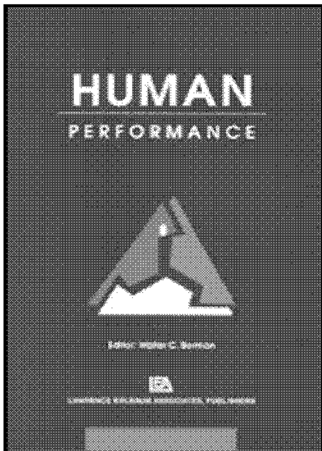

Human Performance
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653650

Subgroup Differences in Situational Judgment Test Performance: A Meta-Analysis

Deborah L. Whetzel [a]; Michael A. McDaniel [b]; Nhung T. Nguyen [c]
[a] Human Resources Research Organization, Alexandria, Virginia
[b] Virginia Commonwealth University,
[c] Towson University,

Online Publication Date: 01 July 2008

To cite this Article: Whetzel, Deborah L., McDaniel, Michael A. and Nguyen, Nhung T. (2008) 'Subgroup Differences in Situational Judgment Test Performance: A Meta-Analysis', Human Performance, 21:3, 291 — 309

To link to this article: DOI: 10.1080/08959280802137820
URL: http://dx.doi.org/10.1080/08959280802137820

Human Performance, 21:291–309, 2008

ISSN: 0895-9285 print/1532-7043 online
DOI: 10.1080/08959280802137820

Routledge
Taylor & Francis Group

Subgroup Differences in Situational Judgment Test Performance: A Meta-Analysis

Deborah L. Whetzel
Human Resources Research Organization
Alexandria, Virginia

Michael A. McDaniel
Virginia Commonwealth University

Nhung T. Nguyen
Towson University

In this article, we describe a systematic review of mean race and sex differences in situational judgment test (SJT) performance. On average, White test takers perform better on SJTs than Black, Hispanic, and Asian test takers. Female examinees perform slightly better than male test takers on SJTs. We investigate two moderators of these differences: loading of *g* or personality on the SJT, and response instructions. Mean race differences between Black, Hispanic, Asian and White examinees in SJT performance are largely explained by the cognitive loading of the SJT such that the larger the cognitive load, the larger the mean race differences. Regarding the effect of personality loadings of SJTs on race differences, Black–White and Asian–White differences are smaller to the extent that the SJT is correlated with emotional stability and Hispanic–White differences are smaller to the extent that SJTs are correlated with conscientiousness and agreeableness. Cognitive loading has minimal effect on male–female SJT score differences; however, SJT score differences are larger, favoring women, when SJTs are correlated with conscientiousness and agreeableness. Concerning response instructions, knowledge response instructions showed greater race differences than behavioral tendency instructions. The mean correlations show that these differences are largely because of the greater g loading of knowledge instructions. A second study showed that when used in hiring, SJTs are likely to have adverse impact by race but not by sex.

Situational judgment tests (SJTs) assess an applicant's judgment regarding situations encountered in the work place (McDaniel, Hartman, Whetzel, & Grubb, 2007;

Correspondence should be sent to Deborah L. Whetzel, HumRRO, 66 Canal Center Plaza, Suite 400, Alexandria, VA 22314. E-mail: dwhetzel@humrro.org

McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). In such tests, respondents are presented with work-related scenarios and potential responses to each scenario. Respondents are to select or rate the potential responses for likelihood of performing the action or effectiveness of the action. Research on SJTs indicates that these tests are useful and are becoming popular selection tools in both the United States and Europe (McDaniel et al., 2007; McDaniel et al., 2001; Salgado, Viswesvaran, & Ones, 2001). There are at least two reasons for the increasing popularity of SJTs. First, meta-analytic findings show that SJTs have useful levels of validity for predicting job performance (McDaniel et al., 2007; McDaniel et al., 2001). Second, because SJTs describe work-related situations, these measures are often viewed as having both face and content validity (Chan & Schmitt, 1997; Motowidlo, Hanson, & Crafts, 1997; Salgado et al., 2001). In spite of their popularity, a concern in using SJTs is the potential for mean subgroup differences. Although the potential for such differences in SJTs is less likely than in cognitive measures, Black–White difference estimates range from as high as almost 1 standard deviation (Chan & Schmitt, 1997) to as low as near zero (e.g., Olson-Buchanan et al., 1998). Subgroup differences in SJT performance are troublesome because they increase the likelihood of discrimination claims and adverse publicity, regardless of whether they are related to true differences in job performance.

One way to estimate the magnitude of subgroup differences in SJTs is by conducting a meta-analysis of such differences. Hough, Oswald, and Ployhart (2001) conducted a meta-analysis documenting subgroup differences in SJTs as part of their review of a variety of selection devices. However, only seven studies were included in the meta-analysis and the subgroup differences reported were limited to sex and Black–White differences.

This meta-analysis complements previous reviews of SJTs (e.g., Hough et al., 2001; McDaniel et al., 2001; McDaniel & Nguyen, 2001) by addressing four issues that may affect performance on SJTs. First, we estimate mean differences in SJTs for four racial groups (White, Black, Hispanic, and Asian). Second, we estimate mean differences in SJTs by sex. Third, we explore whether observed mean differences in SJTs can be explained by one moderator, the extent to which SJTs correlate with cognitive ability and personality. We refer to such correlations as the cognitive loading and personality loading of the SJTs. Fourth, we investigate the extent to which race and sex differences on SJTs are influenced by a second moderator, the response instruction (knowledge vs. behavioral tendency).

