Intelligence 38 (2010) 336–344

Contents lists available at ScienceDirect

Intelligence

The issue of power in the identification of "g" with lower-order factors

Dora Matzke *, Conor V. Dolan, Dylan Molenaar

University of Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 28 August 2009
Received in revised form 3 February 2010
Accepted 8 February 2010
Available online 4 March 2010

Keywords:
Covariance structure modeling
General intelligence
Log-likelihood difference test
Perfect correlation
Statistical power

ABSTRACT

In higher order factor models, general intelligence (g) is often found to correlate perfectly with lower-order common factors, suggesting that g and some well-defined cognitive ability, such as working memory, may be identical. However, the results of studies that addressed the equivalence of g and lower-order factors are inconsistent. We suggest that this inconsistency may partly be attributable to the lack of statistical power to detect the distinctiveness of the two factors. The present study therefore investigated the power to reject the hypothesis that g and a lower-order factor are perfectly correlated using artificial datasets, based on realistic parameter values and on the results of selected publications. The results of the power analyses indicated that power was substantially influenced by the effect size and the number and the reliability of the indicators. The examination of published studies revealed that most case studies that reported a perfect correlation between g and a lower-order factor were underpowered, with power coefficients rarely exceeding 0.30. We conclude the paper by emphasizing the importance of considering power in the context of identifying g with lower-order factors.

1. Introduction

The positive intercorrelation among scores on cognitive ability tests is a well-established phenomenon, which is often explained by positing a general intelligence factor (g). The g-factor and the positive manifold of correlations may be viewed as synonymous, i.e., the positive manifold guarantees a dominant factor in principal component analyses (Basilevsky, 1983). However, we view the g-factor as a strong hypothesis, as the positive manifold may be attributable to causes other than a general factor (Bartholomew, Deary, & Lawn, 2009; van der Maas et al., 2006).

The g-factor is supposed to reflect the operation of a process that is common to all cognitive tasks (Jensen, 1998). There is, however, considerable disagreement as to what the nature of this general process might be. For instance, some researchers argued that g is in large part a reflection of frontal lobe functions such as inhibitory and control processes (e.g., Duncan, Burgess, & Emslie, 1995; Embretson, 1995; Dempster, 1991), while others stressed the importance of the speed of information processing (e.g., Demetriou et al., 2005; Kail & Salthouse, 1994; Jensen, 1998) or the efficiency of working memory (e.g., Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Süß, Oberauer, Wittmann, Wilhelm & Schulze, 2002).

One source of information concerning the nature of g is higher-order factor modeling, in which g features as the highest order factor (Jensen, 1998). In these models, g is often found to correlate perfectly with a lower-order common factor, suggesting that g and some well-defined broad cognitive ability (represented by the lower-order factor) may be identical. However, the results of such studies are inconsistent. First, g has been found to be perfectly correlated with a wide variety of cognitive abilities, such as fluid reasoning (Gustafsson, 1984;

* Corresponding author. University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. Tel.: +31 205258862.
E-mail address: d.matzke@uva.nl (D. Matzke).

0160-2896/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2010.02.001

Undheim & Gustafsson, 1987; Johnson & Bouchard, 2005); nonverbal reasoning (Dunham, McIntosh, & Gridley, 2002), perceptual reasoning (Johnson & Bouchard, 2005), verbal/mathematical ability (Stauffer, Ree, & Carretta, 1996), and working memory (Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004; Stauffer et al., 1996; Kyllonen & Christal, 1990; Kyllonen, 1993; (Colom, Escorial, Shih & Privado, 2007). The interpretation of such a perfect correlation varies from study to study: some discuss the theoretical implications (e.g., Gustafsson, 1984; Undheim & Gustafsson, 1987), while others merely note the results, but do not interpret them theoretically (e.g., Johnson & Bouchard, 2005). Second, several studies have failed to support the equivalence of general intelligence and the proposed cognitive abilities. For example, there is considerable evidence that g is highly correlated with both fluid reasoning and working memory, but cannot be considered identical with either of these constructs (e.g., Bickley et al., 1995; Ackerman, Beier, & Boyle, 2002; Ackerman, Beier, & Boyle, 2005; Kane, Hambrick, & Conway, 2005).

We suggest that the inconsistency in the results of studies that have addressed the equivalence of g and lower-order factors may in part be attributable to a lack of statistical power. In an underpowered study, the probability of rejecting the hypothesis that two highly correlated (e.g., 0.8 or 0.9) factors are in fact perfectly correlated is low. In such studies, one should be reticent to attach too great a meaning to the supposedly perfect correlation between g and the lower-order common factor. Despite the theoretical importance of the issue of the exact nature of g, we are unaware of any study that has addressed the question of power in the context of identifying g with abilities represented by lower-order factors.

The goal of the present paper is therefore to study the power to correctly reject the hypothesis of perfectly correlated factors in situations, where g and the lower-order factor of interest are strongly, but not perfectly, correlated. To this end, we investigated the power to detect a less than perfect correlation using exact population (summary) statistics, which we constructed on the basis of realistic parameter values and the results of selected publications that reported a perfect correlation between g and a lower-order common factor. We focused exclusively on hierarchical factor models, with g as the single highest order factor, i.e., with g at the apex of the hierarchy (see Jensen, 1998).

