $\boldsymbol{\Psi}^*_{nf} = \boldsymbol{\Psi}^*_{f}$, $\boldsymbol{\Theta}_{nf} = \boldsymbol{\Theta}_{f}$, $\boldsymbol{\Gamma}_{nf} = \boldsymbol{\Gamma}_{f}$, $\boldsymbol{\nu}_{nf} \neq \boldsymbol{\nu}_{f}$, $\boldsymbol{\Phi}_{nf} \neq \boldsymbol{\Phi}_{f}$, $\boldsymbol{\delta} = 0$, $\boldsymbol{\tau} = 0$	A3
B2: Strong SH[b]	B1, but $\boldsymbol{\nu}_{nf} = \boldsymbol{\nu}_{f}$, $\boldsymbol{\tau} \neq 0$	B1, B3–B5
B3:[c] Weak SH I	B2, but I $\boldsymbol{\Psi}^*_{nf} \neq \boldsymbol{\Psi}^*_{f}$, $\boldsymbol{\delta} \neq 0$	B1
B4:[c] Weak SH II	B2, but II $\boldsymbol{\Psi}^*_{nf} \neq \boldsymbol{\Psi}^*_{f}$, $\boldsymbol{\delta} \neq 0$	B1
B5:[c] Weak SH III	B2, but III $\boldsymbol{\Psi}^*_{nf} \neq \boldsymbol{\Psi}^*_{f}$, $\boldsymbol{\delta} \neq 0$	B1

[a] By nesting, we mean that one model may be viewed as a special case of a second model. For instance, model A1 is nested under A2 and model A3 is nested under A2 (and under A1 as A2 and A1 are nested). A2 is nested under A1 because the relaxation of the equality constraints $\boldsymbol{\Lambda}_{nf} = \boldsymbol{\Lambda}_{f}$ results in A1. As a second example of nesting B5, B5 is nested under B1, which is nested under A3, which is nested under A2, etc.

[b] SH: Spearman's hypothesis.

[c] Single element of diagonal covariance matrices $\boldsymbol{\Psi}^*_{nf}$ and $\boldsymbol{\Psi}^*_{f}$ free to vary over the groups; one component of $\boldsymbol{\delta}$ is estimated.

also be constrained to be equal over groups and whether observed mean differences could be accounted for by latent mean differences ($\boldsymbol{\delta}$ in Eq. (9)).

The next set of models that we fitted included the second-order common factor, *g* (B1–B5). In model B1, which did not include means, we retained the equality constraints on $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ and added the constraints concerning the equality of the variances of the first-order factors over groups ($\boldsymbol{\Psi}^*_{f} = \boldsymbol{\Psi}^*_{nf}$) and the constraints concerning the equality of the factor loadings on the second-order factor *g* over groups ($\boldsymbol{\Gamma}_{f} = \boldsymbol{\Gamma}_{nf}$). Next, we introduced the mean structure into the model and tested the strong version of Spearman's hypothesis (model B2). The subsequent models represent the weak versions of Spearman's hypothesis. In these models, a difference in means of the first-order residual is consistently accompanied by a difference in variance of that first-order residual. For further discussion of these models, the reader is referred to Dolan (2000).

3.1. Application 1: Te Nijenhuis and van der Flier (1997) data

The first data set stems from a study by Te Nijenhuis and van der Flier (1997) and comprises data of five groups: 806 native Dutch, 535 Surinamese, 126 Dutch Antilleans, 167 North Africans, and 275 Turks. All had applied for blue-collar jobs at the Dutch Railways between 1988 and 1992. The application process included a psychological examination during which the present data were collected. The native (majority) Dutch group ($n = 806$) was selected from a larger sample to achieve comparability with range of jobs and regions of residence of the members of the minority groups. Most of the subjects were male (between 82.5% and 96.5%). The nonnative Dutch groups all consisted of first-generation immigrants.

Table 2
Pattern of matrix $\mathbf{\Lambda}$ used in MGCFA, Te Nijenhuis and van der Flier (1997) data

	F&C	V	S
V	$\lambda_{1,1}$*	0	$\lambda_{1,3}$
AR	$\lambda_{2,1}$	0	0
C	$\lambda_{3,1}$	0	$\lambda_{3,3}$
NC	0	0	$\lambda_{4,3}$
3D	0	$\lambda_{5,2}$*	$\lambda_{5,3}$
TM	0	$\lambda_{6,2}$	$\lambda_{6,3}$
FM	0	$\lambda_{7,2}$	0
MM	0	0	$\lambda_{8,3}$*

In fitting models, the parameters accompanied by an asterisk are fixed to 1 to ensure the estimation of the covariance matrix of the common factors. Likewise, in the second-order factor model, one element of $\mathbf{\Gamma}$ was fixed to 1.

The data are test scores on the following eight subscales of the General Aptitude Test Battery (GATB; U.S. Department of Labor, 1970): Vocabulary (V), Arithmetic Reasoning (AR), Computation (C), Name Comparison (NC), Three-Dimensional Space (3D), Tool Matching (TM), Form Matching (FM), and Mark Making (MM).The reader is referred to Te Nijenhuis and van der Flier (1997) and the references therein for further details and summary statistics.

