Intelligence 37 (2009) 396–404

Contents lists available at ScienceDirect

Intelligence

The power to detect sex differences in IQ test scores using Multi-Group Covariance and Means Structure Analyses[☆]

Dylan Molenaar [*], Conor V. Dolan, Jelte M. Wicherts

Psychological Methods, Department of Psychology, University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 11 March 2008
Received in revised form 7 January 2009
Accepted 17 March 2009
Available online 19 April 2009

Keywords:
Gender differences
Multi-Group Confirmatory Factor Analysis
Multi-Group Covariance and Mean Structure Analysis
Power
Likelihood ratio test
Measurement invariance

ABSTRACT

Research into sex differences in general intelligence, *g*, has resulted in two opposite views. In the first view, a *g*-difference is nonexistent, while in the second view, *g* is associated with a male advantage. Past research using Multi-Group Covariance and Mean Structure Analysis (MG-CMSA) found no sex difference in g. This failure raised the question whether the *g*-difference is truly absent or whether MG-CMSA lacked statistical power to detect it. The present study used the likelihood ratio test to investigate the power to detect a *g*-difference in the WAIS-III factor structure with MG-CMSA. Various situations were examined including those reported in the literature. Results showed that power varies greatly among different scenarios. The scenarios based on previous results were associated with power coefficients of about 0.5–0.6. Implications of these findings are discussed in the light of research into sex differences in IQ.

The aim of this paper is to study the statistical power in Multi-Group Confirmatory Factor Analysis or Multi-Group Covariance and Mean Structure Analyses (MG-CMSA; Jöreskog, 1971; Sörbom, 1974; Muthen, 1989)[1] to identify the nature of mean differences in multivariate intelligence test scores. This undertaking is motivated by several recent papers, which addressed the issue of sex differences in general intelligence (*g*). The accepted view is that sex differences are located in certain first-order common factors, and not in *g* (Loehlin, 2000, p. 177; Mackintosh, 1998, p. 189; Jensen, 1998, p. 541; Brody, 1992, p. 323). These first-order common factors refer to specific abilities, such as Verbal Ability, Spatial Ability, and Perceptual Speed. Contrary to this view, Lynn (1994, 1999) proposed that there is a small, but appreciable, sex difference in *g*, which emerges in adolescence, and persists throughout adulthood. Both Lynn and Nyborg (2003) presented results in support of this proposition in the form of analysis of sum scores and point bi-serial correlations, respectively. In contrast, Colom, Garcia, Juan-Espinosa, and Abad (2002) and Colom, Juan-Espinosa, Abad, and Garcia (2000), using the method of correlated vectors (Jensen, 1998), failed to find a sex difference in *g*.

Dolan et al. (2006), van der Sluis et al. (2006), and van der Sluis et al. (2008) used MG-CMSA to study sex differences in *g* on the Wechsler tests. In analyzing Spanish WAIS-III data in adults from 18 to 34, Dolan et al. found — in addition to four intercept differences (see the following data for the exact meaning in terms of the confirmatory factor model) — the first-order common factors Perceptual Organization and Working Memory to have mean differences favoring males. No *g*-difference was found in this study. Van der Sluis et al. analyzed Dutch WAIS-III test scores, and found the common factors Perceptual Organization and Working Memory to have mean differences favoring males, and the factor Perceptual Speed to have a mean difference favoring females. However,

[☆] The research by Dylan Molenaar was made possible by a grant from the Netherlands Organization for Scientific Research (NWO). We thank the reviewers, Wendy Johnson and Paul Irwing, for their constructive comments on previous drafts of this manuscript.

[*] Corresponding author. Psychological Methods, Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB, Amsterdam, The Netherlands. Tel.: +31 205256584.
E-mail address: D.Molenaar@uva.nl (D. Molenaar).

[1] Note that this approach is also denoted by Mean And Covariance Structure analysis (MACS; Little, 1997).