The outline of this article is as follows. In the first section, we introduce the concept of statistical power and present a brief overview of power calculations in the context of maximum likelihood estimation. In the second section, we describe the study in which we established the power to correctly reject the equivalence of two related factors under a variety of circumstances using realistic parameter values based on the literature. In the third section, we investigate the power of five published studies by reconstructing the original factor models using reported parameter values. We conclude the paper with a discussion.

2. Statistical power and the log-likelihood difference test

The concept of statistical power plays an important role in formal statistical testing. Statistical power represents the probability of rejecting a hypothesis, given that it is false (e.g., Cohen, 1988, 1992; Kraemer & Thiemann, 1989). We show how power calculations are carried out within the common factor model. In this presentation, it may be useful to note that this approach to power calculation within models for covariance structures is conceptually the same as that to power calculation within a more basic statistical test like the independent samples *t*-test. In case of a *t*-test, power represents the probability of correctly rejecting the null-hypothesis (H_0 no difference between the two samples) in favor of an alternative hypothesis (H_A; a difference between the two samples) using the t-statistic. Here, we are concerned with the power to reject a parsimonious model (H_A) in favor of a more complex model (H_0) using the T_{diff}–statistic (see below). We now present the details of this approach.

Table 1
Probabilities of correct and incorrect decisions in hypothesis testing in the context of covariance structure modeling.

		Statistical decision	
		Reject H_A	Accept H_A
True state of the world	H_A is true ($r=1$)	Type I error (α)	$1-\alpha$
	H_0 is true ($r<1$)	Power ($1-\beta$)	Type II error (β)

Note. H_0 = null-hypothesis; H_A = alternative hypothesis; r = correlation between g and the lower-order factor of interest.

In covariance structure modeling, which includes higher order factor modeling, the H_0 concerns the true model, and the H_A concerns the false model (Satorra & Saris, 1985; Saris, W. Satorra, A. 1993). Here H_A is a special case of H_0, in that H_A can be derived from H_0 by the imposition of constraints on the parameters (e.g., by fixing a given parameter in H_0 to zero): H_A is thus nested under H_0. In the present situation, H_0 is the model which includes a less than perfect correlation between g and a first-order factor, and H_A is the model in which the correlation is fixed to equal one, by the imposition of an appropriate constraint.[1] As shown in Table 1, the present article is thus concerned with the power to correctly reject the H_A of perfectly correlated factors in favor of the H_0 of correlated, but distinct, factors. For any given statistical test, power is a function of the sample size, the Type I error probability (α), and the effect size, i.e., the discrepancy between the value of the parameter(s) of interest under the H_0 (correlation of say 0.8) and the H_A (correlation of 1.00) models. A power of 0.80 is generally considered adequate: i.e., by consensus, a value lower than 0.80 implies an unacceptable risk of a Type II error. Power higher than 0.80 may, ceteris paribus, require unrealistically large sample size (Cohen, 1992).

In the setting of maximum likelihood estimation, the tenability of the parameter constraints associated with the H_A (i.e., the constraints that renders the correlation perfect) can be determined using a number of asymptotically equivalent

[1] Given the positive manifold, and its implication that the common factors are positively correlated (Krijnen, 2004), we do not consider the possibility of a perfect negative correlation.

statistical tests (Azzelini, 1996). Here we focus on the log-likelihood difference test, which can be calculated as

$$T_{\text{diff}} = T_A - T_0, \tag{1}$$

where T_A and T_0 are the likelihood ratios of the H_A and H_0 models, respectively.

To determine the power of the test, one has to consider the distribution of the test statistic T_{diff}. If the two factors of interest were truly perfectly correlated (i.e., if the H_A model was true), T_{diff} would asymptotically follow a central χ^2 distribution,

$$T_{\text{diff}} \sim \chi^2\left(df_{diff}, \lambda = 0\right), \tag{2}$$

where df_{diff} is the difference in the number of estimated parameters under the H_0 and the H_A models and λ is the non-centrality parameter, which equals zero here.

If the correlation between the two factors of interest was truly less than one (i.e., if the H_0 model was true), T_{diff} would follow the non-central χ^2 distribution (Satorra & Saris, 1985; Saris, W. Satorra, A. 1993),

$$T_{\text{diff}} \sim \chi^2(df_{\text{diff}}, \lambda > 0). \tag{3}$$

To put the non-central χ^2 distribution at use, one has to determine its shape. The shape of the non-central χ^2 distribution depends on df_{diff} and the non-centrality parameter λ. To obtain a numerical estimate of λ, one first assigns plausible values to the parameters of the H_0 model and calculates the associated population covariance matrix. The parameters of the H_0 model are chosen such that (inter alia) the correlation between g and the lower-order factor of interest is less than one (say 0.8 or 0.9). As the population covariance matrix is generated according to the H_0 model (i.e., true model), it obtained features as the true population covariance matrix. Second, one expresses the H_A model by assigning plausible values to its parameters. The parameters of the H_A model are chosen such that the correlation between g and the lower-order factor of interest equals one. Note that both the H_0 and the H_A models must be fully specified; *all* model parameters must be assigned plausible values.[2] Third, one chooses a realistic, but arbitrary, sample size N, fits the H_0 and the H_A models to the population covariance matrix and establishes the values of T_{diff} and df_{diff}. Note that the likelihood ratio associated with H_0 (T_0) is zero, because H_0 is the true model, i.e., the model used to obtain the population covariance matrix. The likelihood ratio associated with H_A (T_A) is greater than zero, as the H_A model is false and does not fit the population covariance matrix. The value of T_{diff} (Eq. (1)) equals the non-centrality parameter λ (Satorra & Saris, 1985; Saris, W. Satorra, A. 1993).