Te Nijenhuis and van der Flier (1997) tested Spearman's hypothesis using Jensen's test in four separate analyses in which they compared the reference group with each nonreference group. This resulted in correlations of .42 (Surinamese), .71 (Antilleans), .87 (North Africans), and .87 (Turks) using the estimated *g* loadings of the Dutch group. Based on these results, Te Nijenhuis and van der Flier concluded that *g* is the predominant factor determining the size of the differences between the reference group and each nonreference group (p. 675). Below, we test all aspects of Spearman's hypothesis simultaneously by MGCFA to establish whether the conclusion of Te Nijenhuis and van der Flier can be confirmed.

3.1.1. Results of exploratory factor analyses

The pattern of the matrix of factor loadings $\mathbf{\Lambda}$ is established by carrying out exploratory factor analysis followed by promax (oblique) rotation in each group separately. The exploratory factor analyses were carried out using the program LISREL (Jöreskog & Sörbom, 1999). In view of results in Hunter (1983), we expected three correlated first-order factors: a fluid and crystallized abilities factor (F&C), a visual factor (V), and a speed factor (S). On the basis of the results of the exploratory factor analyses, we specified the matrix $\mathbf{\Lambda}$ as shown in Table 2. The models that we considered subsequently included this matrix of factor loadings. A path diagram of this matrix is displayed in Fig. 1.

3.1.2. Multigroup confirmatory factor models

The results of the fitted models are displayed in Table 3. We started with models that do not include a restrictive model for the means.[4] The χ^2 of model A1, in which $\mathbf{\Lambda}$, $\mathbf{\Theta}$, $\mathbf{\Psi}$, and $\boldsymbol{\nu}$ were free to vary over groups, equaled 217.3 ($df = 65$). The fit is not very good, but it should be remembered that all five groups

[4] In these models, we estimate an independent parameter for each observed means. The model for the means is therefore saturated. Although it might seem that one could just as well leave the means out all together, this cannot be recommended. The presence of the means does affect the values of AIC, CAIC, and ECVI. To ensure comparability of these fit indices, the means have to be included at all stages of model fitting (Wicherts & Dolan, 2004).

