An Evaluation of Spearman's Hypothesis by Manipulating *g* Saturation

Michael A. McDaniel and Sven Kepes

Department of Management, Virginia Commonwealth University, 301 West Main Street, PO Box 844000, Richmond, VA 23284-4000, USA. mamcdani@vcu.edu

Spearman's Hypothesis holds that the magnitude of mean White–Black differences on cognitive tests covaries with the extent to which a test is saturated with *g*. This paper evaluates Spearman's Hypothesis by manipulating the *g* saturation of cognitive composites. Using a sample of 16,384 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 measure. 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 the *g* saturation in a measure is obtained at the cost of lower validity and increased prediction errors. We recommend that *g* tests varying in mean racial differences be examined to determine if the Spearman's Hypothesis is a viable explanation for the results.

1. Introduction

Two questions periodically reappear 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. Regarding the first question, scientific evidence indicates that the validity of a cognitive test is largely a function of the extent to which the test measures *g* (Gottfredson, 2002; Olea & Ree, 1994; Ree, Earles, & Teachout, 1994; Sackett, Schmitt, Ellingson, & Kabin, 2001; Schmidt, 2002; Thorndike, 1986). Spearman is credited with identifying a general factor of intelligence (*g*) that could be derived from any broad set of cognitive measures (Spearman, 1904, 1927), and the research stream began by Spearman is often 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 covaried 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.

Typically, studies concerning Spearman's Hypothesis examine intact test composites for the relation between *g* saturation and the mean racial differences of the tests comprising the composite. In this study, we take a different approach. We manipulate a 9-scale test composite to create multiple measures (composites) that vary in *g* saturation. We evaluate Spearman's Hypothesis by examining how mean Black–White score differences covary with *g* saturation. In addition, we also consider how differences in *g* saturation affect criterion-related validity and prediction errors. Differences across the composites cannot be attributed to sample differences because all composites are based on data from the same sample. Furthermore, because we are manipulating the *g* saturation of the composites, we can have greater confidence that the observed effects are due to *g* saturation and not other factors. In addition, we use two approaches to alter the *g* saturation of the composites to evaluate whether our results are due to a particular method. Thus, we offer these analyses as a unique approach to evaluating Spearman's Hypothesis and argue that it represents a unique contribution to the Spear-

International Journal of Selection and Assessment

man's Hypothesis literature. We conclude the paper with a discussion of the usefulness of Spearman's Hypothesis in understanding the results of *g* tests that appear to have lower than typical Black–White mean differences.

When evaluating data with respect to Spearman's Hypothesis and the determination of the *g* saturation of tests, there are three classes of issues that should be considered (Carroll, 1993; Floyd, Shands, Rafael, Bergeron, & 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 samples were selected such that each sample had at least 25 Whites and 25 Blacks, and summed across samples, we had data for 16,384 individuals. Thus, we have more than an adequate sample size for the precise estimation of statistics. The 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 enhances the likelihood of a representative set of data. Thus, our results should generalize to occupational settings.

The second class of issues relates to the tests used in estimating *g*. One issue is the diversity of tests (Carroll, 1993; Johnson & Bouchard, 2005; McGrew, 2009; Reeve & Blacksmith, 2009). A broad array of tests is typically recommended. Test composites may give too much weight to crystallized ability because fluid ability tests 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 tend to inflate the factor loadings. Furthermore, factor loadings tend to be less reliable with few scales or tests. As a result, Major et al. 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 et al. (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* composite to reduce its *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 dropping scales with high *g* saturation. This approach lowers the *g* saturation of a composite by excluding scales with the best *g* saturation. Second, one can alter the *g* saturation of the composite by adding random or near random variance to the composite.[1] This approach lowers the *g* saturation by reducing the reliability of the composite. In this paper, we use both methods in the evaluation of Spearman's Hypothesis.

When reducing the number of tests in a composite, one might expect the reliability of the resulting composite to be smaller than the reliability of a composite with a larger number of tests. If the reliability drops as the number of tests in the composite drops, reliability decrements may be a cause of the decline in validity and mean group racial differences. Thus, effects attributed to the Spearman's Hypothesis may simply be a result of increases in measurement error. In this paper, we empirically address the credibility of this argument.

