PERSONNEL PSYCHOLOGY
2001, 54

ETHNIC GROUP DIFFERENCES IN COGNITIVE ABILITY IN EMPLOYMENT AND EDUCATIONAL SETTINGS: A META-ANALYSIS

PHILIP L. ROTH
Department of Management
Clemson University

CRAIG A. BEVIER
Clemson University

PHILIP BOBKO
Department of Management
Gettysburg College

FRED S. SWITZER III, PEGGY TYLER
Clemson University

The cognitive ability levels of different ethnic groups have interested psychologists for over a century. Many narrative reviews of the empirical literature in the area focus on the Black–White differences, and the reviews conclude that the mean difference in cognitive ability (g) is approximately 1 standard deviation; that is, the generally accepted effect size is about 1.0. We conduct a meta-analytic review that suggests that the one standard deviation effect size accurately summarizes Black–White differences for college application tests (e.g., SAT) and overall analyses of tests of g for job applicants in corporate settings. However, the 1 standard deviation summary of group differences fails to capture many of the complexities in estimating ethnic group differences in employment settings. For example, our results indicate that job complexity, the use of within job versus across job study design, focus on applicant versus incumbent samples, and the exact construct of interest are important moderators of standardized group differences. In many instances, standardized group differences are *less* than 1 standard deviation. We conduct similar analyses for Hispanics, when possible, and note that Hispanic–White differences are somewhat less than Black–White differences.

Ethnic group differences on measures of cognitive ability have been investigated by some of the earliest social science researchers (e.g., Galton, 1892; Thorndike, 1921) and this topic continues to receive a great

The authors thank Mary-Ann Carter-Taylor, Michael Crino, Gene Galluscio, Tom Lee, Gene Stone-Romero, Frank Schmidt, and two anonymous reviewers for their comments on previous drafts of the manuscript. Special thanks also to Malcolm Ree and John Welsh for their help in providing relevant studies for our analysis.

Correspondence and requests for reprints should be addressed to Philip L. Roth, Department of Management, 101 Sirrine Hall, Clemson University, Clemson, SC 29634-1305; roth@clemson.edu.

deal of attention (Herrnstein & Murray, 1994). A high level of interest in this issue seems warranted given the individual, group, organizational, and social consequences of using measures of cognitive ability in selection for employment and education. For example, measures of cognitive ability are widely believed to be among the most valid predictors of job performance (Ree, Earles, & Teachout, 1995; Schmidt & Hunter, 1998), training performance (Earles & Ree, 1992; Ree & Carretta, 1998) and educational success (Rattan & Rattan, 1987; Willingham, Lewis, Morgan, & Ramist, 1990). However, tests of cognitive ability are also associated with large mean differences between Blacks and Whites (Gottfredson, 1988; Sackett & Wilk, 1994) and hiring proportionally fewer Blacks than Whites (Bobko, Roth, & Potoksy, 1999). As such, individuals in certain ethnic groups may have markedly lower levels of access to better jobs and educational opportunities. There are also substantial legal and job performance related implications of these ethnic group differences.

Although the issue of ethnic group differences has received a great deal of study, the integration and cumulation of this literature remains problematic. Most of the cumulation of the literature has been narrative in nature (e.g., Herrnstein & Murray, 1994). Previous narrative reviews are limited by a number of major factors (Hunter & Schmidt, 1990). Such narratives cannot rigorously investigate the role of sampling error, restriction of range, moderator analyses, and so forth, on ethnic group differences on cognitive ability. Thus, the conclusions from many primary studies have not received the rigorous scientific attention toward cumulation that they deserve.

The purpose of this manuscript is to provide a comprehensive meta-analysis regarding ethnic group differences on measures of cognitive ability in applied psychology. Specifically, we propose to increase our understanding of ethnic group differences by systematically addressing at least four issues often neglected in previous research. First, we focus on the range restriction involved in incumbent versus applicant samples. Second, we examine key moderators such as job complexity. Third, we conduct analyses by various constructs in the area of cognitive ability (e.g., general mental ability vs. verbal and mathematical ability). Fourth, we focus on Hispanic–White differences in addition to cumulating Black–White differences.

We also limit our study in several ways. We address Black–White and Hispanic–White differences only, as these comparisons involve two of the largest protected groups in the United States. We do not analyze Asian–White differences as they tend to be of a smaller magnitude and typically do not lead to the exclusion of Asians from employment or educational opportunities. We focus on ethnic group differences in employment testing, though we also report some results from the education literature. We do not examine issues relating to test bias against minority groups (we refer readers to Schmidt, 1988, for a review of this material).

Before going further, we believe it is critical to remember the nature of standardized group differences. Such analyses compare the *average* scores for two groups (e.g., Blacks and Whites) on tests of cognitive ability. Such analyses are useful to understand the influence of using cognitive ability tests in selection and likely levels of adverse impact. However, such analyses do not suggest uniformly high or low levels of cognitive ability for all individuals in various groups. It is generally acknowledged that the high level of variability within an ethnic group is much larger than the variability between ethnic groups (Vernon, 1979). If general mental ability is normally distributed, the practical amount of variance within an ethnic group is approximately six to eight standard deviations. This strongly supports the notion there are exceptionally intelligent individuals from all ethnic groups.

The Importance of Accurate Estimates of Standardized Ethnic Group Differences

There are a number of reasons why academicians, practitioners, and policy makers need accurate estimates of standardized ethnic group differences. First of all, the previously mentioned groups should care about having high quality estimates of important social phenomenon to maximize understanding of the phenomenon itself (e.g., Hunter & Schmidt, 1990). In the next few paragraphs we focus on more applied issues.

Practitioners in organizations should be interested in an accurate benchmark for standardized ethnic group differences—and the *d* statistic offers an important index to guide decision making about selection systems. The *d* statistic is defined as the difference in means (e.g, White mean vs. Black mean) divided by the sample-weighted average of the group standard deviations. For example, a *d* of .5 means that the White mean was one-half of a standard deviation (averaged across the two groups) higher than the Black mean.

Use of *d* is very helpful to guide decision making because analysis of actual adverse impact or adverse impact potential is also inherently influenced by selection ratio (see Sackett & Ellingson, 1997). Thus, adverse impact may vary from one job to another not because differences between ethnic groups have changed on a predictor, but because different jobs or locations have different selection ratios. Estimates of standardized ethnic group differences (*d*) are free from the influence of selection ratios and they therefore allow organizations to analyze the effects of implementing selection systems across various selection ratios.

Further, many human resource managers have used standardized ethnic group differences of predictors to guide decisions about the construction and evaluation of selection systems. For example, we know of organizations that have developed new predictors of job performance using either video tape technology or an oral response format in hopes of reducing ethnic group differences. Members of organizations are likely to use, and we have seen them use, a benchmark of a *d* of 1.0 for Black–White differences on cognitive ability tests. The decision makers then compute the *d* for their new predictor and make a judgment about whether the new predictor will help hire a more diverse work force.

However, the accuracy of the benchmark could easily be influenced by methods used to arrive at the benchmark as well as variables (moderators) that influence the benchmark. A meta-analytic estimate is the most mathematically accurate way to cumulate the literature and allow an examination of moderators. For example, let us assume that moderately complex jobs are associated with a *d* of approximately .7 (which we demonstrate later). Suppose that the organization above develops a new predictor with a *d* of .65. Comparing .65 to 1.0 suggests that substantial progress was made in the organization's ability to hire a more diverse work force. However, there would appear to be only modest improvement if the accurate benchmark was .7. In such cases of modest improvement, other characteristics of the selection system, such as cost and ease of administration, might also influence decision makers in an organization.

Returning to an area of interest to academics, practitioners *and* policy makers, an accurate measure of standardized ethnic group differences is important for social policy related studies (e.g., Sackett & Wilk, 1994). A high quality estimate of *d* for cognitive ability measures is critical to understand and accurately model hiring rates. For example, assume again that applicant pools for jobs of moderate complexity are associated with a *d* of approximately .7 on general mental ability and applicant pools for jobs of low complexity are associated with a *d* of .9. Decision makers contemplating use of a test with a standardized group difference of 1.0 and a selection ratio of .25 for the majority group would expect to hire 4.7% of the Black applicants (Sackett & Ellingson, 1997). However, a *d* of .9 and the same selection ratio would result in a projected hiring rate of 5.8% for Black applicants. Thus, blanket use of the generally accepted value of 1.0 would underestimate minority hiring by (5.8 – 4.7)/4.7 = 19.0%. Similarly, a *d* of .7 and a selection ratio of .25 is associated with hiring 8.5% of the Black applicants, or a 44.7% underestimate in projected minority hiring. Thus, researchers and policy makers who apply the standardized group difference of 1.0 in decision making and projections may markedly underpredict the proportion of minority hires in their analysis of selection systems.

A Brief History of Research on Ethnic Group Differences

The history of ethnic group differences on cognitive ability has generally focused on Black–White differences and judgments or illustrations of these differences. Such judgments have existed since Galton noted and graphically presented large Black–White mean differences in 1869 (Rushton, 1995). Galton later suggested that the average difference between Blacks and Whites was roughly equivalent to one eighth of the difference between the very brightest individual (e.g., Aristotle) and the least brightest individual in a society (Galton, 1892). Thorndike (1921, p. 222) graphically illustrated the difference in Black versus White "intellectual ability" by overlaying two roughly normally distributed curves in which the Black mean appears to be slightly more than one standard deviation lower than the White mean. Similar results were observed on the early Army Beta tests (Vernon, 1979).