0160-2896/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2009.03.007

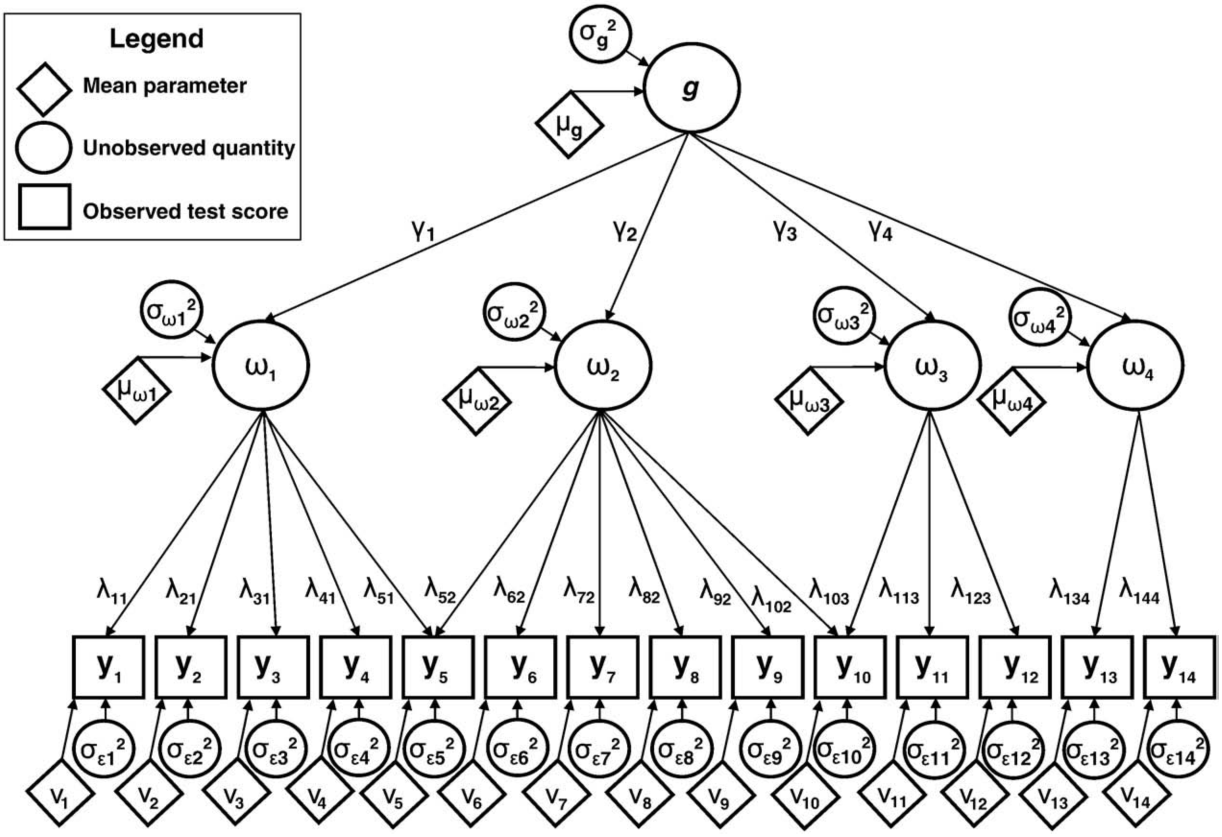

Fig. 1. The WAIS-III factor model as used in the present study. y_k are the observed scores on the kth subtest, $\sigma^2_{\varepsilon k}$ denotes the variance of the residuals of the kth subtest, ν_k is the intercept of the kth subtest, ω_m are the scores on the mth first-order common factor, $\mu_{\omega m}$ denotes the mean of ω_m, $\sigma^2_{\omega m}$ is the residual variance of ω_m, λ_{km} is a factor loading i.e., the regression coefficient in the regression of y_k on ω_m, g are the scores on the general intelligence factor, μ_g denotes the mean of g, σ^2_g is the variance of g, and γ_m denotes a second-order factor loading i.e., the regression coefficient in the regression of ω_m on g.

no *g*-difference was found. Finally, van der Sluis et al. (2008) analyzed WISC-R test scores in Belgium and The Netherlands, and found — besides four intercept differences — no sex differences on the first-order common factors and the *g*-factor.

Dolan et al. (2006), van der Sluis et al. (2006), and van der Sluis et al. (2008) concluded on the basis of their modeling results that males and females do not differ with respect to *g*. The source of discrepancies between these conclusions and those of Lynn (1999) and Nyborg (2003) may lie in the different methodologies used. The former authors employed MG-CMSA, which we consider an appropriate method to investigate group differences in intelligence (e.g., Gustafsson, 1992; Horn, 1997; Millsap, 1997). The latter authors used procedures that suffer from several drawbacks (e.g., Dolan et al., 2006). Thus, it is possible that the demonstrated *g*-differences are artifacts produced by applying a questionable statistical method. However, an alternative explanation of the failure to find the *g*-difference with MG-CMSA is that the method lacks statistical power to detect it. Previous research has demonstrated that the power to detect mean differences in a first-order factor is large (Hancock, Lawrence, & Nevitt, 2000; Kaplan & George, 1995). However, these findings cannot be generalized to differences in *g*, mainly because the models at hand include several first-order factors and the second-order factor, *g*. In addition, previous power analyses did not address the specific details (i.e., parameter values) of research into sex differences in intelligence. Therefore, it remains unclear what the power is to detect a potential *g* effect given that this difference is relatively small.

Fortunately, power analyses in MG-CMSA are well developed for Normal Theory Maximum Likelihood Estimation (Satorra & Saris, 1985; Saris & Satorra, 1993; Dolan, van der Sluis, & Grasman, 2005), and relatively simple to conduct. In the present study we investigate the power to detect sex differences in *g* using MG-CMSA given the results as found in Dolan et al. (2006) and van der Sluis et al. (2006). We do not take into account the results of van der Sluis et al. (2008), as this study involved children, and the *g* effect, as hypothesized by Lynn (1999), is supposed to emerge in adolescence. In the next sections, MG-CMSA is presented and power calculations in these models are discussed. Next, the design of the study is elucidated and results are presented. We end with a general discussion of the implications of our findings. We concentrated mainly on the multi-group model, but we did some additional power calculations in the MIMIC model, i.e., a single group model in which the grouping variable (sex) is treated as a fixed regressor (see Muthen, 1989).

1. MG-CMSA

Fig. 1 depicts the general second-order WAIS-III factor model as used in this study. A matrix algebraic representation

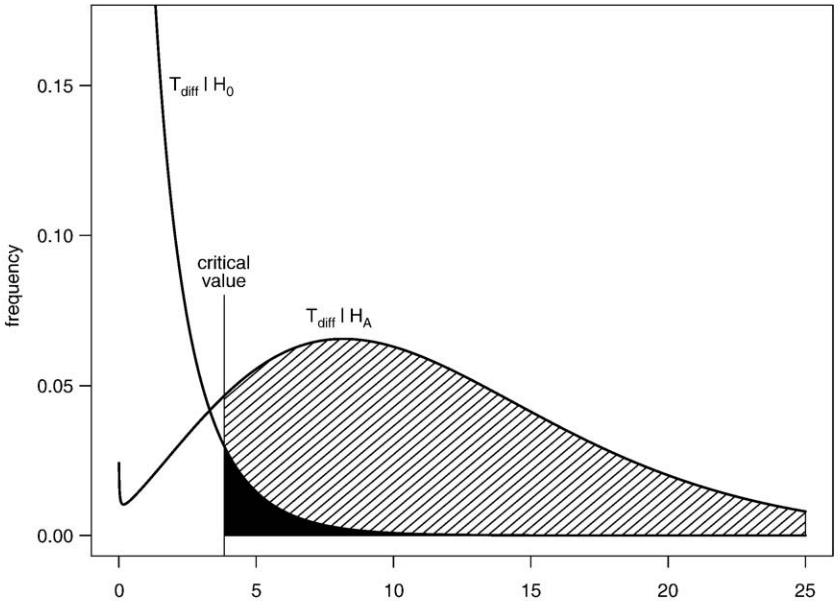

Fig. 2. Example of the distribution of $T_{\mathrm{diff}} | H_0$ (e.g., no sex difference in g), $\chi^2(\mathrm{df}=1, \delta=0)$, and an example of the distribution of $T_{\mathrm{diff}} | H_{\mathrm{A}}$ (e.g., a sex difference g), $\chi^2(\mathrm{df}=1, \delta=10.3)$. In this example, the critical value equals 9.49, the level of significance equals 0.05 (black area) and the power equals 0.89 (shaded area).

of the model is presented in Appendix A. In investigating sex differences in *g* using MG-CMSA, models like that depicted in Fig. 1 are fitted in the male and female sample simultaneously. To ensure a meaningful comparison, measurement invariance has to be established (Meredith, 1993). Measurement invariance involves a series of equality restrictions in the measurement model across males and females. These restrictions concern the following parameters: the factor loadings (λ_{km} in Fig. 1), the error variances ($\sigma^2_{\varepsilon k}$ in Fig. 1), the intercepts (ν_k in Fig. 1) and the second-order factor loadings (γ_m in Fig. 1). Specifically, these parameters are tested for equality across males and females. Note that the first-order residual variances ($\sigma^2_{\omega k}$ in Fig. 1), and the variance of *g* (σ^2_g in Fig. 1) are free to vary over males and females because these are not part of the measurement model (but of the structural model instead).

