Intelligence 57 (2016) 87–95

Contents lists available at ScienceDirect

Intelligence

Testing Spearman's hypothesis with advanced placement examination data

Russell T. Warne *

Department of Behavioral Science, Utah Valley University, United States

ARTICLE INFO

Article history:
Received 8 March 2016
Received in revised form 4 May 2016
Accepted 5 May 2016
Available online 22 May 2016

Keywords:
Spearman's hypothesis
Group differences
Advanced Placement
Academic achievement

ABSTRACT

The nature, source, and meaning of average group score differences between demographic groups on cognitive tests has been a source of controversy for decades. One possible explanation is "Spearman's hypothesis," which states that the magnitude of score differences across demographic groups is a direct function of how strongly the test measures *g*. To test this hypothesis, Jensen (1985, 1998) developed the method of correlated vectors. In this study I used the method of correlated vectors to examine the relationship between racial/ethnic group differences of Advanced Placement (AP) exam scores and the correlation between those AP exam scores and a test of general cognitive ability, the PSAT. Results are consistent with Spearman's hypothesis for White-Black and White-Hispanic comparisons, but not for White-Asian comparisons. Comparisons of White examinees and Native Americans are inconclusive. This study shows that academic achievement tests can be used to test Spearman's hypothesis. Additionally, Spearman's hypothesis is not a unique characteristic of White-Black differences in cognitive test scores, but it may not universal either.

One of the oldest and most consistent findings in intelligence research is the existence of mean score differences among racial and ethnic groups. Generally speaking, examinees descended from East Asians outscore examinees from European populations. These groups' mean scores, in turn, exceed those of populations originating from Latin America, and examinees descended from Sub-Saharan Africans score lowest of all major racial groups (Warne, 2016a; Giessman, Gambrell, & Stebbins, 2013; Gottfredson, 1997; Rushton, 2000). Although not as robust as research on these four racial groups, some researchers have also found an IQ advantage for populations descended from European Jews (Nisbett et al., 2012; te Nijenhuis, David, Metzen, & Armstrong, 2014; Terman, 1926) and lower mean IQ scores for individuals descended from Native Americans (te Nijenhuis, van den Hoek & Armstrong, 2015) and South Asian populations (Lynn & Owen, 1994; Rushton, Čvorović, & Bons, 2007). There is some disagreement over the size of these mean group differences and whether the score gaps are narrowing (Nisbett et al., 2012; Rushton, 2012), but no scholars in the field of human intelligence deny the existence of these score gaps.

However, the cause of these differences is a matter of much discussion. Many members of the public and others unacquainted with modern psychometrics blame the existence of these score differences on test bias (see Warne et al., 2014, for examples), though this explanation is discounted among experts on the topic (Gottfredson, 1997, 2009; Jensen, 1980a, 1980b; Reynolds, 2000). If the tests are not the cause of mean score differences among groups, then the cause possibly arises from differences in ability among groups.

Scholars of human intelligence have long recognized that mean score differences among groups vary from test to test or from subtest to subtest (see Sunne, 1917, for a very early example). To explain this finding, Jensen (1980b, 1985) coined the term "Spearman's hypothesis," which is that the magnitude of group differences on a mental test is a function of the ability of that test to measure *g*. The hypothesis is attributed to Spearman who theorized (1927, pp. 379–380) that differences in mean subtest score differences among Black and White examinees in two prior studies (Derrick, 1920; Pressey & Teter, 1919) were due to the subtests' varying saturation with the *g* factor. According to Spearman—and Jensen—the subtests that were the best measures of *g* would also exhibit the largest Black-White mean score differences. Several studies have shown this to be the case (e.g., Jensen, 1985, 1998; Kane, 2007; te Nijenhuis et al., 2016).

To test Spearman's hypothesis, Jensen (1985, 1998) developed the method of correlated vectors. In its simplest form, the method of correlated vectors "… is one way of testing whether the *g* factor extracted from a battery of diverse tests is related to some variable, *X*, which is external to the battery of tests" (Jensen, 1998, p. 589). In other words, the method of correlated vectors can help researchers determine whether score differences between groups (which is *X* in this case) on cognitive tests are related to the degree that each test measures *g* (as measured by factor loadings or correlation coefficients). If the correlation between *g* saturation and mean score differences is positive, then it is an indication that the data support Spearman's hypothesis. A correlation close to zero

* Department of Behavioral Science, Utah Valley University, 800 W. University Parkway MC 115, Orem, UT 84058., United States.
E-mail address: rwarne@uvu.edu.

http://dx.doi.org/10.1016/j.intell.2016.05.002
0160-2896/

suggests the possibility that the group test score differences and *g* tests' saturation are unrelated and that group score differences are not caused by differences in *g*.

Since Jensen's creation of the method of correlated vectors to examine score differences between White and Black American samples, others have used the procedure to examine whether *g* is at the root of IQ differences among other racial and ethnic groups (e.g., Hartmann, Hye Sun Kruuse, & Nyborg, 2007; Lynn & Owen, 1994; te Nijenhuis et al., 2014; te Nijenhuis, Al-Shahomee, van den Hoek, Grigoriev & Repko, 2015; te Nijenhuis, Willigers, Dragt & van der Flier, 2016; Rushton et al., 2007) and even across species of primates (Fernandes, Woodley, & te Nijenhuis, 2014). Data concerning Black-White differences on cognitive tests generally support Spearman's hypothesis. However, differences among other groups are either inconsistent or there are not enough data to support strong conclusions about whether patterns of test score differences are due to differences in the *g* saturation of the tests (Hartmann et al., 2007). More data concerning the racial and ethnic differences on cognitive test scores is needed to determine the nature of *g* differences among groups and the applicability of Spearman's hypothesis to more demographic groups.