2. Method

2.1. Data source and measures

The 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 samples consistent with past research by the U.S. Employment Service.[2] In each of these samples, the GATB 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 of at least 25 Whites and 25 Blacks with a total of 16,384 individuals.[3]

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 5-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.

2.1.1. Sets of g composites

We created two sets of *g* composites, each constructed to vary in their *g* saturation. In the first set, we created a *g* composite 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):

$$\begin{aligned} g = {} & G^*.91890 + N^*.84574 + V^*.79171 \\ & + P^*.77352 + Q^*.75777 + S^*.70169 \\ & + K^*.53819 + F^*.50443 + M^*.43157 \end{aligned} \quad (1)$$

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* composite created from all nine GATB scales is the most *g* saturated composite in our study. To complete the first set of *g* composites, we created eight *g* composites altered to successively reduce the *g* saturation by removing the scale with the highest loading on the *g* factor from the previous *g* composite. Thus, for the first altered *g* composite, we used the same formula as in Equation (1) but did not include the term: *G**.91890. The second altered *g* composite dropped both the *G* and the *N* terms. The last altered *g* composite was defined as: *M**.43157. Thus, each successive *g* composite has less *g* saturation than the previous *g* composite in the set.[4]

The second set of *g* composites began with the *g* composite defined by Equation (1). We then created 10 more *g* composites that resulted in successively reduced *g* saturation by adding normally distributed random variance to the test scores. Thus, the second *g* composite 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* composites by adding normally distributed random variance in increments of 10%. Note that the *g* composite labeled 100% in our Results section (see Table 3) has twice the variance of the original *g* composite. As is the case with the first set of *g* composites, each successive *g* composite in this second set has less *g* saturation than the previous *g* composites in the set.

2.1.2. Reliability of composites

The second set of *g* composites reduces *g* saturation by increasing measurement error through the introduction of random variance into the composites. However, in the first set of *g* composites, reliability may decline by removing scales from each successive composite. We addressed this by calculating the reliability of each composite. To calculate the reliability of a weighted composite, one needs to know the reliabilities of each scale. The GATB reports two samples that include all nine scales and report reliabilities (U.S. Department of Labor, 1970, Manpower Administration; see tables 15-4 and 15-5; total $N = 1{,}159$). We calculated the sample-weighted mean of the two reliabilities for each scale and used those values in calculating the reliabilities of the composites. We then calculated the reliability of the weighted composites using the Feldt and Brennan (1989) formula as reported in He (2009). We did this for the composites containing between two and nine scales. The last 'composite' consists of only one scale (GATB *M*) and we simply use the reliability of that scale (.73).

In summary, we created two sets of *g* composites. In the first set, *g* saturation was altered by removing the most *g* saturated scales from the previous *g* composite. In the second set, *g* saturation was altered by adding random variance to the *g* composites.

2.2. Analysis approach

We estimated the *g* saturation of each composite by correlating the composite with the *g* composite built from all nine GATB scales. Composites with high correlations with the 9-scale composite have greater *g* saturation than composites with low correlations with the 9-scale composite. We calculated the criterion-related validities of each *g* composite's score and the standardized mean differences between the Whites and the Blacks on the *g* composite's score. For each *g* composite, we estimated regression equations in which the *g* composite score 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 16,384 individuals in the data set. 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.

3. Results

Our first approach to reducing the *g* saturation of the scale composites was to successively drop the most *g* saturated scale. But, were any of the resulting effects partially a function of reducing the reliability of the composite by reducing the number of scales in the composite? A key consideration in evaluating this possibility is the recognition that the reliability of a weighted composite depends on the intercorrelations of the scales. Although all the GATB scales have moderate to high *g* loadings, the GATB uses a broad bandwidth of scales that group into three clusters (GVN, PQS, KFM). GVN is a strong cognitive grouping. PQS is a spatial–perceptual set of scales; it is still a cognitive composite in that the scales have good loadings on *g* (the factor loadings are .77, .76, and .70), but the scales emphasize spatial-perceptual ability. KFM is a set of psychomotor scales. It also has good cognitive loadings (the factor loadings are .54, .50, .43), but the scales emphasize psychomotor ability.

