Running head: Spearman's Hypothesis and Alternative *g* tests

Spearman's Hypothesis Is a Model for Understanding Alternative *g* Tests

Michael A. McDaniel Sven Kepes

Virginia Commonwealth University

Paper to be presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology. San Diego. April, 2012

Author Note: The correspondence regarding this manuscript should be sent to Michael McDaniel, Virginia Commonwealth University, 301 West Main Street, PO Box 844000, Richmond, VA 23284. Electronic mail may be sent to mamcdani@vcu.edu.

March 28. 2012 version distributed to panel members and discussants.

Abstract

Periodically, researchers claim to have developed an alternative general cognitive ability (*g*) test that assesses *g* but exhibits lower than typically found mean racial differences. To the extent that such measures have reduced mean racial differences, we argue that they have less *g* saturation (i.e., they measure *g* less well) and will have lower validity and larger prediction errors than tests with high *g* saturation. Using a sample of 22,728 people from the General Aptitude Test Battery database, we show that one can decrease mean racial differences in a *g* test by altering the *g* saturation of the test. Consistent with Spearman's Hypothesis, the *g* saturation of a test is positively, and strongly, related to the magnitude of White-Black mean racial differences in test scores. We demonstrate that the reduction in mean racial differences accomplished by reducing *g* saturation in a test is obtained at the cost of lower validity and increased prediction errors. We recommend that alternative *g* tests be evaluated in comparison to highly *g* saturated tests with respect to Spearman's Hypothesis, test validity, and prediction errors.

> "Facts are stubborn things; and whatever may be our wishes, our inclinations, or the dictates of our passion, they cannot alter the state of facts and evidence." John Adams, 'Argument in Defense of the Soldiers in the Boston Massacre Trials,' December 1770.

Two questions periodically re-appear in the personnel selection literature: (1) What causes one cognitive ability test to be more predictive of job performance than another? and (2) What causes one cognitive ability test to have smaller mean racial differences than another? The first question is substantially settled for those who base their opinions on research findings. The validity of a cognitive test is largely a function of the extent to which the test measures *g* (Olea & Ree, 1994; Ree, Earles, & Teachout, 1994; Thorndike, 1986). Spearman is credited with identifying a general factor of intelligence (*g*), which could be derived from any broad set of cognitive measures (Spearman, 1904, 1927), and the research stream begun by Spearman is labeled as the "psychometric *g*" literature. The second question was directly addressed by Spearman (1927, p. 379), who noted that the magnitude of mean White-Black differences co-varied with the extent to which a test was "saturated with *g*." This positive relationship between the *g* saturation of tests and the magnitude of the tests' White-Black mean differences became known as "Spearman's Hypothesis." Jensen (1985, 1998) reviewed many studies supporting Spearman's Hypothesis. Thus, to accept Spearman's Hypothesis is to adopt the position that one cannot develop a *g* test that measures *g* well (i.e., a test that has a high *g* saturation) and has low White-Black mean differences.

Some researchers have sought to develop measures of *g* that have high *g* saturation *and* low mean racial score differences. We refer to such measures as "alternative *g* tests" to

distinguish them from *g* measures common to the psychometric *g* literature. Researchers of alternative *g* tests seek to find exceptions to Spearman's Hypothesis. Efforts to develop measures of *g* with small mean racial differences are important because mean racial differences in measured abilities are problematic for employers and society (Ployhart & Holtz, 2008).

Attempts to build alternative *g* tests have a long history and have been reviewed by both Jensen (1980, chapter 14) and, more recently, by Goldstein, Scherbaum, and Yusko (2010). These efforts and the reporting of such efforts typically, but not always, share several characteristics. First, the researchers seek to develop alternative *g* measures that both measure *g* and have lower White-Black mean differences than psychometric *g* measures. Second, the researchers often argue that the traditional *g* model consisting of a general factor of intelligence is not the only model of intelligence, but merely "one point of view" (Goldstein et al., 2010, p 102). Third, arguments are made about the possibility of multiple *g* factors (e.g., Cattell, 1963; Sternberg, 1985), the definition of *g* (Sternberg & Detterman, 1986; Wechsler, 1975), the perceived construct narrowness of psychometric *g* measures (Alfonso, Flanagan, & Radwan, 2005; Chen & Gardner, 2005), and the reliance on measures that are asserted to be culturally biased (Fagan, 1992, 2000; Helms-Lorenz, Van de Vijver, & Poortinga, 2003; Jensen, 1980; Sternberg, 1981). Fourth, the alternative *g* measures are seldom empirically compared to *g* with respect to their *g* saturation (i.e., the extent to which they measure *g*), or the relations between their *g* saturation and mean racial differences, consistent with the Spearman's Hypothesis.

Jensen (1980) has argued that efforts to assess *g* without yielding mean racial differences have been unsuccessful. For example, he reviewed the Davis-Eells games, which were developed on the premise that the lower performance of some demographic groups on *g* measures is due to the verbal, abstract, or school-related content of the tests. The Davis-Eells items were cartoons

that required the respondent to apply reasoning to provide an interpretation of the event depicted in the cartoon. Jensen summarized the research on this test by stating that it was "remarkably unsuccessful" at reducing White-Black mean racial differences (1980, p. 643).

