Intelligence 60 (2017) 26–38

Contents lists available at ScienceDirect

Intelligence

Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results)☆

Jelte M. Wicherts

Tilburg University, Netherlands

ARTICLE INFO

Article history:
Received 5 July 2015
Received in revised form 7 November 2016
Accepted 7 November 2016
Available online 18 November 2016

Keywords:
Group differences
Race differences
Psychometrics
Differential Item Functioning
Measurement invariance

ABSTRACT

Spearman's hypothesis stating that ethnic group differences on cognitive tests are most pronounced on the most highly *g* loaded tests has been commonly tested with Jensen's method of correlated vectors (MCV). This paper illustrates and explains why MCV applied to item-level data does not provide a test of measurement invariance and fails to provide accurate information about the role of *g* in group differences in test scores. I focus on studies that applied MCV to study group differences on items of Raven's Standard Progressive Matrices (SPM). In an empirical illustration of the psychometric problems with this method, I show that MCV applied to 60 SPM items incorrectly yields support for Spearman's hypothesis (so-called Jensen Effects suggesting that the group difference is on *g*) even when the items in the second group are not from the SPM but rather from a test composed of 60 items measuring either anxiety and anger or the big five personality traits. This shows that MCV applied to item level data does not accurately reflect the degree to which item bias or *g* plays a role in group differences. I conclude that MCV applied to items lacks both sensitivity and specificity.

1. Introduction

Spearman's hypothesis states that ethnic group differences on cognitive tests are due to *g* (Jensen, 1985), and hence that observed ethnic group differences on these tests cannot be attributed to lower-order cognitive ability factors or measurement bias at the test or item level. Twelve recent studies used the method of correlated vectors (Jensen, 1998) to test Spearman's hypothesis with scores of different ethnic groups on various versions of Raven's Progressive Matrices (Díaz, Sellami, Infanzón, Lanzón, & Lynn, 2012; Rushton, 2002; Rushton, Bons, Vernon, & Cvorovic, 2007; Rushton, Cvorovic, & Bons, 2007; Rushton & Skuy, 2000; Rushton, Skuy, & Bons, 2004; Rushton, Skuy, & Fridjhon, 2002, 2003; te Nijenhuis, Al-Shahomee, van den Hoek, Allik, et al., 2015; te Nijenhuis, Al-Shahomee, van den Hoek, Grigoriev, & Repko, 2015; te Nijenhuis, Bakhiet, et al., 2016; te Nijenhuis, Grigoriev, & van den Hoek, 2016). The goal of these studies was to test whether ethnic group differences were most pronounced on Raven's items that showed the highest loading on *g*. To this end, vectors of ethnic group differences on Raven's items were correlated with vectors representing the degree to which these Raven's items correlated with the *g* factor. Significant correlations from this method of correlated vectors are called Jensen Effects (Rushton, 1998). Jensen Effects are seen as supporting Spearman's hypothesis and are taken to mean that ethnic differences are "not explainable in terms of test bias or in terms of differences in types of item content or other formal or superficial characteristics of the tests" (te Nijenhuis, Al-Shahomee, van den Hoek, Allik, et al., 2015, p.119). Jensen Effects are accorded a central role in the debate on the nature and nurture of ethnic group differences in cognitive ability test performance (Jensen, 1998; Rushton, 2002; Rushton & Jensen, 2005), and are often invoked as evidence in favor of a genetic component to ethnic differences (Rushton, Bons, et al., 2007). Moreover, finding a Jensen Effect is considered relevant for use of the tests in practice because it appears to suggest that the test at hand can be safely used to make inferences about test-takers' latent ability regardless of their ethnic background.