Table 1 shows the reliabilities of the nine composites. Note that the reliabilities do not decline in a monotonic fashion as might be expected by reducing the number of scales per composite. This nonmonotonic pattern of reliabilities is due to the intercorrelations of the variables retained in the composite and the intercorrelations are highest for variables in the same cluster and lower for variables across clusters. The reliability of the 9-scale composite is .92. When the *G* scale is dropped from the composite resulting in the 8-scale composite, the reliability drops to .91. When the *N* scale is removed (in addition to the *G* scale removed earlier), resulting in the 7-scale composite, the reliability drops to .90. However, when the *V* scale is removed (in addition to the *G* and *N* scales, which were removed earlier) resulting in the 6-scale composite, the reliability returns to .92.

This increase in reliability reflects the changing composition of the battery and the intercorrelations of the remaining scales when the *G*, *V*, and *N* scales are dropped. *G*, *V*, *N* are scales designed to measure *G* (a scale of general intelligence), *V* (verbal aptitude), and *N* (numerical aptitude). In the 6-scale composite, *G*, *V*, and *N* have been dropped; this removes the most *g* loaded cluster of the GATB, and leaves the spatial–perceptual scale cluster (PQS) and psychomotor scale cluster (KFM). When *V* is dropped, the resulting 6-scale composite has a higher reliability because the average intercorrelations (mean correlation: .43) are larger than the correlations within the 7-scale composite (mean correlation: .42). The same scenario occurs at the 3-scale composite. At this point, we have dropped scales *P*, *Q*, and *S*. The reliability increases because the resulting composite is entirely comprised of psychomotor ability tests, which results in the composite having larger average intercorrelations (mean correlation: .48) than in the 4-scale composite (mean correlation: .37).

In brief, the reliability does not meaningfully drop until the last composite which contains only the GATB scale *M*, the least *g* saturated scale in the GATB. More importantly, the changes in reliability are not monotonic and thus cannot explain the monotonic decline in validity and group differences, or the monotonic increase in prediction errors shown in Table 2.

Table 2 contains the results of analyses for the *g* composites 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* composite score composed of all nine GATB scales. This composite has the most *g* saturation. The second row displays the results of analyses with the *g* composite containing eight GATB scales. This second composite has less *g* saturation than the composite 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 composite has the least *g* saturation. The first column of the table shows the number of scales in the *g* composite. The second column indicates which scale(s) were dropped from the original 9-scale *g* composite. The third column displays the correlation of each resulting *g* composite with the 9-scale composite (i.e., the *g* composite with the highest *g* saturation). This correlation is an indicator of the *g* saturation of each respective composite.

The fourth column presents the correlation between each individual *g* composite score and the job performance criterion. These correlations were based on all 16,384 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 2), we only discuss the statistics from the 16,384 individuals

Table 1. Reliabilities of nine composites

Number of scales in the composite	Dropped scale(s)	Reliability of the composite
9		.92
8	*G*	.91
7	*N* (and *G*)	.90
6	*V* (and *G*, *N*)	.92
5	*P* (and *G*, *N*, *V*)	.90
4	*Q* (and *G*, *N*, *V*, *P*)	.89
3	*S* (and *G*, *N*, *V*, *P*, *Q*)	.90
2	*K* (and *G*, *N*, *V*, *P*, *Q*, *S*)	.90
1	*F* (and *G*, *N*, *V*, *P*, *Q*, *S*, *K*)	.73

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

Number of scales in *g* composite	Scale(s) dropped from composite	Correlation with most saturated *g* composite	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*	.99	.210 (.198)	.785 (.784)	−.060 (−.059)	.129 (.127)
7	*G, N*	.97	.197 (.182)	.722 (.718)	−.067 (−.066)	.144 (.141)
6	*G, N, V*	.94	.186 (.170)	.645 (.631)	−.075 (−.073)	.160 (.156)
5	*G, N, V, P*	.92	.187 (.170)	.621 (.601)	−.076 (−.075)	.162 (.159)
4	*G, N, V, P, Q*	.84	.165 (.150)	.563 (.522)	−.083 (−.081)	.178 (.174)
3	*G, N, V, P, Q, S*	.69	.136 (.119)	.298 (.283)	−.100 (−.094)	.214 (.201)
2	*G, N, V, P, Q, S, K*	.61	.122 (.113)	.334 (.311)	−.100 (−.094)	.214 (.200)
1	*G, N, V, P, Q, S, K, F*	.50	.106 (.104)	.251 (.219)	−.104 (−.097)	.223 (.207)