COGNITIVE LOADING OF SJTs

We define the cognitive loading of a SJT as the extent to which the test is correlated with cognitive ability. Thus, a SJT with high cognitive loading has a larger correlation with a cognitive ability test than a SJT with low cognitive loading.

Although race differences on cognitive ability tests have been assessed extensively and are generally accepted as reflecting real population differences in cognitive ability (Roth, Bevier, Bobko, Switzer, & Tyler, 2001), differences on SJTs have not been comprehensively summarized. According to the latest census data, there were more than 24 million Hispanics (there were 23 million Blacks) and close to 9 million Asians of working age (18–64) as of July 1, 2003, and these numbers will continue to rise faster than the overall population (U.S. Census Bureau News, June 14, 2004). Given the increasing representation of the Hispanic and Asian population in the workforce, it is important to include these two subgroups in our examination of race differences in SJT performance.

Many theories have been proposed to explain why subgroup differences exist in test performance, such as differential item functioning theory, which states that people from different cultures perceive the same test item differently and hence obtain different test scores (see Hough et al., 2001, for an extensive review). In this study, we propose that race differences in SJTs are the result of cognitive ability measured in most SJTs. In Roth et al.'s (2001) meta-analysis, Black–White and Hispanic–White differences in general cognitive ability were reported to be 1.1 and .7 standard deviations, respectively, favoring Whites. Because SJTs are correlated with both cognitive and noncognitive traits (e.g., personality; McDaniel et al., 2007), we expect that the magnitude of race differences in SJTs will be smaller than that found in cognitive ability tests and that the cognitive loading of SJTs will moderate mean race differences in SJTs. In other words, to the extent that specific SJTs consist of situations that are cognitively demanding, race differences will be large (Chan & Schmitt, 2002, 2005).

Because mean sex differences in general cognitive ability are reported to be near zero (Hough et al., 2001; Jensen, 1998) or small (Lynn & Irwing, 2004; d = .33), we do not anticipate the cognitive loading of the SJTs to moderate mean sex differences in SJT scores.

PERSONALITY LOADING OF SJTs

We define the personality loading of a SJT as the extent to which the SJT correlates with a measure of personality. Thus, if a SJT has a high conscientiousness loading, the SJT correlates with conscientiousness more so than a SJT with a low conscientiousness loading.

Two lines of research guide our thinking on the extent to which the personality loading of SJTs may moderate mean race differences. First, research on race differences in personality, although based on a small number of studies, shows that few differences exist among race groups in the Big 5 personality dimensions with the largest Black–White difference in Openness to Experience (d = .21) favoring Whites (Hough et al., 2001). Second, McKay and McDaniel (2006) examined mean race differences in job performance as a function of the cognitive and personality loading of job performance criteria. The cognitive loading of the criteria moderated the mean race differences in job performance with larger mean race differences associated with greater cognitive loading of criteria. Personality loading of criteria moderated mean race differences in job performance with larger mean race differences associated with smaller personality loading of criteria. McKay and McDaniel viewed these cognitive and personality moderating effects as expressions of the same phenomena. They speculated, but could not empirically evaluate, that criteria high in cognitive loading would be low in personality loading and vice versa. As such, criteria that correlate more highly with cognitive ability are less likely to correlate with personality or any other variable uncorrelated with cognitive ability. Based on this logic, we anticipate that personality loading will moderate mean race differences such that the larger the personality loading of the SJT, the smaller the mean race differences.

Several studies have reported that women, on average, perform slightly better on SJTs than men. Thus, we anticipate that our meta-analysis will show a small sex difference favoring women. McDaniel and Nguyen (2001) and McDaniel et al. (2007) found that the primary personality correlates of SJTs are agreeableness, conscientiousness, and emotional stability. Costa, Terracciano, and McCrae (2001) reported that women score higher than men on agreeableness ($d = .4$) and conscientiousness ($d = .2$) and lower than men on emotional stability ($d = -.2$). To the extent that the agreeableness and conscientiousness advantage of women can outweigh their emotional stability disadvantage, these personality differences might explain mean sex differences in SJT performance.

RESPONSE INSTRUCTIONS

There are two kinds of response instructions given to SJT examinees. Behavioral tendency instructions ask respondents to report how they typically respond (e.g., What would you most/least likely do?). Knowledge instructions ask respondents to assess the effectiveness of responses (e.g., What is the best/worst response?). In a meta-analysis of the SJT literature, McDaniel et al. (2007) found that SJTs with knowledge instructions had larger correlations with general cognitive ability than SJTs with behavioral tendency instructions. Thus, relative to SJTs with behavioral tendency instructions, we expect that SJTs with knowledge instructions will have, on average, larger mean race differences and larger correlations with general cognitive ability.

To provide information about the four research issues posed earlier in the arricle, we conducted two studies. The first study presents a meta-analysis of race and sex differences in SJTs. The second is a simulation to estimate the degree of adverse impact that can be expected when using SJTs as a selection device.

STUDY 1

Method

Literature search. References cited in the McDaniel and colleagues (McDaniel et al., 2007; McDaniel et al., 2001) and McDaniel and Nguyen (2001) meta-analyses provided this study with the bulk of its effect size data. McDaniel et al. (2001) conducted an extensive literature search and identified data from a wide variety of SJTs useful for their analyses. Data were obtained on measures which are, or have been, commercially marketed, including the Business Judgment Test, Test of Practical Judgment, Supervisory Index, Supervisory Inventory on Human Relations, Social Judgment Test for Supervisors, Supervisor's Opinionaire, Supervisory Practices Test, Supervisory Problems Test, Supervisory Profile Record, Tacit Knowledge Inventory for Managers, Teamwork-KSA Test, and Test of Supervisory Judgment. They also located data on measures used by the federal government. The remaining measures are neither government owned nor commercially marketed and primarily were developed by consulting firms for specific clients. The McDaniel and Nguyen (2001) data set examining the construct validity of the Big 5 personality traits provided the majority of the data for the personality analyses.