Another alternative *g* effort was practical intelligence, also known as tacit knowledge. Sternberg et al. (2000) made two broad claims regarding the nature of practical intelligence. First, they asserted that there exists a general factor of practical intelligence that is distinct from *g*. Second, they declared that practical intelligence predicts success in various domains as well as or better than *g*. These claims were shown to be false: the practical intelligence items do not form a general factor, they are moderately correlated with *g*, and they do not predict criteria better than *g* (Gottfredson, 2003; McDaniel & Whetzel, 2005). Despite the lack of research support, practical intelligence (tacit knowledge) received substantial attention as a "promising" concept in the education, psychology, and management literatures.

Fagan (2000) has contributed to the alternative *g* literature in the educational testing domain. He argued that *g* was best defined as the ability to process information. Further, he argued that psychometric *g* tests depend not solely on processing ability but on what one has been taught. Fagan and Holland (2002, 2007, 2009) reported that when Whites and Blacks had similar exposure to the language (e.g., words, sayings, similarities, and analogies) used in the test, there were only negligible mean racial differences in the processing of the information. For example, in Fagan and Holland (2002), the students were taught the meaning of obscure words (e.g., "venter") and then, shortly later, tested on the meaning of the words that were taught as well as words on which no training was received. Mean racial differences were small for the words for which training was conducted and larger for words for which training was not conducted. These researchers concluded that when there is equal opportunity to learn, there are

no mean racial differences. Although the studies report positive correlations between the tests of learned material and other cognitive measures, one cannot tell from these data the level of *g* saturation in the Fagan and Holland tests. Specifically, the Fagan and Holland tests are not administered in a battery with a sufficient number of *g* scales or tests to derive a *g* factor. Thus, one cannot determine the extent to which Spearman's Hypothesis can explain the findings of the Fagan and Holland studies.

We contrast the research by Fagan and colleagues (Fagan 2000; Fagan & Holland 2002, 2007, 2009) with research on miniature training and evaluation tests (Harris, 1987), also called trainability tests (Roth, Buster & Bobko, 2011). In such tests, applicants receive training concerning skills and knowledge needed for the job for which they are applying. Applicants are then assessed on the trained material. Both Harris (1987), in a set of primary studies, and Roth et al. (2011), in a broader range of studies (that also incorporated the Harris data), reported that such measures show high correlations with *g* and mean racial differences comparable to those found on *g* tests. One possible explanation for the differences with the Fagan studies is that the training component of the Fagan studies tends to be shorter than employment-related trainability tests and targets a narrower domain (e.g., word knowledge). We speculate that the Fagan tests might best be characterized as recall or memory tests. We also speculate that more intensive training, associated with trainability tests in employment settings, is more cognitively demanding and increases the *g* saturation of the trainability tests.

The newest effort in alternative *g* tests is the Siena Reasoning Test (Yusko, Goldstein, Oliver, & Hanges, 2010). We could locate no publications or a test manual on this test. Also, the test publisher did not provide us with copies of past conference presentations on the test or of an alleged technical report that summarizes the results of the test. Material cited here was located on

the internet. The Siena Reasoning Test has been offered as a *g* test that shows smaller mean racial differences than previous measures of *g*. Yusko (2011) asserted that "We have an opportunity to develop cognitive measures that have high validity, high utility, and low adverse impact" (p. 21). Yusko et al. (2010) argued that the Siena Reasoning Test measures cognitive ability and shows reduced mean racial differences because it seeks to reduce reliance on prior knowledge, reduces the use of language, and incorporates graphical stimuli.

Yusko (2011) reported 11 correlations with other cognitive tests that range from .21 to .60 (median r = .35; no sample sizes reported). As summarized by Ones, Dilchert, and Viswesvaran (2012), many employment assessments have mean correlations with *g* near .35, including work samples, assessment centers, situational judgment tests, interviews, biodata, and assessments of openness to experience. We also observe that these measures typically have White-Black mean differences that are lower than *g* tests. We suggest that few would label these types of employment assessments as *g* tests. Also, one may consult Carroll (1993) to identify types of cognitive measures (e.g., memory, processing speed, retrieval, and visual perception) that are not strongly *g*-loaded and, as such, will have correlations with *g* at modest levels. Given this, some may not find it reasonable to label the Siena Reasoning Test as a measure of *g* with substantial *g* saturation.

Some may also question the rationale offered by Yusko et al. (2010) for why the Siena Reasoning Test may have lower White-Black mean differences than psychometric *g* tests (i.e., reductions in reliance on prior knowledge and the use of language, and the incorporation of graphical stimuli). We note that the Davis-Eells tests also sought to limit verbal content and used graphical items but did not substantially reduce White-Black mean racial differences. Likewise, tests such as the Raven's Progressive Matrices and the Advanced Raven's Progressive Matrices

(Raven, Court & Raven, 1994; Raven, Raven, & Court, 1998) do not rely on prior knowledge or language and their items are graphical. The Raven's tests typically show large White-Black mean differences.

