

# Racial/Ethnic Differences in the Criterion-Related Validity of Cognitive Ability Tests: A Qualitative and Quantitative Review

Christopher M. Berry  
Texas A&M University

Malissa A. Clark  
Auburn University

Tara K. McClure  
Wayne State University

The correlation between cognitive ability test scores and performance was separately meta-analyzed for Asian, Black, Hispanic, and White racial/ethnic subgroups. Compared to the average White observed correlation ( $\bar{r} = .33$ ,  $N = 903,779$ ), average correlations were lower for Black samples ( $\bar{r} = .24$ ,  $N = 112,194$ ) and Hispanic samples ( $\bar{r} = .30$ ,  $N = 51,205$ ) and approximately equal for Asian samples ( $\bar{r} = .33$ ,  $N = 80,705$ ). Despite some moderating effects (e.g., type of performance criterion, decade of data collection, job complexity), validity favored White over Black and Hispanic test takers in almost all conditions that included a sizable number of studies. Black–White validity comparisons were possible both across and within the 3 broad domains that use cognitive ability tests for high-stakes selection and placement: civilian employment, educational admissions, and the military. The trend of lower Black validity was repeated in each domain; however, average Black–White validity differences were largest in military studies and smallest in educational and employment studies. Further investigation of the reasons for these validity differences is warranted.

**Keywords:** differential validity, racial/ethnic subgroups, cognitive ability tests, performance

Differential validity refers to differences between subgroups in the correlation between a predictor and a criterion (e.g., Linn, 1978). Research on differential validity has commonly examined differences between racial/ethnic subgroups in the relationship between cognitive ability tests (predictors) and measures of job/training performance or of academic achievement (criteria), with the implicit assumption that cases in which criterion-related validities are lower for traditionally disadvantaged subgroups (i.e., racial/ethnic minorities) are of primary concern. There is a divergence of opinion regarding the existence of differential validity, with some reviews documenting widespread evidence of validity differences (e.g., Young & Kobrin, 2001) and other reviews finding no consistent evidence (e.g., Schmidt, 1988; Schmidt & Hunter, 1981). Given such divergence of opinions, the present study quantitatively summarized existing differential validity evidence by separately meta-analyzing the criterion-related validity of cognitive ability tests for the four racial/ethnic subgroups for which differential validity research is most prevalent: Asians, Blacks, Hispanics, and Whites.

## Cognitive Ability Testing in High-Stakes Selection and Placement

A number of cognitive ability tests have commonly been used for selection and placement in civilian employment (e.g., General Aptitude Test Battery, Wonderlic), military (e.g., Armed Services Vocational Aptitude Battery), and educational admissions settings (e.g., SAT, ACT, GRE). The use of cognitive ability tests in such settings is often referred to as “high-stakes testing” (e.g., Sackett, Borneman, & Connelly, 2008) to reflect that scores on these tests at least partially determine whether applicants will receive a desired job, be allowed into the military or to specific jobs within the military, or be accepted into the college of their choice. Although these tests are not exactly the same, they have in common an attempt to measure cognitive ability, operationalized as developed ability or knowledge. Thus, the present study focuses on this broad range of cognitive ability tests used for selection and placement purposes.

Because there are such high stakes surrounding the use of cognitive ability tests for selection and placement, the validity of the inferences made from these tests is of paramount importance. Given the use of such inferences to predict future performance, a critical form of validity for cognitive ability tests has been their criterion-related validity. Because cognitive ability tests are designed to measure developed ability and knowledge, the most appropriate job/college/military performance criteria are those that depend on technical knowledge, skills, and abilities (i.e., task performance, as opposed to contextual or other performance behaviors that depend less on ability). Therefore, the relationship between cognitive ability test and criterion scores (where criteria

---

This article was published Online First March 28, 2011.

Christopher M. Berry, Department of Psychology, Texas A&M University; Malissa A. Clark, Department of Psychology, Auburn University; Tara K. McClure, Department of Psychology, Wayne State University.

We thank In-Sue Oh for thoughtful comments on earlier versions of this paper. We also thank Mingzhu Yu for assistance with data collection and coding.

Correspondence concerning this article should be addressed to Christopher M. Berry, Department of Psychology, Texas A&M University, 4235 TAMU, College Station, TX 77843. E-mail: cmberry@tamu.edu

are typically some measure of task performance, performance in a job-training program, or college grades) is typically used for indexing the criterion-related validity of cognitive ability tests. Although the absolute magnitude of test–criterion relationships certainly is important, one could argue that a greater concern has been whether cognitive ability tests represent an equivalent and fair assessment for each racial/ethnic subgroup (e.g., Jencks & Phillips, 1998; Steele & Aronson, 1995). One specific fairness issue has been whether cognitive ability test scores relate to performance criteria equally for each subgroup. This issue has been investigated with two complementary methods: differential validity and differential prediction.

### Differential Validity and Differential Prediction

Differential validity focuses on the differences between test–criterion correlation coefficients across subgroups. Differential prediction focuses on differences between unstandardized regression slopes and intercepts relating the test and criterion across subgroups. The differential prediction approach is generally preferred for comparison of predictor–criterion relationships for majority and minority groups in a given sample, for a variety of reasons. First, differential prediction most directly addresses whether test scores predict equivalent criterion scores for different subgroups. Second, unlike the correlation coefficient, the unstandardized regression coefficient is not affected by direct range restriction on the predictor (although it is affected by indirect range restriction; Mendoza & Mumford, 1987). Third, by including separate comparisons of slopes and intercepts, the differential prediction approach is more informative, as the correlation coefficient contains no information relevant to differences in intercepts.

Nonetheless, differential validity remains of interest for at least two reasons. First, although the differential prediction approach is appropriate when examining a specific data set, it is generally not applicable to meta-analysis. Meta-analysis requires a common metric across studies. The correlation coefficient is typically the metric of choice, as it is the standardized covariance between predictor and criterion. The unstandardized regression coefficient can only be meta-analyzed if all studies of interest use exactly the same predictor and criterion measure. However, when addressing broad questions such as “do measures of general cognitive ability relate to job performance equally well for minority and majority job applicants,” one is typically confronted with studies of the ability–performance relationship that employ a variety of ability and performance measures, each using a different metric. As a result, the unstandardized regression slopes are in different metrics for differing studies and are therefore not amenable to meta-analysis. Especially given the low statistical power of differential prediction’s test of slope differences between subgroups (Aguinis, Culpepper, & Pierce, 2010), the inability to meta-analyze slope differences limits the broad conclusions about test–criterion relationships that can be drawn from differential prediction analyses alone. Although one can reframe the question in terms of the frequency with which unstandardized slopes and intercepts differ across subgroups, such an approach does not address the actual magnitude of differences between subgroups. In sum, although the differential prediction approach is appropriately used to examine differences in predictor–criterion relationships across subgroups

for individual applications, differential validity remains of interest if one wishes to combine predictor–criterion relationships by subgroup across multiple studies.

Second, an examination of differential validity and of differential prediction evidence has the potential to provide more information than an examination of differential prediction alone. A predictor–criterion correlation could be the same for two subgroups, but the regression equations relating the predictor and criterion could differ for those subgroups and vice versa (Linn, 1978). For our purposes in the present study, the most important comparison is between the correlation coefficient and the regression coefficient, as the regression coefficient is differential prediction’s analogue to differential validity’s correlation coefficient. For example, the formula relating the regression coefficient to the correlation coefficient is

$$b_{xy} = r_{xy} \left( \frac{s_y}{s_x} \right),$$

where  $b_{xy}$  is the regression coefficient for regressing  $y$  on  $x$ ,  $r_{xy}$  is the correlation between  $x$  and  $y$ , and  $s_x$  and  $s_y$  are the standard deviations of  $x$  and  $y$ , respectively.

As the formula illustrates, the regression coefficient is a function of the correlation and ratio of the standard deviations of the predictor and criterion. Therefore, if one finds differential validity for two groups but the regression coefficients do not differ, this must be due to differences in variability of the predictor, the criterion, or both between subgroups. A number of mechanisms could cause such a difference. For instance, there might be true differences between subgroups in predictor or criterion variability. Differing degrees of range restriction across subgroups could also cause such an effect. More error of measurement affecting scores of one subgroup could cause such an effect. There are doubtless other possibilities. The key point is that it is premature to dismiss differential validity as an unimportant phenomenon. Even in the face of a lack of differential prediction, a finding of differential validity should signal the need for additional investigation, as such a disconnect could be due to a true lack of differential prediction, to artifacts, or even to what some might construe as test bias (e.g., more error in the predictor for one subgroup). Further, the factors that might cause differential validity in the first place (regardless of whether these factors contribute to a disconnect between differential validity and differential prediction evidence) each represent interesting theoretical and practical phenomena. Some of these possible causes are reviewed below.

### Possible Causes of Differential Validity

There are at least four categories of factors that could differentially affect test and/or criterion scores of minority and majority subgroup test takers, thus causing criterion-related validity estimates to differ between subgroups: range restriction, psychometric characteristics of tests or criteria (i.e., measurement error/bias), contextual influences (e.g., stereotype threat), or true differences between subgroups in the role cognitive ability plays in determining performance. These are, of course, not the only possible factors; however, they are particularly plausible ones and are illustrative of the types of things that could cause validity to differ between subgroups. Because instances in which the minority sub-

group's validity is lower are of most concern, the following discussion is focused on factors that could cause criterion-related validity estimates to be lower for the minority subgroup.

Greater amounts of range restriction in test scores of the minority subgroup could cause observed validity to be lower, even if the true validity of the test did not differ by subgroup. Given that White subgroup members, on average, score higher than Black and Hispanic subgroup members on cognitive ability tests (Roth, Bevier, Bobko, Switzer, & Tyler, 2001), one explanation for instances of lower Black or Hispanic validity is that minority groups have greater amounts of range restriction (cf. Boehm, 1972; Hunter, Schmidt, & Hunter, 1979). However, empirical evidence is mixed. First, due to affirmative action, similar cut scores for minority and White subgroups are not always used in practice. The use of different cut scores is perhaps most likely in college admissions, where recent court cases have upheld some forms of racial preference (see the Supreme Court cases *Gratz v. Bollinger*, 2003, and *Grutter v. Bollinger*, 2003, for illustrative examples), and least likely in the military, where there is no federal executive order mandating affirmative action. The sparse available empirical evidence supports this trend, with White samples often being more restricted in range than Black and Hispanic samples, especially in college admissions settings (Berry, 2007). Thus, relative amounts of range restriction are not always in the direction that would be needed to account for lower minority validity. Regardless, differences between subgroups in range restriction would influence variance, which would in turn affect validity estimates.