To supplement our reference list, we contacted several researchers working in this area seeking studies that we were missing. In addition, we reviewed recent journals and programs of recent conferences. Additional studies were added to the data set.

Analysis method. We conducted a psychometric meta-analysis of standardized mean differences (Hunter & Schmidt, 1990, 2004). Analyses were performed in SAS using code adapted from Arthur, Bennett, and Huffcutt (2001). A "bare bones" psychometric meta-analysis was conducted, that is, no corrections were made for measurement error or range restriction. Thus, the magnitude of the mean effect sizes are downwardly biased estimates of the population effect sizes.

Some studies that provided a *d* effect size also reported the correlation between a measure of cognitive ability and the SJT. From studies that contained such data, we created two vectors. One vector is the *d* effect size summarizing mean race or sex differences on the SJT. The second vector is the correlation between the cognitive ability loading and the SJT. The magnitude of the correlation between the two vectors is an indication of the extent to which the cognitive loading of the SJT is correlated with the magnitude of the mean SJT score differences. In Table 1, this vector correlation is reported as the "vector r between *d* and *g* loading of SJT." Because vector correlations may be new to some readers, we note that there is a small literature on vector correlation analysis. Vector correlations are discussed in Jensen (1998) in the context of intelligence research. Hunter and Schmidt (2004,

TABLE 1
Race and Sex Differences in SJT Scores by Response Instructions

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Distribution	*k*	*N*	*Vector r Between d and g Loading of SJT*	*Mean Correlation Between SJT and g (k)*	*Vector r Between d and Cons Loading of SJT*	*Mean Correlation Between SJT and Cons (k)*	*Vector r Between d and Agr Loading of SJT*	*Mean Correlation Between SJT and Agr (k)*	*Vector r Between d and ES Loading of SJT*	*Mean Correlation Between SJT and ES (k)*	*d*	σ_d	*% Variance due to Sampling Error*
Black–White comparison													
All effect sizes	62	42,178	.77	.15 (30)	–.01	.17 (30)	.14	.21 (29)	–.20	.14 (23)	.38	.21	14
Knowledge	45	36,348	.71	.33 (19)	–.05	.18 (25)	.00	.22 (24)	–21	.14 (18)	.39	.20	13
Behavioral tendency	17	5,830	.95	.17 (11)	—	.15 (5)	—	.15 (5)	—	.14 (5)	.34	.25	18
Hispanic–White comparison													
All effect sizes	43	15,195	.49	.29 (21)	–.45	.17 (25)	–.24	.21 (24)	.07	.14 (21)	.24	.26	16
Knowledge	32	11,069	.45	.36 (14)	–.46	.19 (20)	–.33	.23 (19)	.05	.14 (16)	.28	.27	16
Behavioral tendency	11	4,126	—	.17 (7)	—	.13 (5)	—	.14 (5)	—	.13 (5)	.16	.21	24
Asian–White comparison													
All effect sizes	25	16,515	.40	.26 (14)	–.11	.14 (19)	.14	.19 (18)	–.37	.13 (17)	.29	.26	10
Knowledge	19	14,255	.23	.32 (10)	.01	.15 (16)	.12	.20 (15)	–.34	.13 (14)	.30	.25	8
Behavioral tendency	6	2,260	—	.14 (4)	—	.12 (3)	—	.14 (3)	—	.13 (3)	.27	.28	13
Male–female comparison													
All effect sizes	63	37,829	.08	.27 (33)	–.37	.16 (35)	–.49	.19 (35)	.06	.10 (29)	–.11	.19	18
Knowledge	47	32,135	.38	.33 (24)	–.34	.16 (29)	–.47	.20 (28)	.08	.13 (22)	–.12	.19	17
Behavioral tendency	16	5,694	—	.12 (9)	—	.16 (6)	—	.16 (7)	—	.12 (7)	–.08	.20	27

Note. Positive *d* indicates that Whites/Male score higher on average. SJT = situational judgment test.

pp. 294–295) addressed the topic as "study characteristic correlations." Vector correlation analysis is also similar to the use of a single predictor in a meta-regression (Thompson & Higgins, 2002). The vector correlations were sample size weighted.

A procedure for estimating the confidence intervals for the vector correlation is not known. The estimation process is hampered by both of the vectors (the *d* mean race effect size and the correlation between the SJT and another measure, such as cognitive ability) having their own standard errors. These standard errors need to be taken into account to estimate the vector correlation standard error needed for confidence intervals and statistical tests. Because the standard error of the vector correlation is not known, we chose to be cautious and report vector correlations only when at least 10 samples contributed data to the vector correlations.

For each distribution that contained a correlation between the SJT and a measure of cognitive ability, we reported the sample size weighted mean correlation with cognitive ability. The mean correlation can be interpreted as a measure of the mean cognitive loading of the SJT in the distribution. For example, if the mean cognitive ability correlation for behavioral tendency response instruction SJTs is smaller than the mean cognitive ability correlation for knowledge response instruction SJTs, one would conclude that behavioral tendency response instruction SJTs have, on average, less cognitive loading than the knowledge response instruction SJTs. Such data are useful in interpreting the differences in race and sex means across distributions. In Table 1, these means are reported as the "Mean correlation between SJT and *g*."