Given the failure of past efforts to develop *g* measures that eliminate or minimize mean racial differences, assertions about alternative *g* tests, such as the Siena Reasoning Test and the Fagan and Holland measures (2002, 2007, 2009), are intriguing. If these measures have predictive validity comparable to highly *g* saturated measures *and* lower mean racial differences than highly *g* saturated measures, then they are a major scientific breakthrough. They would also serve as important exceptions to Spearman's Hypothesis.

Following Spearman's Hypothesis, we suggest that alternative *g* tests show smaller mean racial differences than traditional psychometric *g* tests because they are tests with lower *g* saturation. That is, alternative *g* tests should measure *g* less well. Thus, we argue that Spearman's Hypothesis explains the results found for alternative *g* tests more credibly than the explanations offered by the proponents of the alternative *g* tests. In this paper, we evaluate the credibility of Spearman's Hypothesis as an explanation for results common to alternative *g* tests. We do this by creating *g* tests with varying levels of *g* saturation.

When evaluating data with respect to Spearman's Hypothesis and the determination of the *g* loading of tests, there are three classes of issues that should be considered (Carroll, 1993; Floyd, Shands, Rafel, Bergerson, & McGrew, 2009; Major, Johnson, & Bouchard, 2011). The first class of issues concerns sample characteristics. One issue in this class is the size of the sample. In this study, the sample was comprised of 22,728 individuals who were either White or Black. Thus, we have more than an adequate sample size for the precise estimation of statistics. A second issue concerning samples is whether the samples are drawn from occupational settings.

Our data were drawn from the General Aptitude Test Battery (GATB; U.S. Department of Labor, 1970) database, and all data are from occupational settings. The large amount of data collected enhances the likelihood of a representative set of data. Thus, our results should generalize to occupational settings.

The second class of issues relate to the tests used in estimating *g*. One issue is the diversity of tests (Carroll, 1993; Johnson & Bouchard, 2005; McGrew, 2009; Major, Johnson, & Bouchard, 2011, Johnson, te Nijenhuis, & Bouchard, 2008; McGrew, 2009; Reeve & Blacksmith, 2009). A broad array of tests is typically recommended. Test batteries may give too much weight to crystallized ability because fluid abilities tend to be more narrowly defined tasks (e.g., number series), which may have more unique variance (Ashton & Lee, 2005; Kvist & Gustafsson, 2008). This can result in the crystallized components defining more of the common variance of *g* than other components (e.g., fluid intelligence). The nine GATB scales used in this study are drawn from a diverse set of 12 tests: Name comparison, Computation, Three-dimensional space, Vocabulary, Tool matching, Arithmetic reason, Form matching, Mark making, Place, Turn, Assemble, Disassemble. Of these tests, Vocabulary appears to be the sole test that is clearly identifiable as crystallized. Arithmetic reason expresses problems verbally and may have some crystallized variance. Computation is addition, subtraction, multiplication, and division of whole numbers. Tool matching and Form matching are perceptual measures. Name comparison is a speeded perception test. The remaining tests assess psychomotor abilities. Thus, the GATB incorporates a broad range of ability scales. Also in this class of issues is the number of tests used in estimating a *g* factor. Major et al. (2011) reported that a small number of scales or tests tends to inflate the factor loadings. Furthermore, factor loadings tend to be less reliable with few scales or tests. As a result, Major et al. (2011) encouraged the use of at least six to seven

indicators (i.e., scales or tests) per factor. In our study, we have nine scales based on 12 separate tests. Thus, our factors can be well defined and the factor loadings can be well estimated.

The third class of issues concerns the choice of factor extraction method. Both Floyd, Shands, Rafel, Bergerson, and McGrew (2009) and Major et al. (2011) reported that principal components analysis tends to overestimate general factor loadings relative to principal factor analysis. Jensen and Weng (1994) also recommended principal factor analysis. Consistent with these findings and recommendations, we used principal factor analysis.

There are two likely scenarios for building a *g* test with less *g* saturation. Both involve altering the measurement of *g* so that a test assesses *g* less well. First, one can alter the assessment of *g* by building a test with items or scales that have low *g* saturation. Second, one can alter the *g* test by adding random or near random variance to the *g* scores.[1] In this paper, we examine the effectiveness of both methods for reducing mean racial differences in *g* tests. We also assess the extent to which *g* tests with varying levels of *g* saturation predict job performance. Finally, we examine prediction errors as a function of *g* saturation.