If internal, psychometric characteristics of the tests, such as measurement error or measurement bias, differed between subgroups, this could also cause validity differences between subgroups. Measurement error is typically assessed via reliability estimates. Measurement bias refers to instances in which individuals who are identical on the construct measured by the test/criterion but who are from different subgroups have different probabilities of attaining the same observed score. Measurement bias has mostly been investigated from two perspectives: differential item functioning (DIF) and measurement invariance. The DIF literature investigates measurement bias associated with individual test/criterion items, typically in the form of item discrimination or difficulty differences between subgroup members matched on a latent trait (e.g., Humphreys, 1986). In contrast, the measurement invariance literature investigates measurement bias at the level of the scale or test, typically in the form of differences between subgroups in the factor structure of the test. Of course, these three concepts (measurement error, DIF, and measurement invariance) are related, such that the presence of one can lead to the manifestation of the others. For instance, if DIF is present and items do not discriminate as well for one subgroup, this could cause differences in the factor structure or reliability of the test. However, these three concepts are not the same; the presence of one does not mandate the others (e.g., if items are more difficult for one subgroup, this does not necessarily mean the factor structure of the test is different for those subgroups). Therefore, the literatures on subgroup differences in each of these three psychometric characteristics are reviewed separately below but with the acknowledgment that these are related concepts.

One way that measurement error could differ between groups is via differential guessing rates. Hunter and Schmidt (1978) noted that because Blacks, on average, score lower than Whites on

cognitive ability tests, Black test takers will be more likely to guess answers. This increased guessing would lower reliability and would in turn lower Black validity. A similar argument could be made for the Hispanic subgroup (which, as a group, has lower test scores than the White subgroup; Roth, Bevier, et al., 2001) but probably does not apply to the Asian subgroup (which as a group has similar test scores to the White subgroup; e.g., Herrnstein & Murray, 1994). The few studies reporting separate reliability estimates for racial/ethnic subgroups have not supported the idea that test reliabilities differ across subgroups (e.g., Domino & Morales, 2000; Jensen, 1977; Stark, Chernyshenko, & Drasgow, 2004). Subgroups could also differ in criterion reliability. The few studies reporting reliabilities separately for Black and White subgroups do support a trend for criterion reliability to be slightly lower (i.e., usually a few reliability points) for Blacks than for Whites (Berry & Sackett, 2008a; Kraiger & Ford, 1990; Willingham, Pollack, & Lewis, 2000). Empirical evidence for other subgroups is lacking. However, it is worth noting that it would take very large differences in test or criterion reliability between groups to account for relatively small differences in validity. For example, if the reliability for majority test takers was .90 and observed validity was .25 and .35 for minorities and majorities, respectively, reliability for minority test takers would have to be .45 to account for this validity difference.

To date, the literature is lacking studies investigating DIF of performance criteria. DIF studies investigating cognitive ability tests have often uncovered many individual items for which minority (typically Black or Hispanic) and White test takers with similar trait/construct standing do not have equal probability of answering correctly (which suggests those items are "biased"). Yet, it should be noted that these differences are usually small and evenly distributed in favor of and against minority test takers, suggesting a lack of bias at the test total score level (Hough, Oswald, & Ployhart, 2001; O'Neill & McPeck, 1993). For instance, O'Neill and McPeck reviewed the DIF literature and found that although there were a number of characteristics that were consistently associated with DIF against minorities (e.g., Black and Hispanic test takers perform less well than White test takers on concrete analogies), there were just as many characteristics associated with DIF in favor of minorities (e.g., Black and Hispanic test takers perform better than White test takers on abstract analogies). In a more recent review, Hough et al. (2001) came to similar conclusions. Comparable trends of positive and negative DIF findings canceling out at the total score level have been found in more recent DIF research (e.g., Stark et al., 2004).

In the present setting, cognitive ability test measurement invariance is typically operationalized as the degree to which the factor structure of a given cognitive ability test is equivalent for minority and majority subgroups (i.e., factorial invariance). If the factor structure differs between subgroups, the psychological meaning of test scores is not the same for each subgroup, which could affect the degree to which test scores are predictive of performance criteria. However, the empirical factorial invariance research to date has not been supportive of differences between subgroups in test factor structure. Jensen (1980) reviewed a number of studies demonstrating that the factor structure of cognitive ability tests was similar across subgroups, although these studies mostly focused on Black and White subgroups and used exploratory factor analysis (multigroup confirmatory factor analysis has since be-

come the more accepted measurement invariance methodology). A few large-sample measurement invariance studies have been carried out in the military literature, with each finding the factor structure of cognitive ability tests (the Armed Services Vocational Aptitude Battery or Air Force Officer Qualifying Test) to be invariant across the racial/ethnic subgroups examined (Asian, Black, or Hispanic subgroups compared to Whites; Carretta & Ree, 1995; Drasgow, Nye, Carretta, & Ree, 2010; Ree & Carretta, 1995). Similar factorial invariance results have been found for child samples with the Wechsler Intelligence Scale for Children (Dolan, 2000; Pandolfi, 1997; Reed, 2000).

Formal factorial invariance research investigating the degree to which performance criteria are invariant across minority and majority subgroups has not been conducted. However, a sizable amount of research exists on other factors relevant to the cross-race comparability of the construct validity of supervisor ratings of job performance. To the degree that the psychological meaning of performance ratings differs for minorities and this difference is a function of things other than true performance differences (e.g., racial discrimination), cognitive ability test scores may not predict performance ratings as well for minorities. For instance, three large-sample studies have investigated the degree to which Black employees receive similar performance ratings when their performance is rated by their Black versus White supervisors (Pulakos, White, Oppler, & Borman, 1989; Rotundo & Sackett, 1999; Sackett & DuBois, 1991). In each study, Black employees received lower ratings from their White supervisors than their Black supervisors (*ds* ranging from 0.10 to 0.27). Because the employees are held constant (i.e., a single employee was rated by both a Black and a White supervisor), differences in the ratings suggest that race plays a role in performance ratings relatively independent of true levels of performance. Similar trends have been outlined in laboratory studies (see, e.g., the December 2008 special issue of *Industrial and Organizational Psychology: Perspectives on Science and Practice*). Other studies have investigated the degree to which Black employees' performance ratings correlate with more objective measures of performance (e.g., production, output, error rates), finding that performance ratings of Black employees are more related to these objective measures than are performance ratings of White employees (Kraiger & Ford, 1990; Oppler, Campbell, Pulakos, & Borman, 1992). This suggests that the psychological meaning of performance ratings differs for Black and White employees, which could affect the comparative predictability of performance ratings for Black and White employees. Comparable research for other racial/ethnic minorities is sparse to nonexistent. In sum, the empirical evidence to date does not support the idea of internal psychometric characteristics of tests (i.e., measurement error/bias) differing between subgroups, although there is some evidence of such differences for performance criteria.

Contextual influences in the testing situation could also affect the validity of minority and majority test scores if their effects differ systematically for the groups. Perhaps the most common example of such a contextual influence is stereotype threat. Stereotype threat refers to the idea that the testing situation can cause minority test takers to feel the threat of confirming a negative stereotype about their racial/ethnic subgroup and that this threat will lead to reduced test performance (Steele & Aronson, 1995). Because minority test takers' scores would reflect true cognitive

ability plus variance due to stereotype threat, stereotype threat would act as construct-irrelevant variance that could cause test scores of minority test takers to be less related to true criterion performance. For instance, Wicherts, Dolan, and Hessen (2005) demonstrated in three lab samples that the factor structure of cognitive ability tests changed in stereotype threat manipulation conditions, which suggests that the psychological meaning of test scores can change as a function of stereotype threat. Such measurement bias in test scores as a function of stereotype threat could cause test scores to be differentially related to performance for minority and majority test takers.

Up to this point in the review of possible causes of differential validity, only statistical or measurement artifacts have been considered. Another possibility is that differential validity reflects true differences between groups in the role of cognitive ability in the determination of performance. For instance, process models of the determinants of job performance suggest that cognitive ability has its most direct effect on acquisition of job knowledge, which in turn is a direct determinant of job performance (Schmidt & Hunter, 1992). To the degree that there are differences between subgroups in the roles that major predictors (e.g., ability, knowledge) play in determining performance, this could cause cognitive ability to be less related to performance for one group. For example, in the educational domain, it is reasonably well established that standardized test scores underpredict grades of women, in part because their grades tend to be driven more by effort and planfulness than are grades of men (Ramist, Lewis, & McCamley, 1990; Stricker, Rock, & Burton, 1993). In this case, the cognitive ability test is accurately capturing true differences between groups in the determinants of performance. If such is the case with racial/ethnic subgroups, these true differences in the determinants of performance could cause differences in the validity of cognitive ability tests.

This section has made the case that there are a number of factors that could differ between racial/ethnic subgroups and affect relevant properties of cognitive ability test or criterion scores, thereby influencing the relative validity of tests for minority and majority subgroups. Of course, not all of the factors listed above need to be present for validities to differ between subgroups; if any of the above factors (or other relevant factors not reviewed here) differ systematically between minority and majority subgroups, this could result in differential validity. Observed validity differences are almost inevitable when thought of from this perspective. Only if true validities are equal for each subgroup and all of the above discussed factors have no net effect on test or criterion scores would observed validity be equivalent for minority and majority test takers. However, to this point in time, at least in industrial/organizational psychology, it has been almost a foregone conclusion that observed validities do not differ between racial/ethnic subgroups (e.g., Schmidt, 1988; Schmidt & Hunter, 1981), with the implicit extrapolation that true validities do not differ either. Therefore, in the following section the empirical evidence regarding the existence of differential validity is reviewed. Additionally, the present meta-analysis stands as a comprehensive quantitative summary of the available empirical evidence. If observed validities are found to differ between subgroups, this suggests that more attention toward understanding the effects of the above-reviewed factors affecting validity is warranted.