If measurement invariance holds, a meaningful comparison of the first- and second-order mean differences is possible. To do so, the means of the first-order residuals ($\mu_{\omega 1}$ to $\mu_{\omega 4}$ in Fig. 1) and the *g*-factor (μ_g in Fig. 1) are fixed to 0 in one sample (say, the males). Then, in the other sample (females), 4 mean parameters can be estimated freely (e.g., $\mu_{\omega 1}$ to $\mu_{\omega 4}$, or $\mu_{\omega 1}$, $\mu_{\omega 2}$, $\mu_{\omega 3}$ and, μ_g). The sex differences in observed means are then modeled by these parameters.

Having presented the second-order factor model that both Dolan et al. (2006) and van der Sluis et al. (2006) found tenable, we now briefly outline the procedure to calculate power in this kind of model.

2. Maximum likelihood and power

If the data, (i.e., realization of the variable y_1–y_{14} — see Fig. 1) are independently and identically distributed according to the multivariate normal distribution, and if the number of subjects is sufficiently large, estimates of the parameters in the model depicted in Fig. 1 (corresponding to Eqs. (A7)–(A9)) can be derived using Maximum Likelihood estimation (ML; Azzelini, 1996). In ML estimation, the log-likelihood ratio function (F) is minimized (Lawley & Maxwell, 1971, p. 25; Bollen, 1989, p. 107). F is a function of $\boldsymbol{\theta}$, a vector of parameters to be estimated.[2] The values of the parameters in $\boldsymbol{\theta}$ that minimize F given the specified model are the ML estimates of the parameters in the model. This minimum of F is denoted F_{MIN}. Now, the value T that is given by

$$T = (N-1)F_{\mathrm{MIN}} \tag{1}$$

has a χ^2 distribution, and could be taken as a measure of goodness-of-fit (i.e., it is the value of χ^2 outputted by standard software like LISREL, Mplus, and Mx).

As the parameter estimation is based on ML, we can use the log-likelihood difference test (see Satorra & Saris, 1985; Saris & Satorra, 1993) to calculate power to find a group difference in *g*. We will show how power calculations are carried out within the common factor model using this procedure. It may be useful to note the similarities of this approach with that of power calculation within a more basic statistical test like the independent samples *t*-test. In case of a *t*-test, power is determined to reject the null-hypothesis (no difference between the samples) in favor of an alternative hypothesis (a difference between the samples) using the *t*-statistic. Here, we determine the power to reject model H_0 in favor of model H_{A} using the T_{diff} statistic. Details are discussed next.

In the log-likelihood difference test, two models are considered; model H_{A} (which corresponds to the alternative

[2] These are the unknown parameters from Fig. 1 (which corresponds to the parameter matrices in Eqs. (A7)–(A9)).

Table 1
Power and level of significance.

	Accept H_0: Conclude that there is no g-difference	Reject H_0: Conclude that there is a g-difference
H_0 is true, there is no g-difference	$1-\alpha$ (0.95)	Level of significance, α (0.05)
H_A is true, there is a g-difference	β (0.11)	Power, $1-\beta$ (0.89)

Note. The probabilities associated with the example in the text are shown in brackets.

hypothesis in the t-test example) with parameter vector $\hat{\boldsymbol{\theta}}_A$ and degrees of freedom df_A, and a more parsimonious model H_0 (corresponding to the null-hypothesis) with parameter vector $\boldsymbol{\theta}_0$ and degrees of freedom df_0. In these models, H_0 is nested under H_A i.e., the parameter vector $\boldsymbol{\theta}_0$ is a constrained version of $\hat{\boldsymbol{\theta}}_A$. Possible constraints include fixed parameter constraints (e.g., fixing a component of $\hat{\boldsymbol{\theta}}_A$ to zero) and equality constraints (e.g., constraining 2 or more components of $\hat{\boldsymbol{\theta}}_A$ to be equal). In the present study of sex differences in g, these two models represent scenarios in which a difference in g is assumed to be absent, i.e., the mean difference is constrained to equal 0 (model H_0), or in which a mean difference in g is present, and is estimated (model H_A). Now, let T_0 be the goodness-of-fit according to Eq. (1) of the more parsimonious H_0, and let T_A be the goodness-of-fit of some composite H_A. The test statistic (that corresponds to the t-statistic in the t-test example) is calculated as follows:

$$T_{\mathrm{diff}} = T_0 - T_A. \tag{2}$$

If H_0 is true, i.e., if the constraint(s) imposed on $\hat{\boldsymbol{\theta}}_A$ hold, T_{diff} asymptotically follows a *central* χ^2-distribution with degrees of freedom equal to

$$\mathrm{df}_{\mathrm{diff}} = \mathrm{df}_0 - \mathrm{df}_A. \tag{3}$$

and with non-centrality parameter, δ, equal to 0. Thus

$$T_{\mathrm{diff}}\,|H_0{\sim}\chi^2(\mathrm{df}_{\mathrm{diff}}, \delta = 0), \tag{4}$$

where the tilde denotes 'distributed as'.

In the case that H_A is true (i.e., the constraint(s) in $\boldsymbol{\theta}_0$ do not hold), H_A is misspecified and the test statistic T_{diff} follows the *non-central* χ^2-distribution (Satorra & Saris, 1985; Saris & Satorra, 1993). The shape of this distribution depends on the degrees of freedom and δ,

$$T_{\mathrm{diff}}\,|H_A{\sim}\chi^2(\mathrm{df}_{\mathrm{diff}}, \delta > 0). \tag{5}$$

To obtain a numerical estimate of δ, we choose a sample size N, and assign specific parameter value(s) to express H_0 and H_A. That is to say, both H_0 and H_A are fully specified in advance to calculate power. For instance, H_A may include a mean difference due to g, whereas in H_0 this parameter is fixed to zero. Given these parameter values, we calculate summary statistics under H_A, i.e., we calculate the exact population values of the means and covariance matrices associated with the choice of parameters. Finally, we fit H_0 and H_A to these statistics, and calculate T_{diff}. Now, δ asymptotically equals this difference. The value of δ depends on the magnitudes of the differences between the parameters in $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_A$ (effect sizes) and the sample size, N. Because all other parameters in the model may also affect the value of δ, we must provide a completely specified model, which includes magnitudes of the parameters of immediate interest, as well as values for all other parameters.