Similarly, some studies that provided a *d* effect size also reported the correlation between a measure of personality (i.e., conscientiousness, agreeableness, and emotional stability) and the SJT. Thus, we calculated the sample size weighted correlation between the mean effect size (*d*) vector and each of the three personality correlation vectors (SJT correlated with conscientiousness, agreeableness, and emotional stability). This vector correlation is reported as "vector r between *d* and [Cons, Agr, or ES] loading of SJT." As with the vector correlation addressing the moderating effect of SJT cognitive loading, the vector correlations for personality loading were reported only when at least 10 samples contributed data to the correlation. Also, for each distribution that contained a correlation between the SJT and a measure of any of the three personality scales, we reported the mean correlation with the personality scale. In Table 1, these means are reported as the "Mean correlation between SJT and [Cons, Agr, or ES]."

McDaniel, Rothstein, and Whetzel (2006) and others (McDaniel, Hurtz, & Donovan, 2006; McDaniel, McKay, & Rothstein, 2006) have raised the issue of publication bias (Rothstein, Sutton, & Borenstein, 2005) in meta-analyses within industrial and organizational psychology. Publication bias exists to the extent that not all studies are available to researchers and that the missing studies have systematically different results from those that are available. Because publication bias has been found in some industrial/organizational research areas (e.g., the validity of interviews; Duval, 2005), and because publication bias analyses are required in meta-analyses published in *Psychological Bulletin* (Cooper, 2003), we applied these methods to our data.

Decision rules. Studies whose participants were employees, applicants, or students were included in our analyses. We included both written and video-based SJTs. By far, most studies used incumbents who completed written SJTs. As a result, the number of studies available to compare written versus video SJTs or applicant versus incumbent versus student samples were too small to conduct credible comparisons.

For some of the studies, we computed an r to d transformation (Swander, 2001a, 2001b; Weekley & Jones, 1997). We did not include studies using the *How Supervise?* test (File, 1943) in the analyses because the items were substantially different than those found in other SJTs (see McDaniel et al., 2007). We also excluded studies where the SJT response instructions could not be coded because of lack of information (Jones, Dwight, & Nouryan, 1999).

We reported one coefficient per sample. For example, if a single study reported two effect sizes for two different SJTs from the same sample, we included only one effect size, which was the mean of the two effect sizes. However, an exception was made with Weekley and Jones's (1997) study whereby two effect sizes were computed from the same sample. We included both effect sizes in the meta-analysis because one coefficient was from the empirically keyed version of the test and the other coefficient was from the rationally keyed version of the test. Because the correlation between the two tests was not very high ($r = .53$), we considered the empirically and rationally keyed tests to be different. Similarly, Weekley's (2006) sample 14 contained two different SJTs—one intended to measure sales performance and one intended to measure leader performance. Because they were two different SJTs, we coded them separately. Waugh and Russell (2005a) provided data from the same sample for two sets of SJT items. These also were coded separately. Nguyen and McDaniel (2003) administered the same SJT twice to the same sample, once with behavioral tendency instructions and once with knowledge instructions. Because response instruction was a moderator examined in this study, we included data from both sets of scores.

The results of several studies were supplemented with personal communications with the researchers. For the Mullins and Schmitt (1998) study, we obtained information needed to interpret the race and sex effect size. For Bess (2001), we obtained means and standard deviations not available in her original master's thesis. Errors in the Oswald, Schmitt, Kim, Ramsey, and Gillespie (2004) study were corrected through personal communication. We obtained race and sex means and standard deviations from Grubb's (2003) dissertation data and from Nguyen (2004) from the authors. We received sample size information for Juraska and Drasgow (2001) and MacLane, Barton, Holoway-Lundy, and Nickels (2001). We verified a discrepancy in a Black–White *d* for Sample 1 in Weekley and Jones (1999; the *d* is .78 not .85). We also verified a discrepancy in Waugh and Russell (2005a). The mean for female participants was 4.72 rather than 4.53 for Mini-form *a*, and the *d* for the Hispanic–White difference was .11 rather than –.11. Also, they differentiated between White and White, non-Hispanic. Thus, the reference group for the Hispanic–White comparison was different than the reference group for the Black–White comparison.

We used the same data from multiple publications only once. For example, the data reported in Pulakos and Schmitt (1996) are also reported in several other publications and were coded only once.

In one study, Leeds, Griffith, and Frei (2003), the items were not easily categorized as knowledge or behavioral tendency. Their sample item in Appendix B asks, "How do you respond to this incident? Choose the four (4) best responses." The question suggests a behavioral tendency item and the statement to choose the best suggests a knowledge item. We coded the study as knowledge SJT, as the last instruction to the respondent was to choose the best responses.

Results and Discussion

Table 1 presents meta-analytic results for Black–White, Hispanic–White, Asian–White, and male–female comparisons. In the table, the first column indicates the distribution analyzed, the next two columns present the number of samples (*k*) and the total sample size across samples (*N*). The fourth column is the sample size weighted vector correlation between the *d* that summarizes the magnitude of the race or sex differences in SJT and the cognitive loading of the SJT. The fifth column provides the sample size weighted mean correlation between the SJT and general cognitive ability. Not all samples provided data for the fourth and fifth columns, so the number of samples for which the vector correlations were calculated (column 4) and the mean correlation was calculated (column 5) is provided in parentheses in the fifth column. Thus, the (*k*) reported in the fifth column is the same (k) as used in the fourth column. Columns 6 to 11 provide the same information as columns 4 and 5 for conscientiousness, agreeableness and emotional stability, respectively. Column 12 is the standardized mean difference (*d*) summarizing the sample size weighted mean race or sex difference in SJT performance. For the effect size estimates (*d*) reported by race, a positive *d* indicates that White examinees' mean test performance exceeded that of the other group. For effect size estimates (*d*) by sex, a positive *d* indicates that the male examinees' means were higher than the female examinees' means. Column 13 is the standard deviation of the standardized mean difference effect size (*d*) distribution. The last column is the percentage of observed variance attributed to sampling error.