### Empirical Evidence for and Against Differential Validity

There is currently a divergence of opinion across the civilian employment, military, and educational admissions fields regarding the existence of differential validity. The civilian employment literature (mostly under the purview of industrial/organizational psychology) has all but completely dismissed differential validity of cognitive ability tests as a nonissue (e.g., Schmidt, 1988; Schmidt & Hunter, 1981). This is at least partially due to a number of influential reviews from the 1970s in the civilian employment literature that demonstrated that statistically significant differences between pairs of minority and White validity coefficients were found only at chance levels (e.g., Hunter & Schmidt, 1978; Hunter et al., 1979; O'Connor, Wexley, & Alexander, 1975; Schmidt, Pearlman, & Hunter, 1980). In contrast, lower criterion-related validity of cognitive ability tests for some minority subgroups (most notably, Black and Hispanic subgroups) is generally an accepted phenomenon in the educational admissions literature, due to a preponderance of empirical evidence (e.g., Linn, 1982; Young & Kobrin, 2001). We are not aware of strong statements in the military selection and placement literature regarding the existence of differential validity, although an examination of the empirical evidence from military studies demonstrates that the existence of differential validity is at least plausible (e.g., Houston & Novick, 1987; McLaughlin, Rossmeissl, Wise, Brandt, & Wang, 1984; Valentine, 1977).

One possible explanation for this divergence of opinion is the statistical approach used in major studies within each domain. A key distinction between the differential validity evidence in the civilian employment literature versus the educational admissions and military literatures is a focus on statistical significance testing versus comparisons of effect sizes. In particular, the educational admissions and military literatures have relied on comparisons of effect sizes in large samples when investigating the differential validity of cognitive ability tests. The typical differential validity study design in educational admissions or military settings entails the collection of large minority and majority subgroup samples and the calculation of test–criterion correlations separately within each racial/ethnic subgroup. The relative magnitude of minority versus majority correlations is used as evidence for or against differential validity. The general trend in these domains has been for test–criterion correlations to be between .02 and .15 correlation points smaller for Black and Hispanic samples than for White samples (e.g., Breland, 1979; Bridgeman, McCamley-Jenkins, & Ervin, 2000; Duran, 1983; Houston & Novick, 1987; Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008; McLaughlin et al., 1984; Morgan, 1990; Valentine, 1977; Young & Kobrin, 2001) and for test–criterion correlations to be relatively comparable for Asian and White samples (Young & Kobrin, 2001).

Within the civilian employment literature, the typical differential validity study design has been to calculate the significance of the difference between minority and majority correlations within a number of samples and then determine whether the frequency of statistically significant minority–majority correlation differences exceeds what one would expect due to chance alone (for representative examples, see Hunter & Schmidt, 1978; Hunter et al., 1979; O'Connor et al., 1975; Schmidt et al., 1980). The general trend in these statistical significance studies was a lack of statis-

tically significant differences between Black, Hispanic, and White correlations (Asian comparisons were never included) more often than what would be expected by chance (Schmidt, 1988). However, the small sample sizes of the minority subgroups (e.g., the average Black sample size in Schmidt, Berner, & Hunter's 1973 review was 49) combined with the relatively small-to-moderate size of the typical minority–majority validity differences (e.g., typically about .02 to .15 correlation points lower for Black and Hispanic subgroups in the studies above that reported effect sizes) make conclusions from the statistical significance studies less interpretable.

Many of the above-mentioned criticisms of the statistical significance studies of differential validity could be addressed through the use of modern meta-analytic techniques. It is instructive to examine the limited meta-analytic evidence regarding differential validity in the civilian employment literature. In one of the statistical significance studies reviewed earlier, Hunter et al. (1979, p. 727) reported that "the overall mean racial difference in validity was .02 (Whites higher)" across 34 validity studies. Schmidt et al. (1980) reported two Hispanic–White meta-analytic comparisons based on 19 studies including 28 samples. In the seven educational admissions samples, Hispanic validity was .21 and White validity was .28. In the 21 employment samples, Hispanic validity was .18 and White validity was .20. Hartigan and Wigdor (1989) reported the results of meta-analytic investigations of differential validity carried out by Synk and Swarthout (1987) in their study of the validity of the General Aptitude Test Battery (GATB), a cognitive ability test that was heavily used by the U.S. Employment Service prior to Hartigan and Wigdor's report. The meta-analysis included validity coefficients (correlations between GATB scores and mostly supervisor ratings of job performance) for Blacks and Whites drawn from 113 validity studies (Black  $N = 7,854$ ; White  $N = 15,768$ ). The sample-size-weighted mean observed validity was .19 for Whites and .13 for Blacks. Meta-analytic evidence for subgroups other than Blacks, Hispanics, and Whites is not available in the industrial/organizational psychology literature.

It is noteworthy that the statistical significance testing evidence in employment settings generally found no evidence of differential validity, and the meta-analytic evidence in employment settings found evidence of lower observed validity for minorities. The differences in magnitude found by meta-analytic studies are relatively similar to those found in educational and military contexts. Upon review of the meta-analyses, the evidence in employment contexts aligns more with the evidence in educational and military contexts.

Therefore, across all three broad fields that commonly use cognitive ability tests for high-stakes selection and placement, there is evidence that lower criterion-related validity for Black samples is relatively common. Evidence for Hispanic samples is similar but is mostly available in educational admissions settings, with some small amount of concurring evidence in civilian employment settings. Evidence for Asian samples is available only in educational admissions settings and suggests that Asian–White validity is relatively comparable. Therefore, differential validity appears to be a common phenomenon, at least for Black–White and Hispanic–White comparisons. This is not to say that there is not strong conflicting evidence. For instance, although there were a number of methodological issues making the results of the

statistical significance studies less interpretable, the results of such studies in the civilian employment literature did not support differential validity (Hunter et al., 1979; Schmidt et al., 1980). Even among the effect size studies from the educational admissions and military literatures, there are a number of studies, some incorporating very large samples, that have not found support for lower criterion-related validity for Black and/or Hispanic subgroups (e.g., Bridgeman et al., 2000; Carretta, 1997; Roberts & Skinner, 1996; Wightman & Muller, 1990). Given this conflicting evidence, a meta-analytic estimate of racial/ethnic subgroup criterion-related validity differences, both across and within these three broad literatures, would be especially useful for summarizing the research to date and highlighting areas for future research. The present study is the first such comprehensive meta-analysis.

### Present Study

The present meta-analysis explores whether the evidence to date is or is not supportive of the existence of differential validity between four racial/ethnic subgroups: Asian, Black, Hispanic, and White. In addition to investigating the overall average magnitude of differential validity, the present study investigated the influence of four potentially important moderators of minority-majority validity differences. The first moderator was study domain. Studies included in the present meta-analysis came from the three broad research literatures, or domains, that most commonly use cognitive ability tests for high-stakes selection purposes: educational admissions, employment, and military. Although the three domains use similar tests for similar purposes, there are substantive differences between the three domains. For instance, the jobs of college student, civilian employee, and soldier have only relatively superficial similarities. Also, although performance likely has a somewhat similar meaning in each domain (e.g., demonstrating technical skill and knowledge), there are clear differences as well (e.g., there are aspects of being a good soldier that probably do not translate to being a good college student). Additionally, although the tests used by each domain are generally similar (Beaujean et al., 2006; Drasgow, 2003; Frey & Detterman, 2004; Kuncel, Hezlett, & Ones, 2004), each domain tends to use different cognitive ability tests. Given such systematic differences between domains in the specific types of jobs and tests used, study domain was investigated as a moderator of subgroup validity differences.

The second potential moderator was type of criterion. Performance criteria can differ in the degree to which they are objective versus subjective. For instance, dollar volume of sales is a relatively objective criterion in that it is countable and verifiable. Supervisor ratings of performance, on the other hand, represent a subjective opinion. To the degree that the performance criterion represents a subjective judgment, the possibility of racial/ethnic bias or discrimination affecting criterion scores increases. If this is the case, racial/ethnic bias could act as construct-irrelevant variance in criterion scores. Thus, subjective performance ratings of minorities may be less related to cognitive ability test scores than are subjective performance ratings of Whites, suggesting there may be a larger cognitive ability test validity gap between minority and majority members if the criterion is relatively subjective as opposed to relatively objective. Thus, type of performance criteria may moderate subgroup validity differences.

The third potential moderator was the decade in which the cognitive ability test scores were collected. The studies used in this meta-analysis spanned more than 40 years. It is possible that over the course of 40 years, the average cognitive ability test may have changed substantially. For instance, researchers and practitioners have been concerned about the possibility of test bias since at least the 1960s (e.g., Cleary, 1968), and over time more sophisticated methods for detecting biased items have been developed (e.g., item response theory, DIF). Thus, test publishers may have increased efforts over the years to detect and eliminate any possible bias in their tests. If such were the case, the minority-majority validity gap may have decreased over time. Thus, the effect of time on meta-analytic results was investigated.

The fourth potential moderator examined was job complexity. Job complexity refers to the information processing demands of a job, with more complex jobs entailing greater information processing demands (e.g., Hunter, Schmidt, & Judiesch, 1990). Some past research has found that job complexity moderates the validity of cognitive ability tests, although conflicting results exist. For instance, Hunter's (1980) meta-analysis of 515 GATB validity studies (total  $N = 38,620$ ) found that job complexity moderated the relationship between GATB scores and supervisor ratings of job performance, such that the GATB-job performance relationship became stronger for jobs of higher complexity. On the other hand, Hartigan and Wigdor (1989) did not find evidence of such a moderating effect of job complexity in their meta-analysis of 264 separate GATB validity studies (total  $N = 38,521$ ). Regardless, if job complexity does moderate cognitive ability test validity, this could confound the results of the present differential validity meta-analysis if minority and majority employees cluster into jobs of different complexity. Thus, the present study investigated whether job complexity might moderate subgroup validity differences.

### Method

#### Search for Primary Data

First, a keyword search of the PsycINFO and ERIC databases was performed, with a combination of the following keywords: *race*, *ethnic*, *ethnicity*, *African American*, *Black*, *Hispanic*, and *Asian*, each combined with *differential validity*, *cognitive ability*, *intelligence + performance* (the keyword performance subsumes other keywords such as job performance, military performance, and academic performance), *intelligence + grades*, *incremental validity*, *adverse impact*, and *differential prediction*). Thus, 49 different keyword combinations were used in both the PsycINFO and ERIC searches. Second, major reviews of differential validity were consulted (i.e., Aguinis et al., 2010; Boehm, 1972; Breland, 1979; Duran, 1983; Hunter et al., 1979; Schmidt et al., 1980; Young & Kobrin, 2001), and all references that appeared to contain relevant data were obtained. Third, the websites of the owners of college admissions tests (e.g., Educational Testing Service, College Board) were searched for any relevant studies. Fourth, calls for unpublished studies were posted on the Society for Industrial and Organizational Psychology's online bulletin board and on the Academy of Management's Research Methods and Human Resources listservs.