More recent narrative reviews have echoed similar judgments and increased the precision of what we label the "generally accepted effect size" (GAES) between Whites and Blacks. Based on analysis of both industrial selection data and educational studies (e.g., studies of the SAT), the GAES of approximately one standard deviation (or about 15 IQ points) between Whites and Blacks began to coalesce in the 1960s and 1970s (Dreger & Miller, 1960, 1968; Jensen, 1973; Loehlin, Lindzey, & Spuhler, 1975; Nichols, 1987; Shuey, 1966; Tyler, 1965; Vernon, 1979). By the 1980s, the language of the literature converged on a GAES of 1.0 (Arvey et al., 1994; Herrnstein & Murray, 1994; Neisser et al., 1996; Sackett & Wilk, 1994; Williams & Ceci, 1997). For example, Hunter and Hunter's review (1984, p. 73) simply states that "Blacks score, on the average, one standard deviation lower than Whites" on the General Aptitude Test Battery (when this instrument was used to measure g). There is also primary study evidence that the Black–White standardized difference may vary across more specific cognitive ability. For example, Loehlin et al. showed that Black–White differences were largest on spatial ability.

One other attempt to summarize Black–White differences was Herrnstein and Murray's (1994) graphical analysis (p. 277). They note the mean d for cognitive ability tests was 1.08 and verbally note there was substantial variability around this mean. Unfortunately, they provide little detail on a variety of subjects related to this analysis (e.g., the nature of studies that were included in this analysis). We believe that the literature could benefit from rigorous methods to obtain precise estimates of ethnic group differences.

Schmitt, Clause, and Pulakos (1996) conducted a meta-analysis of ethnic group differences for a variety of predictors of job performance. They included 30 years of three major journals in their analysis and found that the standardized group difference for Blacks versus Whites on measures of cognitive ability was .83 (K = 16, N = 7,590). Unfortunately, the authors could not code results for general mental ability versus verbal ability, and so on, given reporting limitations in several primary studies. Hispanic–White standardized differences were .48 (K = 2, N = 1,331) for general mental ability, .45 for mathematical ability (K = 2, N = 849), and .58 for verbal ability (K = 1, N = 259). The researchers explicitly noted reservations (p. 720) about the amount of data they found and how conclusive the data were. They also called for more research that reports results by construct such as general mental ability, verbal ability, and so forth.

The Schmitt et al. meta-analysis also highlights two other issues. First, there has been little attempt in this literature to address range restriction. Range restriction can be a *very* important consideration in this area, as samples may be drawn from applicant or incumbent populations. Both applicant populations and incumbent populations are often used in selection research (as per Schmitt, Gooding, Noe, & Kirsch, 1984). Given this state of affairs, we would expect to observe that applicant samples as a whole are not as likely to be as restricted on mental ability when compared to similar incumbent populations. That is, we would expect to see nontrivial differences in standardized ethnic group differences when comparing applicant and incumbent populations across a group of samples. This is because incumbent samples in many organizations have been selected based on cognitive ability tests or other measures, such as interviews, that are correlated with cognitive ability (Huffcutt, Roth, & McDaniel, 1996). As such, cognitive ability differences (*d*s) are reduced by organizational processes (in addition to self-selection, which operates on both applicant and incumbent samples). Combining results from incumbent and applicant populations may result in downwardly biased measures of effect sizes and increased levels of within-cell variance, thereby obscuring important information about ethnic group differences.

Second, we also note that the previous meta-analysis (Schmitt et al., 1996) reported Hispanic–White differences from only two studies. This minimal number of studies is noteworthy given the number of Hispanics in the workforce. There were 14.1 million Hispanics and 15.2 million Blacks in the workforce in 1998 (U.S. Bureau of Labor Statistics, 2000) and the number of Hispanics in the workplace are expected to exceed Blacks in absolute numbers by 2006 (U.S. Census Bureau, 1999).

The small number of studies in Schmitt et al. (1996) is consistent with the state of the Hispanic–White cognitive differences literature. Jensen (1998, p. 352) notes the lack of attention to Hispanics as he writes that Black–White differences are the only set of racial differences for which we have "massive and definitive data." The literature that is available does suggest that Hispanic–White average differences are less than Black–White differences (Herrnstein & Murray, 1994; Neisser et al., 1996). There is some disagreement on the amount of differences as Gottfredsen (1988) suggests an difference of approximately .5 standard deviations and Sackett and Wilk (1994) suggest the difference is likely between .6 and .8 standard deviations. A major meta-analysis should be helpful in accurately estimating cognitive ability differences in this major, and growing, segment of the U.S. working population.

Theoretical Foundation for Analyzing Different Constructs

The structure of cognitive ability is an area with a rich history replete with debate. We examine two major schools of thought on this issue in order to establish how to conceptualize and code various constructs. A synopsis of the history of models of cognitive ability is found in Jensen (1998) or Carroll (1993).

Currently accepted models of cognitive ability appear to share some features. First, the major theories suggest there is a common factor underlying human cognitive abilities. For example, Spearman empirically demonstrated that a general factor was common to all specific measures of cognitive ability (Spearman, 1904). He noted that even though tests contained many different kinds of items (e.g., verbal, mathematical, etc.), scores on these subdomains tended to be highly correlated and that a sizeable portion of variance in test scores could be attributed to a "general factor" or *g* (Spearman, 1927). Empirical evidence suggests that all tests of cognitive ability share common variance (Kranzler, 1997; Neisser et al., 1996).

Second, several major theories agree that there are somewhat more specific abilities such as verbal ability, mathematical ability, spatial ability, and so on, that also exist. They are usually viewed as a second stratum of constructs that load on *g*. However, there is considerable debate about the names and nature of these more specific abilities (Carroll, 1993). Finally, there are thought to be much more specific mental abilities (e.g., short term memorization) at the base of the three levels of cognitive ability. There is considerable support for such a conceptualization of mental ability (e.g., Ree & Caretta, 1994, 1995).

Current theories diverge in how to conceptualize (and analyze) *g* and its subfactors. One school suggests that *g* should be conceptualized and analyzed as a higher order factor (e.g., Jensen, 1998). These researchers extract the first order factors (e.g., verbal ability, mathematical ability) and then estimate the loadings of the first order factors on *g*. A second school suggests that *g* is the primary factor that accounts for mental functioning (Ree & Carretta, 1998; Vernon, 1961, 1979). These theorists suggest that the "general factor" should be extracted first and then more specific abilities such as verbal ability, mathematical ability, spatial ability should be allowed to account for residual variance in cognitive ability scores. Readers may find more information about methods to extract the general factor in Ree and Earles (1991).

Both the "*g* as a second or third order factor" and the "*g* as a more primary factor" theoretical approaches suggest there is a hierarchy of abilities with *g* (or general mental ability) at the apex (Neisser et al., 1996; Ree & Carretta, 1998; Vernon 1961, 1979). In terms of the hierarchy, Vernon (1979) and Carroll (1993) both suggest roughly a 3-stratum theory. Stratum III (the apex) contains *g*, and Stratum II contains cognitive abilities such as verbal ability, or mathematical ability. Stratum I contains about 70 narrow and specific abilities such as inductive reasoning and memory span.

This theoretical structure has important implications for the study of ethnic group differences on cognitive ability—both in terms of hypotheses and methodological strategies. Methodologically, this structure suggests a substantial commonality between intelligence tests and achievement tests. According to Jensen (1980), there is little distinction between the two types, as *g* is predominantly used for both and no clear operational distinction can be made because some form of achievement must always be the vehicle for measurement.

One implication of this reasoning is that a general aptitude test that is designed to measure how much knowledge a student has acquired (e.g., the SAT) is measuring the same latent construct(s) as selection-based intelligence tests designed to measure learning ability or general mental ability (e.g., the Wonderlic). This logic is supported by evidence that academic achievement tests such as the ACT, GRE, SAT, and MCAT correlate highly with many IQ tests (Neisser et al., 1996). Thus, we suggest that an analysis of ethnic group differences can, and should, include both intelligence and achievement tests. Further, it might be theoretically meaningful to report some analyses with both types of tests aggregated together.

A second implication is that there are a series of related constructs in the area of cognitive ability. Although the construct of *g* is at the top of the three strata, some selection systems use only verbal or mathematical abilities. Thus, a precise understanding of standardized ethnic group differences on these facets or second level constructs is also important. Within analyses of constructs such as verbal or mathematical abilities, we use the term "facet" or "subfactor" to describe these constructs. Such terms suggest that one cannot assess constructs such as verbal or mathematical ability without also capturing a substantial portion of g related variance.

Spearman's hypothesis (Spearman, 1927) suggested that the average Black–White differences tended to be higher on tests that were more heavily saturated with g, and empirical evidence strongly supports this hypothesis (Jensen, 1985). As such, we predict that a meta-analysis of ethnic group differences will show increasing levels of group differences as tests have increased covariation with g. For example, tests designed to measure g should show larger Black–White differences than tests designed to measure verbal, mathematical, or other constructs.

Our Approach to a Comprehensive Meta-Analysis

Range restriction. We selectively consider the role of range restriction in our analyses. In some cases, range restriction is a natural, rather than artifactual, process for defining populations. For example, we examine ethnic group differences for the Graduate Record Exam (GRE). We do not correct these findings for range restriction because the population is graduate school applicants and the restriction of range in GRE scores due to the presence of a college degree is a natural part of defining that population. We do, however, investigate the influence of range restriction when we conduct moderator analyses on applicants versus incumbents within the realm of industrial tests. We discuss this issue and our general approach to range restriction in greater detail below as we describe our methods.

Moderator Variables and Analytic Strategy

We examine several potential moderator variables to help account for variability in average ethnic group differences. Specifically, we first discuss more general moderators such as sample types, and then discuss analyses of more specific moderators within each sample type when possible. We conduct separate analysis for Blacks versus Whites and Hispanics versus Whites.

Sample types. The GAES of one standard deviation difference between Blacks and Whites may vary across sample types (i.e., industrial, educational, and military). Samples that are chosen from industry or educational institutions may not be representative of the entire, or targeted, population (Campbell, 1996; Jensen, 1980). The analysis of sample types at the macro level (e.g., educational vs. industrial) involves comparisons which are likely confounded by other moderators such as educational level, applicants versus incumbent status, and so on. We suggest that this is an important place to begin analysis; further analyses then proceed within the broad categories of sample types (e.g., analysis of hires vs. applicants within industrial samples).