Before discussing the results in detail, we note that McDaniel, Rothstein, and Whetzel (2006) called for conducting publication bias analyses on all meta-analyses to determine the extent to which the results are influenced by systematic bias. Using the "trim and fill" procedure for detecting publication bias (Duval, 2005; Duval & Tweedie, 2000a, 2000b), we found no evidence of publication bias in any of the race or sex distributions.

Next we describe the results of this research by posing six questions about subgroup differences on SJT scores suggested earlier and providing responses based on our data analyses. We use .20 as a cutoff for meaningfulness of the correlations. Cohen (1988) described .10 as small and .30 as medium. We chose the middle value of those two numbers.

1) Are there mean race differences in SJT performance?

Yes.

White test takers obtained higher mean scores on SJTs than Black (d = .38), Hispanic (d = .24), and Asian respondents (d = .29). This result is consistent in the distributions subset by response instructions.

2) Are the cognitive and personality loadings of the SJT moderators of the mean race differences in SJT performance?

Yes.

Table 1 presents the correlations between two vectors (column 4). The first vector is the standardized mean race score differences on the SJT and the second vector is the correlation between the SJT and a measure of cognitive ability. We use the correlation between the SJT and cognitive ability to define the cognitive loading of the SJT. A positive vector correlation demonstrates that as the cognitive loading of the SJT increases, the standardized mean race differences on the SJT increases. For the Black–White difference, the vector correlation is .77. The corresponding value for Hispanic–White difference is .49 and .40 for the Asian–White difference. These correlations are quite large and present a compelling case that mean race differences in SJT scores are largely due to mean cognitive ability differences. One would expect that the cognitive loading of SJTs would either not affect Asian–White differences or show a negative correlation because Asian examinees tend to have somewhat higher mean levels of general cognitive ability than White test takers (Hough et al., 2001). However, this was not the case with these data. We note that the Asian advantage may be less than expected due to a slightly lower verbal IQ (Flynn, 1991).

We note that the vector correlations reported likely underestimate the population relationship between the mean race differences in SJT performance and the cognitive loading of the SJTs. This is because both vectors are effects sizes (ds for mean race differences and correlations for cognitive loading). The effect sizes in each vector are subject to simple random sampling error that serves to attenuate the magnitude of vector correlations. This is because the random error variance in one vector will be uncorrelated with the random variance in the other variable. This causes the vector correlation to be an underestimate of population value. Thus, the moderating effect of cognitive loading is likely greater than that shown in Table 1.

McKay and McDaniel (2006) found that mean race differences in job performance criteria covaried with cognitive and personality loading of the criteria. Specifically, mean race differences in job performance were the largest when the cognitive loading was the highest or when the personality loading was the lowest. Our findings concerning Black–White mean race differences are similar with respect to the cognitive and emotional stability loadings of SJTs. Specifically, the mean Black–White race differences are larger when the SJT has higher cognitive loading and a lower emotional stability loading. A related effect is found for Hispanic–White mean difference in that the differences are larger when the SJT has a larger positive cognitive loading and lower conscientiousness and agreeableness loadings.

For the Asian–White comparisons, the correlations between the *d*s and vectors of correlations between the personality constructs suggest that to the extent that SJTs measure emotional stability, the Asian–White mean SJT score difference is smaller ($r = -.37$).

3) Are mean race differences in SJT performance moderated by response instructions?

Yes.

For each comparison of *d* effect sizes, the SJTs with knowledge instructions had slightly higher *d*s than SJTs with behavioral tendency instructions (.39 vs. .34 for Black–White, .28 vs. .16 for Hispanic–White, .30 vs. .27 for Asian–White). We suggest that the response instruction difference is related to the differential cognitive loading of SJTs with the different response instructions. For the race comparisons, the mean cognitive loading is higher for the knowledge instructions than for the behavioral tendency instructions (Black–White comparison: .33 vs. .17; Hispanic–White comparison: .36 vs. .17; Asian–White comparison: .32 vs. .14). In brief, knowledge instruction SJTs have higher mean race differences than behavioral tendency instruction SJTs because the former are more cognitively loaded than the latter.

4) Are there mean sex differences in SJT scores?

Yes.

The sex differences in SJT scores are shown in Table 1. A negative *d* indicates that on average, female examinees outperform male test takers on SJTs. Consistent with our expectations, the female advantage in SJT performance was small (d = –.11).

5) Are the cognitive and personality loadings (i.e., conscientiousness, agreeableness, and emotional stability) of the SJT moderators of mean sex differences in SJT performance?

No, for cognitive ability. Yes, for conscientiousness and agreeableness.

The vector correlation between the cognitive loading of the SJTs and the mean sex differences is .08, indicating that cognitive loading has minimal impact on sex differences in SJTs. For both conscientiousness and agreeableness, the higher the loading of the construct, the higher the scores achieved by women (–.37 for conscientiousness and –.49 for agreeableness. These results show that sex differences can be explained, in part, by the loading of personality in the SJT. That is, the more a SJT is positively correlated with conscientiousness and agreeableness, the greater the sex differences favoring women.