In order to be included in the meta-analysis, the study had to provide separate correlations between some form of cognitive ability test and some type of performance criterion (e.g., job performance, training performance, academic performance) for a minority (i.e., Asian, Black, or Hispanic) and White adult sample (e.g., samples using children or high school students were excluded). The overwhelming majority of cognitive ability tests included in this meta-analysis were multifacet measures providing an overall score that is a composite of the multiple facets (e.g., the SAT total score is a composite of the facet-level Verbal and Mathematical subtests). Performance measures focusing on core, technical aspects of the role of employee/student/soldier (e.g., measures of employees' task performance, measures of students' grades in college courses, and measures of soldiers' grades in technical training programs) constituted the vast majority of criteria included.

The final sample included 166 studies that were used in at least one of three meta-analyses: an Asian-White comparison meta-analysis (including 60 White and 60 Asian correlations), a Black-White meta-analysis (including 405 White and 392 Black correlations), and a Hispanic-White meta-analysis (including 97 White and 97 Hispanic correlations). See the Appendix for tables listing information for each of the primary studies included in the meta-analysis. Across the three meta-analyses, there was considerable overlap in the White samples included (e.g., most samples that included a White-Asian comparison also included White-Black and White-Hispanic comparisons, so a given White correlation might be used in all three meta-analyses). Each of the correlations was drawn from independent samples, meaning each Asian, Black, Hispanic, and White sample contributed only one correlation coefficient to any meta-analysis.

Of the 166 primary studies, 113 were drawn from Synk and Swarthout's (1987) meta-analysis of GATB validity studies. All 113 studies in Synk and Swarthout were in the form of unpublished technical reports. Although every effort was made to locate all 113 unpublished technical reports, there were many that could not be located. Thus, instead of coding each of the 113 individual studies, we used the five meta-analytic estimates (one for each of five job families) reported by Synk and Swarthout in their Table 4. Further, to save space in the References section, we listed Synk and Swarthout (1987) instead of each of the 113 primary studies. As a result, there are only 54 primary studies listed in the reference section as having been included in the present meta-analysis ( $166 - 113 + 1 = 54$ ). This has no effect on the present study's mean meta-analytic correlation estimates. However, it does affect the estimates of variance, as the variance of correlations across 113 primary studies should be greater than the variance across five meta-analytic correlations.

Although a large number of Black and White correlations were located within the civilian employment, educational admissions, and military domains (a breakdown of the number of Black and White correlations within each domain is listed in Table 1), very few Hispanic or Asian correlations were located outside of the educational admissions domain (8 and 1, respectively). As a result, the Asian-White and Hispanic-White differential validity meta-analyses included data only from the educational admissions domain. Although Schmidt et al. (1980) included 1,323 Hispanic-White civilian employment validity pairs, we could not locate enough Hispanic-White differential validity studies for a civilian

employment meta-analysis. There are two reasons for this. First, Schmidt et al.'s 1,323 validity pairs were actually drawn from only 19 studies and did not represent 1,323 independent samples. The 1,323 validity pairs reflected the same set of samples providing perhaps hundreds of validity pairs each (e.g., if one sample completed 10 cognitive tests and 10 performance criteria, this was treated as 100 validity pairs). Second, the 19 studies included in Schmidt et al. were almost all unpublished technical reports carried out between 1969 and 1980. Every effort was made to obtain these unpublished technical reports, but they were all unobtainable.

Another note is necessary regarding the number of samples included in the present meta-analysis as compared to Hunter et al.'s (1979) and Schmidt et al.'s (1980) reviews. For instance, Hunter et al. included 866 Black-White validity pairs. At first glance, this appears to be a greater number than the present study's 797 Black-White correlations. However, Hunter et al.'s 866 validity pairs were based on only 34 independent samples (11 of which were GATB studies included in Synk & Swarthout, 1987). If Hunter et al.'s method had been used in the present study and all possible test-criterion correlation combinations (i.e., validity pairs) were treated as independent samples, it is unknown how many Black-White validity pairs there would be in the present study. Given that the total number of Black-White primary studies (166) in the present meta-analysis was roughly five times larger than that in Hunter et al. (34), it is likely that the current study would have had approximately five times as many validity pairs. The same is true if one compares Schmidt et al.'s Hispanic-White validity pairs to those included in the present study.

### Coding of Study Characteristics

For each independent sample, the correlation between the cognitive ability test and performance criterion was coded, along with the racial/ethnic subgroup and sample size. If multiple cognitive ability tests (e.g., SAT Verbal and SAT Mathematical scores) and/or multiple related performance criteria (e.g., subjective performance ratings and an objective performance index) were included within a single sample, composite formulas (Ghiselli, Campbell, & Zedeck, 1981, pp. 163-164) were used to estimate the correlation between a composite of the multiple tests and/or the multiple criterion measures when intercorrelations among multiple predictors and/or criteria were provided. If intercorrelations were not provided, the multipredictor-criterion correlations were combined by averaging predictor-criterion correlations across the multiple tests/criteria.

Efforts were made to code for statistical artifacts (e.g., range restriction information, reliability information), but these variables were not reported separately for each subgroup in primary studies frequently enough for inclusion in the present meta-analysis. Whether the cognitive ability test was designed to measure a single facet of ability, or multiple facets of ability, or a single higher-order ability factor was coded, there was virtually no variability on this variable (i.e., all but a couple of samples used multiple facet-level measures that were combined into an overall composite score, such as the SAT), so it was not included in moderator analyses. The four moderator variables that were included are described below.

**Study domain.** Study domain was included as a moderator with three levels: educational versus employment versus military

Table 1  
*Black-White Differential Validity Meta-Analysis Results*

Variable	<i>N</i>		<i>k</i>		$\bar{r}$		<i>SD<sub>r</sub></i>		% var		95% CI	
	White	Black	White	Black	White	Black	White	Black	White	Black	White	Black
Overall meta-analyses across and within domains												
All studies	903,779	112,194	405	392	.33	.24	0.06	0.11	6.89	21.16	[.32, .34]	[.23, .25]
Education	759,462	60,096	169	156	.34	.30	0.05	0.08	5.74	33.22	[.33, .35]	[.29, .31]
Employment	20,399	10,350	143	143	.19	.16	0.07	0.10	30.53	31.58	[.18, .20]	[.14, .18]
Military	123,918	41,748	93	93	.34	.17	0.07	0.08	12.63	37.01	[.33, .35]	[.15, .19]
Moderator analyses												
Employment												
Objective	2,819	1,864	15	15	.24	.31	0.15	0.13	20.91	39.52	[.16, .32]	[.24, .38]
Subjective	18,990	9,277	136	136	.19	.14	0.06	0.07	40.17	65.82	[.18, .20]	[.12, .16]
Decade of data collection												
Education												
1960s	5,985	3,946	25	32	.42	.30	0.09	0.13	34.35	38.16	[.38, .46]	[.25, .35]
1970s	118,545	6,223	18	10	.39	.34	0.05	0.06	4.34	34.85	[.37, .41]	[.30, .38]
1980s	229,577	13,281	71	61	.35	.30	0.05	0.08	8.37	56.78	[.34, .36]	[.28, .32]
1990s	190,947	16,215	51	50	.33	.33	0.06	0.10	6.06	26.88	[.31, .35]	[.30, .36]
2000s	214,408	20,431	4	3	.30	.26	0.00	0.00	100	100	[.30, .30]	[.25, .27]
Employment												
1960s	1,151	616	9	9	.26	.27	0.15	0.14	29.61	64.89	[.16, .36]	[.18, .36]
1970s	2,997	1,643	20	20	.22	.27	0.13	0.18	37.19	33.24	[.16, .28]	[.19, .35]
1980s	15,769	7,854	113	113	.19	.13	0.03	0.03	29.62	80.75	[.16, .22]	[.10, .16]
Military												
1970s	31,779	10,595	40	40	.33	.17	0.06	0.07	31.07	67.63	[.31, .35]	[.15, .19]
1980s	92,139	31,153	53	53	.34	.17	0.07	0.08	8.76	27.59	[.32, .36]	[.15, .19]
Job complexity												
Employment												
Low	7,001	4,615	56	56	.19	.14	0.07	0.08	36.38	46.75	[.17, .21]	[.11, .17]
Medium	12,082	5,325	76	76	.21	.19	0.06	0.10	29.27	26.76	[.19, .23]	[.16, .22]
High	1,123	303	7	7	.11	.03	0.09	0.08	29.62	100	[.01, .21]	[-.08, .14]
Military												
Low	1,864	627	4	4	.35	.14	0.04	0.10	100	59.59	[.31, .39]	[.04, .24]
Medium	47,862	11,276	53	53	.38	.20	0.07	0.10	15.96	46.04	[.36, .40]	[.17, .23]
High	17,387	2,653	13	13	.35	.25	0.04	0.10	42.62	43.75	[.33, .37]	[.20, .30]

Note. *N* = total sample sizes; *k* = number of correlations;  $\bar{r}$  = mean sample-size-weighted observed correlation; *SD<sub>r</sub>* = sample-size-weighted observed standard deviation of correlations; % var = percentage of variance attributable to sampling error; CI = confidence interval.

studies. Asian and Hispanic data were available only within the educational domain, so this moderator analysis is relevant only for Black-White validity comparisons. Further, because these three domains are so broad and because a great amount of variability within each of these three broad domains was expected, all of the other moderator analyses were carried out only within each of these three domains. Thus, the study domain moderator can be thought of as a "supermoderator" in the Black-White meta-analysis, with the following moderators as "submoderators" nested within the study domain moderator analysis, as applicable.

**Type of criterion.** This moderator was not relevant for educational or military samples, as these samples virtually always used grades (in college courses and military training courses, respectively) as criteria. For Black-White employment samples, performance criteria were coded into two categories for use in moderator analyses: subjective and objective criteria. Subjective criteria referred to subjective performance ratings of participants made by some person other than the participants themselves (e.g., supervisor ratings of job performance). Objective performance criteria included typical objective measures using verifiable, countable units (e.g., dollar volume of sales,

number of units produced, error rates), as well as work samples and job knowledge tests. Although this objective-subjective distinction is of course not perfect, subjective performance ratings likely make it more possible for idiosyncratic rater biases to affect performance criteria than the criteria included in the "objective" category.

**Decade of data collection.** For the Asian-White, Black-White, and Hispanic-White meta-analyses, the decade in which the study was conducted was coded into one of five categories: 1960s, 1970s, 1980s, 1990s, and 2000s. In most cases, the decade of data collection corresponded to the decade in which the study was published. Some studies' data collection spanned decades (e.g., Moffatt, 1993); in these instances, studies were coded into the decade in which the majority of data were collected. Two studies (Baggaley, 1974; Tracey & Sedlacek, 1985) had samples in which data were collected equally across decades; in these cases the study was coded based on the decade in which the study was published.