To calculate power, we first have to set a critical value on the distribution of the T_{diff} statistic that provides a cut-off for the choice between models H_0 and H_A. If the observed value of T_{diff} exceeds this cut-off value, we will choose H_A, if the observed value of T_{diff} is smaller than this value, we will choose H_0. In Fig. 2, the distribution of $T_{\mathrm{diff}}\,|\,H_0$ and $T_{\mathrm{diff}}\,|\,H_A$ is depicted for $\mathrm{df}_{\mathrm{diff}}=1$ and $\delta=10.3$. Let H_0 be a model without sex difference in g and H_A a model with a sex difference in g. Commonly, the critical value is set in such a way that the probability of incorrectly rejecting H_0 equals 0.05. The choice of a level of significance, α, to 0.05 (that is, the black area in Fig. 2 corresponds to $\alpha=0.05$) is associated with a critical value of 9.48. Note that if we observe $T_{\mathrm{diff}}>9.48$, we will accept H_A, and conclude that there is a g-difference. Therefore, the probability that we accept H_A (given that it is true) equals the probability that T_{diff} exceeds 9.48. This probability is the power to reject H_0 in favor of H_A, i.e., the power to detect the g-difference in model H_A. This probability corresponds to the shaded area in Fig. 2 and equals 0.89. The usual 2×2 table of possible outcomes for this example is depicted in Table 1.

3. Design of the study

The power to detect a g-difference depends on a number of factors, including the size of the g effect, and the number of first-order residual differences. To see how these effects influence the power to detect a g-difference, we manipulated the parameters corresponding to these effects. The effects that we considered were: 1) the size of the g-mean difference, 2) the number of intercept differences, 3) the number of first-order residual mean differences, and 4) the strength of the positive manifold. These four effects were manipulated by considering 3 levels. We therefore obtained $3\times3\times3\times3=81$ conditions. In each condition, population covariance matrices and mean vectors were generated according to Fig. 1 (Eqs. (A7)–(A9)) for two groups (males, females) according to model H_A. Remember that model H_A includes a g-difference between males and females, thus, in order to generate the data, we had to specify this difference (which values we used exactly is discussed below). To these data with the g-difference, the model H_0 without the g-difference was fitted. Notice that we fitted the wrong model, as the data were generated according to H_A. The goodness-of-fit of model H_0 (i.e., T from Eq. (1), which is provided by software like Lisrel, Mplus and Mx) could then be taken as an approximation for the non-centrality parameter.[3] The power to detect the g effect specified in H_A was then calculated using this non-centrality parameter and $\alpha=0.05$. Note that the degrees of freedom equaled 1, as the difference between model H_0 and model H_A concerned only a single parameter (i.e., μ_g in Fig. 1 and $\Delta\boldsymbol{\mu}_g$ in Eq. (A9)). How the parameters were manipulated specifically is described next.

[3] In doing so, the fit of model H_0 is perfect ($T_0=0$), as the data is generated according to this model. As a consequence, $T_{\mathrm{diff}}=T_A$.

Table 2
Power to detect a *g*-difference.

		incpt	Positive manifold		
			Weak	Medium	Strong
a) g effect size −0.2					
lat	0	0	0.41 (2.88)	0.52 (4.03)	0.61 (5.03)
		1	0.40 (2.87)	0.52 (4.01)	0.61 (5.00)
		4	0.38 (2.77)	0.51 (3.90)	0.60 (4.89)
	2	0	0.27 (1.84)	0.41 (2.98)	0.54 (4.29)
		1	0.27 (1.81)	0.40 (2.93)	0.54 (4.22)
		4	0.26 (1.74)	**0.39 (2.81)**	0.52 (4.05)
	3	0	0.18 (1.04)	0.30 (2.08)	0.48 (3.63)
		1	0.17 (1.00)	**0.29 (1.99)**	0.46 (3.48)
		4	0.16 (0.90)	0.27 (1.79)	0.42 (3.12)
b) g effect size −0.3					
lat	0	0	0.72 (6.47)	0.85 (9.02)	0.92 (11.24)
		1	0.72 (6.44)	0.85 (8.98)	0.92 (11.20)
		4	0.70 (6.21)	0.84 (8.73)	0.91 (10.95)
	2	0	0.53 (4.12)	0.73 (6.67)	0.87 (9.62)
		1	0.52 (4.06)	0.73 (6.56)	0.87 (9.45)
		4	0.51 (3.90)	**0.71 (6.30)**	0.85 (9.07)
	3	0	0.34 (2.35)	0.58 (4.66)	0.81 (8.13)
		1	0.32 (2.24)	**0.56 (4.47)**	0.80 (7.80)
		4	0.30 (2.02)	0.52 (4.02)	0.75 (7.00)
c) g effect size −0.4					
lat	0	0	0.92 (11.45)	0.98 (15.94)	0.99 (19.84)
		1	0.92 (11.39)	0.98 (15.87)	0.99 (19.76)
		4	0.91 (10.99)	0.98 (15.43)	0.99 (19.32)
	2	0	0.77 (7.31)	0.93 (11.81)	0.98 (16.99)
		1	0.77 (7.20)	0.93 (11.62)	0.98 (16.70)
		4	0.75 (6.92)	0.92 (11.16)	0.98 (16.04)
	3	0	0.53 (4.15)	0.82 (8.26)	0.97 (14.38)
		1	0.51 (3.98)	0.80 (7.93)	0.96 (13.80)
		4	0.47 (3.58)	0.76 (7.13)	0.94 (12.39)

Note. *lat*: number of latent mean differences; *incpt*: number of intercept differences; non-centrality parameters are in brackets; boldface corresponds to findings of Dolan et al. and van der Sluis et al. and are further explored in Fig. 3.

In the WAIS-III factor structure from Fig. 1, we manipulated the magnitude of the *g*-mean difference. That is, in the design, μ_g in Fig. 1 (or $\boldsymbol{\Delta\mu_g}$ in Eq. (A9)) was nonzero and favored males, with effect sizes of either −0.2, −0.3 or −0.4. These effect sizes (i.e., in standard deviation units) corresponded to IQ differences of 3, 4.5, and 6 points, respectively. We chose these differences to favor males in accordance with Lynn's (1994, 1999) hypothesis.

Second, we manipulated the number of intercept differences (i.e., the number of υ from Fig. 1 that were unequal across sex). An intercept difference suggests that the mean sex difference on the corresponding subtest could not be explained in terms of the sex difference on the specific ability factor that the subtest purports to measure. When intercept differences are present, the corresponding intercepts are estimated in both groups separately. That is, the mean differences on these subtests are no longer attributable to mean differences with respect to the common factors. An intercept difference of a subtest is problematic in the comparison of groups with respect to a common factor, as such a difference suggests that the meaning of the common factor may differ over the groups. For instance, an intercept difference renders group differences in sum scores, a proxy of the common factor, hard to interpret. One reviewer remarked that an intercept difference suggests that the meaning of the first-order common factor, and thus of *g* (modeled as a second-order factor) differs over the groups. We agree with this. In an analysis of common factor mean differences using covariance and mean structure modeling, one can delete the subtest, or retain the subtest, but allow the intercept to vary over groups. In past applications, generally the latter was chosen (e.g., Dolan et al., 2006; van der Sluis et al., 2006), assuming that there were still an adequate number of measurement invariant subtests, to arrive at a good estimate of the common factor mean difference. In this case, deleting the subtest should not greatly affect this estimate. Thus, given the limited objective of analyzing the structure of mean differences, we adopted this pragmatic strategy. Of course, the intercept difference itself poses a more general psychometric problem, which generally (i.e. outside the domain of modeling mean differences) cannot be solved by simply allowing intercepts to differ. It suggests that the subtest is not a suitable indicator of the common factor (e.g., the Information subtest of the WAIS as an indicator of Verbal cognitive abilities in an analysis of sex differences; see Dolan et al., 2006).