Notes: Nine *g* composites were calculated based on a factor analysis of the nine GATB scales. Scales were weighted by their factor loadings. The *g* composite based on all nine scales was iteratively altered by dropping the highest loading GATB scale from the previous composite, thus making each successive composite less *g* saturated than the previous composite. All statistics were calculated in two ways. The first way was to treat the 16,384 observations as one sample. The second yields the results in parentheses. In this second approach, the data were grouped into 101 samples based on their SATB number from the GATB data. The statistics were calculated in each of the 101 samples. The sample-size-weighted mean of these statistics yielded the statistics in parentheses.

considered as one sample. The fifth column shows the White–Black standardized mean difference in the *g* composite. 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 2, the validities of the *g* composites drop from .216 to .106, a reduction of .11 or 51%, as the *g* saturation of the composites is successively reduced by dropping the highest *g* saturated scale from the previous *g* composite. Accompanying the drop in validity is a drop in the standardized mean differences (*d*) between Whites and Blacks on each *g* composite. Specifically, the *d* is .837 for the most *g* saturated composite and .251 for the least *g* saturated composite, a decrease of .586 or 70%. The drop in *g* saturation is accompanied by an increase in prediction errors; prediction errors increase as the *g* saturation of the composite is reduced. The White prediction errors are negative, indicating that the common regression line underpredicts the job performance of Whites, on average. The Black prediction errors are positive, indicating that the common regression line overpredicts the job performance of Blacks, on average. Note that the overprediction of job performance for Blacks is larger than the underprediction of job performance for Whites.

Table 2 displays the results for composites in which the *g* saturation was reduced by adding random variance to the *g* composites. This random variance reduced the ratio of true to observed variance and thus reduces reliability and thereby also the *g* saturation. This table has the same format as Table 2. Note that the first rows of Tables 2 and 3 show the same results because the first *g* composite is the same in both tables (the composite calculated based on Equation 1). As the percentage of random variance added to the *g* composite increases, the validity drops from .216 to .152, a reduction of .064 or 30%. Consistent with Table 2, as the *g* composites are reduced in their *g* saturation, the validity and the White–Black mean differences decline, but the prediction errors increase.

A reviewer requested that we address whether the results in Tables 2 and 3 can reasonably be attributed to sampling error. First, we note that statistics based on a sample size of 16,384 have very little random sampling error. Second, Tables 2 and 3 show monotonic relationships across composites, and random sampling error would cause random variations in statistics and not show monotonic relationships. On the other hand, adjoining composites (e.g., a 9-scale composite and an 8-scale composite) tend to show small differences. For example, in Table 2, the validity of the 9-scale composite is .216 and the validity of the 8-scale composite is .210.

Table 3. 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* composite (%)	Correlation with most saturated *g* composite	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* composites were calculated based on a factor analysis of the nine GATB scales. The *g* composite based on all nine scales was iteratively altered by adding increasing amounts of random variance, thus making each successive composite less *g* saturated than the previous composite. All statistics were calculated in two ways. The first way was to treat the 16,384 observations as one sample. The second yields the results in parentheses. In this second approach, the data were grouped into 101 samples based on their SATB number from the GATB data. The statistics were calculated in each of the 101 samples. The sample-size-weighted mean of these statistics yielded the statistics in parentheses.

Table 4. Confidence intervals for validities from Table 1

Scales in composite	Validity	Standard error	Lower CI	Upper CI
9	.216	.0074	.201	.231
8	.210	.0075	.195	.225
7	.197	.0075	.182	.212
6	.186	.0075	.171	.201
5	.187	.0075	.172	.202
4	.165	.0076	.150	.180
3	.136	.0077	.121	.151
2	.122	.0077	.107	.137
1	.106	.0077	.091	.121

One can approach this sampling error question from the perspective of overlapping confidence intervals.[5] Table 4 shows the nine composites, the validity coefficients from Table 1, the standard error of the validity coefficient, and the confidence intervals for the validity coefficients. Given the sample size of 16,384, the validities are estimated with substantial precision. Stated another way, their sampling error expressed as a standard error is relatively small, resulting in confidence intervals that are quite small in range. Given that the sample size is constant, the standard error varies solely as a function of the magnitude of the correlation, which is used as the estimate of rho in the standard error calculation. Finally, we note that the confidence intervals are not perfectly symmetrical around the validity due to the asymmetry of the sampling error distribution of correlation coefficients. With rounding, most of the confidence intervals appear symmetrical.