**Job complexity.** This moderator analysis was relevant only for employment and military studies, as all educational studies used college students as participants. Job complexity was coded



into low, medium, or high complexity based on job titles, according to a three-level framework developed by Hunter et al. (1990).

### Accuracy Checks

The second and third authors and a graduate assistant each independently coded a common set of 17 articles, including 106 independent samples, to calculate interrater agreement. Agreement was calculated for the coding of racial group, correlations, sample sizes, name of cognitive ability test, study domain, and the three other moderator variables. Across all variables, agreement was quite high. Overall, average agreement across all raters and variables was 98.84%, and for no variable or rater combination did agreement fall below 92%. All coding disagreements were minor and were resolved via discussion as needed. Once adequate agreement was obtained, the second and third authors and the graduate assistant divided up the remaining studies to be coded.

### Analyses

Meta-analytic mean correlations, standard deviations, estimated percentage of variance due to sampling error, and confidence intervals were calculated separately for each racial/ethnic subgroup. Formulas presented by Hunter and Schmidt (2004) were used to calculate meta-analytic mean correlations, standard deviations, and percentage of variance due to sampling error. Additionally, confidence intervals around mean correlations were calculated with formulas provided by Whitener (1990). Moderator analyses were carried out with the same techniques as with the full sample meta-analyses.

## Results

### Black–White Differential Validity Meta-Analysis

**Overall Black–White results.** Overall meta-analytic validities reported separately for Blacks and Whites across all studies, regardless of study domain, are listed in the first row of results in Table 1. Cognitive ability test criterion-related validity for the White sample ( $\bar{r} = .33$ ,  $k = 405$ ,  $N = 903,779$ ) was .09 higher than for the Black sample ( $\bar{r} = .24$ ,  $k = 392$ ,  $N = 112,194$ ). Ninety-five percent confidence intervals for the Black and White samples did not overlap. Although this is evidence for differential validity, the percentage of variance accounted for was very small, suggesting the presence of moderator variables. Thus, the within-study-domain moderator analyses were carried out next.

**Overall Black–White results within study domains.** The second through fourth rows in the first section of Table 1 list the overall test validities separately for Blacks and Whites within the educational admissions, civilian employment, and military domains. Within the educational domain, criterion-related validity was .04 higher for the White sample ( $\bar{r} = .34$ ) than for the Black sample ( $\bar{r} = .30$ ). Within the employment domain, criterion-related validity was .03 higher for the White sample ( $\bar{r} = .19$ ) than for the Black sample ( $\bar{r} = .16$ ), although confidence intervals overlapped. Within the military domain, criterion-related validity was .17 higher for the White sample ( $\bar{r} = .34$ ) than for the Black sample ( $\bar{r} = .17$ ). Thus, criterion-related validity of cognitive ability tests was greater for Whites than for Blacks within each of the three

study domains. However, there were sizable differences between the three domains in the magnitude of these validity differences, with the validity differences being only .03–.04 in the education and employment domains but much larger (.17) in the military domain.

**Black–White moderator analyses.** The remaining moderator analyses were carried out within study domains. Meta-analytic results for the moderator analyses are also listed in Table 1.

**Type of criterion.** Within the employment studies, 136 samples used subjective criteria (e.g., supervisor ratings of performance), and a relatively small number of samples (15) used criteria falling into the objective criterion category. Thus, the type of criterion moderator analysis within employment samples should be considered tentative, as these results may be affected by second-order sampling error (Hunter & Schmidt, 2004). The average ability–performance correlation was .05 higher for Whites ( $\bar{r} = .19$ ) than for Blacks ( $\bar{r} = .14$ ) in samples using subjective criteria. However, the average ability–performance correlation was .07 higher for Blacks ( $\bar{r} = .31$ ) than for Whites ( $\bar{r} = .24$ ) in samples using objective criteria. Thus, although these results were based on very few samples, type of criterion moderated Black–White validity differences within employment samples.

**Decade of data collection.** Decade moderated validity differences within the educational admissions samples, with validity differences of .12, .05, .05, .00, and .04 in favor of White samples in the 1960s, 1970s, 1980s, 1990s, and 2000s, respectively. At least two trends are noteworthy. First, there has been a general trend within the educational studies for Black–White validity differences to reduce over time. From the 1960s to the 1990s this validity difference reduced from .12 to .00, although the average validity difference was .04 in the few large differential validity samples carried out since 2000. Second, despite the overall reduction in the Black–White validity gap over time, average validity was lower for Black samples than for White samples in each decade except for the 1990s.

A different pattern of results is apparent in the employment samples. In the 1960s and 1970s there were validity differences of .01–.05 in favor of Black samples. However, in the 1980s, average validity was .06 higher for White samples than for Black samples. This highlights at least three noteworthy points. First, the present study reaches a conclusion similar to those of the statistical significance studies from the 1970s in the employment domain regarding the lack of evidence of Black–White validity differences up through the 1970s. Second, the results of studies conducted in the 1960s and 1970s were overwhelmed in the meta-analytic averages by the much larger set of GATB validity studies (e.g., Hartigan & Wigdor, 1989; Synk & Swarthout, 1987) that were carried out in the 1980s. Thus, the meta-analytic averages in the employment domain were heavily influenced by the GATB validity studies. Third, despite the evidence of differential validity from the GATB studies in the 1980s, almost no studies reporting separate Black and White validities after the 1980s were located. This may be a function of the strong statements made regarding the inexistence of differential validity in a number of impactful differential validity reviews in the industrial/organizational psychology literature (e.g., Linn, 1982; Schmidt, 1988; Schmidt & Hunter, 1981).

Military studies were located from only two decades: the 1970s and 1980s. Decade of data collection did not moderate

Black–White validity differences in the military studies. Cognitive ability test validity was .16 lower for Black samples ( $\bar{r} = .17$ ) than for White samples ( $\bar{r} = .33$ ) in the 1970s and .17 lower for Black samples ( $\bar{r} = .17$ ) than for White samples ( $\bar{r} = .34$ ) in the 1980s.

**Job complexity.** In the civilian employment studies, average Black validity ( $\bar{r} = .14$ ) was .05 lower than White validity ( $\bar{r} = .19$ ) in low-complexity jobs, average Black validity ( $\bar{r} = .19$ ) was .02 lower than White validity ( $\bar{r} = .21$ ) in medium-complexity jobs, and average Black validity ( $\bar{r} = .03$ ) was .08 lower than White validity ( $\bar{r} = .11$ ) in high-complexity jobs. There were very few high-complexity job samples, so these estimates are likely affected by second-order sampling error. Regardless, although job complexity did moderate the magnitude of Black–White validity differences (differences ranging from .02 to .08), average cognitive ability test validity was still lower for Black samples in each of the three job complexity categories in the civilian employment studies.

A similar pattern emerged in the military samples, where the magnitude of the Black–White validity gap differed depending on level of job complexity, but average validity was consistently lower for Black samples. Average validity for Black samples ( $\bar{r} = .14$ ) was .21 lower than for White samples ( $\bar{r} = .35$ ) in low-complexity jobs, .18 lower for Black samples ( $\bar{r} = .20$ ) than for White samples ( $\bar{r} = .38$ ) in medium-complexity jobs, and .10 lower for Black samples ( $\bar{r} = .25$ ) than for White samples ( $\bar{r} = .35$ ) in high-complexity jobs. Very few military validity studies included low-complexity jobs, so the low-complexity category results are likely affected by second-order sampling error.

### Hispanic–White Differential Validity Meta-Analysis

**Overall Hispanic–White results.** Overall meta-analytic validities reported separately for Hispanics and Whites across all studies are listed within the first row in Table 2. Cognitive ability

test criterion-related validity for the White sample ( $\bar{r} = .34$ ,  $k = 97$ ,  $N = 725,915$ ) was .04 higher than for the Hispanic sample ( $\bar{r} = .30$ ,  $k = 97$ ,  $N = 51,205$ ). Ninety-five percent confidence intervals for the Hispanic and White samples did not overlap. Although this is evidence for differential validity, the percentage of variance accounted for was very small, suggesting the presence of moderator variables.

**Hispanic–White decade of data collection moderator analysis.** There were only sufficient data to carry out the decade of data collection moderator analysis for the Hispanic–White meta-analysis. Similar to the results for the Black–White meta-analysis, the size of the Hispanic–White validity gap has been shrinking over time. White validity was .12, .05, .01, and .01 higher than Hispanic validity in the 1970s, 1980s, 1990s, and 2000s, respectively.

### Asian–White Differential Validity Meta-Analysis

**Overall Asian–White results.** Overall meta-analytic validities reported separately for Asians and Whites across all studies are listed within the first row of the second section in Table 2. Cognitive ability test criterion-related validity for the White sample ( $\bar{r} = .34$ ,  $k = 60$ ,  $N = 673,303$ ) was .01 higher than for the Asian sample ( $\bar{r} = .33$ ,  $k = 60$ ,  $N = 80,705$ ). Ninety-five percent confidence intervals for the Asian and White samples overlapped. Although this is evidence against differential validity, the percentage of variance accounted for was very small, suggesting the presence of moderator variables.

**Asian–White decade of data collection moderator analysis.** Once again, there were only sufficient data to carry out the decade of data collection moderator analysis for the Asian–White meta-analysis. The Asian–White validity gap (or the lack thereof) has remained relatively constant over time, with Asian and White validity never differing by more than one correlation point from the 1970s through 2000s.

Table 2

*Hispanic–White and Asian–White Differential Validity Meta-Analysis Results*

Variable	<i>N</i>		<i>k</i>		$\bar{r}$		<i>SD<sub>r</sub></i>		% var		95% CI	
	White	Hispanic	White	Hispanic	White	Hispanic	White	Hispanic	White	Hispanic	White	Hispanic
Overall meta-analysis	725,915	51,205	97	97	.34	.30	0.05	0.06	3.81	47.55	[.33, .35]	[.29, .31]
Decade of data collection												
1970s	116,709	3,758	15	15	.39	.27	0.05	0.07	3.86	62.38	[.35, .45]	[.33, .45]
1980s	204,320	6,576	29	29	.35	.30	0.05	0.09	5.08	50.13	[.34, .38]	[.33, .37]
1990s	190,478	17,488	49	49	.33	.32	0.06	0.07	5.84	48.25	[.31, .35]	[.32, .36]
2000s	214,408	23,383	4	4	.30	.29	0.00	0.02	784.22	56.88	[.30, .30]	[.29, .31]
Variable	<i>N</i>		<i>k</i>		$\bar{r}$		<i>SD<sub>r</sub></i>		% var		95% CI	
	White	Asian	White	Asian	White	Asian	White	Asian	White	Asian	White	Asian
Overall meta-analysis	673,303	80,705	60	60	.34	.33	0.05	0.06	2.73	15.68	[.33, .35]	[.31, .35]
Decade of data collection												
1970s	95,535	3,469	3	3	.40	.39	0.04	0.05	1.70	26.62	[.35, .45]	[.33, .45]
1980s	175,423	12,822	6	6	.36	.35	0.03	0.03	2.56	29.65	[.34, .38]	[.33, .37]
1990s	187,937	35,495	47	47	.33	.34	0.06	0.08	5.63	15.88	[.31, .35]	[.32, .36]
2000s	214,408	28,919	4	4	.30	.30	0.00	0.01	784.22	69.65	[.30, .30]	[.29, .31]

*Note.* *N* = total sample sizes; *k* = number of correlations;  $\bar{r}$  = mean sample-size-weighted observed correlation; *SD<sub>r</sub>* = sample-size-weighted observed standard deviation of correlations; % var = percentage of variance attributable to sampling error; CI = confidence interval.