Job complexity. Job complexity can be viewed as a moderator variable within the industrial sample type. We use the term "job complexity" to refer to the information processing demands within a given job that individuals in that job will experience as they function in the job. We note that job complexity is an important moderator for the *validity* of cognitive ability tests predicting performance (Hunter & Hunter, 1984). We also believe it may moderate ethnic group differences, as individuals may "self-select" or gravitate toward certain types of jobs (Wilk & Sackett, 1996). For example, some individuals with lower cognitive ability may not apply for medium and high complexity jobs because such jobs may require higher levels of job experience, greater training requirements, or higher levels of education (Gottfredson, 1988). Thus, the range of ability may be (more) restricted in the medium and high complexity jobs relative to low complexity jobs (Wilk & Sackett, 1996).

The association between job complexity and standardized group differences received attention in one previous meta-analysis (Huffcutt & Roth, 1998). The researchers examined the relationship between job complexity and ethnic group differences in interview evaluations. Their Black–White analysis resulted in standardized group differences of .43 (K=12, N=5,148) for low complexity jobs, .22 (K = 13, N = 4,093) for medium complexity jobs, and –.09 (K = 5, N = 768) for high complexity jobs. A similar pattern existed for Hispanic–White comparisons. One reason for this pattern of results may be that interview ratings correlate moderately with measures of cognitive ability (Huffcutt et al., 1996; Walters, Miller, & Ree, 1993).

Not all researchers agree with the suggestion that there will be increased range restriction on cognitive ability with increasing job complexity. In fact, one researcher describes the dominant view among most other researchers in the field as being that the GAES of 1.0 applies across all complexity levels—as shown by the work of Jensen (Frank Schmidt, personal communication, May 1999). Jensen (1980) argues that job applicants are selected from the same relative position in their own racial, ethnic, or cultural group's distribution of aptitude and that a standardized group difference of 1.0 should be found for all types of jobs. Jensen bases his arguments on a narrative review of a database by Wonderlic and Wonderlic (1972). We retrieve this database and include it our meta-analysis.

Examining both sets of arguments, we hypothesize that Black–White and Hispanic–White standardized group differences will be lower for medium and high complexity jobs than for low complexity jobs.

Employment Status. As noted in some depth above, employment status may also be a moderator such that studies of applicants would be associated with much larger standardized ethnic group differences than studies of incumbents.

Educational Level. This moderator can be thought of as a factor within educational samples. Samples can be obtained from high school populations in general, from high school students applying to college, or from applicants to graduate school. We believe that there is likely to be a nonlinear pattern for standardized difference statistics across these situations. We hypothesize that mean ethnic group differences will be relatively large on tests of high school students as a whole because there will be little restriction of range relative to the general population. The differences will narrow for college-bound students as they "self-select" based on academic and cognitive ability. Finally, we hypothesize that differences will again increase for graduate bound students because tests such as the GRE are designed to maximize differences between applicants for graduate school; that is, test difficulty will be primarily responsible for larger mean differences for this test.

Construct. The mental construct measured may also be important. Our analysis uses a 2-level taxonomy that mirrors the Level II and Level III tests noted above in Carroll's (1993) 3-stratum theory and the work of Vernon (1979) (there were not enough measures of Level I tests available for analysis). The highest level (III, which will be referred to as *g*) will include tests or batteries that measure multiple aspects of mental abilities or tests that measure *g* (e.g., SAT, GRE, Wonderlic). The second level will include the facets of *g* of verbal ability and mathematical ability. As suggested earlier, we predict larger group differences for *g* tests than verbal or mathematical ability tests.

Methodology

Literature Review: Data Sources

Articles on Black–White and Hispanic–White differences on tests of cognitive abilities were gathered from several sources including PsychLit of the American Psychological Association, Educational Resources Information Center (known as ERIC), Dissertation Abstracts International, and Abstracted Business Information (known as ABI Inform).

Reference lists and studies used by several narrative literature reviews and meta-analyses of related concepts were also examined (Dreger & Miller, 1968; Herrnstein & Murray, 1994; Jensen, 1980; Osborne & McGurk, 1982; Schmitt et al., 1996; Toquam, Corpe, & Dunnette, 1989). Attempts were made to overcome the "file drawer problem" by contacting test publishers and researchers active in the field. Letters were written to 6 major publishers of cognitive ability tests and 16 prominent researchers working in the area.

Criteria for Inclusion

There were five criteria that the studies had to meet in order to be included in the meta-analysis. First, the individuals in the sample could be no younger than 14 years of age (i.e., ninth grade high school students). This cutoff was chosen to represent subjects of an employable age (both parttime and fulltime workers) and allow for IQ to stabilize after childhood (Herrnstein & Murray, 1994). Second, data must have been gathered at the individual level rather than group level. This would insure that the subjects' ethnic status could be specifically categorized rather than being estimated by the racial composition of their school or workgroup. An example of data not meeting this criteria is illustrated by several studies using the Project TALENT information (Humphreys, Fleishman, & Linn, 1977). Third, data must have been in its primary format (i.e., raw data). Data that was transformed (e.g., expert rating of test results) was not included. Fourth, data had to have been obtained from normal (nonclinical) populations. For example, populations including mental patients would not qualify. Fifth, studies had to include means and standard deviations or an appropriate statistic to calculate/derive standardized differences (e.g., an F statistic).

Coding the Data

Two individuals independently coded the variables in the studies using a standardized coding form. After examining and coding approximately 25% of the articles, the two coders met to resolve discrepancies through consensus. Reliabilities of all ratings before consensus were above 95% for categorical variables or .95 for continuous variables.

The sample type was coded as industrial, military, or educational. Samples are further broken down by several sample characteristics within each main type (e.g., applicants vs. incumbents, high school vs. college).

Job complexity was coded at three levels, reflecting the framework developed and used by Hunter, Schmidt, and Judiesch (1990). Their framework is based on ratings of Data and Things from the *Dictionary of*

Occupational Titles (U.S. Department of Labor, 1977). The three levels within the framework refer to the amount of information processing that is required to successfully do the job. Thus, at low levels of complexity, we would include machine operators, line workers, and telephone operators. Examples of medium complexity jobs include first line supervisors and middle managers. Examples of high complexity jobs include upper management and high level technicians.

Ethnic group was coded as White, Black, or Hispanic as reported in the document.

Regarding the strata or level of cognitive ability, tests were coded as representing g, verbal, or mathematical ability. In order to be considered a test of g, a test had to measure overall or general mental ability rather than only one part of it (e.g., verbal).

Computing the Minority–White Difference

We used the d statistic to summarize ethnic group differences. We computed d by subtracting the minority mean from the majority mean. This difference was divided by the sample weighted average standard deviation of the two groups. Again, we note it is important to recall that standardized ethnic group differences are an index of the *average* difference between groups. As such, it does not suggest that there are not high scoring individuals in both groups.

We also formed composite scores of g for some tests that reported group differences only at the subtest level. For example, data on ethnic group differences on the GRE were only reported for verbal, mathematical, and analytical tests, but not for overall differences. We formed a composite d by unit weighting each of the three parts to represent an overall test of general mental ability. We only formed composite scores for tests that reported differences for both mathematical and verbal abilities, or both of these subtests and additional areas of mental ability. We formed the composites so that adequate data were available to examine g (Level III) as well as the constructs of verbal and mathematical ability.

Meta-Analytic Procedure

The Hunter–Schmidt (1990) approach was used to analyze the data. It involved computing sample size weighted observed means and standard deviations. The program VGBOOT (Switzer, 1992) was used to calculate the meta-analysis results.

We chose not to correct for measurement error in our analyses as we were interested in the operational impact of tests of cognitive ability rather than "true" estimates of ethnic group differences. Finally, we did not correct the variable of "ethnic status" for attenuation as the reliability probably approaches 1.0.

We considered the effects of range restriction in our meta-analyses. Unfortunately, arithmetically correcting for range restriction was not possible, as studies did not report the relevant statistics. However, we did control for range restriction by generally treating incumbent and applicant data separately (what we called "employment status"). We explicitly incorporated this distinction into our analyses such that employment status and the primary moderator variable could be conceptualized as separate "factors." For example, we allowed both employment status and construct of interest to vary in analyses. In such cases, measuring and understanding variance due to employment status allowed us to focus on variance in the primary moderator. In some other cases we focused only on applicants as sufficient incumbent samples were not available. Overall, we believe this approach provides maximally accurate measurement and increases the sensitivity of our search for other moderators.

Results

The results of this study are broken into two parts for ease of interpretation. First we present Black–White sample results and then we turn to Hispanic–White sample results.

Black–White Samples

Overall analyses. Table 1 reports the overall results and results by sample type for the Black–White samples. The overall uncorrected d score was 1.10, somewhat higher than the GAES of 1.0. Only a very small percentage of the observed variance was accounted for by sampling error (i.e., .8%) and the variability in d across studies strongly suggests moderators.

Before examining further results, it is important to note our conventions for coding number of studies (K). In general, we coded the smaller of two numbers when there are two ways to code the data. For example, we coded the data from the Wonderlic as $K = 3$ because there were three waves of data collection. We followed this convention in all tables that do not explicitly refer to within job analyses (e.g., Table 1). We could have coded the Wonderlic data as $K = 82$ because there were two sets of data that were available only across jobs and one dataset where data ($K = 79$) were also available within jobs. We coded the number of studies conservatively for two reasons. First, the across job analysis more closely mirrors current practices in examining cognitive ability scores in which they were examined without a within job focus (e.g., Herrnstein

TABLE 1

Overall Results for Black–White Samples[1]

Sample	d	K	N	95% Conf. int.	Observed variance	Sampling error
Overall g	1.10	105	6,246,729	1.06 – 1.15	.0013	.0000
Education g	1.12	48	5,378,539	1.09 – 1.17	.0008	.0000
Education, no GRE	1.00	38	3,007,284	.98 – 1.06	.0013	.0000
Industrial g	.99	34	464,201	.88 – 1.11	.0024	.0001
Military g	1.10	22	387,705	.56 – 1.19[2]	.0028	.0000

[1] K is defined as the number of studies in the analysis. All Ks are coded conservatively such that we chose the smaller K when there was any judgment involved. Examples include coding the Wonderlic K as 3 because we had aggregate data for many studies in 1970, 1983, and 1992. Similarly, Graduate Record Exam K was coded as 10 because we obtained all GRE scores for each year from 1988–1997.