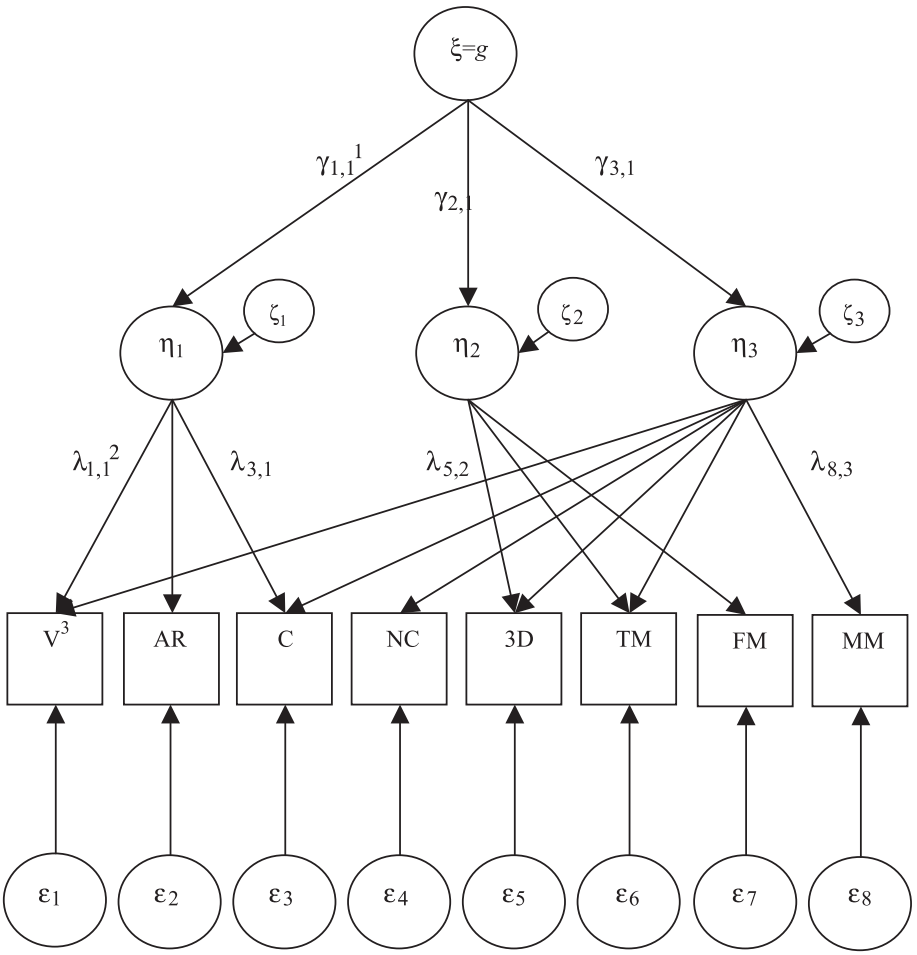

Fig. 1. Path diagram of factor model for the Te Nijenhuis and van der Flier (1997) data. [1]For example, $\gamma_{1,1}$ is an element of $\mathbf{\Gamma}$, and represents the regression of η_1 on g. In fitting the models, $\gamma_{1,1}$ was fixed to 1.0 to identify the variance of g, the second order factor. [2]For example, $\lambda_{1,1}$ is an element in $\mathbf{\Lambda}$, and represents the regression coefficient in the regression of v on η_1. In fitting the models, $\lambda_{1,1}$, $\lambda_{5,2}$, and $\lambda_{8,3}$ were fixed to 1.0. to identify the variances of residuals ζ_1, ζ_2, and, ζ_3. To avoid clutter, only four factor loadings as shown in this figure. [3]V=Vocabulary, AR=Arithmetic Reasoning, C=Computation, NC=Name Comparison, 3D=Three Dimensional Space, TM=Tool Matching, FM=Form Matching, and MM=Mark Making.

were analyzed simultaneously and that the total sample size is large ($n = 1909$). In addition, we are especially interested in the changes of fit as we add substantive constraints associated with strict factorial invariance. Therefore, model A1 is accepted as reasonable point of departure.

In models A2–A4, we subsequently introduced the constraints that comprise the hypothesis of strict factorial invariance. First, in A2, we tested the equality of $\mathbf{\Lambda}$ over groups. This resulted in a $\chi^2(101)$ of 268.7. Compared with A1, A2 had better values of χ^2/df, AIC, and CAIC, and we therefore concluded that constraining $\mathbf{\Lambda}$ to be equal over groups is acceptable. Next, in A3, the testing of the equality of $\mathbf{\Theta}$ over groups resulted in a $\chi^2(133)$ of 406.5. According to the CAIC, A3 must be preferred to A2. However, the χ^2/df and the AIC indicated that A3 fits less well than A2, and we therefore concluded

Table 3
Goodness-of-fit indices of the fitted models on the Te Nijenhuis and van der Flier (1997) data

	df	χ^2	χ^2/df	RMSEA	ECVI	AIC	CAIC	NNFI
A1	65	217.3	3.34	0.077	0.27	521.9	1537.8	0.94
Strict factorial invariance								
A2	101	268.7	2.66	0.065	0.26	500.3	1280.3	0.96
A3	133	406.5	3.06	0.074	0.31	582.4	1152.7	0.95
A4	153	1167.1	7.63	0.130	0.70	1326.7	1765.8	0.84
Strong factorial invariance								
A4-**Θ**	121	1053.4	8.71	0.140	0.65	1233.3	1882.2	0.82
B1-**Θ**	121	323.9	2.67	0.065	0.27	511.9	1160.8	0.96
B2-**Θ**	149	1241.9	8.33	0.140	0.73	1392.0	1857.4	0.82
B3-**Θ**[a]	141	1196.2	8.48	0.140	0.72	1372.0	1889.8	0.82
B4-**Θ**[b]	141	1193.4	8.46	0.140	0.71	1357.8	1875.6	0.82
B5-**Θ**[c]	141	1128.3	8.00	0.130	0.65	1238.7	1756.5	0.83

[a] Testing weak version of Spearman's hypothesis: *g* and F&C factor account for observed nr−r differences.
[b] Testing weak version of Spearman's hypothesis: *g* and V factor account for observed nr−r differences.
[c] Testing weak version of Spearman's hypothesis: *g* and S factor account for observed nr−r differences.

that the equality of **Θ** over groups could not be established unambiguously. Finally, by fitting model A4, we tested whether the means could be included to the model and whether they could be constrained to be equal over groups. The resulted $\chi^2(153)$ of 1167.1 and the unacceptable value of the NNFI (0.84) clearly indicated rejection of A4 and therefore rejection of strict factorial invariance. This conclusion seemed also inescapable when taking into consideration the chances in fit going from A1 to A4, which are illustrated by the increase of the χ^2/df (from 3.34 in A1 to 7.63 in A4), the AIC (from 521.9 in A1 to 1326.7 in A4), and the CAIC (from 1537.8 in A1 to 1765.8 in A4). Furthermore, the dramatic increase of the standardized residuals of the covariance matrices, which ranged from −5.44 to 6.95 in A1 and from −15.68 to 21.50 in A4, clearly indicated the failure of strict factorial invariance.

Because we wanted to establish that the failure of A4 was not simply due to the equality constraint on **Θ**, we repeated the analyses without this constraint and tested the hypothesis of strong factorial invariance in A4-**Θ** (i.e., the covariance matrices of the residuals were free to vary over the groups). As the fit indices of A4-**Θ** [$\chi^2(121) = 1053.4$, $\chi^2/df = 8.71$, AIC = 1233.3, CAIC = 1882.2, and NNFI = 0.82] still indicated a serious deterioration of fit compared with those of A2, we also rejected strong factorial invariance and concluded that some or all of the tests are biased (Meredith, 1993; Millsap, 1997; Muthén & Lehman, 1985; Oort, 1996).