As the number of intercept differences increases, power to detect a *g*-difference was expected to decrease, because there remain fewer indicators that depend on *g*. In the design, either 0, 1, or 4 intercepts were unequal across sex. In the case of zero differences, all intercepts were modeled as equal (in accordance with the model). In the case of 1 intercept difference, there was a male advantage with respect to υ_5 (see Fig. 1). This corresponded to the results of van der Sluis et al. (2006). In the case of 4 intercept differences, there was a male advantage with respect to υ_3, υ_5 and υ_8, and a female advantage with respect to υ_1 (see Fig. 1). This latter configuration was in accordance with the results of Dolan et al. (2006).

Third, we manipulated the number of first-order residual mean differences. A residual mean difference suggests that there is a sex difference on the corresponding ability that could not be explained in terms of the mean difference in *g*. More residual mean differences were expected to decrease power as the presence of first-order residual mean differences renders the *g* effect less influential in the model of means. In the design, either 0, 2, or 3 first-order residual mean differences were present. In the case of 0 first-order residual mean differences, all μ_ω were 0. In the case of 2 first-order residual mean differences, the $\mu_{\omega 2}$ and $\mu_{\omega 3}$ favored males, as Dolan et al. found (effect sizes were −0.3 and −0.3, respectively). In addition, in the case of 3 first-order residual mean differences, $\mu_{\omega 2}$ and $\mu_{\omega 3}$ favored males and $\mu_{\omega 4}$ favored females (effect sizes were −0.3, −0.3, and 0.65, respectively) in accordance with the van der Sluis et al. results.

Fourth, the strength of the positive manifold was manipulated. As the strength of the positive manifold increases, the intercorrelations among the first-order common factors increase, indicating that *g* explains more of the variance in these factors. Note that as these first-order correlations increase, the correlation among the observed variables increase as well. In the design we distinguished three levels of the strength of the positive manifold, which we denoted weak, medium, and strong. In the weak condition, first-order residuals (i.e., the diagonal elements of $\boldsymbol{\Sigma}_{\omega i}$ from Eq. (A7) corresponding to σ^2_ω in Fig. 1) were chosen in such a way that *g* explained 20, 30, 20 and 30% of the variance in the four

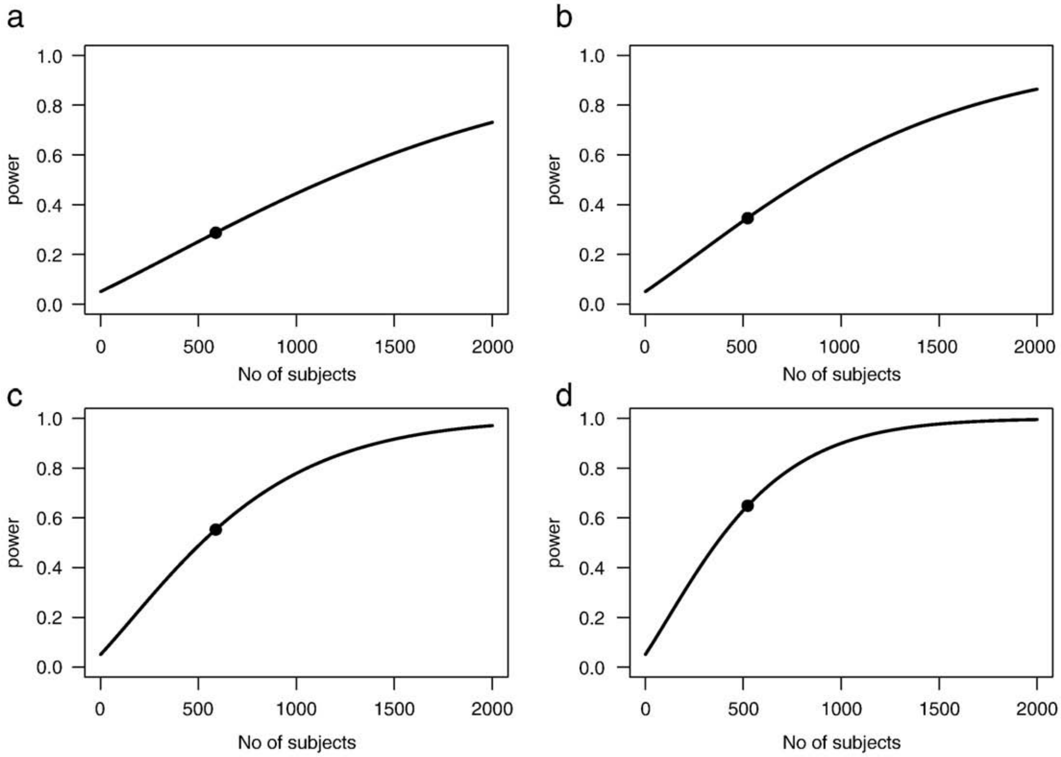

Fig. 3. The power to detect a *g*-difference plotted against number of subjects. Panels a and c correspond to the findings of van der Sluis et al. in case of a *g* effect size of −0.2 and −0.3 respectively. Panels b and d correspond to the findings of Dolan et al. in case of a *g* effect size of −0.2 and −0.3 respectively. Dots denote the actual sample size in that particular study.

common factors of the WAIS-III factor structure, respectively.[4] In the medium condition, these percentages were 40, 50, 40, and 50, and in the strong condition: 70, 80, 70, and 80. Increasing the positive manifold was expected to be associated with larger power to detect the *g*-difference, as *g* became proportionately more influential in the model.

A factor not manipulated in the design was sample size. In sex difference research, sample sizes are about equal. We therefore set $N_1 = 300$ and $N_2 = 300$ in the male and female sample, respectively. However, results from the power analyses can easily be generalized to other sample sizes as well, provided that the sample sizes are equal. As we will present the non-centrality parameters for all analyses in the study, power can be calculated for any sample size using:

$$\delta_{\text{new}} = (\delta_{\text{indicated}} / 600)^* N_{\text{new}}, \tag{6}$$

where N_{new} is the newly chosen total sample size (that is, $N_{\text{new}}/2$ is the sample size in each group, females and males). We emphasize that our results cannot be used to calculate power for unequal group sizes. This requires actually refitting the model with the desired unequal sample sizes. Analyses were carried out in the freely available software packages Mx (Neale, Boker, Xie, & Maes, 2002) and *R* (R development Core Team, 2007). The syntax of the analyses including all parameter choices is available upon request. Note that this syntax can be used, with minimal adaptation, to calculate power in the case of unequal group sizes.