[2] The wide confidence interval may be the result of including a large number of applicant samples and a large number of incumbent samples in the same cell for analysis.

TABLE 2

Analysis of Black–White Applicant Studies by Job Complexity for Industrial Samples

Sample	d	K	N	95% Conf. int.	Observed variance	Sampling error
Within job studies						
Low complexity	.86	64	125,654	.80 – .93	.0056	.0004
Moderate comp.	.72	18	31,990	.55 – .90	.0104	.0005
High comp.	.63	2	4,884	.61 – .63	.0000	.0004
Wonderlic within job studies						
Low complexity	.86	62	124,527	.80 – .93	.0056	.0004
Moderate comp.	.73	15	28,391	.55 – .93	.0098	.0005
High comp.	.63	2	4,884	.61 – .63	.0000	.0004
Non-Wonderlic within job studies						
Low complexity	.86	2	1,127	.65 – .86	.0021	.0013
Moderate comp.	.60	3	1,530	.15 – .87	.0218	.0014

& Murray, 1994). As such, this practice allows our initial estimates to be more easily integrated with existing literature. Second, our coding conventions also allow a meta-analytic benchmark comparison of coding standardized ethnic group differences *across jobs* to coding them *within jobs* within this meta-analysis. Please note that one implication of our coding decision is that there appear to be a relatively small number of studies in Table 1 relative to Table 2. In fact, there is no mathematical discrepancy; the values reflect our logic of how to code K to maximize integration with previous literature and allow both within job and across job comparisons.

Sample type. Results vary somewhat by sample type. The educational sample *d* of 1.12 is larger than the *d* of .99 for industrial samples. Within the educational samples, the Graduate Record Examination (GRE) sample has a large influence on *d,* and when removed, the overall *d* is reduced to 1.0. However, this is not surprising because we believe that analysis of college application tests such as the SAT were important data for many researchers who adopted the GAES of 1.0.

The overall military *d* is 1.10. It is, however, interesting to note that the largest sample within this analysis was for 212,238 applicants taking the Armed Forces Vocational Aptitude Battery (ASVAB) and the *d* was 1.19. It is also important to note that tests such as the ASVAB (and the Armed Forces Qualification Test score derived from it) were developed with a great deal of psychometric rigor (e.g., Berger, Gupta, & Berger, 1990; Caretta & Ree, 1997; Cowen, Barrett, & Wegner, 1989; Jensen, 1980; Skinner & Ree, 1987), and they were designed for the purpose of selecting employees (as opposed to selecting college students). As such, this test provides important insights into the magnitude of Black–White differences for selection.

Job complexity. Results for job complexity, which appears to be an important moderator, are presented in Table 2. As noted earlier, we divided studies that reported results for individual jobs into three levels of complexity based on the amount of information processing required for the job. We report results in three different ways. First, we report results for data from applicant samples. Second, we report results just for the Wonderlic. Reporting results just for the Wonderlic is particularly important as such an analysis controls for variance across tests, applies to applicants only (Jensen, 1980; Sackett & Ostgaard, 1994), and allows for more straightforward comparisons across complexity levels. Third, we report analyses for non-Wonderlic studies. Although the number of these studies is quite small, there is no across-organization variance in these analyses so that they complement the Wonderlic analyses.

Table 1 suggested that the *d* for industrial jobs is .99, close to the GAES of 1.0. However, the pattern of results in Table 2 for within job studies suggests the value of *d* is lower. Overall, low complexity jobs are associated with a *d* of .86 and moderate complexity jobs are associated with a *d* of .72. Specific analysis of the Wonderlic applicant data shows a similar trend. For low job complexity, the *d* is .86 and the *d* for medium complexity jobs is .73. Further, although there were only two jobs coded as high complexity, they resulted in a *d* of .63. Analyses of the very small number of applicant, non-Wonderlic studies also shows standardized group difference decreasing as complexity increases.

In all three analyses we note substantially overlapping confidence intervals across the complexity levels. However, we believe that the differences for job complexity are important as researchers and practitioners may overstate the standardized ethnic group differences, and consequently underestimate minority hiring rates for particular types of jobs. That is, ethnic differences for jobs of low complexity are somewhat less than the GAES of 1.0. However, ethnic group differences for jobs of greater complexity are *markedly* less than the GAES of 1.0. We refer the reader to one of our early examples which demonstrated that the blanket use of the GAES of 1.0 may not be optimal for decision making and it may also substantially underestimate expected minority hiring rates.

It should also be clear that our analysis of the same Wonderlic set of data used by Jensen (1980) appears to contradict his conclusions that the GAES of 1.0 applies to all employment situations. A meta-analytic approach to the data shows an effect due to job complexity.

Within versus across job studies. The moderator analysis of complexity led us to examine a previously neglected issue in the ethnic group differences literature. It appears that we (and possibly others) have not considered the importance of reporting ethnic group differences *within* a given job, versus reporting ethnic group differences *across* jobs (a cogent analysis of such factors is presented by Ostroff & Harrison, 1999). Logically, standard deviations are often less for within versus across job analyses (Sackett, & Ostgaard, 1994). In addition, researchers conducting across job analyses are implicitly averaging many White or Black job means together to get the "overall" means. The effect of this practice is unknown. Thus, values of d computed within each job and then averaged are likely to be different than values of d computed on data that cut across a variety of jobs. Unfortunately, the issue of job specific analyses has not received a great deal of attention.

The analyses in Table 2 with the Wonderlic helped coalesce this issue in our minds. The ds for the Wonderlic across jobs was $d = 1.00$ ($K = 3$, $N = 355{,}587$). We contrast this figure with the values above of $d = .86$ for jobs of low complexity and $d = .73$ for jobs of medium complexity for the Wonderlic. The differences in such ds appear to be important to us as we believe that much of the evidence researchers used to develop the GAES of 1.0 relied on large databases of across jobs studies (e.g., GATB studies).

Table 3 shows our results for this issue. As suggested by a reviewer, we report applicant results for the three different levels of complexity from Table 2 in order to incorporate this important factor into our within versus across job analyses. We note that the ds for all three levels of complexity, is/are smaller than the d of 1.23 for non-Wonderlic industrial tests and d of 1.00 for the Wonderlic. An even stronger contrast is noted

TABLE 3

Analysis of Black–White Within Job and Across Job Study Results for Industrial Samples

Sample	d	K	N	95% Conf. int.	Observed variance	Sampling error
Within job studies						
Applicants						
Low complexity	.86	64	125,654	.80 – .93	.0056	.0004
Moderate comp.	.72	18	31,990	.55 – .90	.0104	.0005
High comp.	.63	2	4,884	.61 – .63	.0000	.0004
Incumbents	.38	6	2,006	.28 – .67	.0060	.0028
Across job studies[1]						
Applicants w/o Wonderlic	1.23	4	18,028	.87 – 1.26	.0012	.0002
Applicants using Wonderlic	1.00	3	355,587	.82 – 1.07	.0009	.0000
Incumbents	.92	6	48,638	.47 – .97	.0011	.0001

[1] No military samples are included in this or any industrial test category. We note that the across job d for military studies is given in Table 1 and is 1.10.

when comparing incumbents on within job studies and across job studies (d = .38 vs. d = .92). It appears that ethnic group differences using a within jobs focus may be overstated by the GAES of 1.0.

Employment status. Although we have already hinted at the incumbent/applicant distinction, Table 4 further explicates results for this issue. Employment status (applicant samples vs. incumbent samples) appears to have moderated the observed ethnic group difference. We were somewhat surprised by the relatively small amount of data available in this regard, relative to the voluminous amount of data on the relationship of mental ability to job performance. It appears that much of our understanding of Black–White differences on cognitive ability tests for employee selection has come from a comparatively small number of large sample studies from the Wonderlic and GATB.

Interpreting the results for applicant versus incumbent samples is difficult for two reasons. First, results of all studies are dominated by large sample studies of the Wonderlic in the applicant category and the GATB databases in the incumbent category. Second, level of construct (g, math, or verbal) could also confound analysis. We therefore report analyses in Table 4 with and without large samples and by constructs.

The overall d for industrial applicant g is 1.00 and the overall d for incumbent g is .90. Results eliminating the two large databases show that applicant samples are associated with a d of .99 and incumbent samples were associated with a d of .41 for measures of g. The math ds are .74 for applicants and .54 for incumbents. The ds for verbal ability fol-

TABLE 4

Analysis of Black–White Applicant and Incumbent Samples

Sample	*d*	*K*	*N*	95% Conf. int.	Observed variance	Sampling error
Industrial samples						
Applicants *g*	1.00	11	375,307	.87 – 1.11	.0016	.0000
Incumbents *g*	.90	13	50,799	.38 – .96	.0035	.0002
Industrial samples—large samples excluded[1]						
Applicant *g*	.99	8	6,169	.57 – 1.23	.0181	.0009
Applicant verbal	.83	6	8,633	.74 – .91	.0017	.0005
Applicant math	.74	5	4,556	.33 – .95	.0134	.0008
Incumbent *g*	.41	11	3,315	.29 – .58	.0088	.0030
Incumbent verbal	.63	5	1,471	.29 – 1.00	.0264	.0029
Incumbent math	.54	4	1,150	.42 – .67	.0032	.0031
Military studies						
Applicant *g*	1.46[2]	1	245,036	–	–	–
Incumbent *g*	1.05	22	133,488	.69 – 1.08	.0016	.0001
Incumbent *g* w/o large samples[3]	.53	19	21,081	.30 – .86	.0122	.0008

[1] Large samples included data from the Wonderlic and the GATB tests.

[2] The single large study uses an across jobs approach to calculating *d*.

[3] All samples within this "cell" are within job samples as the large sample studies tended to be used across job samples.

low a similar pattern in which the applicant $d = .83$ and the incumbent $d = .63$. The pattern from military samples is similar. The applicant *d* of 1.46 refers to all test takers. The incumbent *d* of 1.05 is for all studies of selected individuals (e.g., those in training). The incumbent *d* of .53 refers to all studies with three particularly large samples removed. Unfortunately, military data were not available to analyze other constructs such as mathematical and verbal ability.