6) Are mean sex differences in SJT performance moderated by response instructions?

Probably.

Knowledge instruction SJTs show somewhat larger sex differences than behavioral tendency SJTs (–.12 vs. –.08).

STUDY 2

Method

The previous reported mean racial differences in SJTs may result in adverse impact if the SJT is used in selection decisions. To estimate the amount of adverse impact, we conducted a simulation. First, we used 2005 U.S. Census population estimates based on individuals who declare a single race to identify reasonable estimates of percentages of White, Black, Hispanic, and Asian adults in the U.S. population (U.S. Census Bureau, 2006). Accordingly, we observed that 69% of the population is White, 13% is Black, 14% is Hispanic, and 4% is Asian.

Normally distributed populations of 10,000 individuals were created for each race group. The means of the distributions were changed to reflect the mean racial differences in SJT scores shown in Table 1. We use the Black–White analysis to describe the simulation method. The Black examinees' mean was set to be .38 standard deviations lower than the White examinees' mean. We then drew random samples of 84 (69/(69 + 13)) White examinees and 16 (13/(69 + 13)) Black examinees to form a sample of 100 applicants. We sorted the applicants by descending score and counted how many Black and White applicants were in the top 10% to simulate a .10 selection ratio. We then calculated the rates of Black and White applicants selected. We did this 5,000 times and used the mean rates across the 5,000 samples for Black and White applicants as the expected hiring rates and compared them using the 80% rule of the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978).

The aforementioned procedure was repeated for varying selection ratios to determine the maximal selection ratio that can be used to meet the 80% rule. The hiring rates and thus the 80% rule statistic (the ratio of the minority hire rate to the majority hire rate) varied across the samples. We calculated the frequencies of the various 80% rule statistic across the 5,000 samples and report the 2.5 percentile and the 97.5 percentile value as the lower and upper confidence interval for the 80% rule statistic. We followed the same process for Hispanic–White comparisons and Asian–White comparisons. We also conducted the same analysis by sex assuming a sample of 50 male and 50 female applicants.

Results and Discussion

Table 2 displays the results of the adverse impact simulation. Column 1 shows the comparison being made. Column 2 shows the maximal selection ratio that can be used to meet the 80% rule for no adverse impact. Column 3 shows the 95% confidence interval for the 80% rule statistic (the mean of the 80% rule statistic is always .80). Table 2 does not show any results for male and female applicants because there is no selection ratio that causes sex-based adverse impact. This is in part because of the very small sex difference in SJT means (.11 favoring female) and in part due to male and female applicants having the same sample size in the analyses.

TABLE 2
Adverse Impact Simulation Results

Comparison	*Maximum Selection Ratio That Can Be Used to Meet the 80% Rule*	*95% Confidence Interval for the 80% Rule Statistics (the Mean is Always .80)*
White Black	.70	.49 to 1.20
White Hispanic	.46	.34 to 1.35
White Asian	.58	.33 to 1.41

Table 2 shows that racial adverse impact on SJTs is the rule not the expectation when selection ratios are small to medium. One would need to hire 70 of 100 Black applicants to avoid adverse impact, 46 of 100 Hispanic applicants to avoid adverse impact, and 58 of 100 Asian applicants to avoid adverse impact. We note that these results have substantial variance as indicated by the wide confidence intervals. Still, the results are correct, on average, for samples of 100 where minorities are represented in a manner that mirrors the U.S. population.

The *Uniform Guidelines* imbed a false assumption that there should be no substantial mean race differences in personnel selection instruments. Thus, in an adversarial situation, the typical SJT race difference would likely cause an enforcement agency to require the employer to produce job-related documentation on the SJT. Until the *Uniform Guidelines* are revised to remove the false assumption of equal racial means in job-related abilities, employers are advised to maintain job-related documentation for any SJT used to select employees.

CONCLUSIONS

Although SJTs have been in existence for more than 70 years and have received increasing research attention in recent years (McDaniel et al., 2007; McDaniel et al., 2001), this meta-analysis is the first comprehensive quantitative review of subgroup differences using this popular selection method. Our results reflect the best estimates of race and sex differences in SJT performance.

On average, White examinees obtain higher SJT scores than Black, Hispanic, and Asian examinees. The vector correlations between the mean race differences between Black, Hispanic, Asian, and White respondents and the magnitude of the cognitive loading of the SJTs are large (.77, .49, and .40 respectively). Regarding personality, the Black–White and Asian–White comparisons were most affected by emotional stability such that the higher the loadings on SJT, the lower the score differences. After cognitive ability, the Hispanic–White differences were most affected by conscientiousness and agreeableness such that the higher the loadings, the lower the score differences. The most parsimonious interpretation of these results is that mean race differences in cognitive ability cause the mean race differences in SJT performance and that different personality constructs affect the correlations differently by race. Given the magnitude of relationships, we are much more confident of the cognitive moderator conclusions than the personality moderator conclusions.

The female advantage on SJTs is minimal and is unrelated to the *g* loading of SJTs. The female advantage on SJTs increases to the extent that the SJTs are correlated with conscientiousness and agreeableness.

Our results are most clear for the Black–White comparisons because of both the larger amount of data and the large magnitude of the effects. Black–White mean differences in SJT performance are substantially a function of the cognitive loading of the SJT. The more cognitively loaded the test, the larger the magnitude of the mean racial differences. Some authors (Hough et al., 2001) have speculated on a variety of variables that might explain adverse impact. A review of a variety of causes of adverse impact is certainly warranted. However, based on the Hough et al. review, some might conclude that the causes of adverse impact are unknown. This is not the case for Black–White differences in SJTs. The most credible and data supported explanation for Black–White mean racial difference in SJTs is simply that there are large, and to date relatively intractable, Black–White differences in general cognitive ability and that these differences manifest themselves in SJTs to the extent that the SJTs measure general cognitive ability.