Because bias invalidates group comparisons, we provide the results of the remaining models for reasons of comparison only. In these models, we included the second-order factor *g* and tested the strong and the weak versions of Spearman's hypothesis. We did not impose the restriction concerning the equality of the error covariance matrices. The χ^2 of B1-**Θ**, which did not include means, equaled 323.9 ($df = 121$). Although the strong version of Spearman's hypothesis is not expected to fit (Te Nijenhuis & van der Flier, 1997), it was tested in B2-**Θ** [$\chi^2(149) = 1241.9$]. The weak versions of Spearman's hypothesis were tested in B3-**Θ**, B4-**Θ**, and B5-**Θ**. From the results, it appeared that the observed B-W differences could not be accounted for by *g* and the F&C factor [$\chi^2(141) = 1196.2$], by *g* and the V factor

[$\chi^2(141) = 1193.4$], or by g and the S factor [$\chi^2(141) = 1128.3$]. We conclude that all versions of Spearman's hypothesis must be rejected. Of course, this conclusion already followed from the rejection of A4-$\boldsymbol{\Theta}$: the second-order models are restrictive versions of that model (i.e., nested under A4-$\boldsymbol{\Theta}$).

3.1.3. Conclusion

From these results, the GATB clearly is not factorially invariant. Although group comparisons are difficult to interpret in case of bias, we tested the weak and strong versions of Spearman's hypothesis and found no version of the hypothesis to be tenable. Hence, the statement that g is the predominant factor determining the size of the differences between the reference group and each nonreference group (Te Nijenhuis and van der Flier, 1997, p. 675) cannot be confirmed.

3.2. Application 2: Lynn and Owen (1994) data

The second data set that we analyzed is published in Lynn and Owen (1994).[5] The samples consist of an equal number of male and female adolescents aged 15–16. The sample sizes are 1056 Whites, 1093 Blacks, and 1063 Indians. The subjects attended high schools and were tested in 1985 or 1986 by school psychologists in the school setting. The Blacks attended schools in Pretoria-Witwatersrand-Vereeniging (PWV) area (3 schools) and representative schools in Black areas in KwaZulu adjacent to urban areas in Natal (25 schools). The Whites attended schools in the PWV area (20) and in the Cape Peninsula (10). The Indians attended schools in and around Durban (30 schools). Judging by these locations, all subjects apparently attended high schools in urban or suburban areas.

The data include scores on the following 10 subscales of the Junior Aptitude Test (JAT): Classification (CL), Reasoning (R), Number (N), Synonyms (S), Comparisons (CO), Spatial 2D (S2D), Spatial 3D (S3D), Memory 1 (M1), Memory 2 (M2), and Mechanical Insight (Mi). The reader is referred to Lynn and Owen (1994) and the references therein for further details and the summary statistics.

Lynn and Owen (1994) reported a significant correlation of .624 for the B-W differences using the corrected g factor loadings in the Black sample and concluded that this correlation confirms the weak version of Spearman's hypothesis (p. 31). Below, we test whether this conclusion can be corroborated by MGCFA.

3.2.1. Results: exploratory factor analyses

We first established the pattern of the matrix of factor loadings $\boldsymbol{\Lambda}$ by carrying out exploratory factor analysis followed by promax (oblique) rotation in each group separately. The results of these analyses indicated that a three-factor model provided a reasonable fit. The loadings in the White and the Indian groups were comparable and interpretable. However, in the Black group, a different and less interpretable pattern was found. According to Lynn and Owen (1994), this can be ascribed to the Blacks' lack of proficiency in the language of the test (see also Jensen, 1998, p. 388). This raises the question whether a group comparison makes any sense at all. Regardless of statistical modeling, it does not seem reasonable to expect groups to be comparable with respect to their test scores if one group lacks

[5] Lynn and Owen's table of summary statistics contained some typographical errors. The missing correlation between Comparisons and Spatial 2D in the Indian group was estimated by a multiple regression analysis. For this correlation, a value of .29 was found. Further, we assumed that the correlations of Memory 1 (paragraph) in the White group were shifted one place to the right.

proficiency in the language of the test. In view of this, we performed our analyses with and without the Black group. We used the pattern of factor loadings that is displayed in Table 4. A path diagram of this matrix is displayed in Fig. 2.