One large study assessing *g* reported both applicant and incumbent standardized difference scores (Carretta, 1997). This study is particularly noteworthy given it is based on a large sample and it is the only study that directly contrasts applicant and incumbent samples. The military sample using the AFOQT resulted in *d*s for applicants and incumbents of 1.19 and .46, respectively. The drop in *d*s is probably a function of direct range restriction on the test.

An important implication of our analyses is that one must be very cautious about using incumbent samples of cognitive ability to make inferences about applicant samples or populations. Results are likely to be different. A second implication was pointed out by a reviewer. He or she noted that moderately large differences in cognitive ability persist even after selection.

TABLE 5

Analysis of Black–White Samples for Educational Level

				95%	Observed	Sampling
Sample	*d*	*K*	*N*	Conf. int.	variance	error
High School	.95	5	18,104	.86 – 2.05	.0075	.0001
College applicants[1]	.98	13	2,911,312	.95 – .99	.0000	.0000
College students	.69	7	1,953	.55 – .85	.0066	.0034
Graduate school applicants[2]	1.34	10	2,371,255	1.32 – 1.36	.0000	0000
Other graduate applicant samples	1.17	13	11,604	.72 – 1.34	.0097	.0007

[1] Data is reported for the SAT and ACT for each year for all individuals taking this test. Thus, some individuals would argue we have virtually the population of data and confidence intervals are not necessary. We report confidence intervals because we only have the data for the ACT from 1991–1997 and the SAT from 1970–1998.

[2] Graduate School Applicants include all individuals taking the Graduate Record Exam from 1988 to 1997.

Educational level. Table 5 presents the results for different educational levels. The high school sample shows a similar *d* (.95) for overall *g* to college applicant samples (.98). However, we could find few high school samples ($K = 5$) that fit our rigorous criteria for inclusion. The actual college student samples standardized group difference was .69, and the graduate school samples standardized group difference was 1.34 for the GRE and 1.17 for other available graduate school tests. Such a pattern only partially supports our hypothesis of nonlinear changes in standardized differences as educational level increases. We had expected to see a difference between all high school students and college applicants, but this did not appear. Results did suggest that analyses of college students are likely to yield different results than analyses of the population of college applicants (though sample size for actual college students is small). One finding of particular interest was a *d* of .69 for college students. The size of this *d* may be a function of both selection from all applicants and within-school analysis (that parallels our within job analysis). Again, a helpful reviewer pointed out that, even after selection, a sizeable difference exists between Black mean and White mean levels of cognitive ability.

Construct. Table 6 presents our analysis by construct in which *g* results are contrasted with results for verbal and mathematical abilities. We expected slightly larger *d*s for *g*. We aggregate across applicant and incumbent samples to provide a straightforward analysis.

For industrial tests, the hypothesized pattern is supported. The *d* for tests of *g* was .99, and the *d*s for verbal and math are .76. Results change somewhat when the GATB is removed to .76 and .71, respectively. Although these results are hardly surprising, they illustrate an important

TABLE 6

Analysis of Black–White Samples by Construct of Interest

Sample	*d*	*K*	*N*	95% Conf. int.	Observed variance	Sampling error
Industrial tests						
Measures of *g*	.99	33	464,046	.88 – 1.11	.0024	.0001
Verbal	.76	15	34,957	.66 – .83	.0029	.0003
Math	.76	11	28,337	.51 – .85	.0027	.0003
Verbal w/o GATB	.76	14	13,410	.63 – .86	.0076	.0008
Math w/o GATB	.71	10	6,790	.46 – .89	.0106	.0012
Educational tests						
SAT total[1]	.97	6	2,412,651	.95 – .98	.0000	.0000
SAT verbal	.84	5	241,462	.81 –.85	.0000	.0000
SAT math	.90	5	241,462	.88 – .91	.0000	.0000
ACT total	1.02	7	498,661	.99 – 1.05	.0001	.0000
ACT verbal	.92	7	498,661	.89 – .94	.0001	.0000
ACT math	.82	7	498,661	.78 – .86	.0002	.0000
GRE total	1.34	10	2,371,255	1.32 – 1.36	.0000	.0000
GRE verbal	1.10	10	2,371,255	1.08 – 1.11	.0000	.0000
GRE math	1.08	10	2,371,255	1.06 – 1.10	.0000	.0000
GRE analytical	1.23	10	2,371,255	1.20 – 1.26	.0000	.0000

[1] We include SAT data from 1970 to 1998 in this analysis. Previous data on the SAT is available, but revisions to the test make its generalizability to current forms of the tests problematic.

point. A GAES of 1.0 does not necessarily reflect cognitive ability differences for verbal or mathematical ability. Thus, researchers might expect lower levels of ethnic group differences if they focus on a more specific stratum of cognitive abilities (e.g., Pulakos & Schmitt, 1996). The hypothesized pattern is also supported by data from educational samples. The overall *d* for educational samples was 1.10 (from Table 1) and overall verbal and math *d*s were .95 and .96 (computed from Table 6), respectively. Similar patterns are found for major tests such as the SAT, ACT, and GRE. Overall, the results for both industrial and educational samples provide support for Spearman's hypothesis. That is, Black–White differences on measures of cognitive ability tended to increase with the saturation of *g* in the measure of ability. Finally, the hypothesized pattern of results was also supported by one military sample of 4,462 (not in Table 6), which was chosen to be fairly representative of the U.S. population (Nyborg & Jensen, 2000). These researchers reported *d*s of 1.46 for *g*, 1.01 for verbal ability, and 1.15 for mathematical ability.

TABLE 7

Overall Results for Hispanic–White Samples[1]

Sample	d	K	N	95% Conf. int.	Observed variance	Sampling error
Overall g	.72	39	5,696,519	.60 – .88	.0034	.0000
Education g	.71	22	5,131,886	.58 – .89	.0037	.0000
Education, no GRE	.73	12	2,840,649	.55 – .95	.0000	.0000
Industrial g	.83	14	313,635	.74 – .97	.0005	.0000
Industrial g w/o Wonderlic	.58	11	6,133	.40 – .74	.0066	.0018
Military g[2]	.85	1	221,233	–[3]	–	–

[1] K is defined as the number of studies in the analysis. All Ks are coded conservatively such that we chose the smaller K when there is any judgment involved. Examples include coding the Wonderlic K as 3 because we had aggregate data for many studies in 1970, 1983, and 1992. Similarly, Graduate Record Exam K is coded as 10 since we obtained all GRE scores for each year from 1988–1997.

[2] The military sample g is from a single very large sample study. The results are from the Armed Forces Qualifying Test (AFQT) that is administered to assess g across a wide variety of military jobs.

[3] It is not possible or meaningful to compute confidence intervals, variance across studies, or estimate sampling error with only one study.

Hispanic Samples

Tables 7 and 8 present the available results for the Hispanic samples. There was much less data relevant to Hispanic–White differences than Black–White differences. Thus, we were not able to conduct many moderator analyses. The overall d for Hispanics is .72. The d for industrial samples is .83 compared to the educational d of .71. The overall industrial figure of .83 and the figure from a large military sample converge on a d in the middle .80's. In both cases, studies report a standardized group difference across multiple jobs.

The overall figures exceed the one half standard deviation estimate of Gottfredson (1988), but are generally in (and occasionally exceed) the range of .6 to .8 provided by Sackett and Wilk (1994). Even though our analyses are close to the previously noted range, they provide important point estimates for Hispanic–White standardized differences via rigorous cumulation of the literature.

For Hispanic samples, we were able to conduct moderator analyses comparing standardized differences of g to verbal and mathematical ability (see Table 8). As predicted, verbal and mathematical abilities had smaller standardized group differences (d = .28 for math and d = .40 for verbal) for the few industrial studies available. A similar, though weaker, trend is also present for educational tests.

We also note two interesting analyses of individual tests not reported in the tables. For Hispanics, the Wonderlic is associated with a d of .84

TABLE 8

Analysis of Hispanic–White Samples by Construct of Interests

Sample	*d*	*K*	*N*	95% Conf. int.	Observed variance	Sampling error
Industrial tests						
Industrial *g* w/o Wonderlic	.58	11	6,133	.40 – .74	.0066	.0018
Verbal	.40	7	5,590	.27 – .63	.0067	.0012
Math	.28	7	2,375	−.06 – .51	.0084	.0017
Educational tests						
SAT total[1]	.77	3	2,362,216	.63 – .90	.0033	.0000
SAT verbal	.70	3	2,362,216	.60 – .80	.0018	.0000
SAT math	.69	3	2,362,216	.53 – .83	.0036	.0000
ACT total	.56	7	471,516	.53 – .59	.0001	.0000
ACT verbal	.61	7	471,516	.59 – .63	.0000	.0000
ACT reading	.53	7	471,516	.51 – .55	.0000	.0000
ACT math	.35	7	471,516	.31 – .40	.0002	.0000
ACT science	.58	7	471,516	.54 – .63	.0002	.0000
GRE total	.72	10	2,291,237	.70 – .74	.0000	.0000
GRE verbal	.60	10	2,291,237	.58 – .62	.0000	.0000
GRE math	.51	10	2,291,237	.50 – .53	.0000	.0000
GRE analytical	.71	10	2,291,237	.69 – .73	.0000	.0000

[1] We include SAT data from 1970 to 1998 in this analysis. Previous data on the SAT is available, but revisions to the test make its generalizability to current forms of the tests problematic.

(K = 3, N = 307,502). The Armed Forces Qualification Test score from the ASVAB for applicants is associated with a *d* of .85 (N = 212,233) while the incumbents (in military training) are associated with a *d* of .40 (N = 12,819). Similar figures for Black–White comparisons are *d*s of 1.00 for the Wonderlic, 1.19 for AFOQT applicants and .46 for incumbents on the AFOQT. Thus, there appears to be a reduction of *d* between applicant samples and incumbent samples due to direct range restriction. This difference again highlights a concern about analyzing incumbent samples to estimate expected ethnic group differences for applicants in selection systems.