Suggestions for Future Research

First, there is a clearly a need for more data. More primary studies on differences in SJT performance would yield more stable estimates of the mean racial and sex differences and more stable estimates of the moderators of the differences.

Second, a method to determine the standard error of the vector correlations is needed. The estimation of the standard error is a complicated issue because the variables being correlated are effect sizes with standard errors that vary across the samples. We limited our vector correlations to analyses where there were at least 10 samples contributing data to the vector correlation. We did this to help ensure stable estimates of the vector correlations. Our conclusions regarding the magnitude of the vector correlation would be more credible if a method to estimate the standard error of the vector correlation were available.

Third, we recommend better reporting of data. Editorial practice should require mean sex and race data be reported for all predictors and criteria. Although we found no evidence for publication bias in this literature, we would not be surprised if authors would want to report mean demographic differences when they are small and omit reporting them when they are large. Editorial policy mandating the reporting of means and standard deviations by demographic subgroups would remove this potentially biasing discretion of authors.

REFERENCES

*References marked with an asterisk indicate studies included in the meta-analysis.

Arthur, W., Bennett, W., & Huffcutt, A. (2001). *Conducting meta-analysis using SAS.* Mahwah, NJ: Erlbaum.

*Bess, T. L. (2001). *Exploring the dimensionality of situational judgment: Task and contextual knowledge.* Unpublished masters' thesis, Virginia Polytechnic Institute and State University, Blacksburg. Supplemented with personal communication from M. E. Mullins to M. A. McDaniel on July 25, 2006.

*Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in SJTs: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143–159.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15,* 233–254.

Chan, D., & Schmitt, N. (2005). Situational judgment tests. In A. Evers, O. Smit-Voskuijl, & N. Anderson (Eds.), *Handbook of personnel selection* (pp. 219–242). Oxford, UK: Blackwell.

*Clevenger, J., Jockin, T., & Morris, S. (1999). *A situational judgment test for engineers: construct and criterion related validity of a less adverse alternative*. Paper presented at the 14th Annual Conference of the Society for Industrial & Organizational Psychology, Atlanta, GA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooper, H. M. (2003). Editorial. *Psychological Bulletin, 129*, 3–9.

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*, 322–331.

Duval, S. (2005). The "Trim and Fill" method. In H. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (pp. 127–144). New York: Wiley.

Duval, S., & Tweedie, R. (2000a). A nonparametric "Trim and Fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89–98.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463.

Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register*, 38290–393150.

File, Q. W. (1943). *How supervise?* (Questionnaire Form B). New York: Psychological Corporation.

Flynn, J. R. (1991). *Asian Americans: Achievement beyond IQ*. Hillsdale, NJ: Erlbaum.

*Grubb, W. L. (2003). *Situational judgment and emotional intelligence tests: Constructs and faking*. Unpublished doctoral dissertation, Virginia Commonwealth University, Richmond. Supplemented with personal communication from W. L. Grubb to M. A. McDaniel on January 4, 2005.

*Hanson, M. A. (1994). *Development and construct validation of a situational judgment test of supervisory effectiveness for first-line supervisors in the U.S. Army*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis, MN.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9,* 152–194.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.

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

*Joiner, D. (2004). Personal communication from D. Joiner to M. A. McDaniel August 24, 2004.

Jones, M. W., Dwight, S. A., & Nouryan, T. R. (April, 1999). *Exploration of the construct validity of a situational judgment test used for managerial assessment.* Paper presented at the 14th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, Georgia.

*Juraska, S., & Drasgow, F. (April, 2001). *Faking in a situational judgment test*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA. Supplemental information on sample sizes was provided in personal communication from S. Juraska to M. A. McDaniel on March 19, 2005 and March 28, 2005.

*Leaman, J. A., & Vasilopoulos, N. L. (1997). *Development and validation of the U.S. Immigration Officer Applicant assessment* (Rep. No. 97–4, Research and Development Division, Office of Human Resources and Development). Washington, DC: U.S. Immigration and Naturalization Service.

*Leaman, J. A., Vasilopoulos, N. L., & Usala, P.D. (August, 1996). *Beyond integrity testing: Screening border patrol applicants for counterproductive behaviors*. Paper presented at the 104th Annual Convention of the American Psychological Association.

*Leeds, J. P., Griffith, R., & Frei, R. L. (2003). The development and validation of a situational judgment test for security officers. *Security Journal*, pp. 63–78.

*Lievens, F., Coetsier, P., & Decaesteker, C. (April, 2000). *Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity.* Paper presented at the 15th Annual Conference of the Society for Industrial & Organizational Psychology, New Orleans, LA.

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matricies: A meta-analysis. *Intelligence, 32*, 481–498.

*MacLane, C. N., Barton, M. G., Holoway-Lundy, A. E., & Nickels, B. J. (April, 2001). *Keeping score: Empirical vs. expert weights on situational judgment responses.* Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA. Supplemented with personal communication from C. N. MacLane to M. A. McDaniel October 11, 2004.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L. & Grubb, W. L., III (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60*, 63–91.

McDaniel, M. A., Hurtz, G. M., & Donovan, J. J. (2006, May). *An evaluation of publication bias in Big 5 validity data.* Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas.