We note that the confidence intervals of validities from adjacent composites overlap, but as one moves to composites that differ more in the number of scales included in the composite, the confidence intervals do not overlap. For example, the confidence intervals for the 4-scale composite do not overlap with the confidence intervals of the composite with seven scales or the composite with two scales. Thus, the monotonic decline in validities and mean group differences across the nine composites cannot credibly be attributed to sampling error.

4. Discussion

Spearman's Hypothesis is well supported by these results. Black–White differences are largest in the most *g* saturated composites but these composites also have the largest validity and the smallest prediction errors. Because Spearman's Hypothesis has undesirable societal consequences, psychology has a long history of attempts to develop alternative *g* measures that have lower mean racial differences. These attempts are admirable as they seek to address the diversity–validity dilemma (De Corte, Sackett, & Lievens, 2010; De Corte, Sackett, & Lievens, 2011; Ployhart & Holtz, 2008) as well as general social and political concerns (Gottfredson, 1997). We applaud such efforts. However, given the cumulative evidence in support of Spearman's Hypothesis (e.g., Gottfredson, 1997; Hunt, 2011; Jensen, 1980, 1998; National Academy of Sciences, 1982; Reeve & Hakel, 2002; Rushton & Jensen, 2005; Sackett, Borneman, & Connelly, 2008; Sackett et al., 2001), assertions that a test is both an excellent measure of *g* and has lower than typical mean Black–White differences will be received with substantial skepticism.

Recently, two tests have been offered as measures of *g* with lower than typical Black–White mean differences. Fagan (2000) argued that *g* is 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 a similar exposure to the language (e.g., words, sayings, similarities, analogies) used in the test, there were only negligible mean racial differences in the processing of the information. We contrast the Fagan and Holland research with research on miniature training and evaluation tests (Harris, 1987), also called trainability tests (Roth, Buster, & Bobko, 2011). In this line of research, as with the Fagan and Holland tests, applicants receive training and then are evaluated on their knowledge of the trained material. Both Harris (1987) in a set of primary studies and Roth et al. (2011) in a broader range of studies, which 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. The discrepancy in these two research streams has not been resolved.

The second test is the Siena Reasoning Test (Goldstein, 2008). The Siena Reasoning Test has been offered as a test of *g* that shows smaller mean racial differences than previous measures of *g*. Yusko, Goldstein, Oliver, and Hanges (2010) argued that the Siena Reasoning Test measures cognitive ability and shows smaller mean racial differences than typical *g* tests because it reduces reliance on prior knowledge, reduces the use of language, and incorporates graphical stimuli. We contrast this research with a test called the Davis–Eells Games that also sought to reduce subgroup differences by limiting verbal content and using graphical items. The Davis–Eells Games test did not yield substantially reduced White–Black mean racial differences (Jensen, 1980, p. 643). Likewise, tests such as the Raven's Progressive Matrices and the Advanced Raven's Progressive Matrices (Raven, Raven, & Court, 1998; Raven, Court, & Raven, 1994) do not rely on prior knowledge or language and their items are graphical. The Raven's tests typically show large White–Black mean differences near a full standard deviation in magnitude. One might infer that the explanations offered by Siena Reasoning Test researchers for why that test purportedly yields smaller mean differences are inconsistent with cumulative research in intelligence.