Discussion

The results can best be understood by discussing (a) answers to our research question and implications of those answers, (b) limitations of the study, and (c) future research.

Answers to the Research Questions

The answer to the research question of "what are the standardized difference scores between ethnic groups?" is now clearer than the GAES of "1.0 SD for Blacks versus Whites" and "somewhere between .5 and .8

SDs for Hispanics versus Whites." If one simply examines the aggregate data and ignores the moderators, the overall *d*s for *g* are 1.10 for the Black–White difference and .72 for the Hispanic–White difference. However, there are a number of important moderators that merit discussion. We discuss the Black–White moderators first.

Sample type (educational vs. industrial) did moderate the size of the Black–White standardized difference, but such comparisons confound other moderators and are therefore difficult to interpret. For example, as expected, job complexity was a very important moderator. Although the GAES of 1.0 was somewhat close to the low complexity *d* of .86, the GAES was markedly different than the moderate complexity *d* of .72. The high complexity *d* of .63 was limited in its interpretability by having only two studies in that cell. This pattern of results is consistent with the results found by Huffcutt and Roth (1998) in their analysis of ethnic differences in the employment interview. We illustrated the importance of these results in the introduction and noted that use of the GAES of 1.0 could underestimate projected Black hiring by approximately 19% relative to a *d* of .9 and approximately 44% for a *d* of .7. We suggest such differences in standardized group differences for low versus medium complexity jobs are clearly important for understanding the likely minority hiring rates from both academic, practitioner, and policy-maker perspectives.

The nature of study design (within job or across jobs) was also an important moderator (Ostroff & Harrison, 1999). Results suggest that the GAES of 1.0 overestimates likely differences within jobs, although it is more accurate for across job comparisons. Thus, researchers need to be very careful regarding to what population they wish to generalize their results. Researchers designing studies to assess or model group differences for single jobs should examine the job complexity and relevant construct for accurate modeling.

As expected, employment status (applicant vs. incumbent) also moderates the standardized White–Black difference. Applicant samples are associated with higher standardized group differences than incumbent samples. We attribute these differences to direct and indirect range restriction within the hiring process. This finding has potential implications that extend far beyond the realm of applicant versus incumbent populations in cognitive ability tests. As noted earlier, researchers examining selection devices sometimes report and compare standardized group differences across a variety of "alternative" selection devices (e.g., situational judgment, interviews, biodata, etc.) in concurrent studies. Our analyses suggest there may be substantial differences for incumbent versus applicant populations. Further, reporting concurrent *d*s and generalizing a similar pattern to applicant populations also assumes there is no differential restriction across variables or predictors associated with these *d*s—an unlikely assumption. We call for more research on this topic to facilitate our understanding of incumbent versus applicant standardized group differences across a wide variety of selection devices and situations, and to reassess some previous work (e.g., Bobko et al., 1999; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997).

Educational level also moderated the standardized group differences. We noted standardized differences close to 1.0 for high school students and college applicants. The differences were smaller for college students, but quite large for graduate school applicants on the GRE. However, our focus is on tests used in industrial selection, so we will not dwell on this issue.

The facet of overall cognitive ability (construct) was also examined. As expected, tests assessing *g* were generally associated with larger differences than verbal or mathematical abilities. These differential values help refine our understanding of Black–White differences beyond the GAES of 1.0 in selection and provide further support for Spearman's hypothesis. However, researchers and practitioners should note that even when focusing on a facet of *g* (e.g., verbal ability), they are inherently capturing a substantial portion of *g* related variance given the structure of cognitive ability.

Implications for Practitioners

There are two primary sets of implications of this research for practitioners. First, practitioners should be sure to use the most appropriate benchmark *d* for cognitive ability. Previously, the GAES of 1.0 for tests of cognitive ability was the primary benchmark. As noted above, the GAES is limited by a number of factors and now more precise estimates of applicant level *d*s are available. We urge researchers to focus on within job *d*s and find the appropriate complexity level of the job in question in order to estimate *d* and the potential level of adverse impact for a given job (e.g., .86 for low complexity jobs, .72 for medium complexity). We also urge caution in using certain types of organizational data to estimate effect sizes. Practitioners should not use data on incumbents in a current job to estimate *d* and adverse impact in future hiring, as organizational processes will restrict the range of cognitive ability scores relative to applicant populations. That is, incumbent estimates will be downwardly biased. The results of this investigation indicate that the downward bias can be quite large.

The second set of implications for practitioners is that organizations may be able to reduce the amount of adverse impact by focusing only on the cognitive abilities required by the job. Data from this investigation suggests that measures of category II constructs such as verbal or quantitative ability are associated with somewhat lower levels of d (as they are only partial facets of g). Choosing only the relevant abilities could reduce adverse impact somewhat. It is also possible to use category I measures of cognitive ability that are closely linked to job relevant knowledge, skills, and abilities. Unfortunately, we found it very difficult to find such information in our analysis, but an example is that Verive and McDaniel (1996) estimate that d for tests of short term memory is .54 for adults ($K = 16$, $N = 8{,}891$).

The above strategy might reduce adverse impact. However, we urge two cautions in its use. First, adverse impact is still likely for a d of .5 at selection ratios from .1 through .5 (e.g., Sackett & Ellingson, 1997). Thus, practitioners are still likely to have adverse impact, but somewhat less of it (depending upon what selection ratio is in use). Second, using several category I measures in concert is most likely to indirectly "recreate" a measure of g (as one would predict from the factorial structure of various three stratum theories). Nevertheless, for certain jobs, the use of facets of g may be a viable strategy.

Hispanic–White Differences

It is unfortunate that we could not complete as many analyses on Hispanic–White differences given the lack of data. The overall analyses suggest that the standardized group difference is near the top of the range suggested by some researchers (Sackett & Wilk, 1994). However, the difference is also, as anticipated, lower than comparable values for White–Black standardized differences. We also found that educational samples without the GRE ($d = .73$) were associated with different results than industrial samples ($d = .83$). Overall, our results suggest that cognitive ability differences are somewhat larger for Hispanics versus Whites than previously thought. It is also interesting to note that the standardized group differences are smaller for math than for verbal abilities. Neisser et al. (1996) noted that a large percentage of Puerto Ricans, Mexican Americans, and Cubans do not speak English well.

Limitations

There are some limitations to consider when interpreting results from this meta-analysis. First, we found relatively few industrial samples, although the values of N were often large. We speculate that this is partially due to the diffuse nature of the literature on ethnic group differences. We found studies in a variety of fields and journals. The relatively small number of industrial studies led to somewhat large confidence intervals. We note that the confidence intervals in many of our moderator analyses did overlap. For example, the confidence intervals associated with low and medium complexity jobs overlapped considerably. Although our focus was on obtaining the best mean estimates in many cases, we do note this limitation.

Another limitation is the influence of studies with a large sample size, in that they had a substantial effect on the results of the meta-analysis. For example, the GRE is associated with a large Black–White *d* score in the overall and educational samples. In addition, studies using the Wonderlic contributed a large portion of data and may have a large influence on our results. When appropriate, we analyzed data with and without such large samples.

A third limitation is that there may be a number of latent variables associated with our moderators. For example, there may be some socioeconomic variables that correlate with job complexity which partially obscure the interpretation or causality of the exact effect of job complexity on standardized group differences. We encourage basic research into this issue below.

A fourth limitation is that we were unable to assess the influence of time on standardized ethnic group differences. A significant body of research has suggested that average scores on mental abilities are rising and this trend may narrow the Black–White group difference (e.g., Flynn, 1999). This research is not without its methodological problems (e.g., Jensen, 1998) or data contradicting it (Nyborg & Jensen, 2000). Although we had originally coded date of publication in our meta-analysis, we found that there was such a large influence of extraneous factors such as varying sample sizes by time, various tests across time, and so on, that we simply did not put much faith in this analysis. Instead, we tried to control for the influence of time by choosing the most recent studies when there was an option. For example, we chose to include only the last few years of tests such as the SAT, GRE, and ACT because they have been revised to reduce ethnic group differences and they provide the most recent data available. Within our analyses we did find three longitudinal studies that addressed this trend using the same test(s) across time. Without devoting a great deal of time to this debate, we refer the interested reader to the following sources (Lynn, 1998; Nyborg & Jensen, 2000; Wonderlic & Wonderlic, 1972). As a whole, these studies suggest that there are observed gains for both groups, but the reduction in the between-group difference is either small, potentially a function of sampling error (Lynn, 1998), or nonexistent for highly *g* loaded instruments (Nyborg & Jensen, 2000).

Future Research

There are a host of avenues for future research on ethnic group differences in cognitive ability that include testing strategies (e.g., Chan & Schmitt, 1997; Schmitt & Chan, 1998) and reasons for ethnic group differences. We mainly confine ourselves to discussing issues that arose from our analysis. First, we encourage more research on the influence of job complexity on average ethnic group differences. Such research should focus on establishing more precise effect sizes for moderate and high complexity jobs and examining the reasons (e.g., self-selection) for such average ethnic group differences.

Second, we call for research to understand the differences between applicant and incumbent populations on ethnic group differences. The investigation of the sources and degrees of range restriction are clear initial candidates for research. Until a substantial body of such research appears, we urge caution in using incumbent populations to estimate applicant population ethnic group differences. We also call for caution in comparing standardized group differences across various predictors in incumbent populations due to differential range restriction for each predictor.

Third, we encourage more research into the standardized group differences for Hispanics versus Whites. Overall, we have comparatively little data at the present time when we use Black–White comparisons as a benchmark. We suggest that one additional avenue for future research is to examine if Hispanic–White comparisons really represent a heterogeneous group of comparisons. We wonder if Hispanic–White comparisons might vary as a function of subgroups such as Hispanics of Puerto-Rican descent versus those of Mexican descent, and so forth (Wightman, 1997).