McDaniel, M. A., McKay, P. & Rothstein, H. (2006, May). *Publication bias and racial effects on job performance: The elephant in the room.* Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology. Dallas.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using Situational Judgment Tests: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103–113.

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case of four test vendors. *Personnel Psychology, 59*, 927–953.

McKay, P., & McDaniel, M. A. (2006). A re-examination of White-Black mean differences in work performance: More data, more moderators. *Journal of Applied Psychology, 91,* 538–554.

*Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*, 640–647.

Motowidlo, S. J., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations. In D.L. Whetzel & G. R. Wheaton (Eds.) *Applied measurement methods in industrial psychology* (pp. 241–260). Palo Alto, CA: Davis Black.

*Motowidlo, S. J., & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology, 66,* 337–344.

*Mullins, M. E., & Schmitt, N. (April, 1998). *Situational judgment testing: Will the real constructs please present themselves?* Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology. Supplemented with personal communication from M. E. Mullins to M. A. McDaniel October 19, 2004.

*Nguyen, N. T. (February, 2004). *Response instructions and construct validity of a Situational Judgment Test.* Proceedings of the 11th Annual Meeting of the American Society of Business and Behavioral Sciences, Las Vegas, NV. Supplemented by personal communication from N. Nguyen to M. A. McDaniel on October 21, 2004.

*Nguyen, N. T., & McDaniel, M. A. (2003). Response instructions and ethnic differences in a situational judgment test. *Applied H.R.M. Research, 8*, 33–44.

*O'Connell, M. S., McDaniel, M. A., Grubb, W. L., Hartman, N. S., & Lawrence, A. (April, 2002). *Incremental validity of situational judgment tests for task and contextual performance.* Paper presented at the 17th Annual Conference of the Society for Industrial & Organizational Psychology, Toronto.

*Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51,* 1–24.

*Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89,* 187–207. Supplemented with personal communication from F. Oswald to M. A. McDaniel, October 19, 2004.

*Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement, 65*, 70–89.

*Pereira, G. M., & Harvey, V. S. (April, 1999). *Situational judgment tests: Do they measure ability, personality or both?* Paper presented at the 14th Annual Conference of the Society for Industrial & Organizational Psychology, Atlanta, GA.

*Phillips, J. F. (1992). Predicting sales skills. *Journal of Business & Psychology, 7*, 151–160.

*Phillips, J. F. (1993). Predicting negotiation skills. *Journal of Business & Psychology, 7,* 403–411.

*Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9,* 241–258.

*Pulakos, E. D., Schmitt, N., & Chan, D. (1996). Models of job performance ratings: An examination of rate race, rate gender, and rater level effects. *Human Performance, 9*, 103–119.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297–330.

Rothstein, H., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta analysis: Prevention, assessment and adjustments*. New York: Wiley.

*Sacco, J. M., Scheu, C. R., Ryan, A. M., Schmitt, N., Schmidt, D. B., & Rogg, K. R. (2000). *Reading level and verbal test scores as predictors of subgroup differences and validities of situational judgment tests*. Unpublished manuscript.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs, methods, and techniques. In N. R. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, & organizational psychology: Vol. 1* (pp. 165–199). London: Sage.

*Smith, K. C., & McDaniel, M. A. (April, 1998). *Criterion and construct validity evidence for a situational judgment measure.* Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

*Strong, M. H., & Najar, M. J. (April, 1999). *Situational judgment versus cognitive ability tests: Adverse impact and validity.* Paper presented at the 14th Annual Conference of the Society for Industrial & Organizational Psychology, Atlanta, GA.

*Swander, C. (2000). *Video-based situational judgment test characteristics: Multidimensionality at the item level and impact of situational variables*. Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.

*Swander, C. (April, 2001a). *Exploring the criterion validity of two alternate forms of a situational judgment test*. Paper presented at the 16th Annual Meeting of the Society for Industrial & Organizational Psychology, San Diego, CA.

*Swander, C. (April, 2001b). *Video-based and written situational judgment tests: Method effects and construct validity.* Paper presented at the 16th Annual Meeting of the Society for Industrial & Organizational Psychology, San Diego, CA.

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21*, 1559–1573.

U. S. Census Bureau. *Table 3: Annual Estimates of the Population by Sex, Race and Hispanic or Latino Origin for the United States: April 1, 2000 to July 1, 2005* (NC-EST2005–03). http://www.census.gov/popest/national/asrh/NC-EST2005-srh.html accessed August 1, 2006.

*Watley, D. J., & Martin, H. T. (1962). Prediction of academic success in a college of business administration. *Personnel and Guidance Journal, 41*, 147–154.

*Waugh, G. W., & Russell, T. L. (2005a). Criterion situational judgment test (CSJT). In D. J. Knapp, C. E. Sager, & T. R. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria*. (pp. 59–66). Arlington, VA: US Army Research Institute for the Behavioral and Social Sciences.

*Waugh, G. W., & Russell, T. L. (2005b). Predictor situational judgment test (PSJT). In D. J. Knapp, C. E. Sager, & T. R. Tremble (Eds.), *Development of experimental Army enlisted personnel selection and classification tests and job performance criteria*. (pp. 135–154). Arlington, VA: US Army Research Institute for the Behavioral and Social Sciences.

*Weekley, J. A. (2006, June). Personal communication (Samples 1–20) from J. Weekley to M. A. McDaniel on June 7, 2005 to July 24, 2006.

*Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50*, 25–49.

*Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52*, 679–700.

*Weekley, J. A., Ployhart, R. E., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity, measurement and subgroup differences. *Human Performance, 17*(4), 433–461.