Once the value of λ is obtained, one can calculate the power of the test as follows. Choose the Type I error rate (α) and the associated critical value (C) based on the central χ^2 distribution. Next, determine the Type II error rate (β) by calculating the probability of observing a value of the test statistic T_{diff} that is smaller than the chosen critical value C, given the non-central χ^2 distribution,

$$\beta = P\left[\chi^2(df_{\text{diff}}, \lambda) < C\right]. \tag{4}$$

The power of the test is then given by $1-\beta$ (see Table 1). Note that this approach to power calculation does not rely on Monte Carlo simulations, rather it uses analytic methods to determine the probability to correctly reject the H_A.

Thus the steps towards analytical power calculation are 1) choose the parameter values of the H_0 model (i.e., the true model in which the correlation between g and the first-order factor is less than perfect) and calculate the associated population covariance matrix; 2) choose an arbitrary sample size N; 3) fit the H_0 model to obtain T_0, which should equal zero (a useful check); 4) fit the H_A model (i.e., the false model in which the correlation between g and the first-order factor equals one) to obtain T_A, which will assume a value greater than zero; 5) calculate λ and df_{diff}; 6) choose the Type I error rate (α) and calculate the associated critical value C; 7) calculate the power given, λ, N and α. As noted below, the power for other values of N can be calculated easily, i.e., does not require refitting the model.

3. Power analysis

The objective of the power study was to establish the power to correctly reject the H_A of perfectly correlated factors under a variety of circumstances using realistic parameter values. The power to detect the distinctiveness of g and a lower-order factor depends on a number of aspects of the factor model, including the number and the reliability (i.e., explained variance in the factor model) of the indicators, and the effect size (i.e., the discrepancy between the correlation of g and the lower-order factor under the H_0 and the H_A models). Therefore, we systematically manipulated these features of the factor models to study their influence on power.

3.1. Design

We focused exclusively on hierarchical factor models with five first-order factors and with g as a single second-order factor. First, we specified the H_0 models in which the correlations between g and all first-order factors were less than one. The parameter values of the models were chosen to span the range of values reported in published studies. In each model, the correlations (i.e., standardized second-order factor loadings) between g and the first-order factors were set to 0.95, 0.90, 0.85, 0.80, and 0.75, with corresponding explained variances of 0.90, 0.81, 0.72, 0.64, and 0.56.[3] To study the effects of the number and the reliability of the indicators, the H_0 models were created by systematically manipulating these aspects of the models. The H_0 models featured either two, three, or four indicators per first-order factor, where the reliabilities of the indicators were either relatively high (ranging from 0.55 to 0.80) or relatively low (ranging from

[2] See Hancock (2006) for a simplified approach to power calculation that does not require the specification of all parameter values.

[3] The explained variances of the first-order factors are given by the squares of the standardized second-order factor loadings. For example, the explained variance of η_1 in Fig. 1 is given by $0.95^2 = 0.9$.

0.20 to 0.45). This design resulted in $3 \times 2 = 6$ different H_0 models. Fig. 1 presents an example of a hierarchical factor model used in the study. Note that each first-order factor had the same number of indicators and that the residuals of the indicators as well as the residuals of the first-order factors were uncorrelated.

Second, we specified the H_A models that implied perfect correlations between g and a given first-order factor. Each H_0 model had five corresponding H_A models, where each H_A model was created by constraining the correlation between g and one of the five first-order factors to one. This design resulted in $6 \times 5 = 30$ different H_A models. Because each first-order factor correlated differently with g, this approach allowed us to investigate the influence of various effect sizes, ranging from 0.05 to 0.25.

Third, we obtained the non-centrality parameter λ for each H_A model. To this end, we first calculated the population covariance matrices associated with the six H_0 models. Next, we fitted each H_A model to the data generated according to its corresponding H_0 model, using an arbitrary sample size of $N = 200$. To obtain perfectly correlated factors, each H_A model was fitted by constraining the residual variance (i.e., unexplained variance) of the first-order factor of interest to equal zero (van der Sluis, Dolan, & Stoel, 2005). Discarding subject indices, let

$$\eta_i = \gamma_i g + \zeta_i \tag{5}$$

denote the regression of the *ith* first-order factor (η) on g, with ζ_i representing the residual. Scaling the variance of g to equal one, the correlation between η_i and g equals

$$\rho_{\eta_i g} = \frac{\gamma_i}{\sqrt{\gamma_i^2 + \sigma_{\zeta_i}^2}}. \tag{6}$$

If $\sigma_{\zeta_i}^2$ is constrained to equal zero, then clearly the correlation equals

$$\rho_{\eta_i g} = \frac{\gamma_i}{\sqrt{\gamma_i^2}} = 1. \tag{7}$$

The goodness-of-fit statistics of the H_A models (T_A) were used as approximations for the non-centrality parameters.