Method

Data Source and Measures. The General Aptitude Test Battery (GATB) is a set of nine cognitive scales that are used in various employment contexts. The nine scales are: G – general learning ability, V – verbal aptitude, N – numerical aptitude, S – spatial aptitude, P – form perception, Q – clerical perception, K – motor coordination, F – finger dexterity, and M – manual dexterity. From the U.S. Employment Service, U.S. Department of Labor, we obtained a data set containing GATB scores and job performance data. Data were formed into multiple

[1] This second approach is similar to what is accomplished by adding a measure with low *g* saturation and low validity (e.g., a resume review) to a selection composite containing *g*.

samples consistent with past research by the U.S. Employment Service.[2] In each of these samples, the GATB test battery was administered and job performance data were collected. We retained all samples that contained at least 25 Whites and 25 Blacks, used an identical supervisory performance rating form as the criterion, and had no missing data on any GATB scale or the job performance criterion. This screening yielded 101 samples containing a total of 22,728 individuals who were either White or Black. We did not examine other racial groups due to limited data availability for such groups.

The job performance criterion, labeled "Descriptive Rating Scale," provided one page of instructions to the supervisor(s) who completed the ratings, followed by six rating items with five point anchored rating scales. The six rating items were: quantity of work, quality of work, accuracy of work, knowledge about the job, the variety of tasks that the worker can perform efficiently, and an overall rating of the worker's job performance.

We created two sets of *g* measures, each constructed to vary in their *g* saturation. In the first set, we created a *g* measure based on a factor analysis of the nine GATB scales and used the factor loadings on the first factor to weight the scales, yielding a measure of *g*. Specifically, *g* was defined as shown in equation (1):

$$g = \text{G}*0.91890 + \text{N}*0.84574 + \text{V}*0.79171 + \text{P}*0.77352 + \text{Q}*0.75777 + \text{S}*0.70169 + \text{K}*0.53819 + \text{F}*0.50443 + \text{M}*0.43157 \quad (1)$$

We note that although the GATB scales measure a very diverse set of abilities, all scales loaded on the first factor with more than adequate factor loadings, with the smallest factor loading being .43157. This *g* measure created from all nine GATB scales is the most *g* saturated measure in our study. To complete the first set of *g* measures, we created eight *g* scales altered to

[2] Samples in the GATB database are identified by the variable SATBNO, where SATB is an acronym for "Special Aptitude Test Battery" and NO presumably stands for number.

successively reduce the *g* saturation by removing the scale with the highest loading on the *g* factor from the previous *g* measure. Thus, for the first altered *g* measure, we used the same formula as in equation 1 but did not include the term: G*0.91890. The second altered *g* measure dropped both the G and the N terms. The last altered *g* scale was defined as: M*0.43157. Thus, each successive *g* measure has less *g* saturation than the previous *g* measures in the set.[3]

The second set of *g* measures began with the *g* variable defined by equation 1. We then created 10 more *g* measures that resulted in successively reduced *g* saturation by adding normally distributed random variance to the test scores. Thus, the second *g* measure in this set had its variance increased by 10%, and this additional variance was from a normally distributed random variable. We then created additional *g* measures by adding normally distributed random variance in increments of 10%. Note that the *g* measure, labeled 100% in our results section (see Table 2), has twice the variance of the original *g* measure. As is the case with the first set of *g* measures, each successive *g* measure in this second set of measures has less *g* saturation than the previous *g* measures in the set.

In summary, we created two sets of *g* measures. In the first set, *g* saturation was altered by removing *g* loaded scales from the *g* measure. In the second set, *g* saturation was altered by adding random variance to the *g* measures. These two approaches to reducing *g* saturation are consistent with our assertions concerning two primary ways of reducing mean racial differences in *g* measures.

[3] Removing any scale from equation 1 reduces the *g* saturation of the resulting measure. By removing the highest *g* loaded scale from the *g* factor expressed in equation 1 from the previous *g* measure, we are producing the largest possible decline in *g* saturation between each *g* measure. We note that we could have dropped the *g* saturation of the successive *g* variables by dropping the least *g* saturated scale from the previous *g* measure. However, that would have reduced *g* saturation in successive measures much less effectively. For example, the last remaining *g* scale would have consisted of the GATB G scale which has a correlation of .89 with the most *g* saturated scale. As seen in Table 1, removing the highest *g* loaded scale results in a set of measures with substantial variability in *g* saturation.

Analysis. We estimated the saturation of each individual *g* measure by correlating the measures with the *g* measure consisting of the weighted composite of all nine GATB scales. Measures with high correlations with the nine GATB scale composite have greater *g* saturation than measures with low correlations with the nine GATB scale composite, the *g* measure with the highest *g* saturation. We calculated the validities of the *g* measures and the standardized mean differences between the Whites and the Blacks on the *g* measures. For each *g* measure, we estimated regression equations where the *g* measure is the independent variable, and job performance is the dependent variable. These regressions used the White and Black data combined, yielding the common regression lines. We then calculated the amount of error of prediction by race and expressed it in standard deviation units of the criterion.

All analyses were conducted twice. In the first set of analyses, we formed one sample based on all 22,728 individuals. In the second set of analyses, we conducted the analyses separately for each of the 101 individual samples and then calculated the sample-size-weighted mean of the statistics across samples.