## Discussion

### Summary of Findings

The present study represented the largest test to date of racial/ethnic differential validity for cognitive ability tests, with studies including more than one million participants aggregated across and within the educational admissions, civilian employment, and military literatures. The present meta-analysis found evidence of lower criterion-related validity of cognitive ability tests for racial/ethnic minorities both across and within each of these three broad domains. Observed validity for the Black subgroup was lower than the White subgroup across all three domains and almost all moderator categories. Observed validity for the Hispanic subgroup was also lower than the White subgroup, although data were available only in educational settings. Similar to mean test score comparisons between Black, Hispanic, and White subgroups (Roth, Bevier, et al., 2001), the Hispanic-White observed validity difference was smaller than for Black-White comparisons. Asian-White validity data were available only in educational settings, and the observed Asian-White validity gap was small to nonexistent.

Despite differences between subgroups and study domains in the exact size of validity gaps, the fact that validity consistently favors Whites (almost regardless of the moderators examined in the present research) is striking and calls into question claims by previous researchers that differential validity does not exist (e.g., Schmidt, 1988; Schmidt & Hunter, 1981). The results of the present study, at the very least, demonstrate that differential validity cannot be ruled out. Perhaps because of some of the strong past statements regarding the inexistence of differential validity, research on the factors causing validity to differ between subgroups has been sparse at best. The present study demonstrates that the evidence to date is supportive of differential validity. This highlights the need for future research investigating these causal factors.

Although the present meta-analysis could not exhaustively test all possible causes of differential validity, a number of possible explanations were tested. For instance, criterion bias in the form of racial/ethnic discrimination in performance ratings could cause differential validity; therefore, the present meta-analysis tested whether differential validity findings differ across subjective versus objective performance ratings. Differential validity findings did differ across these types of criteria, suggesting future research should investigate this possibility in more detail. Additionally, differential validity could be an artifact of differences between subgroups in the types of jobs typically held; this was tested (and not supported) by carrying out job complexity moderator analyses. Thus, although the present meta-analysis cannot provide a comprehensive answer to the question of why differential validity may exist, it does shed some light on a number of possible explanations that should be of value to future research.

However, the main contribution of this meta-analysis is in documenting (a) that the existing evidence is supportive of differential validity and (b) the average magnitude of these validity differences. Regarding this second point, the sizes of validity differences in the present meta-analysis were generally quite appreciable. Although the absolute magnitude of validity differences in some domains may at first seem relatively small (e.g., validity .04 higher for Whites than for Blacks and Hispanics in educational

admissions, validity .03 higher for Whites than for Blacks in civilian employment), the absolute magnitude of such differences can be misleading. In percentage terms, these validity differences are quite sizable. Black and Hispanic validity was 11.8%  $(.34 - .30)/.34 = .118$  lower than validity for Whites in the educational admissions domain, and Black validity was 15.8%  $(.19 - .16)/.19 = .158$  lower than White validity in the civilian employment literature. These percentages represent considerable differences in the validity of tests, especially when thought of in terms of test utility (e.g., Roth, Bobko, & Mabon, 2001; Schmidt, Hunter, McKenzie, & Muldrow, 1979). In particular, utility is a function of the validity of a test, not  $r^2$ . So reductions of 11.8–15.8% in the validity of a test means that, holding all other factors influencing utility constant (e.g., average predictor score, standard deviation of performance), utility of the test (as measured by output, dollars, mean performance, etc.) is 11.8–15.8% lower for these minority subgroups than for the White subgroup. It might prove difficult to explain to an organization or college considering using a cognitive ability test that it is inconsequential that validity and utility is 11.8–15.8% lower for minority test takers. Further, Aguinis and Smith (2007) demonstrated that even very small differences in test validity between subgroups (e.g., differences even as small as one correlation point), can cause there to be substantial differences between groups in the rate of false positive and false negative hires, which some believe affects the fairness of selection. Thus, the minority-majority validity differences found in the present meta-analysis are quite noteworthy.

This is not to say that there were no meaningful moderators of the validity differences. One of the most noticeable moderators was study domain, with validity differences much more pronounced in military studies than in civilian employment or educational admissions studies. The average Black-White validity difference in military studies was .17, and no variables examined in the present study meaningfully moderated this Black-White validity gap. This is in contrast to the civilian employment and educational admissions studies, for which the average validity gaps were .03 and .04, respectively. It is noteworthy that the Asian-White and Hispanic-White validity gaps in educational admissions settings were of similar size. The reason for this pronounced difference of the military validity gap relative to education and employment settings is not necessarily clear. Perhaps the most likely possibility is that range restriction affects Black test-criterion correlations more in military settings than in the other two domains. This implies that selection is more directly based on cognitive ability test scores in the military than in postsecondary education, perhaps as a function of affirmative action. Empirical evidence regarding this supposition would be useful.

Although none of the variables examined in the present study meaningfully moderated Black-White validity differences in military settings, there were noteworthy moderators within the educational admissions and civilian employment settings. One such moderator was decade of data collection, which moderated the observed Black-White and Hispanic-White validity differences in educational admissions settings. There has been a general trend toward smaller Black-White and Hispanic-White validity gaps over time, although the most recent large-sample educational studies carried out since 2000 suggested the Black-White validity gap has not disappeared. Regardless, the general trend of decreasing validity differences over time in educational settings is certainly

noteworthy. The reason for this general trend is not known, although we offer one observation. The reduction in validity differences is a function of the validity for Whites decreasing over time instead of the validity for Blacks or Hispanics increasing over time. This pattern makes it less likely that decreases in test bias or criterion bias over time are accounting for the reduction in the validity gap.

Another clear need for future research is to investigate whether this trend of reductions in the magnitude of differential validity generalizes to civilian employment and military settings. Unfortunately, publicly available differential validity research carried out in civilian employment or military settings since the 1980s is almost nonexistent. Therefore, given the current evidence, it is not possible to know whether the sizable amounts of differential validity in employment and military settings, outlined by the present meta-analysis, may have reduced over time. We hope that the results of the present meta-analysis act as a call for future research on differential validity in these settings.

Type of criterion also had a noteworthy moderating effect on Black–White differential validity evidence. In civilian employment settings, mean Black validity was lower than mean White validity in the subjective criterion (i.e., supervisor ratings) samples but not in the objective criterion samples. Such a result is what would be expected if racial/ethnic bias or discrimination in criterion ratings were a determining factor in differential validity evidence. Although this is certainly a possible explanation, the small number of objective criterion samples and the overrepresentation of GATB validity samples in the subjective criterion category confound this conclusion. An alternative explanation is that the GATB was simply less valid for Blacks, and the appearance of greater differential validity in the subjective criterion category had nothing to do with the objective/subjective nature of the criteria. Regardless, the finding of greater differential validity in samples using subjective performance ratings highlights the idea that evidence of differential validity is perhaps as likely to be a function of criterion bias as of test bias.

Another point regarding the influence of the GATB validity studies deserves attention. Because the GATB validity studies made up such a large percentage of the employment samples in the present meta-analysis, this begs the question of whether results would be greatly affected if the GATB studies were removed. Table A4 in the Appendix lists meta-analytic results including and excluding the GATB validity studies both for (a) overall analyses including all samples and for (b) just employment samples. Overall results including all studies are virtually unchanged when the GATB validity studies are excluded. However, the pattern of results changes markedly in the employment samples analysis when the GATB studies are excluded; in this case validity favors Blacks and the Black and White confidence intervals overlap. This highlights the sensitivity of the employment sample analyses and demonstrates how much the employment sample results rely on the GATB validity studies. Thus, it remains an open question whether the differential validity found for the GATB generalizes to other cognitive ability tests used in employment settings. All but one (Gardner & Deadrick, 2008) of the non-GATB employment studies were carried out prior to the GATB validity studies (almost all in 1960s and 1970s), so it also remains an open question whether the Black validity

advantage with non-GATB tests in employment settings generalizes to present times. Clearly, future research with modern tests in employment settings is needed.

### Limitations and Directions for Future Research

Further investigation regarding the existence or magnitude of differential validity is needed. Although differential validity evidence is available for Black and White subgroups in each of the three study domains, there is very little evidence for Asian and Hispanic subgroups (as well as other racial/ethnic subgroups) in the civilian employment and military domains. Especially given the evidence that Hispanic subgroup validity is lower in the educational admissions domain, this lack of evidence in employment and military studies seems to be a large oversight. Even for Black–White comparisons, in each of the three study domains, questions remain regarding the existence and magnitude of differential validity. For instance, despite a long history of consistently documenting the existence and magnitude of Black–White differential validity in educational admissions settings, the question of whether these validity differences are shrinking over time remains. Thus, it would be fruitful to collect new data due to the dynamic nature of the magnitude of Black–White validity differences in educational admissions. In civilian employment and military settings, on the other hand, there is a need for future research documenting the existence and magnitude of Black–White differential validity simply because there has been a dearth of research on the phenomenon for the past 20 to 30 years. Although the present meta-analysis demonstrated that criterion-related validity is lower for Black samples in the data available to date in employment and military settings, the amount of available data is small (relative to that for educational admissions settings) and dated.