Finally, we calculated the power to reject each H_A model using the obtained non-centrality parameter and $df_{\text{diff}} = 1$. Note that because the H_A models were obtained by constraints on the variance components, the Type I error probability (α) was doubled to correct for the violation of the admissible parameter space (e.g., Dominicus, Skrondal, Gjessing, Pedersen, & Palmgren, 2006; Stoel, Garre, Dolan, & Van den Wittenboer, 2006). The power was therefore computed given $\alpha = 0.05 \times 2 = 0.1$. Note also that the results reported in the following section are based on non-centrality parameters computed with $N = 200$. However, the non-centrality parameters and the power can easily be computed for any other sample size using:

$$\lambda_{\text{new}} = \left(\frac{\lambda_{\text{original}}}{200}\right) \times N_{\text{new}}, \tag{8}$$

where $\lambda_{\text{original}}$ is the value of the non-centrality parameter reported in the present article (based on $N = 200$) and N_{new} is the new sample size of interest. The models were fitted using LISREL (Jöreskog & Sörbom, 2001). The data generation and the power calculations were carried out using the R package (R Development Core Team, 2006). The R code for the power calculation is presented in the Appendix.

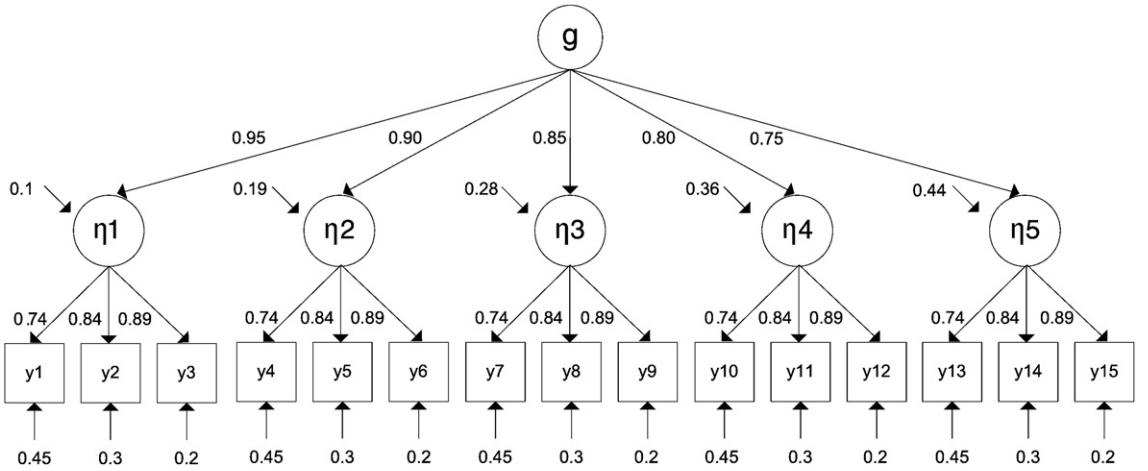

Fig.1. Example of a H_0 model used in the study. The model features three relatively reliable indicators per first-order factor η. The values corresponding to the arrows beneath the indicators and the values corresponding to the arrows left above the first-order factors represent residual variances. Standardized parameters are shown.

3.2. Results

Table 2 summarizes the results of the power calculations. Over the H_A models that we considered, power varied from 0.17 to 1. In general, the three manipulations (i.e., the number and the reliability of the indicators and the effect size) all influenced the power to detect the distinctiveness of g and the lower-order factors.

We first consider the effects of the reliability of the indicators. The results indicated that power increased as the reliability of the indicators increased. The increase in power was, however, more pronounced when the number of indicators and the effect sizes were relatively low. For low effect sizes, power coefficients varied by as much as 0.6 across the two reliability conditions. This difference was reduced for higher effect sizes, especially when the number of indicators was relatively high. With respect to the sample size, the results followed the same pattern. The sample size required to reach sufficient power (i.e., 0.80) was substantially lower for reliable indicators than for unreliable indicators; increases in the reliability of the indicators resulted in an average decrease of 90% in the necessary sample size.

Turning to the effects of the number of indicators, the results indicated that for reliable indicators power was high regardless of the number of indicators. For unreliable indicators, power increased as the number of indicators increased, with power coefficients varying by about 0.15 over the levels of this factor. Similarly, for reliable indicators, the necessary sample size was relatively low regardless of the number of indicators. For unreliable indicators, an increase in the number of indicators was accompanied with an average decrease of 45% in the required sample size.

Lastly, with respect to the influence of the effect size, the results indicated that for reliable indicators power was high regardless of the value of the effect size. For unreliable indicators, power increased as the effect size increased. Note, however, that the increase in power was generally more pronounced – reaching nearly 0.40 – for lower effect sizes, especially when the number of indicators was high. Similarly, for reliable indicators, the necessary sample size was relatively low regardless of the value of the effect size. For unreliable indicators, the necessary sample size decreased as the effect size increased. This decrease varied between 70% and 30% over the levels of this factor.