In sum, we suggest a renewed focus on determining and studying average ethnic group differences on cognitive ability tests. The current study has systematically provided a unique meta-analysis regarding accurate estimation of Black–White effect sizes for cognitive ability tests. We also clarified the important role of variables such as job complexity, study design, employment status, and saturation of g in selection tests. We look forward to future analyses that expand and build on these foundations.

REFERENCES

Studies marked with an asterisk contained data used in the meta-analysis.

*Angoff WH, Ford SF. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95–106.

Arvey RD, et al. (1994, December 13). Mainstream science on intelligence. *The Wall Street Journal*, 356.

*Baehr ME, Saunders DR, Froemel EC, Furcon JE. (1971). The prediction of performance for Black and for White police patrolmen. *Professional Psychology, 2*, 46–57.

*Barrett GV, Miguel RF, Doverspike D. (1997). Race differences on a reading comprehension test with and without passages. *Journal of Business and Psychology, 12*, 19–24.

Berger FR, Guptea WB, Berger RM. (1990). *Air force officer qualifying test Form P: Test manual* (ARHRL TR-89-56). Brooks AFB, TX: Air Force Systems Command.

Bobko P, Roth PL, Potosky D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. PERSONNEL PSYCHOLOGY, *52*, 561–589.

*Boone JA, Adesso VJ. (1974). Racial differences on a Black intelligence test. *Journal of Negro education, 43*, 429–436.

*Brigham CC. (1923). *A study of American intelligence*. Princeton, NJ: Princeton University Press.

Campbell JP. (1996). Group differences and personnel decisions: Validity, fairness, and affirmative action. *Journal of Vocational Behavior, 49*, 122–158.

*Carretta TR. (1997). Group differences on U.S. Air Force pilot selection tests. *International Journal of Selection and Assessment, 5*, 115–126.

Caretta TR, Ree MJ. (1997). *The best retest score is the average: Findings and implications* (AL/HR-TP-1996-0021). Brooks AFB, TX: Human Resources Directorate.

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

*Centra JA, Linn RL, Parry ME. (1970). Academic growth in predominately Negro and predominately White colleges. *American Educational Research Journal, 7*, 83–98.

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

*Chan D, Schmitt N, Sacco JM, DeShon RP. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology, 83*, 471–485.

*Cleary A. (1966). *Test bias: Validity of the scholastic aptitude test for Negro and White students in integrated colleges*. Research Bulletin RB-66-31. Princeton, NJ: Educational Testing Service.

*Confidential Technical Report (1977). *Selecting process operator and laboratory technician trainees*. U.S.A.

*Cook LR. (1990). The predictive validity of traditional and nontraditional admissions measures for college performance in students grouped by sex, race, age, and academic risk (Doctoral dissertation, Fordham University, 1990). *Dissertation Abstracts International, 51*, 841 (291).

*Cornwell GG. (1983). Test-wiseness, test anxiety and racial bias in employment testing. (Doctoral dissertation, University of South Florida, 1983). *Dissertation Abstracts International, 44*, 2928 (111).

Cowen DK, Barrett LE, Wegner TG. (1989). *Air force reserve officer training corps selection system validation* (ARHRL-TR-88-54). Brooks AFB, Texas: Air Force Systems Command.

*DeShon RP, Smith MR, Chan D, Schmitt N. (1998). Can racial differences in cognitive test performance be reduced by presenting problems in a social context? *Journal of Applied Psychology, 83*, 438–451.

*Distefano MK Jr, Pryer MW, Craig SH. (1976). Predictive validity of general ability tests with Black and White psychiatric attendants. PERSONNEL PSYCHOLOGY, *29*, 197–204.

*Distefano MK Jr, Pryer MW, Craig SH. (1980). Job-relatedness of a posttraining job knowledge criterion used to assess validity and test fairness. PERSONNEL PSYCHOLOGY, *33*, 785–793.

Dreger RM, Miller KS. (1960). Comparative psychological studies of Negros and White in the United States. *Psychological Bulletin, 57,* 361–402.

Dreger RM, Miller KS. (1968, September). Comparative psychological studies of Whites and Negros in the United States: 1959–65. *Psychological Bulletin Monograph Supplement,* 1–58.

Earles JA, Ree MJ. (1992). The predictive validity of the ASVAB for training grades. *Educational and Psychological Measurement, 52,* 721–726.

*Educational Testing Service (1989). *Sex, race, ethnicity, and performance on the GRE general test: A technical report.* Princeton, NJ: Educational Testing Service.

*Educational Testing Service (1990). *Sex, race, ethnicity, and performance on the GRE general test: A technical report.* Princeton, NJ: Educational Testing Service.

*Educational Testing Service (1991). *Sex, race, ethnicity, and performance on the GRE general test: A technical report.* Princeton, NJ: Educational Testing Service.

*Educational Testing Service (1992). *Sex, race, ethnicity, and performance on the GRE general test: A technical report.* Princeton, NJ: Educational Testing Service.

*Educational Testing Service (1993). *Sex, race, ethnicity, and performance on the GRE general test: A technical report.* Princeton, NJ: Educational Testing Service.

*Educational Testing Service (1994). *Sex, race, ethnicity, and performance on the GRE general test: A technical report.* Princeton, NJ: Educational Testing Service.

*Educational Testing Service (1995). *Sex, race, ethnicity, and performance on the GRE general test: A technical report.* Princeton, NJ: Educational Testing Service.

*Educational Testing Service (1996). *Sex, race, ethnicity, and performance on the GRE general test: A technical report.* Princeton, NJ: Educational Testing Service.

*Educational Testing Service (1997). *Sex, race, ethnicity, and performance on the GRE general test: A technical report.* Princeton, NJ: Educational Testing Service.

*Farr JL, O'Leary BS, Bartlett CJ. (1971). Ethnic group membership as a moderator of the prediction of job performance. PERSONNEL PSYCHOLOGY, *24,* 609–636.

Flynn JR. (1999). The discovery of IQ gains over time. *American Psychologist, 54,* 5–20.

*Fox H, Lefkowitz J. (1974). Differential validity: Ethnic group as a moderator in predicting job performance. PERSONNEL PSYCHOLOGY, *27,* 209–223.

*Gael S, Grant DL. (1972). Employment test validation for minority and nonminority telephone company service representatives. *Journal of Applied Psychology, 56,* 135–139.

*Gael S, Grant DL, Ritchie RJ. (1975a). Employment test validation for minority and nonminority clerks with work sample criteria. *Journal of Applied Psychology, 60,* 420–426.

*Gael S, Grant DL, Ritchie RJ. (1975b). Employment test validation for minority and nonminority telephone operators. *Journal of Applied Psychology, 60,* 411–419.

Galton F. (1892). *Hereditary genius.* London: Macmillan.

*Gordon ME, Arvey RD, Daffron WC, Umberger DL. (1974). Racial differences in the impact of mathematics training at a manpower development program. *Journal of Applied Psychology, 59,* 253–258.

Gottfredson LS. (1988). Reconsidering fairness: A matter of social and ethical priorities. *Journal of Vocational Behavior, 33,* 293–319

*Grant DL, Bray DW. (1970). Validation of employment tests for telephone company installation and repair occupations. *Journal of Applied Psychology, 54,* 7–14.

Herrnstein RJ, Murray C. (1994). *The bell curve: Intelligence and class structure in American life.* New York: Free Press.

Huffcutt AI, Roth PL. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83,* 179–189.

Huffcutt AI, Roth PL, McDaniel MA. (1996). A meta-analytic investigation of cognitive ability in employment interviews: Moderating characteristics and implications for incremental validity. Journal of Applied Psychology, 81, 459-437.

Humphreys LG, Fleishman AI, Lin P. (1977). Causes of racial and sociographic differences in cognitive tests. *Journal of Research in Personality, 11*, 191–208.

Hunter JE, Hunter RF. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Hunter JE, Schmidt FC. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park: Sage.

Hunter JE, Schmidt FC, Judiesch MK. (1990.) Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75*, 28–42.

Jensen AR. (1973). *Educability and group differences*. London: Methuen.

Jensen AR. (1980). *Bias in mental testing*. New York: Free Press.

Jensen AR. (1985). The nature of Black–White differences on various psychometric tests. *Behavioral and Brain Sciences, 8*, 193–263.

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

*Kassinove H, Rosenberg E, Trudeau P. (1970). Cross validation of the Environmental Participation Index in a group of economically deprived high school students. *Journal of Clinical Psychology, 26*, 373–376.

*Kaufman AS, Wang J. (1992). Gender, race, and education on the K-BIT at ages 4 to 90 years. *Journal of Psychoeducational Assessment, 10*, 219–229.

*Kaufman JC, McLean JE, Kaufman AS, Kaufman NL. (1994). White–Black and White–Hispanic differences on fluid and crystallized abilities by age across the 11 to 94-year range. *Psychological Reports, 75*, 1279–1288.

*Kaufman AS, McLean JE, Kaufman JC. (1995). The fluid and crystallized abilities of White, Black, and Hispanic adolescents and adults, both with and without an education covariate. *Journal of Clinical Psychology, 51*, 636–647.

*Kaufman JC, Chen TH, Kaufman AS. (1995). Ethnic group, education, and gender differences on six horn abilities for adolescents and adults. *Journal of Psychoeducational Assessment, 13*, 49–65.

Kranzler JH. (1997). Educational and policy issues related to the use and interpretation of intelligence tests in the schools. *School Psychology Review, 26*, 150–162.

*Lawrence W, Brown D. (1976). An investigation of intelligence, self-concept, socioeconomic status, race, and sex as predictors of career maturity. *Journal of Vocational Behavior, 9*, 43–52.

*Lefkowitz J. (1972). Differential validity: Ethnic group as a moderator in predicting tenure. PERSONNEL PSYCHOLOGY, *25*, 223–240.

Loehlin JC, Lindzey G, Spuhler JN. (1975). *Race differences in intelligence*. New York: Freeman.

Lynn R. (1996). Racial and ethnic differences in intelligence in the United States on the differential ability scale. *Personality and Individual Differences, 20*, 271–273.