Results

Table 1 contains the results of all analyses for the *g* measures that were altered by successively removing the scale with the highest loading on the *g* factor. Thus, the first row of the results shows the findings for the *g* measure composed of all nine GATB scales. This measure has the most *g* saturation. The second row displays the results of analyses with a *g* measure containing eight GATB scales. This second measure has less *g* saturation than the measure based on nine GATB scales (the G scale was dropped). The last row of the table shows a *g* variable composed solely of the M scale of the GATB, and this measure has the least *g* saturation. The first column of the table shows the number of scales in the *g* measure. The

second column indicates which scale(s) was (were) dropped from the *g* measure. The third column shows the correlation of each resulting *g* measure with the nine GATB scale composite (i.e., the *g* measure with the highest *g* saturation). This correlation is an indicator of the *g* saturation of each respective measure. The fourth column presents the correlation between each individual *g* measure and the job performance criterion. These correlations were based on all 22,728 individuals being considered as one sample. The sample-size-weighted mean correlations based on the 101 individual samples are shown in parentheses. The same practice of showing the estimates for the overall sample and the 101 individual samples is followed for the remaining columns in the table. Because the results are nearly identical (see Table 1), we only discuss the statistics from the 22,728 individuals considered as one sample. The fifth column shows the White-Black standardized mean difference in the *g* measures. A positive *d* indicates that Whites scored higher than Blacks, on average. The last two columns show the prediction error in criterion standard deviation units with one column showing the mean prediction errors for Whites and the other showing the mean prediction errors for Blacks.

As seen in Table 1, the validities of the *g* measures drop from .216 to .106, a reduction of .11 or 51%, as the *g* saturation of the measures is successively reduced by dropping the highest *g*-loaded scale from the previous *g* measure. Accompanying the drop in validity is a drop in the standardized mean differences between Whites and Blacks on each *g* measure. Specifically, the *d* is .837 for the most *g* saturated measure and .251 for the least *g* saturated measure, a decrease of .586 or 70%. Prediction errors increase as the *g* saturation of the measure is reduced. The White prediction errors are negative, indicating that the common regression line under-predicts the job performance of Whites, on average. The Black prediction errors are positive, indicating that the common regression line over-predicts the job performance of Blacks, on

average. Note that the over-prediction of job performance for Blacks is larger than the under-prediction of job performance for Whites.

Table 2 displays the results for measures in which the *g* saturation was reduced by adding random variance to the *g* measures. This table has the same format as Table 1. Note that the first row of Tables 1 and 2 show the same results because the first *g* measure is the same in both tables (the measure calculated based on equation 1). As the percentage of random variance added to each *g* measure increases, the validity drops from .216 to .152, a reduction of .064 or 30%. Consistent with Table 1, as the *g* measures are reduced in their *g* saturation, the validity and the White-Black mean differences decline, but the prediction errors increase.

Table 3 shows the correlations of the variables from Tables 1 and 2. The bottom triangle of Table 3 presents the intercorrelations of variables in Table 1, and the top triangle presents the intercorrelations from the variables in Table 2. This correlation table is at the unit of a test, and thus correlations in the bottom triangle are based on a sample of nine, the nine different *g* measures from Table 1, and the upper triangle correlations are based on a sample of 11, the 11 different *g* measures from Table 2. Consistent with Spearman's Hypothesis, as the *g* saturation of the *g* measures decrease so do the mean racial differences on *g*. The magnitude of this correlation is near perfection (all correlations in the table are in the absolute value range of .95 to 1.00). The correlations between validity and prediction errors for Whites are positive. This means that as validity increases, the negative prediction errors for Whites move toward zero (i.e., the prediction errors get smaller). The correlations between validity and prediction errors for Blacks are negative, indicating that as the validity increases, the positive prediction errors for Blacks move toward zero (i.e., the prediction errors get smaller). In brief, one can reduce the magnitude

of White-Black mean differences in *g* by lowering the *g* saturation of the measure but at the cost of lower validity and larger prediction errors.

Discussion

Psychology has a long history of attempts to develop alternative *g* measures that have low mean racial differences. All tests associated with these claims, which have been carefully evaluated, have failed in their goal of being both excellent *g* measures (measures with high *g* saturation) and having reduced mean racial differences. We have offered two ways to reduce the *g* saturation of tests so as to yield lower mean racial differences. One method is to reduce *g* saturation by removing *g*-relevant variance from the measure. The second way is to reduce *g* saturation by adding random variance to the *g* measure. We recognize that few organizations would add random variance to a *g* measure to reduce its *g* saturation. However, many organizations do add measures, such as resume reviews or poorly developed and administered interviews, with relatively low correlations with job performance to supplement their *g* measure's prediction of job performance. Such practices can reduce the *g* saturation of the composite measure and they may have similar effects in reducing mean racial differences as adding random variance to a *g* measure.

Our assertions concerning the reduction of *g* saturation in alternative tests that purport to measure *g* can be empirically evaluated. When factor analyzed with a set of scales of varying *g* saturation, the alternative *g* test(s) can be expected to have lower loadings on the first factor (the *g* factor) than highly *g* saturated tests. If criterion data are available, the validity of the alternative *g* test(s) should be lower than the validity of a highly *g* saturated test. Likewise, the prediction errors for the highly *g* saturated tests should be smaller than for the alternative *g* tests. The loading of the various *g* tests can be correlated with White-Black mean score differences, and the

mean differences should co-vary with the *g* saturation of the test(s). Similarly, the prediction errors of the *g* test can be evaluated as a function of their *g* saturation.