To summarize, the results of the analyses indicated that power was substantially influenced by the effect size and the number and the reliability of the indicators. For reliable indicators (i.e., from 0.55 to 0.80), power was high regardless of the effect size and the number of indicators. For less reliable indicators (i.e., from 0.20 to 0.45), power increased as the effect size and the number of indicators increased, with a corresponding decrease in the sample size required to reach sufficient power.

4. Power analysis of selected case studies

In this section, we investigate the approximate power of five published studies in order to highlight the importance of power in the context of identifying g with lower-order factors. We reconstructed the original factor models using reported parameter values and determined the power to reject the equivalence of g and the proposed lower-order factors. We focused on studies that used hierarchical factor models – with g as a second or third-order factor – and reported a (near) perfect correlation between g and a lower-order factor.

Table 2
Power to detect the distinctiveness of g and the lower-order factors.

Effect size		2 indicators		3 indicators		4 indicators	
		Reliable (0.75)	Unreliable (0.38)	Reliable (0.68)	Unreliable (0.32)	Reliable (0.70)	Unreliable (0.33)
0.05	Power at $N=200$	0.77	0.17	0.87	0.21	0.96	0.31
	N at power $=0.8$	221	3016	160	1846	106	974
	λ	5.60	0.41	7.76	0.67	11.72	1.27
0.10	Power at $N=200$	1.00	0.33	1.00	0.47	1.00	0.69
	N at power $=0.8$	66	853	46	509	30	268
	λ	18.76	1.45	26.98	2.43	41.24	4.62
0.15	Power at $N=200$	1.00	0.55	1.00	0.74	1.00	0.94
	N at power $=0.8$	35	401	23	237	15	125
	λ	36.23	3.09	54.39	5.22	84.20	9.97
0.20	Power at $N=200$	1.00	0.72	1.00	0.90	1.00	0.99
	N at power $=0.8$	24	249	16	146	10	77
	λ	52.84	4.98	82.42	8.51	129.61	16.23
0.25	Power at $N=200$	1.00	0.85	1.00	0.97	1.00	1.00
	N at power $=0.8$	18	172	12	100	7	53
	λ	69.28	7.23	111.77	12.45	179.93	23.74

Note. The effect size reflects the discrepancy between the correlation of g and the lower-order factor under the H_A (i.e., correlation of one) and the H_0 models. The average reliability of the indicators is shown in brackets. The reliabilities of the indicators in the H_0 models were chosen as follows. In the two H_0 models that featured two indicators per first-order factor, the reliabilities of the indicators were set to 0.7 and 0.8 in the reliable condition and to 0.3 and 0.45 in the unreliable condition. In the two H_0 models that featured three indicators per first-order factor, the reliabilities of the indicators were set to 0.55, 0.7 and 0.8 in the reliable condition and to 0.2, 0.3 and 0.45 in the unreliable condition. In the two H_0 models that featured four indicators per first-order factor, the reliabilities of the indicators were set to 0.55, 0.7, 0.75 and 0.8 in the reliable condition and to 0.2, 0.3, 0.35 and 0.45 in the unreliable condition.

Table 3
Power and the details of the factor models used in the case studies.

Article		Lower-order factor of interest	Mean reliability of indicators	Number of indicators	Effect size	N	λ	Power	N at power = 0.80
Colom et al. (2004)	Model 1 (p.284) [a]	Working memory	0.37	12 (3) [b]	0.05	198	0.24	0.14	5101
	Model 2 (p.285)	Working memory	0.36	15 (3)	0.05	203	1.20	0.29	1046
	Model 3 (p.286)	Working memory	0.39	15 (3)	0.07	193	0.69	0.21	2010
Dunham et al. (2002)	Model 1 (p.159)	Nonverbal reasoning	0.57	6 (2)	0.05	130	0.24	0.14	3349
	Model 2 (p.159) [c]	Nonverbal reasoning	0.49	9 (2)	0.05	130	0.40	0.17	2010
	Model 3 (p.159) [c]	Memory	0.49	9 (2)	0.05	130	0.21	0.14	3828
Gustafsson (1984)	Model 1 (p.192)	Fluid reasoning	0.67	18 (2)	0.05	981	3.60	0.60	1685
Johnson and Bouchard (2005)	Model 1 (p.403)	Fluid reasoning	0.47	42 (7)	0.05	436	22.01	0.99	123
	Model 2 (p.408)	Perceptual reasoning	0.52	42 (5)	0.01	436	0.30	0.15	8985
Undheim and Gustafsson (1987)	Model 1 (p.155)	Fluid reasoning	0.52	26 (3)	0.05	144	0.40	0.17	2226
	Model 2 (p.157)	Fluid reasoning	0.53	10 (3)	0.05	144	0.25	0.14	3561
	Model 3 (p.161)	Fluid reasoning	0.49	28 (3)	0.05	149	0.69	0.21	1336
	Model 4 (p.163)	Fluid reasoning	0.54	13 (3)	0.05	149	2.50	0.48	369
	Model 5 (p.166)	Fluid reasoning	0.51	18 (3)	0.03	148	0.36	0.16	2542

[a] The page number of the original factor model is shown in brackets.
[b] The average number of indicators per first-order factor is shown in brackets.
[c] The original factor model featured perfect correlations between g and both the nonverbal reasoning and memory factors. In the present analysis, we examined the power to reject the equivalence of g and the two first-order factors using two separate factor models.