3.2.2. Multigroup confirmatory factor models

The results of the fitted models are displayed in Table 5. We started with the analyses on all three groups. We use the suffix 3g to indicate that the models are three group models. The χ^2 of A1-3g, in which the model for the means is saturated (see footnote 3) and in which $\mathbf{\Lambda}$, $\mathbf{\Theta}$, $\mathbf{\Psi}$, and $\boldsymbol{\nu}$ were free to vary over groups, equaled 381.5 with 84 *df*. Although measurement invariance is not supposed to hold due to the Black's lack of command of the language of the test, we tested the constraints that are associated with the hypothesis of strict factorial invariance. First, in A2-3g, the constraint that $\mathbf{\Lambda}$ is equal over groups resulted in a $\chi^2(106)$ of 719.9. Next, in A3-3g, constraining both $\mathbf{\Lambda}$ and $\mathbf{\Theta}$ to be equal over groups resulted in a $\chi^2(126)$ of 1723.5, and finally, in A4-3g, constraining $\mathbf{\Lambda}$, $\mathbf{\Theta}$, and $\boldsymbol{\nu}$ to be equal over groups resulted in a $\chi^2(df = 140)$ of 2801.2. In a comparison of the fit indices of A1-3g and A4-3g (e.g., χ^2/df: 4.54 vs. 20.01, AIC: 604.5 vs. 2925.9, and CAIC: 1389.8 vs. 3315.0 in A4-3g), strict factorial invariance clearly must be rejected. As Lynn and Owen (1994) already reported that the JAT is probably biased in the Black group, we excluded this group from the analyses and proceeded to analyze the data of the Whites and the Indians.

Again, we started by fitting models that do not include a restrictive model for the means. The χ^2 of A1, in which $\mathbf{\Lambda}$, $\mathbf{\Theta}$, $\mathbf{\Psi}$, and $\boldsymbol{\nu}$ were free to vary over groups, equaled 286.5 with 56 *df*. The fit is still not very good, but we accepted it as point of departure. Next, in A2, we introduced the constraint concerning the equality of $\mathbf{\Lambda}$ over groups. This resulted in a $\chi^2(67)$ of 463.3. As all fit indices indicated that A2 fit worse than A1, we conclude that constraining $\mathbf{\Lambda}$ over groups is not tenable. In A3, we tested the equality of $\mathbf{\Theta}$ over groups. This resulted in a $\chi^2(77)$ of 881.5. Again, all fit indices suggested that this constraint is not acceptable. Finally, the inclusion of the means into the model (A4) resulted in a $\chi^2(84)$ of 1352.5 and clearly indicated the failure of model A4. Comparing models A1 and A4, the failure of A4 is also evident in the χ^2/df (5.12 vs. 16.10), the AIC (431.7 vs. 1404.6), and the CAIC (924.4 vs. 1710.9). Furthermore, the standardized residuals (ranging from -5.83 to 8.23 in A1 and from -14.80 to 10.24 in A4) support the conclusion that the hypothesis of strict factorial invariance is untenable.

Table 4
Pattern of matrix $\mathbf{\Lambda}$ used in MGCFA, Lynn and Owen (1994) data

	1	2	3
CL	$\lambda^*_{1,1}$	0	$\lambda^*_{1,3}$
R	0	$\lambda^*_{2,2}$	0
N	0	$\lambda_{3,2}$	0
S	0	$\lambda_{4,2}$	0
CO	0	$\lambda_{5,2}$	$\lambda_{5,3}$
S2D	$\lambda_{6,1}$	0	0
S3D	$\lambda_{7,1}$	0	0
M1	$\lambda_{8,1}$	0	$\lambda_{8,3}$
M2	0	0	$\lambda_{9,3}$
Mi	$\lambda_{10,1}$	$\lambda_{10,2}$	0

In fitting all models, the parameters accompanied by an asterisk are fixed to 1 to ensure the estimation of the covariance matrix of the common factors. Likewise, in the second-order factor model, one element of $\mathbf{\Gamma}$ was fixed to 1.