4. Results

4.1. MG-CMSA

Table 2a to c summarizes the results of the power study. In general, power was as low as 0.16 or as high as 0.99 depending on the circumstances. From the table, it appears that the power depended largely on the size of the *g*-difference. This is not surprising as power is a direct function of this effect size. The number of first-order residual mean differences was an important factor as well. Power coefficients varied by 0.2 to 0.4 over the levels of this factor, i.e., power coefficients across the number of first-order residual differences differed from each other by about 0.2 to 0.4. More first-order residual mean differences were associated with lower power. However, as the strength of the positive manifold increased, the number of first-order residual mean differences became less influential. This is particularly evident in the case of a *g*-effect size of 0.4. The strength of the positive manifold was also influential. Power coefficients varied by about 0.1 over the levels of this factor. The number of intercept differences hardly influenced the power to detect the *g*-effect. Power coefficients only varied by 0.01 to 0.04 over the levels of this factor.

A factor not manipulated but obviously relevant is the total sample size. In Fig. 3, power is plotted against sample size for four specific situations from the design. These situations are boldfaced in Table 2. Fig. 3a and c represents the findings of van der Sluis et al. (2006), i.e., three first-order residual mean differences, one intercept difference and a medium strength

[4] Specifically, we chose the first-order residual variances, σ^2_ω, in such a way that $(\gamma^2\ \sigma^2_g)/(\gamma^2\ \sigma^2_g + \sigma^2_\omega)$ equalled 0.2, 0.3, 0.2, or 0.3.

of the positive manifold. The hypothesized g-difference was either 0.2 (Fig. 3a) or 0.3 (Fig. 3c). The dots in the figures correspond to the sample size in the study of van der Sluis et al. ($N=522$). The dots are located at power coefficients of 0.29 and 0.55, respectively. Fig. 3b and d represent the findings from Dolan et al. (2006), i.e., two first-order residual mean differences, four intercept differences, and a medium strength of the positive manifold. The hypothesized g-difference was either 0.2 (Fig. 3b) or 0.3 (Fig. 3d) standard deviations units. The dots in the figures represent the sample size in the study of Dolan et al. ($N=588$). The dots are located at power coefficients of 0.35 and 0.65, respectively.

4.2. MIMIC

An alternative to MG-CMSA is MIMIC modeling (Irwing, 2007). In this approach, the data of males and females are pooled and analyzed in a single group model, with sex included as a fixed regressor (dummy coded 0 and 1). The factors of interest (e.g., g or some first-order common factors) are regressed on sex. The mean differences are expressed in the corresponding regression coefficients. Note that equal covariance matrices between males and females are assumed in this approach, i.e., homoscedasticity. That is, the assumptions of the MIMIC model are stricter than the assumptions underlying measurement invariance. Besides equality of the parameters in the measurement model, the parameters in the structural model are also assumed to be equal across sexes. If this assumption holds, MG-CMSA and MIMIC produce identical results.[5] In that case, power in MIMIC models will equal the power of MG-CMSA. However, our power calculations were not based on equal covariance matrices between the groups, as factor mean differences are often accompanied by differences in variance. We modeled variance differences in g and the first-order common factors as these differences were found by Dolan et al. and van der Sluis et al.

We carried out additional analyses to determine how power differs for the MIMIC approach. A problem with the standard MIMIC model (i.e., the standard MIMIC model, see Muthen, 1989) is that it cannot readily accommodate intercept differences between sexes (Bollen, 1989, p. 321). The misfit due to these intercept differences (which largely favour males) will be apparent in the g-difference, making this difference either greater (when males are favored in g) or smaller (when females are favored in g). This results in power coefficients that are larger or smaller than those obtained for MG-CMSA (where the intercepts differences are included in the model). Note that these power coefficients correspond to the power to detect a g-effect that is biased (i.e., biased by the intercept differences). To render the g-difference unbiased, the subtests that display intercept differences should be regressed on the dummy coded sex variable (this is not straightforward, see Bollen, 1989, p. 395, for a possible solution; a LISREL input file with our — simpler — solution is available upon request). When fitting a MIMIC model, the modification indices that LISREL produced (see Sörbom, 1989) served as diagnostic tools to identify sex differences in intercepts.

We found that violation of homoscedasticity due to group differences in factor variances hardly affected power in the MIMIC approach, i.e., the power was almost identical to that obtained in MG-CMSA under all circumstances presented in this study. Note that these variance differences we considered were not particularly large. It is possible that greater differences in variance will result in a divergence in the power of the MIMIC model and MG-CMSA. This remains to be investigated (see Irwing, 2007).

5. Discussion

This study was inspired by the recurrent failure to detect a sex difference in the general intelligence factor, g (Dolan et al., 2006; van der Sluis et al., 2006; van der Sluis et al., 2008), while a small difference has been hypothesized by Lynn (1999), and Nyborg (2003).

In this study, we showed how the power to detect a mean male advantage in g using MG-CMSA depended on four factors, i.e., the number of intercept differences, the number of first-order residual mean differences, the strength of the positive manifold, and the effect size of the g effect. Specifically, we found that the power to detect a g-difference of 3 IQ points (i.e., 0.2 SD units) given the samples of Dolan et al. and van der Sluis et al. was 0.3–0.35. A g-difference of 4.5 IQ points (i.e., 0.3 SD units) was associated with a power coefficient of 0.55–0.65. This difference was somewhat larger than the hypothesized IQ difference of 4 IQ points. We therefore conclude that the power to detect the effect postulated by Lynn is at most 0.5–0.6. Commonly, a power coefficient of 0.80 is viewed as acceptable, so the 0.5–0.6 level is not nearly sufficient.

We have only considered power in MG-CMSA, as we consider this the best way to address hypotheses concerning mean differences in multidimensional IQ data (e.g., Gustafsson, 1992; Horn, 1997; Millsap, 1997). Other methods which have been considered (e.g., Jensen's method of correlated vectors and the point bi-serial correlation) are known to be problematic because they do not involve explicit model fitting or competing model comparison, and they do not involve a complete or unambiguous test of measurement invariance (Dolan & Hamaker, 2001; Dolan, 2000). In addition the method of correlated vectors is shown to be insensitive to model violations (Lubke, Dolan, & Kelderman, 2001; Dolan, Roorda, & Wicherts, 2004) and is not judged to be suitable in case of small mean differences (e.g., sex differences; see Nyborg, 2003). Therefore we did not consider these methods in the present study.