Our results indicate that one can build a *g* test, albeit with less *g* saturation, that has lower mean racial differences than a highly *g* saturated test. However, the reduction of mean racial differences is at the cost of lower validity and larger prediction errors. We also show that one can easily build a *g* measure with lower *g* saturation by removing subtests with the most *g* saturation or by adding random variance to the scores.

These results have implications for U.S. employment regulations, which mandate that if two selection procedures have the same validity, one should use the selection procedure with the lower mean racial differences. Our large sample results suggest that one would not find two *g* tests with equal validity where one has lower mean racial differences. Thus, barring large sample credible research to the contrary, there is unlikely to be a situation in which there is a legal requirement to use an alternative *g* test. Employers could of course use an alternative *g* test to reduce mean racial differences and its use would result in lower validity and greater prediction errors.

The ease with which one can alter the saturation of a *g* test limits the need to purchase commercially available tests with low *g* saturation. One can simply take a highly *g* saturated test and damage (i.e., reduce) its *g* saturation. Or, one could build a test using less *g* saturated scales. For example, one can build less *g* saturated measures by considering Carroll's (1993) Three Stratum theory of intelligence. Carroll's (1993, p. 627) figure 15.1 graphically displays the *g* saturation of various cognitive abilities such that those most related to *g* are on the left of the graph and those least related to *g* are on the right of the graph. Thus, a *g* measure drawing on abilities to the right of the graph (e.g., processing speed, retrieval, perception) can be expected

to have lower *g* saturation than a *g* measure drawing on abilities from the left side of the graph (e.g., fluid intelligence and crystalized intelligence). For example, Barrett, Carobine, and Doverspike (1999) found smaller mean racial differences for a short-term memory test (d = .39), a less *g* saturated test, than a reading comprehension test (d = .80), a more *g* saturated test. One could also factor analyze a set of *g* items and remove the items with the highest loading on the *g* factor. In addition, one could examine item level mean racial differences and remove the items with the largest mean racial differences. Either of these last two approaches should reduce the *g* saturation of the test. Unfortunately, any approach that reduces the *g* saturation of the test may inevitably reduce validity and increase prediction errors.

Attempts to identify alternative *g* tests with predictive values as large as highly *g* saturated tests and with no or minimal mean racial differences is a laudable goal worthy of research attention. However, there are decades of research suggesting that it is unlikely that one can build a highly *g* saturated test without substantial mean racial differences. Thus, when one argues that an alternative *g* test is as good of a predictor of job performance as highly *g* saturated tests but has smaller mean racial differences, one should expect, based on decades of research, substantial skepticism. Bold claims should be accompanied by substantial research support. To date, no alternative *g* test has offered such research support.

Spearman's Hypothesis is offered as the best model for explaining the lower magnitude mean racial differences that are found or claimed for alternative *g* tests. We recognize that this explanation is not appealing to authors of alternative *g* tests, and we acknowledge that other explanations are possible. Because Spearman's Hypothesis explanation is easy to evaluate empirically, and is consistent with decades of research, we encourage alternative *g* research to

empirically rule out this explanation prior to proceeding with other possible explanations that are typically offered but seldom evaluated empirically.

References

Alfonso, V. C., Flannagan, D. P., & Radwan, S. (2005). The impact of the Cattell-Horn-Carroll theory on test development, and interpretation of cognitive and academic performance. In D.P. Flannagan & P.L. Harrison (Eds.) *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 185-202). New York: Guilford Press.

Ashton, M. C., & Lee, K. (2005). Problems with the method of correlated vectors. *Intelligence, 33*, 431-444. doi: 10.1016/j.intell.2004.12.004

Barrett, G. V, Carobine, R. G., & Doverspike, D. (1999). The reduction of adverse impact in an employment setting using a short-term memory test. *Journal of Business and Psychology, 14*, 373-377. doi: 10.1023/a:1022107611888

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22. doi: 10.1037/h0046743

Chen, J., & Gardner, H. (2005). Assessment based on multiple-intelligence theory. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, test, and issues* (2nd ed., p. 77-102). New York: Guilford Press.