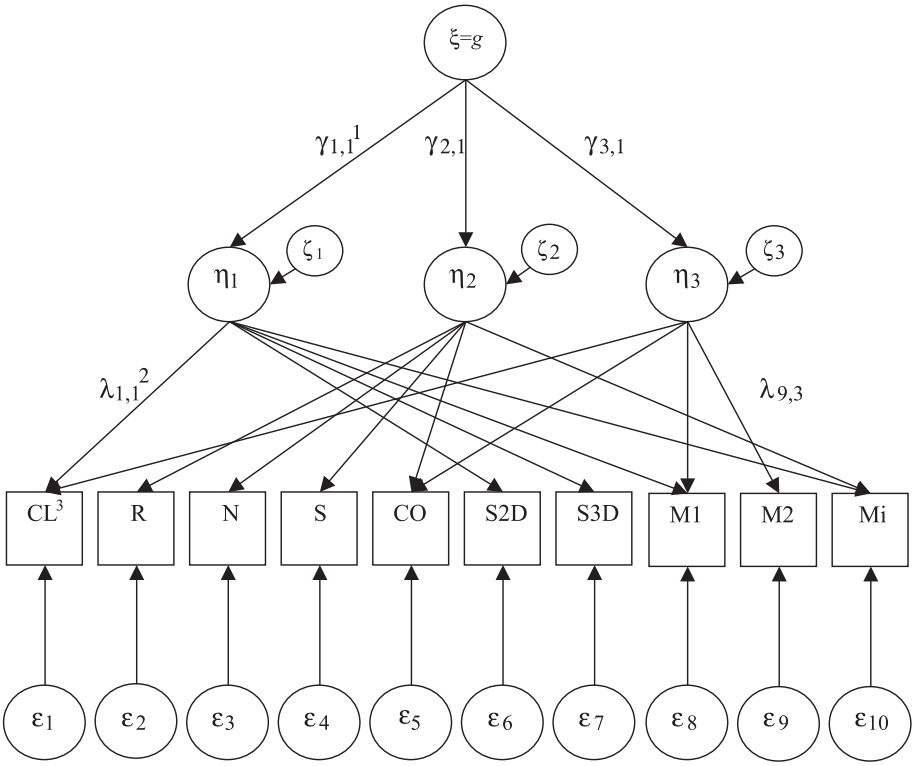

Fig. 2. Path diagram of factor model for Lynn and Owen (1994) data. [1]For example, $\gamma_{1,1}$ is element of $\mathbf{\Gamma}$ and represents the regression of η_1 on ξ. In fitting the models, $\gamma_{1,1}$ was fixed to 1. [2]For example, $\lambda_{1,1}$ is an element of $\mathbf{\Lambda}$ and represents the regression of CL on η_1. In fitting the models, $\lambda_{1,1}$, $\lambda_{2,2}$, and $\lambda_{1,3}$ were fixed to 1. For the sake of clearness, only two names of $\mathbf{\Lambda}$ loadings are displayed in the figure. [3]CL = Classification, R = Reasoning, N = Number, S = Synonyms, CO = Comparisons, S2D = Spatial 2D, S3D = Spatial 3D, M1 = Memory 1, M2 = Memory 2, and Mi = Mechanical Insight.

Because we wanted to be sure that the failure of A4 (strict factorial invariance) was not simply due to the equality of $\mathbf{\Theta}$ over groups, we repeated previous analyses without this constraint (A4-$\mathbf{\Theta}$ to B5-$\mathbf{\Theta}$). First, in A4-$\mathbf{\Theta}$, the inclusion of the means resulted in a $\chi^2(74)$ of 1016.0. The goodness-of-fit indices show some improvement; however, compared with A2, the hypothesis of strong factorial invariance is still untenable. Although testing Spearman's hypothesis is complicated by the presence of bias, we introduced the second-order factor *g* for the sake of comparison. In B2-$\mathbf{\Theta}$, the testing of the strong version of Spearman's hypothesis resulted in a $\chi^2(81)$ of 1298.4. In the next subset of models (B3-$\mathbf{\Theta}$ to B5-$\mathbf{\Theta}$), the weak versions of Spearman's hypothesis were tested. The χ^2 goodness-of-fit indices of these models were 1258.4 ($df = 79$) for B3-$\mathbf{\Theta}$ (*g* + first-order factor 1), 1186.3 ($df = 79$) for B4-$\mathbf{\Theta}$ (*g* + first-order factor 2), and 1264.3 ($df = 79$) for B5-$\mathbf{\Theta}$ (*g* + first-order factor 3). We therefore conclude that all versions of Spearman's hypothesis must be rejected.

3.2.3. Conclusion

The present study supports the previous result that the JAT is not psychometrically satisfactory for Black pupils (Jensen, 1998; Lynn & Owen, 1994; Owen, 1989). Although Lynn and Owen (1994)

Table 5
Goodness-of-fit indices of the fitted models on the Lynn and Owen (1994) data, analyses done on all three groups and on the White and Indian groups only

	df	χ^2	χ^2/df	RMSEA	ECVI	AIC	CAIC	NNFI
Analyses on all three groups, strict factorial invariance								
A1-3g	84	381.5	4.54	0.058	0.19	604.5	1389.8	0.95
A2-3g	106	719.9	6.79	0.073	0.28	895.6	1525.2	0.92
A3-3g	126	1723.5	13.68	0.110	0.58	1851.5	2339.6	0.83
A4-3g	140	2801.2	20.01	0.130	0.91	2925.9	3315.0	0.75
Analyses excluding the Black group, strict factorial invariance								
A1	56	286.5	5.12	0.062	0.20	431.7	924.4	0.95
A2	67	463.3	6.92	0.075	0.28	587.0	1006.6	0.93
A3	77	881.5	11.45	0.098	0.46	964.7	1317.6	0.88
A4	84	1352.5	16.10	0.120	0.66	1404.6	1710.9	0.83
Analyses excluding the Black group, strong factorial invariance								
A4-Θ	74	1016.0	13.73	0.110	0.55	1160.4	1533.3	0.86
B1-Θ	72	563.6	7.83	0.081	0.32	686.9	1073.1	0.92
B2-Θ	81	1298.4	16.03	0.120	0.69	1458.2	1784.5	0.83
B3-Θ^a	79	1258.4	15.93	0.120	0.67	1423.4	1736.0	0.83
B4-Θ^b	79	1186.3	15.02	0.120	0.62	1307.6	1647.2	0.84
B5-Θ^c	79	1264.3	16.00	0.120	0.66	1391.8	1731.4	0.83