As power in MG-CMSA is insufficient in the circumstance considered here, the question arises as to how to overcome this. Obviously, as we showed, power can be raised by increasing the sample sizes. As we presented the non-centrality parameters for all power analyses in the study, sample size required for an adequate power can be calculated a priori for the situation at hand, using Eq. (6). For instance, consider the power of 0.55 in the van der Sluis et al. study; in case of a g effect of 0.3 (see Table 1b). The non-centrality parameter in this case equaled 4.47. Increasing N from $N=600$ to $N=900$ (i.e., 450 males and females in each group respectively) would increase the power to 0.74.

Replication is a second way to deal with the insufficient power. Replicating underpowered studies may shed light on the existence of an effect despite the relatively low power of the individual studies. Consider for example a series of

[5] An unpublished report demonstrating this is available upon request.

underpowered studies that did not reveal the effect (i.e., suppose that power to detect a putative mean difference in *g* was approximately 0.4 in each study). If the effect is never found in the series of studies, notwithstanding the low power, the effect is likely simply to be absent. Regardless, more studies employing MG-CMSA of samples that are sufficiently large and representative of well defined populations, are clearly required to reach a definite conclusion about the role of *g* in sex differences on intelligence test scores.

Finally, we should note that replication and sufficient sample sizes are not the only issues to bear in mind in sex differences research. Having representative samples that adequately reflect the characteristics in the population is clearly as important, as there are many sources of unrepresentative sampling. Another issue is that measurement invariance is not always tenable, i.e., violations of intercept invariance are sometimes evident with respect to some subtests. With relatively few intercept invariant subtests, one could argue that the common factor of interest is not adequately represented. However, in terms of power, we showed that the effect of intercept differences is negligible.

The power analyses presented in this paper are highly dependent on the choices we made. Although these choices are well considered and based on past research, it is possible that one is interested in power in a situation that deviates from those in this study. This should not be a problem, as we showed how to calculate power in any circumstance possible (syntax to do so is available upon request).

Appendix A

In this appendix, we present the matrix algebraic representation of the models used in this study. Let $\mathbf{y}_{ij}$ denote the column vector with observed scores of subject *j* in the *i*th population, on the *p* subtests of any intelligence test. As we followed Dolan et al. (2006) and van der Sluis et al. (2006), *i* is either 1 (male) or 2 (female) and *p* equals 14, which is the number of subtests in the WAIS-III. These scores are entered into the following common factor model (Lawley and Maxwell, 1971; Sörbom, 1974; Bollen, 1989):

$$\mathbf{y}_{ij} = \boldsymbol{\upsilon}_i + \boldsymbol{\Lambda}_i \boldsymbol{\eta}_{ij} + \varepsilon_{ij}, \tag{A1}$$

in which $\boldsymbol{\eta}_{ij}$ denotes the $q \times 1$ vector of unobservable scores of subject *j* from population *i* on the *q* common factors. In this study, *q* equals 4 as the WAIS-III factor structure contains 4 factors (Verbal Comprehension, Working Memory, Perceptual Organization and Perceptual Speed, see WAIS-III, WMS-III Technical Manual, Psychological Corporation, 1997). The $p \times 1$ vector $\boldsymbol{\upsilon}_i$ contains the intercepts of the subtests in the *i*th population and $\boldsymbol{\varepsilon}_{ij}$ denotes the vector with residual scores of subject *j* in population *i* on the *p* subtests. These residuals contain both random and systematic error (Meredith, 1993), and are uncorrelated with the common factors. The $p \times q$ matrix $\boldsymbol{\Lambda}_i$ contains factor loadings; these equal the regression coefficients in the regression of $\mathbf{y}_{ij}$ on $\boldsymbol{\eta}_{ij}$.

Dolan et al. (2006), van der Sluis et al. (2006) and van der Sluis et al. (2008) found that the *g*-model, in which the general intelligence factor is incorporated as a second-order factor, was tenable. In Eq. (A1), $\boldsymbol{\eta}_{ij}$ can thus be modeled in a linear regression on g_{ij}, which is the score of subject *j* in population *i* on the general intelligence factor,

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\Gamma}_i g_{ij} + \boldsymbol{\omega}_{ij}, \tag{A2}$$

where $\boldsymbol{\omega}_{ij}$ is the $q \times 1$ vector of first-order residuals. The vector $\boldsymbol{\Gamma}_i$ contains second-order factor loadings, which can again be interpreted as regression coefficients, in this case in the regression of the first-order common factor scores ($\boldsymbol{\eta}_{ij}$) on the general intelligence factor score (g_{ij}).

Substituting Eq. (A2) in (A1), the second-order common factor model is

$$\mathbf{y}_{ij} = \boldsymbol{\upsilon}_i + \boldsymbol{\Lambda}_i \left(\boldsymbol{\Gamma}_i g_{ij} + \boldsymbol{\omega}_{ij} \right) + \varepsilon_{ij}, \tag{A3}$$

in which the expected mean vector equals

$$\boldsymbol{\mu}_{yi} = \boldsymbol{\upsilon}_i + \boldsymbol{\Lambda}_i \left(\boldsymbol{\Gamma}_i \boldsymbol{\mu}_{\mathbf{gi}} + \boldsymbol{\mu}_{\boldsymbol{\omega}i} \right). \tag{A4}$$

In Eq. (A4), $\boldsymbol{\mu}_{\mathbf{gi}}$ denotes the mean score of population *i* on the general intelligence factor and $\boldsymbol{\mu}_{\boldsymbol{\omega}i}$ is the $q \times 1$ vector of first-order residual means in population *i*.

In Eq (A4), the expected covariance matrix of $\mathbf{y}_{ij}$ equals

$$\boldsymbol{\Sigma}_{yi} = \boldsymbol{\Lambda}_i \left[\boldsymbol{\Gamma}_i \boldsymbol{\sigma}^2_{\mathrm{g}i} \boldsymbol{\Gamma}^t_i + \boldsymbol{\Sigma}_{\omega i} \right] \boldsymbol{\Lambda}^t_i + \boldsymbol{\Sigma}_{\omega i}, \tag{A5}$$

in which $\boldsymbol{\sigma}^2_{gi}$ is the variance of the general intelligence factor in population *i* and $\boldsymbol{\Sigma}_{\varepsilon i}$ is the $p \times p$ covariance matrix of the residuals, $\boldsymbol{\varepsilon}_{ij}$ in that population. The $q \times q$ diagonal matrix $\boldsymbol{\Sigma}_{\omega i}$ contains variances of the first-order residuals. Eq. (A5) is derived under the assumptions that $\mathrm{cov}(\varepsilon_{ij}, g_{ij}) = \mathrm{cov}(\varepsilon_{ij}, \omega_{ij}) = \mathrm{cov}(g_{ij}, \omega_{ij}) = 0$.