Past research into Spearman's hypothesis has been almost exclusively conducted with cognitive ability/intelligence tests. For example, in Jensen's (1985) original test of Spearman's hypothesis, he used data from the Wechsler scales, the Armed Services Vocational Aptitude Battery, SAT, General Aptitude Test Battery, and other tests of aptitude. In recent years many of researchers studying Spearman's hypothesis have used the Raven's matrices tests (te Nijenhuis, Al-Shahomee, et al., 2015). However, academic achievement tests also show substantial mean score differences among racial and ethnic groups (e.g., ACT, Inc., 2014; Warne, Anderson, & Johnson, 2013), and these achievement tests are sometimes used as measures of *g* in research (e.g., Kura, 2013; Lynn & Meisenberg, 2010). The purpose of this study is to apply Jensen's (1985, 1998) method of correlated vectors to determine whether Spearman's hypothesis is supported with publicly available academic achievement test data. The use of educational data in a study of cognitive ability differences is a new characteristic of research on Spearman's hypothesis. There are two advantages of opening up tests of Spearman's hypothesis to educational achievement data. The first advantage is pragmatic: if Spearman's hypothesis can be shown to apply to achievement tests, then more measures and datasets could be used to test the hypothesis in the future. Experts in human intelligence and education would not be confined to explicitly cognitive tests of *g*, as has been the case with studies of Spearman's hypothesis thus far. Additionally, showing the applicability of Spearman's hypothesis to achivement data would connect the theoretical research on intellectual abilities to new practical, real-world measures and outcomes, especially in education.

The second advantage is theoretical: for decades there has been a debate in the psychometric and education communities about the relationship between academic achievement tests and academic aptitude tests (Merwin & Gardner, 1962; Schmeiser & Welch, 2006). Some theorists claim that there is no difference between the two types of tests (e.g., Corno et al., 2002; Schmeiser & Welch, 2006), while others claim that these are two separate families of tests, though with some overlap in content (e.g., Glaser, 1963; Zwick, 2006). A finding that Spearman's hypothesis applies to achievement tests would indicate that both types of tests measure the same construct. Testing Spearman's hypothesis with achievement data can also make contributions to the theory of education policy and curriculum planning. Understanding the degree to which *g* theory applies to educational achievement can help policy makers and school personnel make informed decisions about the likely success of educational interventions. If the degree of success on academic achievement tests is a product of *g*, then courses and services could be targeted to students based on their likelihood of benefiting from remedial or advanced services. At the individual level, academic programs with relatively low *g* loadings could be opened to a high proportion of students in a school, while programs that require high levels of *g* could be targeted to the most academically prepared students. At the school level, a school's demographics and mean test scores could help personnel decide the most cost-effective use of new resources, and school personnel would maximize the probability of success for an intervention (Warne, 2016a).

In sum, in this study I examined whether Spearman's hypothesis is consistent with the pattern of differences in educational test scores among five large racial and ethnic groups in the United States: Whites, Blacks, Hispanics, Asian Americans, and Native Americans. My goals are to (a) add to the data concerning White-Black differences in test scores of *g*, (b) test the applicability of Spearman's hypothesis to other racial and ethnic groups, and (c) determine whether academic achievement tests can be used to Spearman's hypothesis.

1. Methods

1.1. Data sources

1.1.1. Measure of g

The measure of *g* in this study is the PSAT, a standardized test of academic aptitude created and administered by the College Board for students in Grades 9 and 10. PSAT scores have correlations >0.80 with other recognized measures of *g*, such as the SAT (Proctor & Kim, 2010). The PSAT has three sections: math, critical reading, and writing. For this study, I used the combined math and critical reading score as a measure of *g*.

1.1.2. Academic achievement measures

The academic achievement measures in this study were the Advanced Placement (AP) exams, which are standardized tests that students take at the end of an AP course. AP courses are high school classes with an introductory college curriculum taught by high school teachers. AP exams are scored on a scale of 1 to 5, with a score of 3 generally recognized as a passing score. AP exams are created by the College Board, but the College Board does not grant college credit for passing AP exams. Rather, the student reports AP scores to the college they later attend, and the college decides the minimum score necessary to grant college credit for performance on an AP exam (Lichten, 2000, 2010). The College Board offers 35 AP exams in 26 subjects, ranging from foreign languages to physical sciences and from the arts to mathematics.

1.1.3. Vectors

The vector data in this study come from two public data sources. The first source was a College Board study by Zhang, Patel, and Ewing (2014), which provided the vector representing the correlation of AP exam scores and PSAT scores. In this study about 30% of PSAT examinees later took an AP exam. This is in line with independent data which shows that about 36% of high school graduates completed at least one AP course (see Warne et al., 2015). Zhang et al. (2014) did not report a correlation between the PSAT and AP Italian Language and Culture test, so this test was eliminated from all analyses. Content on the PSAT changes from year to year, yet a comparison of the Zhang et al. (2014) data with similar College Board studies (Camara & Millsap, 1998; Ewing, Camara, & Millsap, 2006) shows that there is little variation from year to year in the PSAT-AP exam score correlations.

The source of the vector containing the demographic group score gaps is College Board data from the 2013–2015 administration years (College Board, 2013, 2014a, 2015a). Data from each year were analyzed separately so that each year of data could serve as a replication for the other years. The College Board only reports numbers of students who earned each score (i.e., from 1 to 5). These ordinal scores were converted to effect size estimates (i.e., Cohen's *d*) through a method that can recover group differences from percentages of examinees that fall into ordinal categories (Ho & Reardon, 2012; Reardon & Ho, 2015). In this procedure, the area under the receiver operating curve is converted into an estimate of group differences. Simulations and real data have