Can the results of these newer *g* tests be better understood in the context of Spearman's Hypothesis? To our knowledge, the Fagan and Holland tests have not been evaluated with respect to Spearman's Hypothesis nor do the studies provide sufficient data for such an evaluation. However, Scherbaum, Hanges, Yusko, Goldstein, and Ryan (2012) presented data related to the Siena Reasoning Test and more traditional *g* measures that could be analyzed in the context of the Spearman's Hypothesis. They reported results from two studies. In our analysis of their results, the correlation between the *g* saturation of each test, defined as factor loadings, and mean Black–White differences was .994 for their first study and .620 for the second study. When *g* saturation was defined as the correlation of the test with the most *g* loaded composite, the correlation between *g* saturation of each test and the mean Black–White difference was .996 for the first study and .536 for the second study. Thus, the data offered by scholars associated with the Siena Reasoning Test are largely consistent with the inference that the Spearman's hypothesis is a plausible explanation of the Siena Reasoning Test findings. Specifically, the results are consistent with the inference that the reported lower mean racial differences in the Siena Reasoning Test are due to its lower *g* saturation relative to other *g* tests. If this inference is correct, one could also infer, consistent with the findings of our study, that the apparent lower *g* saturation of the Siena Reasoning Test would be associated with lower validity and larger prediction errors.

We offer several important caveats to these inferences due to limitations of the Scherbaum et al.'s (2012) study. The results offered by Scherbaum et al. are limited by the small number of tests used, which may distort the factor loadings (Major et al., 2011) and may misestimate correlations between *g* saturation and mean Black–White differences, possibly substantially, due to sampling error. Furthermore, the studies used college students as participants and thus there was likely range restriction on all the cognitive measures used. The study is also limited in that it does not examine validity and prediction errors that may be associated with the lower *g* saturation. An additional limitation is the relatively small size of each of the Scherbaum et al.'s samples. Finally, we note that any set of data permits varying inferences and that parties may have different perspectives on the meaning and import of results.

4.1. Implications

Our findings have implications for US federal 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 substantially lower mean racial differences. Thus, barring large sample credible research to the contrary, there is unlikely a situation in which there is a legal requirement to use a *g* test with lower mean racial differences. Em-

ployers could of course use a *g* test with lower mean racial differences and its use will result in lower validity and greater prediction errors.

The ease with which one can alter the *g* saturation of a 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. This could be done by removing the most *g* saturated components of the test, 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) widely accepted Three Stratum theory of intelligence (Deary, 2012). 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 crystallized 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 can also damage the *g* saturation of a test by conducting a factor analysis of *g* items and then removing the items with the highest loading on the *g* factor. In addition, one can damage the *g* saturation of a test by removing the items with the largest mean racial differences. Finally, one could simply add random variance to a *g* test to damage its *g* saturation. Unfortunately, any approach that reduces the *g* saturation of the test may inevitably reduce validity and increase prediction errors.

We strongly support calls for the development of strategies for achieving diversity without sacrificing validity. Sackett et al. (2001) described and evaluated several approaches to achieve this and that, in the interest of equitable treatment across all demographic groups and organizational functioning, any proposed strategy should not sacrifice validity in the interest of reducing subgroup differences (see also, e.g., Sackett et al., 2008; Schmidt, 2002). As Schmidt and Hunter (1998) have shown, a combination of cognitive ability and personality tests, such as conscientiousness or integrity tests, increases the predictive validity of the test composite, on average. Furthermore, McDaniel, Psotka, Legree, Yost, and Weekley (2011) illustrated how mean racial differences on situational judgment tests can be reduced considerably without sacrificing validity. In our view, more research in these areas could be of great benefit because a combination of *g* and personality or situational judgment tests can reduce mean racial differences, on average (Ployhart & Holtz, 2008; Viswesvaran & Ones, 2002). Several other strategies are possible as well (e.g., Sackett et al., 2001, 2008; Schmidt, 2002), and we support research in these areas.

Notes

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*.
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.
3. In a version of the paper presented at the SIOP convention, McDaniel and Kepes (2012) reported the sample size as 22,728. This was in error. That sample size ($N = 22{,}728$) reflects the sample before screening samples to have at least 25 Whites and 25 Blacks.
4. Removing any scale from Equation (1) reduces the *g* saturation of the resulting composite. By removing the highest *g* loaded scale from the *g* factor expressed in Equation (1) from the previous *g* composite, we are producing the largest possible decline in *g* saturation between each *g* composite. We note that we could have dropped the *g* saturation of the successive *g* composites by dropping the least *g* saturated scale from the previous *g* composite. However, that would have reduced *g* saturation in successive composites much less effectively. For example, the last remaining *g* composite would have been consisted of the GATB *G* scale which has a correlation of .89 with the most *g* saturated composite (the 9-scale *g* composite). As seen in Table 2, removing the highest *g* loaded scale results in a set of measures with substantial variability in *g* saturation.
5. The confidence intervals we calculated are correct but conclusions concerning statistical significance of the difference between two correlations are approximate due to two issues. First, any use of confidence intervals, even in independent samples, underestimates the statistical significance of the difference between the correlations because the statistical difference test relies on the standard error of the difference and not the standard errors of the two confidence intervals. Second, our correlations are dependent because they are based on the same sample. The standard error of the difference for dependent correlations is smaller than the standard error for independent correlations (i.e., for a given *N*, the same magnitude difference between two correlation coefficients can be statistically significant for dependent correlations but not for independent correlations). For both of these reasons, the correlations with overlapping confidence intervals can still be statistically significantly different.