A substantial literature addressed the drawbacks of the method of correlated vectors (Ashton & Lee, 2005; Dolan, 1997, 2000; Dolan & Hamaker, 2001; Dolan & Lubke, 2001; Dolan, Roorda, & Wicherts, 2004; Lubke, Dolan, & Kelderman, 2001; Millsap, 1997; Wicherts & Dolan, 2010; Wicherts & Johnson, 2009), but the method continuous to be used commonly. The goal of this paper is to discuss in non-technical terms the method of correlated vectors (MCV) to study Spearman's hypothesis at the item level. MCV applied to items revolves around item-total correlations as measures of items' loadings on the *g* factor, and the group difference in proportions correct on each item, or, in other words, the group differences in items' p-values. I will discuss drawbacks of the use of such *classical test theory* (*CTT*) item statistics that have been known since the 1940s (Ferguson, 1941; Gulliksen, 1950), and inspired the development of modern item response theory or IRT (Embretson & Reise, 2000; Lord, 1980; Lord & Novick, 1968). A fundamental difference between CTT and IRT is that in the former

☆ Author note: The preparation of this article was supported by a VIDI grant (no. 452-11-004) from the Netherlands Organisation for Scientific Research.

http://dx.doi.org/10.1016/j.intell.2016.11.002
0160-2896/

framework item statistics are operationalized on the basis of observed item scores (here: correct or incorrect) while in IRT the item parameters are defined vis-à-vis the latent ability that the test purports to measure. One crucial implication is that CTT item statistics (like p-values and item-total correlations) are necessarily different between groups that differ in latent ability (Embretson & Reise, 2000), whereas IRT item parameters can be meaningfully compared across groups. IRT allows a rigorous test of whether the items in a scale function equivalently across different groups (i.e., display no Differential Item Functioning or DIF), which is a crucial requirement for any meaningful interpretation of group differences in terms of latent variables such as *g*. Because CTT does not offer tests of measurement invariance that involve latent variables, CTT-based methods (such as MCV) are ill equipped to study whether group differences on item performance can be attributed to the targeted latent variable(s) or to measurement bias. Another problem with CTT is that it focuses on the "true score", which cannot be equated with the construct that the test is supposed to measure (Borsboom & Mellenbergh, 2002). Even nonsensical tests composed of heterogeneous items tapping on widely different constructs have a true score as defined in CTT, as I will illustrate below by adding items from different mood and personality scales. Because MCV at the item level uses this true score as means to operationalize the targeted trait (here *g*) and the degree to which items correlate with that targeted trait (here the *g* loading), MCV could lead to incorrect assessments of the role of *g* in group differences on the items when in fact the true score does not accurately reflect *g*.

In this article, I will first introduce MCV by focusing on how it was originally developed (Jensen, 1980, 1998), namely for studying group differences on subtests from a larger cognitive ability (IQ) test battery with linear factor models. Subsequently, I will discuss four problems with MCV applied to item level data (see also: Wicherts & Johnson, 2009), concerning its inability to test measurement invariance, the group-specificity of item-total correlations, the unwarranted interpretation of item-total correlations as *g* loadings, and the complex nonlinear relations between the vectors in MCV. Finally, I present the results of an empirical study of what happens with MCV if we replace cognitive test items with items from entirely unrelated scales measuring anger, anxiety, and personality in one of the two groups that are being compared. These results are valuable in assessing whether MCV is capable of detecting instances in which item bias and DIF can hardly be any more severe simply because items measure different traits across groups.

2. Method of correlated vectors with subtests

Spearman's hypothesis states that ethnic group differences are due to *g*, implying that the degree to which any cognitive subtest shows group differences can be predicted by the degree to which each subtest measures *g*. In its original form, MCV (Jensen, 1980, 1998) uses *g* loadings based on a factor analysis of the subtests within the two groups that are being compared. Subsequently, these *g* loadings are put in a vector that is as long as the number of subtests. Next, the between-group mean differences on each subtest are computed, and some effect size measure (typically Cohen's *d*) will indicate how strongly the two groups differ on each of these subtests. The crucial test of Spearman's hypothesis in MCV is the correlation between the vector of subtests' *g* loadings and the vector of group differences on the same subtests. Significant MCV correlations (tested against a correlation of 0 using as N the number of subtests) are then called Jensen Effects (Rushton, 1998, 2002).