4.1. Design

The selected case studies and some details of the investigated factor models are listed in Table 3. It is important to note that some of these studies (e.g., Gustafsson, 1984; Undheim & Gustafsson, 1987) formally tested the presence of a perfect correlation and clearly emphasized the equivalence of g and the proposed lower-order factors, while others (e.g., Johnson & Bouchard, 2005) merely reported the perfect correlation between the two factors and did not draw further conclusions about their equivalence. Nevertheless, the results of these publications provided interesting case studies to investigate the power to detect the distinctiveness of highly correlated factors.

The power analyses of the case studies were conducted as follows. First, we reconstructed the H_0 models and the corresponding population covariance matrices using the original factor structures and the exact parameter values reported in the articles.[4] Although we attempted to approximate the original factor models as closely as possible, we did not allow for cross-factor loadings and residual correlations. Further, if the reported correlation between g and the lower-order factor of interest equaled one – either because it was fixed to one or estimated to be one – we set the standardized factor loading between the two factors to 0.95. Note also that in some models the residual variance of the lower-order factor of interest was negative (i.e., Heywood case; Heywood, 1931). Although a Haywood case raises important questions related to model mis-specification and/or over-parameterization (Jöreskog & Sörbom, 1988), the issue of the adequacy of these models is beyond the scope of the present article and will not be considered further. For the purposes of the present investigation, if the reported residual variance of a lower-order factor was negative, we set the standardized residual variance to 0.10, corresponding to a standardized factor loading of 0.95. Next, we specified the H_A models of perfect correlation by constraining the residual variances of the lower-order factors of interest to equal zero. The H_A models were then fitted to the generated datasets using the original sample sizes. Lastly, we computed the power to reject the various H_A models using the obtained non-centrality parameters, $df_{\text{diff}} = 1$, and $\alpha = 0.05 \times 2 = 0.1$.

4.2. Results

The results of the case studies are shown in Table 3. Across the various models, power varied from 0.135 to 0.99. In most cases, however, power did not exceed 0.3, indicating that the power to reject the equivalence of g and a given lower-order factor was generally very low. In the light of the results reported above, this result was not unexpected. Most case studies featured extremely low effect sizes and factor models with only a few (two or three) relatively unreliable indicators per first-order factor. Also in line with our results, power was substantially higher for studies that used relatively reliable or a large number of indicators, reaching 0.6 for Gustafsson's (1984) model and exceeding 0.9 for the first model of Johnson and Bouchard (2005). In summary, the results suggested that the selected case studies, with a very few exceptions, were underpowered to detect the distinctiveness of g and the proposed lower-order factors.

5. General discussion

The goal of this study was to determine the power to reject the hypothesis that g and a lower-order factor are perfectly correlated, given that the correlation is relatively high, but

[4] In a few instances, our approximations of the original factor models have failed to converge. In these cases, we have slightly adjusted the factor structure or the parameter values of the models to assure convergence.

less than one. First, we established the power under a variety of realistic circumstances using artificial datasets. Second, we investigated the power of five published studies by reconstructing the original factor models using reported parameter values.

The results of our power analyses revealed that power was substantially influenced by the effect size and the number and the reliability of the indicators. For highly reliable indicators, power was high regardless of the effect size and the number of indicators. For less reliable indicators, power increased as the effect size and the number of indicators increased. In the light of these results, the ideal dataset to investigate the equivalence of g and a lower-order factor would feature a relatively large number (i.e., three or four indicators per first-order factor) of reliable (i.e., from 0.55 to 0.80) indicators. Note, however, that Dolan (2000; see Jensen & Reynolds, 1982 for the summary statistics) found the mean reliability of the indicators of the Wechsler Intelligence Scale for Children — Revised (WISC-R; Wechsler, 1974) to equal approximately 0.44 (SD = 0.15). Similarly, Dolan and Hamaker (2001; see Naglieri & Jensen, 1985 for the summary statistics) reported that the mean reliability of the indicators of the WISC-R and the Kaufman Assessment Battery for Children (K-ABC, Kaufman & Kaufman, 1983) equaled approximately 0.45 (SD = 0.18).

Our examination of published studies revealed that most of our case studies, which reported a perfect correlation between g and a lower-order factor, were underpowered, with power coefficient rarely exceeding 0.3. Consistent with our previous results, power was substantially higher for studies that used relatively reliable (i.e., Gustafsson, 1984) or a large number of indicators (i.e., Johnson & Bouchard, 2005). In the light of these findings, we recommend that one consider the issue of power before concluding that g and a given lower-order factor are perfectly correlated. In a study designed to address the possibly perfect relationship between g and a lower-order factor, one would ideally conduct power calculations beforehand. However, in a study in which one encounters a (possibly) unexpected perfect correlation, post-hoc power calculations can be useful.