Fagan, J. F. (1992). Intelligence: A theoretical viewpoint. *Current Directions in Psychological Science, 1*, 82-86. doi: 10.1111/1467-8721.ep10768727

Fagan, J. F. (2000). A theory of intelligence as processing: Implications for society. *Psychology, Public Policy & Law, 6*, 168-179. doi: 10.1037/1076-8971.6.1.168

Fagan, J. F., & Holland, C. R. (2002). Equal opportunity and racial differences in IQ. *Intelligence, 30*, 361-387. doi: 10.1016/s0160-2896(02)00080-6

Fagan, J. F., & Holland, C. R. (2007). Racial equality in intelligence: Predictions from a theory of intelligence as processing. *Intelligence, 35*, 319-334. doi: 10.1016/j.intell.2006.08.009

Fagan, J. F., & Holland, C. R. (2009). Culture-fair prediction of academic achievement. *Intelligence, 37*, 62-67. doi: 10.1016/j.intell.2008.07.004

Floyd, R. G., Shands, E. I., Rafael, F. A., Bergeron, R., & McGrew, K. S. (2009). The dependability of general-factor loadings: The effects of factor extraction methods, test battery composition, test battery size, and their interactions. *Intelligence, 37*, 453-465. doi: 10.1016/j.intell.2009.05.003

Goldstein, H. W., Scherbaum, C. A., & Yusko, K. (2009). Adverse impact and measuring cognitive ability. In J. Outtz's (Ed.) *Adverse impact: Implications for organizational staffing and high stakes testing* (pp. 95-134). New York: Psychology Press.

Gottfredson, L. S. (2003). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence, 31*, 343-397. doi: 10.1016/s0160-2896(02)00085-5

Harris, P. A. (1987). *A final report on the miniature training and evaluation test*. Washington, DC: U.S. Office of Personnel Management.

Helms-Lorenz, M., Van de Vijver, F. J., & Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance, and Spearman's Hypothesis: *g* or *c*? *Intelligence, 31*, 9-29. doi: 10.1016/s0160-2896(02)00111-3

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jensen, A. R. (1985). The nature of the black-white differences on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8*, 193-263. doi: 10.1017/S0140525X00020392

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Jensen, A. R., &Weng, L-J. (1994).What is a good g? *Intelligence, 18*, 231–258. doi: 10.1016/0160-2896(94)90029-9

Johnson, W., & Bouchard, T. J., Jr. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence, 33*, 393–416.

Johnson, W., te Nijenhuis, J., & Bouchard, T. J., Jr. (2008). Still just 1 g: Consistent results from five test batteries. *Intelligence, 36*, 81-95. doi: 10.1016/j.intell.2004.12.002

Kvist, A. V., & Gustafsson, J.-E. (2008). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's investment theory. *Intelligence, 36*, 422-436. doi: 10.1016/j.intell.2007.08.004

Major, J. T., Johnson, W., & Bouchard, T. J. (2011). The dependability of the general factor of intelligence: Why small, single-factor models do not adequately represent g. *Intelligence, 39*, 418-433. doi: 10.1016/j.intell.2011.07.002

McDaniel, M. A., & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence*, 33, 515-525. doi: 10.1016/j.intell.2005.02.001

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1–10. doi: 10.1016/j.intell.2008.08.004

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology, 79*, 845-851. doi: 10.1037/0021-9010.79.6.845

Ones, D. S., Dilchert, S., & Viswesvaran, C. (2012). Cognitive abilities. In N. Schmitt (Ed). *The Oxford handbook of personnel assessment and selection* (pp. 179-224). New York: Oxford University Press.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*, 153-172. doi: 10.1111/j.1744-6570.2008.00109.x

Potosky, D., Bobko, P., & Roth, P. L. (2005). Forming composites of cognitive ability and alternative measures to predict job performance and reduce adverse impact: Corrected estimates and realistic expectations. *International Journal of Selection and Assessment, 13*, 304-315. doi: 10.1111/j.1468-2389.2005.00327.x

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices*. Oxford, UK: Oxford Psychologists Press Ltd.

Raven, J. C., Court, J. H., & Raven, J. (1994). *Advanced progressive matrices: Sets I and II. Manual for Raven's progressive matrices and vocabulary scales.* Oxford, UK: Oxford Psychologists Press.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79*, 518-524. doi: 10.1037/0021-9010.79.4.518

Reeve, C. L., & Blacksmith, N. (2009). Equivalency and reliability of vectors of g-loadings across different methods of estimation and sample sizes. *Personality and Individual Differences, 47*, 968-972. doi: 10.1016/j.paid.2009.07.025

Roth, P. L., Buster, M. A., & Bobko, P. (2011). Updating the trainability tests literature on Black-White subgroup differences and reconsidering criterion-related validity. *Journal of Applied Psychology, 96*, 34-45. doi: 10.1037/a0020923

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: Macmillan.

Sternberg, R. J. (1981). Intelligence and non-entrenchment. *Journal of Educational Psychology, 73*, 1-16. doi: 10.1037/0022-0663.73.1.1

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R. J., & Detterman, D.K. (1986). *What is intelligence?: Contemporary viewpoints on its nature and definition.* Norwood, NJ: Ablex.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., et al. (2000). *Practical Intelligence in Everyday Life*. New York: Cambridge University Press.

Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior, 29*, 332-339. doi: 10.1016/0001-8791(86)90012-6

U.S. Department of Labor, U.S. Employment Service. (1970). *General Aptitude Test Battery, Section III: Development*. Washington, D.C.: U.S. Government Printing Office.