Jensen (1998) reported that the typical MCV correlation based on cognitive subtests and applied to Black-White differences in the United States is around 0.63. Since that time, a great deal of research addressed the factor analytic technicalities of MCV applied to subtests (Ashton & Lee, 2005; Dolan, 1997, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004; Lubke et al., 2001). The main conclusion from these studies is that MCV applied to subtests might lead to misleading results, or as te Nijenhuis (2013) called it: "The method of correlated vectors is not a strong statistic [sic]" (p. 228).[1]

Several empirical studies (Dolan & Hamaker, 2001; Dolan et al., 2004) used multi-group confirmatory factor analysis (MGCFA) and found that large Jensen Effects with MCV can occur even if *g* is *not* the main (or only) source of the group difference as evidenced by substantial violations of measurement invariance and group differences in first-order factors. These results are problematic because the study of Jensen Effects aims at distinguishing between two alternative hypotheses, one in which *g* explains the group difference (Spearman's hypothesis) and another in which other factors (item bias, subtest-specific abilities, or other non-*g* factors) play a role in the observed group differences in test or item performance. In terms of diagnostics, high sensitivity would imply that if *g* is indeed the only source of the group difference, the MCV correlation should be close to 1. On the other hand, if *g* is not the (only) source of the group difference, the MCV correlation should be close to zero (or perhaps even negative), thereby supporting MCV's specificity. Dolan and colleagues (Dolan & Hamaker, 2001; Dolan et al., 2004) showed both empirically and formally that MCV applied to the subtest level exhibits weak specificity because Jensen Effects can occur even if *g* is not the main source of group differences (a false positive in diagnostic terms). Ashton and Lee (2005) studied scenarios wherein Spearman's hypothesis was true while MCV (at the subtest level) yielded low correlations. In terms of diagnostics, this means that false negatives are likely and hence that MCV applied to the subtest level data has weak sensitivity (which does not mean that true positives or true negatives cannot also occur in MCV; Dolan, 1997; Dolan & Lubke, 2001).

3. MCV does not yield a test of measurement invariance

A comparison of cognitive test scores across groups in terms of latent variables requires that the tests or items are measurement invariant with respect to these groups. Measurement invariance is a core requirement for Spearman's hypothesis stating that groups only differ in the latent variable *g* (Dolan, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004; Lubke, Dolan, Kelderman, & Mellenbergh, 2003a,b; Lubke et al., 2001).

Mellenbergh's (1989) general definition of measurement invariance focuses on the distribution (in his formulation expressed with P) of observed test or item scores *X*, conditional on the latent variable *theta* that the test or item is supposed to measure, and a group indicator *v*. Measurement invariance with respect to groups based on *v* holds if:

$$P(X|theta, v) = P(X|theta). \quad (1)$$

This definition uses conditional distributions (indicated by "P(|)") that describe the distribution of scores on X after we have taken into account the scores on the latent cognitive ability theta within the groups. Specifically, the definition states that the distribution of observed scores X, which is conditional on the latent cognitive ability (theta), does not also depend on the grouping variable *v*. This definition is general as it underlies both tests of measurement invariance in the linear factor model (Meredith, 1993) and tests of measurement invariance at the item level (Holland & Wainer, 1993). When considering items, P in Eq. (1) denotes the chance of answering the item correctly, conditional on the targeted trait (*theta*) and the group indicator *v*. If we replace *theta* with *g*, the definition of invariance offers another way of expressing Spearman's hypothesis for dichotomous items in a scenario where the test measures only *g*. In this hypothetical case, invariance implies that

[1] te Nijenhuis (2013) proposed to combine MCV with psychometric meta-analytic approaches (Hunter & Schmidt, 2004). It is beyond the scope of this paper to discuss these extensions, which have been applied in a number of papers (e.g., te Nijenhuis, Jongeneel-Grimen, & Kirkegaard, 2014), but whose technical specifics have not been studied formally.