One limitation of the present meta-analysis was the inability to account for statistical artifacts, especially range restriction. In the context of differential validity, appropriate range restriction corrections would entail making separate corrections for minority and majority subgroups using subgroup-specific range restriction information. In almost no instances did primary studies report any range restriction information, let alone subgroup-specific information. Given the well-known effects of range restriction on the relationship between cognitive ability tests and performance and given mean differences between subgroups on cognitive ability tests, this makes the effects of range restriction on differential validity a clear need for future research. Although the lack of range restriction corrections in the present study limits interpretations about the true relative validity of tests for each subgroup, it is important to note that the most highly cited differential validity reviews in industrial/organizational psychology (e.g., Hunter et al., 1979; Schmidt et al., 1980) did not account for range restriction either. We hope the consistent findings of lower observed validity in the present meta-analysis act as a catalyst for future research to (a) empirically investigate the role of range restriction in differential validity evidence or (b) at least begin reporting subgroup-specific range restriction information.

Another issue that should be discussed is the lack of a moderating effect of job complexity in the present meta-analysis. A common empirical finding has been for cognitive ability test validity to be stronger for more complex jobs (e.g., Hunter, 1980). Pooling across Black and White subgroups in the present meta-



analysis, validity in employment samples was .18, .19, and .09 in low-, medium-, and high-complexity jobs, respectively; validity in military samples was .30, .35, and .34 in low-, medium-, and high-complexity jobs, respectively. A few points are relevant here. First, the number of samples in some complexity categories was very low, so the lack of a complexity effect could be due to second-order sampling error. Second, the job complexity moderation effect has not always been found in previous research. For instance, although Hunter (1980) found that job complexity moderated validity in 515 GATB validity studies, Hartigan and Wigdor (1989) did not replicate Hunter's job complexity moderation effect in their meta-analysis of 264 subsequent GATB validity studies (with a total sample size almost exactly equal to Hunter's). The results of the present meta-analysis align most with past research that has not found a moderating effect of job complexity. Finally, the more important point from a differential validity perspective is whether differential validity evidence changes if job complexity is held constant. Lower validity for the Black subgroup remained at all levels of job complexity in the present meta-analysis.

Although future research investigating the existence and magnitude of differential validity is warranted, enough evidence currently exists to conclude that it is likely observed test-criterion correlations differ for Black and White subgroups. Thus, the next step is attempting to explain the underlying causes of differential validity. Before the results of this meta-analysis, the strong conclusion, at least in industrial/organizational psychology, was that differential validity did not exist; thus, investigation into the possible causes of differential validity was likely not deemed warranted. As a result of these statements, a necessary first step was documenting that evidence of differential validity does exist, even if the present meta-analysis could not fully account for exactly what was causing it. A number of possible causes of differential validity were outlined in the opening sections of this paper. It is hoped that the present study acts as a call to and guide for future differential validity research.

## References

- \*References marked with an asterisk indicate studies that provided one or more samples for the meta-analysis.
- Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95*, 648–680. doi:10.1037/a0018714
- Aguinis, H., & Smith, M. A. (2007). Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. *Personnel Psychology, 60*, 165–199. doi:10.1111/j.1744-6570.2007.00069.x
- \*American Association of Colleges for Teacher Education. (1992). *Academic achievement of White, Black, and Hispanic students in teacher education programs* (Report No. 0-89333-094-9). Washington, DC: Author.
- \*Baggaley, A. R. (1974). Academic prediction at an Ivy League college, moderated by demographic variables. *Measurement and Evaluation in Guidance, 6*, 232–235.
- Beaujean, A. A., Firmin, M. W., Knoop, A. J., Michonski, J. D., Berry, T. P., & Lowrie, R. E. (2006). Validation of the Frey and Detterman (2004) IQ prediction equations using the Reynolds Intellectual Assessment Scales. *Personality and Individual Differences, 41*, 353–357.
- Berry, C. M. (2007). *Toward an understanding of evidence of differential validity of cognitive ability tests for racial/ethnic subgroups* (Unpublished doctoral dissertation). University of Minnesota.
- Berry, C. M., & Sackett, P. R. (2008a, April). *Black–White differences in the properties of grades*. Poster session presented at the meeting of the Society for Industrial and Organizational Psychology, San Francisco, CA.
- \*Berry, C. M., & Sackett, P. R. (2008b, April). *Toward understanding race differences in validity of cognitive ability tests*. Poster session presented at the meeting of the Society for Industrial and Organizational Psychology, San Francisco, CA.
- Boehm, V. R. (1972). Negro–White differences in validity of employment and training procedures: Summary of research evidence. *Journal of Applied Psychology, 56*, 33–39. doi:10.1037/h0032130
- Breland, H. M. (1979). *Population validity and college entrance measures* (College Board Research and Development Report RDR 78-79 No. 2). Princeton, NJ: Educational Testing Service.
- \*Breland, H., & Griswold, P. (1981). *Group comparisons for basic skills measures* (College Board Report No. 81-6). New York, NY: College Entrance Examination Board.
- \*Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning test* (Research Report No. 2000-1). New York, NY: College Entrance Examination Board.
- \*Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. (1973). *An investigation of sources of bias in the prediction of job performance: A six year study* (Final Report No. PR 73-37). Princeton, NJ: Educational Testing Service.
- Carretta, T. R. (1997). Group differences on U.S. Air Force pilot selection tests. *International Journal of Selection and Assessment, 5*, 115–127. doi:10.1111/1468-2389.00051
- Carretta, T. R., & Ree, M. J. (1995). Near identity of cognitive structure in sex and ethnic groups. *Personality and Individual Differences, 19*, 149–155. doi:10.1016/0191-8869(95)00031-Z
- \*Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124. doi:10.1111/j.1745-3984.1968.tb00613.x
- \*Crawford, P. L., Alferink, D. M., & Spencer, J. L. (1986). *Postdictions of college GPAs from ACT composite scores and high school GPAs: Comparisons by race and gender*. West Virginia State College (ERIC Document Reproduction Service No. ED 326541).
- \*Davis, J. A., & Kerner, S. E. (1971). *The validity of tests and achievement in high school for predicting initial performance in the public universities of North Carolina with special attention to black students* (Publication No. 1973-03-00). Princeton, NJ: Educational Testing Service.
- \*Distefano, M. K., Pryor, M. W., & Craig, S. H. (1976). Predictive validity of general ability tests with Black and White psychiatric attendants. *Personnel Psychology, 29*, 197–204. doi:10.1111/j.1744-6570.1976.tb00408.x
- \*Dittmar, N. (1977). *A comparative investigation of the predictive validity of admissions criteria for Anglos, Blacks, and Mexican Americans* (Unpublished doctoral dissertation). University of Texas at Austin.
- Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35*, 21–50. doi:10.1207/S15327906MBR3501\_2
- Domino, G., & Morales, A. (2000). Reliability and validity of the D-48 with Mexican American college students. *Hispanic Journal of Behavioral Sciences, 22*, 382–389. doi:10.1177/0739986300223007
- Drasgow, F. (2003). Intelligence and the workplace. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Vol. 12. Industrial and organizational psychology* (pp. 107–130). Hoboken, NJ: Wiley.
- Drasgow, F., Nye, C. D., Carretta, T. R., & Ree, M. J. (2010). Factor structure of the Air Force Officer Qualifying Test Form S: Analysis and comparison with previous forms. *Military Psychology, 22*, 68–85. doi:10.1080/08995600903249255

- \*Duran, R. P. (1983). *Hispanics' education and background: Predictors of college achievement*. New York, NY: College Board.
- \*Elliott, R., & Strenta, A. C. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. *Journal of Educational Measurement*, 25, 333–347. doi:10.1111/j.1745-3984.1988.tb00312.x
- \*Farr, J. L., O'Leary, B. S., Pfeiffer, C. M., Goldstein, F. L., & Bartlett, C. J. (1971). *Ethnic group membership as a moderator in the prediction of job performance: An examination of some less traditional procedures* (Rep. No. 151-277). Silver Spring, MD: American Institutes for Research.
- \*Fox, H., & Lefkowitz, J. (1974). Differential validity: Ethnic group as a moderator in predicting job performance. *Personnel Psychology*, 27, 209–223. doi:10.1111/j.1744-6570.1974.tb01529.x
- Frey, M. C., & Detterman, D. K. (2004). Scholastic achievement or g? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science*, 15, 373–378.
- \*Gael, S., & Grant, D. L. (1972). Employment test validation for minority and nonminority telephone company service representatives. *Journal of Applied Psychology*, 56, 135–139. doi:10.1037/h0032659
- \*Gael, S., Grant, D. L., & Ritchie, R. J. (1975a). Employment test validation for minority and nonminority clerks with work sample criteria. *Journal of Applied Psychology*, 60, 420–426. doi:10.1037/h0076908
- \*Gael, S., Grant, D. L., & Ritchie, R. J. (1975b). Employment test validation for minority and nonminority telephone operators. *Journal of Applied Psychology*, 60, 411–419. doi:10.1037/h0076909
- \*Gardner, D., & Deadrick, D. L. (2008). Underprediction of performance for U.S. minorities using cognitive ability measures. *Equal Opportunities International*, 27, 455–464. doi:10.1108/02610150810882305
- Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco, CA: Freeman.
- \*Goldman, R. D., & Hewitt, B. N. (1975). An investigation of test bias for Mexican-American college students. *Journal of Educational Measurement*, 12, 187–196. doi:10.1111/j.1745-3984.1975.tb01021.x
- \*Goldman, R. D., & Hewitt, B. N. (1976). Predicting the success of Black, Chicano, Oriental, and White college students. *Journal of Educational Measurement*, 13, 107–117. doi:10.1111/j.1745-3984.1976.tb00002.x
- \*Goldman, R. D., & Richards, R. (1974). The SAT prediction of grades for Mexican-American versus Anglo-American students at the University of California, Riverside. *Journal of Educational Measurement*, 11, 129–135. doi:10.1111/j.1745-3984.1974.tb00983.x
- \*Grant, D. L., & Bray, D. W. (1970). Validation of employment tests for telephone company installation and repair occupations. *Journal of Applied Psychology*, 54, 7–14. doi:10.1037/h0028648
- \*Haney, R., Michael, W. B., & Martois, J. (1976). The prediction of success of three ethnic groups in the academic components of a nursing-training program at a large metropolitan hospital. *Educational and Psychological Measurement*, 36, 421–431. doi:10.1177/001316447603600222
- Hartigan, J. A., & Wigdor, A. K. (1989). Differential validity and differential prediction. In J. A. Hartigan & A. K. Wigdor (Eds.), *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery* (pp. 172–188). Washington, DC: National Academy Press.
- Herrnstein, R. J., & Murray, C. (1994). *The bell curve*. New York, NY: Free Press.
- Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, 9, 152–194. doi:10.1111/1468-2389.00171
- \*Houston, W. M., & Novick, M. R. (1987). Race-based differential prediction in Air Force technical training programs. *Journal of Educational Measurement*, 24, 309–320. doi:10.1111/j.1745-3984.1987.tb00282.x
- Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology*, 71, 327–333. doi:10.1037/0021-9010.71.2.327
- Hunter, J. E. (1980). *Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB)*. Washington, DC: U.S. Department of Labor, Employment Service.
- Hunter, J. E., & Schmidt, F. L. (1978). Differential and single-group validity of employment tests by race: A critical analysis of three recent studies. *Journal of Applied Psychology*, 63, 1–11. doi:10.1037/0021-9010.63.1.1
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). New York, NY: Sage.
- Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, 86, 721–735. doi:10.1037/0033-2909.86.4.721
- Hunter, J. E., Schmidt, F. L., & Judiesch, M. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology*, 75, 28–42. doi:10.1037/0021-9010.75.1.28
- Jencks, C., & Phillips, M. (1998). *The Black-White test score gap*. Washington, DC: Brookings Institution Press.
- Jensen, A. R. (1977). An examination of culture bias in the Wonderlic Personnel Test. *Intelligence*, 1, 51–64. doi:10.1016/0160-2896(77)90026-5
- Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: Free Press.
- \*Kallingal, A. (1971). The prediction of grades for Black and White students at Michigan State University. *Journal of Educational Measurement*, 8, 263–265. doi:10.1111/j.1745-3984.1971.tb00935.x
- \*Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. (1968). *Testing and fair employment*. New York, NY: New York University.
- Kraiger, K., & Ford, J. K. (1990). The relation of job knowledge, job performance, and supervisory ratings as a function of ratee race. *Human Performance*, 3, 269–279. doi:10.1207/s15327043hup0304\_4
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology*, 86, 148–161.
- \*Lichtman, C. (2008). *ACT scores and college grades at a large inner-city university* (Unpublished technical report). Wayne State University.
- Linn, R. L. (1978). Single-group validity, differential validity, and differential predictions. *Journal of Applied Psychology*, 63, 507–512. doi:10.1037/0021-9010.63.4.507
- Linn, R. L. (1982). Ability testing: Individual differences, prediction, and differential prediction. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies* (pp. 335–388). Washington, DC: National Academy of Sciences.
- \*Lopez, F. M. (1966). Current problems in test performance of job applicants: I. *Personnel Psychology*, 19, 10–18. doi:10.1111/j.1744-6570.1966.tb02430.x
- \*Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential validity and prediction of the SAT* (College Board Research Report No. 2008-4). New York, NY: College Board.
- \*McCornack, R. L. (1983). Bias in the validity of predicted college grades in four ethnic minority groups. *Educational and Psychological Measurement*, 43, 517–522. doi:10.1177/001316448304300220
- \*McLaughlin, D. H., Rossmeissl, P. G., Wise, L. L., Brandt, D. A., & Wang, M. (1984). *Validation of current Armed Services Vocational Aptitude Battery (ASVAB) area composites, based on training and Skill Qualification Test (SQT) information in fiscal year 1981 and 1982* (Report No. ARI-TR-651, AD-A156 807). Alexandria, VA: Army Research Institute.
- Mendoza, J. L., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational Statistics*, 12, 282–293. doi:10.2307/1164688