With respect to the correlations between g and the proposed lower-order factors, there is little doubt that these may be quite large. Such correlations do certainly require an explanation. We believe that the high correlations often found between g and lower-order factors may partly result from using the same instrument to measure the common factors. Specifically, using the same instrument (e.g., WISC-R) to define the lower-order factors may introduce variance that is attributable to the particular measurement instrument (i.e., method variance; Campbell & Fiske, 1959), which in turn may increase the correlations between the measures and ultimately the correlation between g and the lower-order factors. For instance, two tests of a given construct, which employ the same method (e.g., paper and pencil), are likely to correlate higher than two tests that employ different methods (e.g., paper and pencil vs. experimental task). Also, the high correlations may partly be attributable to sampling fluctuations or may result from using heterogeneous samples to assess the equivalence of g and the proposed lower-order factors. For instance, IQ test scores are generally more highly correlated in a sample of participants with a wide age range (say 18 to 76 years of age) than in a sample with a smaller age range (18 to 36 years of age). Lastly, the reported high correlations may result from disattenuation effects associated with the unreliability of the composite scores derived from the aggregation of subtests loading on the lower-order factors (Gignac, 2007).

Nevertheless, it is possible that g may indeed be identical to a certain lower-order factor. However, given the very mixed results and the lack of power of most published studies, we are reluctant to accept this. We point out that if such an identity did truly exist, it would imply, by the application of Occam's razor, the demise of g as a causal factor in the study of individual differences in cognitive abilities: why entertain the notion of a single essentially ill-defined higher order factor, if it is in fact identical to a well-defined lower-order factor (say, working memory)?

We also note that the focus on individual differences, which characterizes studies of g, has its inherent interpretational limitations: the presence of a perfect correlation between two variables is not sufficient to conclude that the variables are identical or that they share a common causal substrate. Suppose, for the sake of argument, that in a sample of children, who vary sufficiently in age, an appropriate statistical test reveals that the correlation between height and weight is equal to one. The perfect correlation between height and weight does obviously not imply that the two variables are identical nor that they necessarily share a common causal substrate.

In conclusion, the goal of the present study was to highlight the importance of considering power in the context of identifying g with lower-order factors. Our results provide useful guidelines on the ideal dataset and the necessary sample size required to reach sufficient power in a variety of realistic situations. Furthermore, the procedure used here to investigate power is easy to implement and the R code presented in the Appendix provides a helpful tool to establish the power and the necessary sample size in the particular situation at hand. As pointed out above, the failure to do so might result in mistakenly concluding that g and a lower-order factor are perfectly correlated and therefore – at least from the perspective of individual differences – can be considered identical.

Acknowledgements

Dylan Molenaar is supported by a grant from the Netherlands Organization for Scientific Research (NWO).

Appendix A

R code for power calculations

This appendix presents the R code that can be used to calculate power for various sample sizes. The code takes as inputs the goodness-of-fit statistic of the H_A model (`TA`), the chosen Type I error probability (`alpha`), df_{diff} (`df`), the sample size (`N`) used to obtain the non-centrality parameter λ, and the minimum (`minN`) and maximum (`maxN`) sample sizes of interest. The output provided by the code consists of the power coefficients corresponding to sample sizes ranging from the minimum and the maximum sample size of interest.

```
#Goodness-of-fitstatistic of HA model (i.e., non-centrality parameter lambda)
TA = 7.23
#Type I error probability
alpha = 0.05*2
#Degrees of freedom (i.e., df.diff)
df = 1
#Sample size used to calculate the non-centrality parameter lambda
N = 200
#Critical value
C = qchisq(alpha, df=df, ncp=0, lower.tail=F)
#Minimum sample size of interest
minN = 100
#Maximum sample size of interest
maxN = 2000
power = matrix(0,maxN-minN+1,2)
#Nnew is new sample size of interest
for (Nnew in minN:maxN){
        #lambda.new is the value of the non-centrality parameter
        #corresponding to the new sample size
        lambda.new = (TA/N)*Nnew
        #calculate power
        power[Nnew-minN+1,1] = pchisq(C, df=df, ncp=lambda.new, lower.tail=F)
        power[Nnew-minN+1,2] = Nnew
}
#power plot
plot(power[,2],power[,1],type='l',xlab="Sample size", ylab="Power")
#power for original N
print(power[power[,2]==N,1])
```

References

Ackerman, P., Beier, M., & Boyle, M. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General*, *131*, 567.

Ackerman, P., Beier, M., & Boyle, M. (2005). Working memory and intelligence: The same or different constructs. *Psychological Bulletin*, *131*, 30—60.

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

Bartholomew, D., Deary, I., & Lawn, M. (2009). A new lease of life for Thomson's bonds model for intelligence. *Psychological Review*, *116*, 567—579.

Basilevsky, A. (1983). *Applied matrix algebra in the statistical sciences.* Amsterdam: Elsevier Science Publishing.