*Lynn R. (1998). Has the Black–White intelligence difference in the United States been narrowing over time? *Personality and Individual Differences, 25*, 999–1002.

*Matarazzo JD, Wiens AN. (1977). Black intelligence test of cultural homogeneity and Wechsler Adult Intelligence Scale scores of Black and White police applicants. *Journal of Applied Psychology, 62*, 57–63.

*McClelland L. (1974). Effects of interviewer–respondent race interactions of household inventory measures of motivation and intelligence. *Journal of Personality and Social Psychology, 29*, 392–397.

*Moore CL, McNaughton JF, Osburn HG. (1969). Ethnic differences within an industrial selection battery. PERSONNEL PSYCHOLOGY, *22*, 473–482.

Neisser U, Boodoo G, Bouchard TJ Jr, Boykin AW, Brody N, Ceci SJ, Halpern DF, Loehlin JC, Perlof R, Sternberg RJ, Urbina S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77–101.

Nichols RC. (1987). Racial differences in intelligence. In Modgil S, Modgil C (Eds.), *Arthur Jensen: Consensus and controversy* (pp. 213–220). New York: Falmer Press.

*Nyborg H, Jensen AR. (2000). Black-White differences on various psychometric tests: Spearman's hypothesis tested on American armed services veterans. *Personality and Individual Differences, 28,* 593–599.

Osborne RT, McGurk FCJ. (1982). *The testing of Negro intelligence (Vol. 2).* Athens, GA: Foundation for Human Understanding.

*Osborne RT, Miele F. (1969). Racial differences in environmental influences on numerical ability as determined by hereditability estimates. *Perceptual and Motor Skills, 28,* 535–538.

Ostroff C, Harrison DA. (1999). Meta-analysis: Level of analysis, and best estimates of population correlations: Cautions for interpreting meta-analytic results in organizational behavior. *Journal of Applied Psychology, 84,* 260–270.

*Pandy RE. (1974). Intellectual characteristics of successful, dropout, and probationary Black and White university students. *Psychological Reports, 34,* 951–953.

*Plamondon KE, Schmitt N. (2000, April). Validity and subgroup differences of combinations of predictors as a function of research design. In Dwight SA (Chair), *An applied look at reducing adverse impact by differentially weighting selection measures.* Symposium conducted at the 15th Annual Meetings of the Society for Industrial and Organizational Psychology, New Orleans, LA.

*Pfeifer Jr, CM. (1976). Relationship between scholastic aptitude, perception of university climate, and college success for Black and White students. *Journal of Applied Psychology, 61,* 341–347.

*Pulakos ED, Schmidt N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9*(3), 241–258

Rattan AI, Rattan G. (1987). A historical perspective on the nature of intelligence. In Dean RS (Ed.), *Introduction to assessing human intelligence* (pp. 5–28). Springfield, IL: Thomas.

Ree MJ, Carretta TR. (1994). Factor analysis of the ASVAB: Confirming a Vernon-like structure. *Educational and Psychological Measurement, 54,* 459–463.

Ree MJ, Carretta TR. (1995). Group differences in aptitude factor structure on the ASVAB. *Educational and Psychological Measurement, 55,* 268–277.

Ree MJ, Carretta TR. (1998). General cognitive ability and occupational performance. In Cooper CL, Robertson IT (Eds.), *International review of industrial and organizational psychology, 13,* 159–171.

Ree MJ, Earles JA. (1991). The stability of *g* across different methods of estimation. *Intelligence, 15,* 271–278.

Ree MJ, Earles JA, Teachout MS. (1995). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79,* 518–525.

*Reynolds CR, Chaistain RL, Kaufman AS, McLean JE. (1987). Demographic characteristics and IQ among adults: Analysis of the WAIS-R standardization sample as a function of the stratification variables. *Journal of School Psychology, 25,* 323–342

*Rotundo M, Sackett PR. (1998, April). *Effect of rater race on conclusions regarding differential prediction in cognitive ability tests.* Presented at the 13th annual meetings for the Society for Industrial and Organizational Psychology, Dallas, TX.

*Ruda E, Albright LE. (1968). Racial differences on selection instruments related to subsequent job performance. PERSONNEL PSYCHOLOGY, *21,* 31–41.

Rushton JP. (1995). *Race, evolution, and behavior: A life history perspective*. New Brunswick: Transaction.

*Ryan AM, Friedel LA. (1998). Using personality testing to reduce adverse impact: A cautionary note. *Journal of Applied Psychology, 83*, 298–307.

Sackett PR, Ellingson J. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. PERSONNEL PSYCHOLOGY, *50*, 707–721.

Sackett PR, Ostgaard DJ. (1994). Job specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *Journal of Applied Psychology, 79*, 680–685.

Sackett PR, Wilk SL. (1994). Within group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49*, 929–954.

*Sahai H. (1989). Relations of sociodemographic variables and cognitive ability: A comparative analysis of the cognitive scores of high school seniors. *Perceptual and Motor Skills, 69*, 1139–1157.

Schmidt FL. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior, 33*, 272–292.

Schmidt FL, Hunter JE. (1998). The validity of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmitt N, Chan D. (1998). *Personnel selection: A theoretical approach*. Thousand Oaks, CA: Sage.

Schmitt N, Clause C, Pulakos E. (1996). Subgroup differences associated with different measures of some common job relevant constructs. In Cooper CL, Robertson IT (Eds.), *International review of industrial and organizational psychology, 11*, 115–137.

Schmitt N, Gooding RZ, Noe RA, Kirsch M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. PERSONNEL PSYCHOLOGY, *37*, 407–422.

*Schmitt N, Hattrup K, Landis RS. (1993). Item bias indices based on total test score and job performance estimates of ability. PERSONNEL PSYCHOLOGY, *46*, 593–611.

Schmitt N, Rogers W, Chan D, Sheppard L, Jennings D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 5*, 719–730.

*Shuey AM. (1942). A comparison of Negro and White college students by means of the American council psychological examination. *Journal of Psychology, 14*, 35–52.

Shuey AM. (1966). *The testing of Negro intelligence (2nd ed.)*. New York: Social Sciences Press.

Skinner J, Ree MJ. (1987). *Air Force officer qualifying test: Item and factor analysis of form O* (AFHRL-TR-86-68). Brooks AFB, TX: Air Force Systems Command.

*Sparks CP, Manese WR. (1970). Interview ratings with and without knowledge of preemployment test scores. *The Experimental Publication System, 4*, 142.

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

Spearman C. (1927). *The abilities of man*. New York: MacMillian.

Switzer, FS. (1992) VGBOOT : A program to derive validity generalization estimates with bootstrapped confidence intervals. *Applied Psychological Measurement, 15*, 360–361.

Thorndike EL. (1921). *Educational psychology*. New York: Teacher's College of Columbia.

Toquam JL, Corpe VA, Dunnette MD. (1989). *Literature review: Cognitive abilities—theory, history, and validity*. U.S. Army, Research Institute for the Behavioral and Social Sciences.

Tyler LE. (1965). *The psychology of human differences (3rd ed.)*. New York: Appleton-Century-Crofts.

*Trottman FK. (1977). Race, IQ, and the middle class. *Journal of Educational Psychology, 69*, 266–273.

U.S. Bureau of Labor Statistics. (2000). *Employment and earnings, 46*(1), 13–14 (Chart A-4).

U.S. Census Bureau. (1999). *Statistical abstract of the United States 1999.* Washington, DC: U.S. Government Printing Office.

U.S. Department of Labor. (1977). *Dictionary of occupational titles.* Washington, DC: U.S. Government Printing Office.

Verive JM, McDaniel MA. (1996). Short-term memory tests in personnel selection: Low adverse impact and high validity. *Intelligence, 23*, 15–32.

Vernon PE. (1961). *The structure of human abilities.* London: Methuen & Co.

Vernon PE. (1979). *Intelligence: Heredity and environment.* San Francisco: WH Freeman.

*Waldman D, Avolio BJ. (1991). Race effects in performance evaluations: Controlling for ability, education, and experience. *Journal of Applied Psychology, 76*, 897–901.

Walters LC, Miller MR, Ree MJ. (1993). Structured interviews for pilot selection: No incremental validity. *International Journal of Aviation Psychology, 3*, 25–38.

*Weekley JA, Jones C. (1997). Video-based situational testing. PERSONNEL PSYCHOLOGY, *50*, 25–49.

*Whitney DJ. (1995). *The influence of racial identity and cultural values on responses to biodata employment items: An investigation of differential functioning.* (Doctoral Dissertation, Michigan State University, 1995). Dissertation Abstracts International, 56/10-B, p. 5813 (298).

*Wilbourn JM, Valentine LD, Ree MJ. (1984). *Relationships of the Armed Services Vocational Aptitude Battery (ASVAB) forms 8, 9, and 10 to Air Force technical school final grades* (AFHRL TR 84-8). Brooks AFB, TX: Manpower and Personnel Division.

Wilk SL, Sackett PR. (1996). Longitudinal analysis of ability–job complexity fit and job change. PERSONNEL PSYCHOLOGY, *49*, 937–967.

Williams WM, Ceci SJ. (1997). Are Americans becoming more or less alike? Trends in race, class, and ability differences in intelligence. *American Psychologist, 52*, 1226–1235.

Willingham WW, Lewis C, Morgan R, Ramist L. (1990). *Predicting college grades: An analysis of institutional trends over two decades.* Princeton: Educational Testing Service.

*Witworth RH, Barrientos GA. (1990). Comparison of Hispanic and Anglo Graduate Record Examination scores and academic performance. *Journal of Psychoeducational Assessment, 8*, 128–132.

*Wonderlic EF, Wonderlic CF. (1972). *Wonderlic Personnel Test: Negro norms.* Northfield, IL: EF Wonderlic & Assoc.

*Wonderlic EF, Wonderlic CF. (1992). *Wonderlic Personnel Test user's manual.* Libertyville, IL: Wonderlic Personnel Test, Inc.

*Wysocki BA, Wysocki AC. (1969). Cultural differences in Wechsler–Bellevue intelligence (WBII) test. *Psychological Reports, 25*, 95–101.