References

Ashton, M. C., & Lee, K. (2005). Problems with the method of correlated vectors. *Intelligence*, 33, 431–444.

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.

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

De Corte, W., Sackett, P. R., & Lievens, F. (2010). Selecting predictor subsets: Considering validity and adverse impact. *International Journal of Selection and Assessment*, 18, 260–270.

De Corte, W., Sackett, P. R., & Lievens, F. (2011). Designing pareto-optimal selection systems: Formalizing the decisions required for selection system development. *International Journal of Selection and Assessment*, 96, 907–926.

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology*, 63, 453–482.

Fagan, J. F. (2000). A theory of intelligence as processing: Implications for society. *Psychology, Public Policy, and Law*, 6, 168–179.

Fagan, J. F., & Holland, C. R. (2002). Equal opportunity and racial differences in IQ. *Intelligence*, 30, 361–387.

Fagan, J. F., & Holland, C. R. (2007). Racial equality in intelligence: Predictions from a theory of intelligence as processing. *Intelligence*, 35, 319–334.

Fagan, J. F., & Holland, C. R. (2009). Culture-fair prediction of academic achievement. *Intelligence*, 37, 62–67.

Feldt, L., & Brennan, R. (1989). Reliability. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: The American Council on Education, MacMillan.

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.

Goldstein, H. (2008, November). *Building cognitive ability tests with reduced adverse impact*. Paper presented to the Mid-Atlantic Personnel Assessment Consortium, New York, NY.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history and bibliography. *Intelligence*, 24, 13–23.

Gottfredson, L. S. (2002). Where and why g matters: Not a mystery. *Human Performance*, 15, 25–46.

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

He, Q. (2009). *Estimating the reliability of composite scores*. The Office of Qualifications and Examinations Regulation Report: Ofqual/10/4703. Available at http://dera.ioe.ac.uk/1060/1/2010-02-01-composite-reliability.pdf (accessed 20 June 2014).

Hunt, E. (2011). *Human intelligence*. New York: Cambridge University Press.

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

Jensen, A. R. (1985). The nature of the black-white difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences*, 8, 193–219.

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.

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

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.

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.

McDaniel, M. A., & Kepes, S. (2012). *Spearman's hypothesis is a model for understanding alternative g tests*. Presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology*, 96, 327–336.

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.

National Academy of Sciences. (1982). *Ability testing: Uses, consequences, and controversies*. (Vol. 1). Washington, DC: National Academy Press.

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

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.

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

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.

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.

Reeve, C. L., & Hakel, M. D. (2002). Asking the right questions about g. *Human Performance*, 15, 47–74.

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.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, 11, 235–294.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, 63, 215–227.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, 56, 302–318.

Scherbaum, C. A., Hanges, P. J., Yusko, K., Goldstein, H. W., & Ryan, R. (2012). *The Spearman Hypothesis cannot explain all racial score differences*. Paper presented at the Annual Meeting of the Society for Industrial and Organizational Psychology.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance*, 15, 187–211.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262–274.

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

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

Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior*, 29, 332–339.

U.S. Department of Labor. (1970). *General Aptitude Test Battery, section III: Development*. Washington, DC: U.S. Government Printing Office.

Viswesvaran, C., & Ones, D. S. (2002). Agreements and disagreements on the role of general mental ability (GMA) in industrial, work, and organizational psychology. *Human Performance*, 15, 212–231.

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 Annual Meeting of the Society for Industrial and Organizational Psychology, Atlanta, GA.