Eqs. (A4) and (A5) are simultaneously fitted in both groups, i.e., the males and females. To ensure a meaningful comparison, measurement invariance constraints are introduced in the covariance and mean model (Mellenbergh, 1989; Meredith, 1993). With these constraints in place, Eqs. (A4) and (A5) become:

$$\boldsymbol{\mu}_{yi} = \boldsymbol{\upsilon} + \boldsymbol{\Lambda} \left(\boldsymbol{\Gamma} \boldsymbol{\mu}_{\mathbf{g}i} + \mu_{\boldsymbol{\omega}i} \right) \tag{A6}$$

$$\boldsymbol{\Sigma}_{yi} = \boldsymbol{\Lambda} \left[\boldsymbol{\Gamma} \boldsymbol{\sigma}^2_{\mathrm{g}i} \boldsymbol{\Gamma}^t + \boldsymbol{\Sigma}_{\omega \mathrm{i}} \right] \boldsymbol{\Lambda}^t + \boldsymbol{\Sigma}_{\varepsilon}, \tag{A7}$$

i.e., the first and second-order factor loadings, residual variances, and intercepts are assumed to be equal across groups. This model satisfies strict factorial invariance, in the terminology of Meredith (1993). A less restrictive model includes group differences in the residuals, i.e., $\boldsymbol{\Sigma}_{\varepsilon}$ is replaced by $\boldsymbol{\Sigma}_{\varepsilon \mathrm{i}}$. This model, which satisfies strong factorial invariance, is slightly less informative, but still admits an interpretation of group differences in observed means in terms of group differences in latent (common factor) means.

Eqs. (A6) and (A7) together are considered the general second-order MG-CMSA model subject to measurement invariance (see also Byrne and Stewart, 2006). In this model, we fix certain elements for reason of identification.

First, to ensure identifiable unobservable (co)variances, we fix certain elements in $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ to be equal to 1 (Bollen, 1989). Second, the mean differences in the first-order common factors are estimated relative to an (arbitrary) reference group, as it is not possible to estimate these means in both groups (Bollen, 1989; Sörbom, 1974). Now, the model of means in the male and female sample equals

$$\boldsymbol{\mu}_{y1} = \boldsymbol{\upsilon} \tag{A8}$$

$$\boldsymbol{\mu}_{y2} = \boldsymbol{\upsilon} + \boldsymbol{\Lambda}\left(\Gamma\Delta\boldsymbol{\mu}_{\boldsymbol{g}} + \Delta\boldsymbol{\mu}_{\boldsymbol{\omega}}\right), \tag{A9}$$

respectively, where $\Delta\boldsymbol{\mu}_{\boldsymbol{g}}$ represents the mean difference in general intelligence between females and males ($\Delta\boldsymbol{\mu}_{\boldsymbol{g}} = \boldsymbol{\mu}_{\boldsymbol{g}2} - \boldsymbol{\mu}_{\boldsymbol{g}1}$) and $\Delta\boldsymbol{\mu}_{\boldsymbol{\omega}}$ represents the $q \times 1$ vector of first-order residual factor mean differences between females and males ($\Delta\boldsymbol{\mu}_{\boldsymbol{\omega}} = \boldsymbol{\mu}_{\boldsymbol{\omega}2} - \boldsymbol{\mu}_{\boldsymbol{\omega}1}$). As it stands, the model of means is not identified. There are $q+1$ unobservable mean differences (i.e., q in $\Delta\boldsymbol{\mu}_{\boldsymbol{\omega}}$ and 1 in $\Delta\boldsymbol{\mu}_{\boldsymbol{g}}$) of which only q could be estimated. Therefore, at least 1 element of $\Delta\boldsymbol{\mu}_{\boldsymbol{\omega}}$ or $\Delta\boldsymbol{\mu}_{\boldsymbol{g}}$ needs to be fixed to zero.

References

Azzelini, A. (1996). *Statistical inference based on the likelihood.* London: Chapman and Hall.

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

Brody, N. (1992). *Intelligence.* San Diego: Academic Press.

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modelling*, *13*, 287−321.

Colom, R., Garcia, L. F., Juan-Espinosa, M., & Abad, F. J. (2002). Null sex differences in general intelligence: Evidence from the WAIS-III. *The Spanish Journal of Psychology*, *5*, 29−35.

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

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of Multi-Group Confirmatory Factor Analysis. *Multivariate Behavioral Research*, *35*, 21−50.

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

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., Roorda, W., & Wicherts, J. M. (2004). Two failures of Spearman's hypothesis: The GATB in Holland and the JAT in South Africa. *Intelligence*, *32*, 155−173.

Dolan, C. V., van der Sluis, S., & Grasman, R. (2005). A note on normal theory power calculation in structural equation modeling with data missing completely at random. *Structural Equation Modeling*, *12*, 245−262.

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

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling*, *7*, 534−556.

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

Irwing, P. (2007, December). A two stage procedure for locating group differences in latent means of higher-order factor models. Paper presented at the meeting of the International Society for Intelligence Research, Amsterdam, The Netherlands.

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.

Loehlin, J. C. (2000). Group differences in intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 176−198). Cambridge: Cambridge University Press.

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

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method.* New York: American Elsevier.

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

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.

Lynn, R. (1994). Sex differences in intelligence and brain size: A paradox resolved. *Personality and Individual Differences*, *17*, 257−271.

Lynn, R. (1999). Sex differences in intelligence and brain size: A developmental theory. *Intelligence*, *27*, 1−12.

Mackintosh, N. J. (1998). *IQ and human intelligence.* Oxford, England: Oxford University Press.

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.

Muthen, B. O. (1989). Latent variable modelling in heterogeneous populations. *Psychometrika*, *54*, 557−585.

Nyborg, H. (2003). Sex differences in g. In H. Nyborg (Ed.), *The scientific study of general intelligence* (pp. 187−222). Amsterdam: Pergamon.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2002). *Mx: Statistical modeling*, 6th ed. VCU, Richmond, VA: Author.

Psychological Corporation (1997). *WAIS-III WMS-III technical manual.* San Antonio: Harcourt Brace & Co.

R Development Core Team (2007). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing URL: http://www.R-project.org

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 181−204). Newbury Park, CA: Sage.

Satorra, A., & Saris, W. E. (1985). The power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83−90.

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. (1989). Model modification. *Psychometrika*, *54*, 371−384.

van der Sluis, S., Derom, C., Thiery, E., Bartels, M., Polderman, T. J. C., Verhulst, F. C., et al. (2008). Sex differences on the WISC-R in Belgium and the Netherlands. *Intelligence*, *36*, 48−67.

van der Sluis, S., Posthuma, D., Dolan, C. V., de Geus, E. J. C., Colom, R., & Boomsma, D. I. (2006). Gender differences on the Dutch WAIS-III. *Intelligence*, *34*, 273−289.