Bickley, P., Keith, T., & Wolfle, L. (1995). The three-stratum theory of cognitive abilities: test of the structure of intelligence across the life span. *Intelligence*, *20*, 309—328.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81—105.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Earlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155—159.

Colom, R., Escorial, S., Shih, P., & Privado, J. (2007). Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents. *Personality and Individual Differences*, *42*, 1503—1514.

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence*, *32*, 277—296.

Conway, A., Cowan, N., Bunting, M., Therriault, D., & Minkoff, S. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*, 163—183.

Demetriou, A., Kui, Z., Spanoudis, G., Christou, C., Kyriakides, L., & Platsidou, M. (2005). The architecture, dynamics, and development of mental processing: Greek, Chinese, or Universal? *Intelligence*, *33*, 109—141.

Dempster, F. (1991). Inhibitory processes: a neglected dimension of intelligence. *Intelligence*, *15*, 157—173.

Dolan, C. (2000). A model-based approach to Spearman's hypothesis. *Multivariate Behavioral Research*, *35*, 21—50.

Dolan, C., & Hamaker, E. (2001). Investigating black-white differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC, and a critique of the method of correlated vectors. In F. Columbus (Ed.), *Advances in psychological research*, *Vol. VI.* (pp. 31—60)Huntington, NY: Nova Science Publishers, Inc.

Dominicus, A., Skrondal, A., Gjessing, H., Pedersen, N., & Palmgren, J. (2006). Likelihood ratio tests in behavioral genetics: problems and solutions. *Behavior Genetics*, *36*, 331−340.
Duncan, J., Burgess, P., & Emslie, H. (1995). Fluid intelligence after frontal lobe lesions. *Neuropsychologia*, *33*, 261−268.
Dunham, M., McIntosh, D., & Gridley, B. (2002). An independent confirmatory factor analysis of the Differential Ability Scales. *Journal of Psychoeducational Assessment*, *20*, 152.
Embretson, S. (1995). The role of working memory capacity and general control processes in intelligence. *Intelligence*, *20*, 169−189.
Gignac, G. (2007). Working memory and fluid intelligence are both identical to g?! Reanalyses and critical evaluation. *Psychology Science*, *42*, 187−207.
Gustafsson, J. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*, 179−203.
Hancock, G. (2006). Power analysis in covariance strucure modeling. In G. Hancock, & R. Mueller (Eds.), *Structural equation modeling: A second course.* (pp. 69−115)Greenwich, CT: Information Age Publishing, Inc.
Heywood, H. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society, Series A*, *134*, 486−501.
Jensen, A. (1998). *The g factor: The Science of Mental Ability*. New York: Praeger.
Jensen, A., & Reynolds, C. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences*, *3*, 423−438.
Johnson, W., & Bouchard, T. (2005). The structure of human intelligence: it is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*, 393−416.
Jöreskog, K., & Sörbom, D. (1988). *LISREL 7. A guide to the program and applications.* Mooresville: Scientific Software, Inc.
Jöreskog, K., & Sörbom, D. (2001). *LISREL 8.50: User's reference guide*. Chicago: Scientific Software International.
Kail, R., & Salthouse, T. (1994). Processing speed as a mental capacity. *Acta Psychologica*, *86*, 199−225.
Kane, M., Hambrick, D., & Conway, A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 66−71.
Kaufman, A., & Kaufman, N. (1983). *Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.
Kraemer, H., & Thiemann, S. (1989). *How many subjects?: Statistical power analysis in research.* Newbury Park: CA: Sage.
Krijnen, W. (2004). Positive loadings and factor correlations from positive covariance matrices. *Psychometrika*, *69*, 655−660.
Kyllonen, P. (1993). Aptitude testing inspired by information processing: a test of the four-sources model. *Journal of General Psychology*, *120*, 375−405.
Kyllonen, P., & Christal, R. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*, 389−433.
Naglieri, J., & Jensen, A. (1985). Comparison of black-white differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence*, *11*, 21−43.
R Development Core Team (2006). *Computer software manual.* Vienna, Austria. http://www.R-project.org.
Saris, W., & Satorra, A. (1993). Power evaluation in structural equation models. In K. Bollen, & J. Long (Eds.), *Testing strucural equation models*. (pp. 181−204) Newburry Park: CA: Sage.
Satorra, A., & Saris, W. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83−90.
Stauffer, J., Ree, M., & Carretta, T. (1996). Cognitive-components tests are not much more than g: An extension of Kyllonen's analyses. *Journal of General Psychology*, *123*, 193−206.
Stoel, R., Garre, F., Dolan, C., & Van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *11*, 439−455.
Süß, H., Oberauer, K., Wittmann, W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence*, *30*, 261−288.
Undheim, J., & Gustafsson, J. (1987). The hierarchical organization of cognitive abilities: restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research*, *22*, 149−171.
van der Maas, H., Dolan, C., Grasman, R., Wicherts, J., Huizenga, H., & Raijmakers, M. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychologival Review*, *113*, 842−861.
van der Sluis, S., Dolan, C., & Stoel, R. (2005). A note on testing perfect correlations in SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*, 551−577.
Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children—Revised.* New York: The Psychological Corporation.