3g means that all three groups are included into the analyses.
[a] Testing weak version of Spearman's hypothesis: *g* and the first first-order factor account for observed nf−f differences.
[b] Testing weak version of Spearman's hypothesis: *g* and the second first-order factor account for observed nf−f differences.
[c] Testing weak version of Spearman's hypothesis: *g* and the third first-order factor account for observed nf−f differences.

conceded the biasedness of the JAT, they concluded that the correlation between the B-W differences and the corrected *g* loadings in the Black sample confirmed Spearman's hypothesis (p. 31). Apparently, they failed to realize that group comparisons of observed score are problematic when the tests are biased and factorial invariance is not tenable. Lynn and Owen emphasized that they focused on determining whether B-W-Indian mean test differences were correlated with the JAT's *g* factor loadings and that they were not directly concerned with the suitability of the JAT for use as a common test battery for Blacks and Whites from the point of view of test and item bias (p. 31). In light of this remark, we think it is useful to stress again the fact that factorial invariance is a necessary condition for testing Spearman's hypothesis. Because the hypothesis of measurement invariance clearly failed in the Black group, no support can be inferred in support of Spearman's hypothesis.

4. Discussion

The objective of the present article was to test Spearman's hypotheses in two previously published data sets using MGCFA. In contrast of previous applications of this method to the issue of Spearman's hypothesis, which were focused on data collected in the United States (Dolan, 2000; Dolan & Hamaker,

2001), the data sets that we analyzed in this article were collected in Holland (Te Nijenhuis & van der Flier, 1997) and in South Africa (Lynn & Owen, 1994).

Two major differences appeared from the analyses of the U.S. B-W data sets and the analyses of the present analyses. The first concerns the hypothesis of factorial invariance. In the U.S. data sets, the hypothesis of factorial invariance appeared to be acceptable, but in the Dutch and South African data sets, the hypothesis of factorial invariance was clearly rejected. This finding is consistent with the consensus that bias is absent in IQ subtest scores of U.S. Blacks and Whites (see Jencks, 1998; Jensen, 1998). The second difference concerns the role of *g*. The analyses of the U.S. data indicated that it is difficult to distinguish between models that do and that do not incorporate *g*. However, the present results indicate that factorial invariance is not remotely tenable (see models A2–A4 in Tables 3 and 5). This makes any conclusion concerning the role of *g* impossible to evaluate. Therefore, the present failure of Spearman's hypothesis has a bearing neither on the issue of the existence of *g* in general nor on the role of *g* in the differences in IQ test scores among other groups.

The reason that our conclusions concerning the present data are diametrically opposed to that of both Lynn and Owen (1994) and Te Nijenhuis and van der Flier (1997) is possibly twofold. One reason lies in the fact that both Lynn and Owen and Te Nijenhuis and van der Flier pass over the fact that bias at the level of the subtests is not strictly compatible with Spearman's hypothesis. In both articles, the authors state that the tests may be biased due to differences between the groups in language proficiency (Lynn & Owen, 1994, p. 31; Te Nijenhuis & van der Flier, 1997, p. 685). Nonetheless, they draw the conclusion that Spearman's hypothesis holds. For instance, Lynn and Owen state: ''The suitability of the JAT for use as a common test battery for Blacks and Whites from the point of view of test and item bias (. . .) does not directly concern us here. What is important is whether B-W-Indian mean test differences are correlated with JAT's *g* loadings'' (p. 31). These authors miss the intimate connection among unbiasedness, factorial invariance, and Spearman's hypothesis (see above and see also Lubke et al., 2003).

A second reason for the discrepancy between the present results and those of Te Nijenhuis and van der Flier and of Lynn and Owen may lie in the fact that the central test statistic in Jensen's method of correlated vectors (viz. the correlation between factor loadings and standardized differences in means) is insensitive to violations of the underlying model and may therefore be difficult to interpret with any confidence (Dolan & Lubke, 2001; Lubke et al., 2001). Any correlation ranging from .3 to .9 can be, and in fact is, interpreted in support of the central role of *g* in determining group differences. It is therefore disconcerting that Jensen's method is gaining widespread acceptance (Colom et al., 2000, 2001; Flynn, 2000; Helms-Lorenz et al., 2003; Must et al., 2003; Nyborg & Jensen, 2000; Rushton, 2001; Te Nijenhuis et al., 2000) and that results obtained with this method are taken at face value (e.g., Roth, Bevier, Bobko, Switzer, & Tyler, 2001, p. 305). In view of its insensitivity, this method cannot trusted to establish the role of *g* (or of any other latent variable; e.g., see Helms-Lorenz et al., 2003) in group differences. The alternative method of MGCFA has the advantage that it allows one to represent (see Eqs. (11)–(14)) Spearman's hypothesis accurately, it allows one to test the model and its assumptions rigorously, and it allows one to compare Spearman's hypothesis with competing models (see Dolan, 2000; Dolan & Hamaker, 2001).