Wechsler, D. (1975). Intelligence defined and undefined: A relativistic appraisal. *American Psychologist, 30*, 135-139. doi: 10.1037/h0076868

Yusko, K. P., Goldstein, H. W., Oliver, L. O., & Hanges, P. J. (2010). *Building cognitive ability tests with reduced adverse impact: Lowering reliance on prior knowledge*. Paper presented at the 25th Annual Conference of the Society for Industrial and Organizational Psychology. Atlanta.

Table 1. Validity, mean racial differences, and prediction errors as a function of reducing *g* saturation by removing scales.

Number of scales in *g* measure	Scale(s) dropped from measure	Correlation with most saturated *g* test	Validity of *g*		White-Black *d* on *g*		Mean White (under) Prediction Error in SD units		Mean Black (over) Prediction Error in SD units	
9	None	1.00	.216	(.205)	.837	(.836)	-.055	(-.055)	.118	(.117)
8	G	0.99	.210	(.198)	.785	(.784)	-.060	(-.059)	.129	(.127)
7	G, N	0.97	.197	(.182)	.722	(.718)	-.067	(-.066)	.144	(.141)
6	G, N, V	0.94	.186	(.170)	.645	(.631)	-.075	(-.073)	.160	(.156)
5	G, N, V, P	0.92	.187	(.170)	.621	(.601)	-.076	(-.075)	.162	(.159)
4	G, N, V, P, Q	0.84	.165	(.150)	.563	(.522)	-.083	(-.081)	.178	(.174)
3	G, N, V, P, Q, S	0.69	.136	(.119)	.298	(.283)	-.100	(-.094)	.214	(.201)
2	G, N, V, P, Q, S, K	0.61	.122	(.113)	.334	(.311)	-.100	(-.094)	.214	(.200)
1	G, N, V, P, Q, S, K, F	0.50	.106	(.104)	.251	(.219)	-.104	(-.097)	.223	(.207)

Notes: Nine *g* measures were calculated based on a factor analysis of the nine GATB scales. Scales were weighted by their factor loadings. The *g* measure based on all nine scales was iteratively altered by dropping the highest loading GATB scale from the previous measure thus making each successive measure less *g* saturated then the previous measures. All statistics were calculated in two ways. The first way was to treat the 22,728 observations as one sample. A second yields the results in parentheses. In this second approach, the data were into 101 samples based on SATB number. The statistics were calculated in each of the 101 samples. The sample-size-weighted mean of these statistics yielded the statistics in parentheses.

Table 2. Validity, mean racial differences, and prediction errors as a function of reducing *g* saturation by adding random variance.

% increase in variance due to adding random numbers to the nine variable *g* measure	Correlation with most saturated *g* test	Validity of *g*		White-Black *d* on *g*		Mean White (under) Prediction Error in SD units		Mean Black (over) Prediction Error in SD units	
0%	1.00	.216	(.205)	.837	(.836)	-.055	(-.055)	.118	(.117)
10%	.95	.205	(.194)	.799	(.788)	-.061	(-.060)	.129	(.128)
20%	.91	.197	(.184)	.766	(.747)	-.065	(-.064)	.139	(.138)
30%	.88	.189	(.176)	.736	(.712)	-.069	(-.068)	.147	(.145)
40%	.85	.182	(.168)	.709	(.681)	-.072	(-.071)	.153	(.152)
50%	.82	.176	(.161)	.686	(.654)	-.074	(-.074)	.159	(.157)
60%	.79	.170	(.156)	.664	(.631)	-.077	(-.076)	.164	(.162)
70%	.77	.165	(.150)	.644	(.609)	-.079	(-.076)	.169	(.166)
80%	.75	.161	(.146)	.626	(.590)	-.081	(-.079)	.173	(.169)
90%	.73	.156	(.141)	.610	(.572)	-.083	(-.081)	.176	(.172)
100%	.71	.152	(.137)	.594	(.556)	-.084	(-.082)	.180	(.175)

Notes: Nine *g* measures were calculated based on a factor analysis of the nine GATB scales. The *g* measure based on all nine scales was iteratively altered by adding increasing amounts of random variance thus making each successive measure less *g* saturated then the previous measures. All statistics were calculated in two ways. The first way was to treat the 22,728 observations as one sample. A second yields the results in parentheses. In this second approach, the data were into 101 samples based on SATB number. The statistics were calculated in each of the 101 samples. The sample-size-weighted mean of these statistics yielded the statistics in parentheses.

Table 3. Correlation table for altered *g* measures.

	1	2	3	4	5
1. Validity		1.00	1.00	1.00	-1.00
2. Correlation with most saturated g measure	.99		1.00	1.00	-1.00
3. White-Black *d* on *g*	.98	.96		1.00	-1.00
4. White prediction error (one sample)	.98	.95	0.99		-1.00
5. Black prediction error (one sample)	-.98	-.95	-0.99	-1.00	

Note. The correlations in the bottom half of the table are based on the nine *g* measures where *g* was altered through the sequential removal of *g*-loaded tests. The correlations in the top half of the table are based on the 11 *g* measures which were altered by sequentially adding random variance to the test scores. The correlations are based on the analysis where all observations were considered one sample.