Intelligence 64 (2017) 67–70

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell

Predicting group differences from the correlation of vectors

Kimmo Sorjonen*, Jon Aurell, Bo Melin

Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, 171 65 Solna, Sweden

A R T I C L E I N F O

Keywords:
Factor loadings
Intelligence
Jensen effect
Sigmoid association
Simulation

A B S T R A C T

It has been proposed that a correlation between a vector of factor loadings of intelligence tests and a vector of group differences on these same tests (= correlation of vectors) indicate that the group difference is mainly in *g*. In the present simulation, we show that there is an inverse sigmoid association between the difference between population means on a latent variable and the correlation of vectors and that the appearance and precision of this association is moderated by sample size and the standard deviation of factor loadings. In high powered studies, a weak correlation of vectors would falsify, while a strong correlation would not be able to verify, a hypothesis about a sizeable difference between population means.

1. Introduction

According to Jensen (1985, 1998), a correlation between a vector of factor loadings of intelligence tests and a vector of group differences on these same tests (Table 1) indicates that the group difference is mainly in *g* rather than in more specific abilities and that the cause of this difference is the same as the cause of variation within populations. Such a correlation between vectors has been found in various comparisons (see te Nijenhuis et al., 2016, for a review).

It has been argued that some of the findings taken to support the claim that group differences are due to the same cause as variation within populations could be due to violation of the assumption of measurement invariance across groups, something that can lead to a positive correlation between factor loadings and the degree of group differences on the tests even when there is no difference in *g* between the groups (Dolan, Roorda, & Wicherts, 2004). The use of Multigroup Confirmatory Factor Analysis (MGCFA), rather than Jensen's method of correlated vectors, has been recommended with the argument that if variance between groups is due to the same factor as variance within groups, e.g. *g*, factorial invariance across groups should be possible to demonstrate (e.g. Dolan, 2000; Lubke, Dolan, & Kelderman, 2001; Lubke, Dolan, Kelderman, & Mellenbergh, 2003). Re-analyzes using MGCFA have sometimes failed to demonstrate measurement invariance (Dolan et al., 2004). On the other hand, Ashton and Lee (2005) demonstrate that the method of correlated vectors also can fail to reveal a correlation between the vector of *g*-loadings and the vector of correlations between subtests and a variable V even when V has a strong correlation with *g*. Ashton and Lee also show that *g*-loadings depend on the included subtests and this can affect the results obtained through the method of correlated vectors.

Schönemann (1989, 1997) performed simulations and argued that a correlation between the vectors of tests' *g*-loadings and the size of differences between populations is a tautological consequence that will arise if (1) these tests are positively correlated, and (2) people in one of the populations tend to score higher on the tests than people in the other population. However, Schönemann's simulations have, in their turn, been criticized to produce trivial results due to using the scores on the first factor as a selection variable, and it has been argued that a correlation between *g*-loadings and the size of the group differences is *not* a mathematical necessity (Dolan, 1997; Dolan & Lubke, 2001).

The objective of the present simulation was to look for a function that can be used to predict the difference between population means on a latent variable from the correlation between the vectors of factor loadings and group differences on items.

2. Method

Using R 3.2.2 statistical software (R Core Team, 2015), a dataset was simulated through the following steps (code and dataset available as supplementary material): (1) Two samples with between 50 and 12,800 (= eight doublings) persons each were created (the same number in each sample); (2) for both samples, values were randomly drawn from a normally distributed variable T with $SD = 1$ and with a defined difference between population means varying between -1.5 and $+1.5$; (3) fifteen normally distributed items with varying correlations with the variable T were created. The range of the correlations was decreased in steps from 0.1–0.9 to 0.475–0.525 (evenly spaced) in order to vary the standard deviation of the factor loadings. The average

* Corresponding author.
E-mail address: kimmo.sorjonen@ki.se (K. Sorjonen).

http://dx.doi.org/10.1016/j.intell.2017.07.008
Received 13 February 2017; Received in revised form 12 July 2017; Accepted 26 July 2017
Available online 02 August 2017
0160-2896/

Table 1
An example where the correlation between the vector of factor loadings and the vector of group differences equals 0.79 and the standard deviation of loadings equals 0.22.

Test	Loading	Group 1[a]	Group 2[a]	Diff[b]
1	0.2	85 (19)	88 (19)	− 0.16
2	0.3	90 (18)	81 (21)	0.46
3	0.4	95 (12)	94 (17)	0.07
4	0.5	110 (13)	99 (12)	0.88
5	0.6	115 (14)	107 (14)	0.57
6	0.7	120 (15)	106 (12)	1.04
7	0.8	125 (19)	111 (17)	0.78

[a] Group mean (*SD*).
[b] Difference between group means in pooled standard deviations.

correlation was thus kept at approximately 0.5 in all of the simulations; (4) using the psych package (Revelle, 2015) the loadings of these fifteen items on one factor were calculated. As a group difference on the latent factor may inflate factor loadings when using the pooled group (Jensen, 1998), the calculation was conducted in one of the two samples; (5) the correlation of the vectors of factor loadings and group differences on items and the standard deviation of factor loadings were calculated and saved, together with the defined sample size and difference between population means, in a data frame. These five steps were run a thousand times for nine sample sizes, 31 defined differences between population means on the variable *T*, and for 16 different ranges of factor loadings, resulting in 4,464,000 simulated datasets.