We conclude this paper with two remarks. First, we emphasize that the failure of Spearman's hypothesis in the present data has no implications for other group comparisons. It is not possible, nor is it desirable, to generalize a negative result. Second, and more generally, we emphasize that

our misgivings concerning the application of Jensen's method to group differences in IQ subtest scores are limited to the method. We consider the parsimonious hypothesis that differences between groups, racial or otherwise, are due to the latent construct "general intelligence" as good as any other hypothesis.

Acknowledgements

The preparation of this article was supported by a grant from the Netherlands Organization for Scientific Research (NWO). The research of Conor Dolan was made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

References

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

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

Bollen, K. A., & Long, J. S. (1993). Introduction. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 1–9). Newbury Park, CA: Sage.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

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.

Colom, R., Juan-Espinosa, M., Abad, F., & García, L. F. (2000). Negligible sex differences in general intelligence. *Intelligence*, *28*, 57–68.

Colom, R., Juan-Espinosa, M., & García, L. F. (2001). The secular increase in test scores is a "Jensen effect". *Personality and Individual Differences*, *30*, 553–559.

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. 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 corrected factors. In F. Columbus (Ed.), *Advances of Psychological Research*, *vol. 6* (pp. 31–59). Huntington: Nova Science.

Dolan, C. V., & Lubke, G. H. (2001). Viewing Spearman's hypothesis from the perspective of multi-group PCA: A comment on Schönemann's criticism. *Intelligence*, *29*, 231–245.

Flynn, J. R. (2000). IQ gains, WISC subtests and fluid *g*: *g* theory and the relevance of Spearman's hypothesis to race. In G. R. Bock, J. A. Goode, & K. Webb (Eds.), *The nature of intelligence: Novartis Foundation Symposium 233*. Chichester, UK: Wiley.

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

Helms-Lorenz, M., van der Vijver, F. J. R., & Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance and Spearman's hypothesis: *g* or *c*? *Intelligence*, *31*, 9–29.

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

Hunter, J. E. (1983). The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of the general factors over specific factors in the prediction of job performance for the U.S. Employment Service (*USES Test Research Rep. No. 44*). Washington, DC: U.S. Employment Service, U.S. Department of Labor.

Jencks, C. (1998). Racial bias in testing. In C. Jencks, & M. Phillips (Eds.), *The Black–White test score gap* (pp. 55–85). Washington: Brookings Institution 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. (1998). *The g factor. The science of mental ability.* Praeger: Westport.
Jensen, A. R., & Reynolds, C. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences*, *3*, 423–438.
Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426.
Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.
Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software International.
Little, T. D. (1997). Mean and covariance structures (MACS) analysis of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53–76.
Long, J. S. (1983). *Confirmatory factor analysis*. Beverly Hills: Sage.
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). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence*, *31*(6), 543–566.
Lynn, R., & Owen, K. (1994). Spearman's hypothesis and test scores differences between Whites, Indians, and Blacks in South Africa. *Journal of General Psychology*, *121*, 27–36.
Marsh, H. W., & Grayson, D. A. (1990). Public/Catholic differences in the high school and beyond data: A multi-group structural equation modeling approach to testing mean differences. *Journal of Educational Statistics*, *15*, 199–235.
Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.
Millsap, R. E. (1997). The investigation of Spearman's hypothesis and the failure to understand factor analysis. *Cahiers de Psychologie Cognitive*, *16*, 750–757.
Millsap, R. E., & Everson, H. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research*, *26*, 479–497.
Must, O., Must, A., & Raudik, V. (2003). The secular rise in IQs: In Estonia the Flynn effect is not a Jensen effect. *Intelligence*, *31*, 461–471.
Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics*, *10*, 133–142.
Naglieri, J. A., & Jensen, A. R. (1987). Comparison and Black–White differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence*, *11*, 21–43.
Neale, M. C. (1997). *Mx: Statistical modeling*. Richmond: Medical College of Virginia.
Nyborg, H., & 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. (1992). Using restricted factor analysis to detect item bias. *Methodika*, *6*, 150–166.
Oort, F. J. (1996). *Using restricted factor analysis in test construction*. PhD thesis, Psychology Faculty, University of Amsterdam.
Owen, K. (1989). *Test and item bias: The suitability of the Junior Aptitude Tests as a common test battery for White, Indian and Black pupils in standard 7*. Pretoria: Human Sciences Research Council.
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.
Roth, P. L., Bevier, C. A., Bobko, P., Switzer III, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*, 297–330.
Rushton, J. P. (2001). Black–White differences on the *g*-factor in South-Africa: A ''Jensen Effect'' on the Wechsler Intelligence Scale for Children-Revised. *Personality and Individual Differences*, *31*, 1227–1232.
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.
Te Nijenhuis, J., Evers, A., & Mur, J. P. (2000). Validity of differential aptitude test for the assessment of immigrant children. *Educational Psychology*, *20*, 99–115.

Te Nijenhuis, J., & 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.

U.S. Department of Labor (1970). *Manual for the General Aptitude Test Battery.* Washington, DC.

Wicherts, J. M., & Dolan, C. V. (2004). A cautionary note on the use of information fit indices in covariance structure modeling with means. *Structural Equation Modeling* (in press).

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Newbury Park, CA: Sage.