3. Results

The difference between population means was found to be an inverted sigmoid function of the correlation of vectors (Fig. 1). A sigmoid curve is given by the following formula:

$$y = \frac{d}{1 + e^{-s \cdot x}} - c$$

In this formula, d = the distance between the floor and the ceiling of the curve; s = the steepness of the curve; c = the ceiling of the curve. In the present case, the function is symmetrical on both sides of zero, giving that $d = 2c$, and the function can be simplified. After inversion we get that the difference between population means (D_P) can be predicted by:

$$D_P = -\frac{1}{s} \cdot ln \left[\frac{2}{\frac{R_V}{c} + 1} - 1 \right]$$

In this formula, R_V stands for the correlation between vectors. An analysis found the values of the s and c parameters in this function to be moderated by the standard deviation of the factor loadings (SD_L) and the sample size (N, in each sample) according to the following:

$$s = 0.883 + 0.000249 \cdot N + 23.8 \cdot SD_L + 0.00853 \cdot N \cdot SD_L$$

$$c = 2.38 - 0.158 \cdot \ln(N) + 1.22 \cdot \ln(SD_L) - 0.142 \cdot \ln(N) \cdot \ln(SD_L)$$

It is also apparent in Fig. 1 that the standard error of the prediction is affected by the standard deviation of the factor loadings, sample size, and the correlation of vectors. In order to avoid negative predictions, the standard error was logarithmized. This $\ln(SE(D_P))$ was found to be a curvilinear function of the correlation of vectors (R_V) and after exponentiation:

$$SE(D_P) = e^{b_0 + b_1 \cdot R_V + b_2 \cdot R_V^2}$$

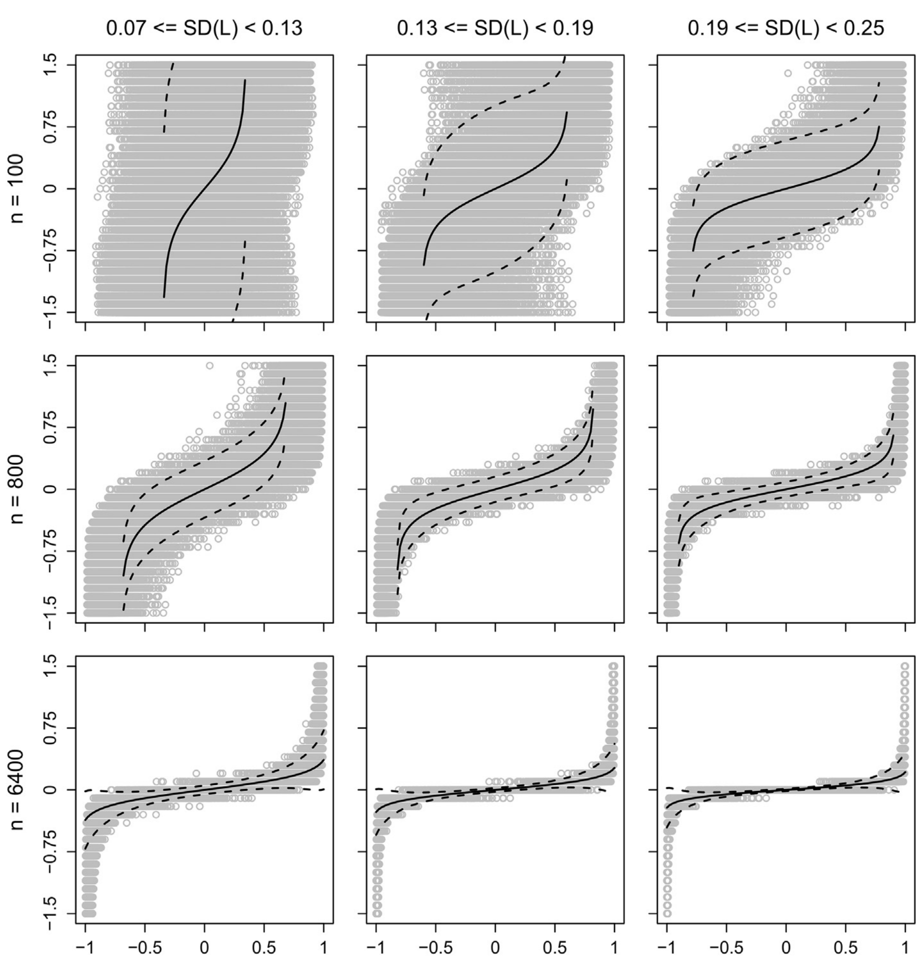

Fig. 1. The association between the difference between population means on a latent variable (y-axis) and the correlation between the vectors of factor loadings and group differences on items (x-axis) for the combinations of three different sample sizes (n in each sample) and three different ranges of standard deviations of factor loadings (*SD*(*L*)). The predicted differences between population means (solid lines) with 95% CI (dotted lines) have been calculated using the formulas presented in the text.