- \*Moffatt, G. K. (1993, February). *The validity of the SAT as a predictor of grade point average for nontraditional college students*. Paper presented at the meeting of the Eastern Educational Research Association, Clearwater Beach, FL.
- \*Morgan, R. (1990). Analyses of predictive validity within student categorizations. In W. W. Willingham, C. Lewis, R. Morgan, & L. Ramist (Eds.), *Predicting college grades: An analysis of institutional trends over two decades* (pp. 225–238). Princeton, NJ: Educational Testing Service.
- \*Noble, J., Crouse, J., & Shulz, M. (1996). *Differential prediction/impact in course placement for ethnic and gender group* (ACT Research Report No. 96-8). Iowa City, IA: ACT.
- O'Connor, E. J., Wexley, K. N., & Alexander, R. A. (1975). Single-group validity: Fact or fallacy? *Journal of Applied Psychology*, 60, 352–355. doi:10.1037/h0076635
- \*O'Leary, B. S., Farr, J. L., & Bartlett, C. J. (1970). *Ethnic group membership as a moderator of job performance* (Tech. Rep. No. 1). Washington, DC: American Institutes for Research.
- O'Neill, K. A., & McPeck, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Erlbaum.
- Oppler, S. H., Campbell, J. P., Pulakos, E. D., & Borman, W. C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: Review, results, and conclusions. *Journal of Applied Psychology*, 77, 201–217. doi:10.1037/0021-9010.77.2.201
- \*Pandey, R. E. (1971). The SCAT and race. *Psychological Reports*, 28, 459–462.
- Pandolfi, V. (1997). *Assessment of factor models underlying the WISC-III in White, Black, and Hispanic subgroups of the standardization sample* (Unpublished doctoral dissertation). Hofstra University.
- \*Patterson, B. F., Mattern, K. D., & Kobrin, J. L. (2009). *Validity of the SAT for predicting FYGPA: 2007 SAT validity sample*. Retrieved from [http://professionals.collegeboard.com/profdownload/pdf/2007\\_Replication\\_Admissions\\_and\\_Differential.pdf](http://professionals.collegeboard.com/profdownload/pdf/2007_Replication_Admissions_and_Differential.pdf)
- \*Pearson, B. Z. (1993). Predictive validity of the Scholastic Aptitude Test for Hispanic bilingual students. *Hispanic Journal of Behavioral Sciences*, 15, 342–356.
- \*Pennock-Román, M. (1990). *Test validity and language background: A study of Hispanic American students at six universities*. New York, NY: College Board.
- \*Pfeifer, C. M., & Sedlacek, W. E. (1971). The validity of academic predictors for Black and White students at a predominantly White university. *Journal of Educational Measurement*, 8, 253–261.
- Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. *Journal of Applied Psychology*, 74, 770–780.
- Ramist, L., Lewis, C., & McCamley, L. (1990). Implications of using freshman GPA as the criterion for the predictive validity of the SAT. In W. W. Willingham, C. Lewis, R. Morgan, & L. Ramist (Eds.), *Predicting college grades: An analysis of institutional trends over two decades* (pp. 253–288). Princeton, NJ: Educational Testing Service.
- Ree, M. J., & Carretta, T. R. (1995). Group differences in aptitude factor structure on the ASVAB. *Educational and Psychological Measurement*, 55, 268–277.
- Reed, C. L. (2000). *An investigation of measurement invariance in the WISC-III: Examining a sample of referred African American and Caucasian students* (Unpublished doctoral dissertation). University of South Florida.
- \*Roberts, H. E., & Skinner, J. (1996). Gender and racial equity of the Air Force Officer Qualifying Test in officer training school selection decisions. *Military Psychology*, 8, 95–113.
- Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, 54, 297–330.
- Roth, P. L., Bobko, P., & Mabon, H. (2001). Utility analysis: A review and analysis at the turn of the century. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesveran (Eds.), *Handbook of industrial, work, and organizational psychology* (pp. 363–384). London, England: Sage.
- Rotundo, M., & Sackett, P. R. (1999). Effect of rater race on conclusions regarding differential prediction in cognitive ability tests. *Journal of Applied Psychology*, 84, 815–822.
- Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, 63, 215–227.
- Sackett, P. R., & DuBois, C. L. Z. (1991). Rater-race effects on performance evaluation: Challenging meta-analytic conclusions. *Journal of Applied Psychology*, 76, 873–877.
- Schmidt, F. L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior*, 33, 272–292.
- Schmidt, F. L., Berner, J. G., & Hunter, J. E. (1973). Racial differences in validity of employment tests: Reality or illusion. *Journal of Applied Psychology*, 58, 5–9.
- Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist*, 36, 1128–1137.
- Schmidt, F. L., & Hunter, J. E. (1992). Development of a causal model of processes determining job performance. *Current Directions in Psychological Science*, 1, 89–92.
- Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology*, 64, 609–626.
- Schmidt, F. L., Pearlman, K., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology*, 33, 705–724.
- \*Scott, C. C. (1976). *Longer-term predictive validity of college admission tests for Anglo, Black, and Mexican American students* (Unpublished doctoral dissertation). University of New Mexico.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, 89, 497–508.
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69, 797–811.
- Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). Sex differences in prediction of college grades from Scholastic Aptitude Test scores. *Journal of Educational Psychology*, 85, 710–718.
- \*Sue, S., & Abe, J. (1988). *Predictors of academic achievement among Asian American and White students* (Research Report No. 88-11). New York, NY: College Board.
- \*Synk, D. J., & Swarthout, D. (1987). *Comparison of Black and nonminority validities for the General Aptitude Test battery* (Research Report No. 51). Washington, DC: U.S. Department of Labor.
- \*Temp, G. (1971). Validity of the SAT for Blacks and Whites in thirteen integrated institutions. *Journal of Educational Measurement*, 8, 245–251.
- \*Tracey, T. J., & Sedlacek, W. E. (1985). The relationship of noncognitive variables to academic success: A longitudinal comparison by race. *Journal of College Student Personnel*, 26, 405–410.
- \*U.S. Department of Labor, Manpower Administration. (1974). *Development of USTES aptitude test battery for drafter* (Tech. Rep. No. S-266R74). Washington, DC: Author.
- \*U.S. Department of Labor, Manpower Administration. (1985). *Development of USTES aptitude test battery for airframe-and-powerplant mechanic* (Tech. Rep. No. S-111R84). Washington, DC: Author.
- \*Valentine, L. D. (1977). *Prediction of Air Force technical training success from ASVAB and educational background* (Report No. AFHRL-



- TR-77-18). Lackland AFB, TX: Air Force Human Resources Laboratory, Personnel Research Division.
- Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology*, 75, 315–321.
- Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, 89, 696–716.
- \*Wightman, L. F. (2000). *Beyond FYA: Analysis of the utility of LSAT scores and UGPA for predicting academic success in law school* (LSAC Research Report No. 99-06). Newtown, PA: Law School Admission Council.
- \*Wightman, L. F., & Muller, D. G. (1990). *An analysis of differential validity and differential prediction for Black, Mexican American, Hispanic, and White law school students* (LSAC Research Report No. 90-03). Newtown, PA: Law School Admission Council.
- \*Wilbourn, J. M., Valentine, L. D., & Ree, M. J. (1984). *Relationships of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10 to Air Force technical school final grades*. Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.
- Willingham, W. W., Pollack, J. M., & Lewis, C. (2000). *Grades and test scores: Accounting for observed differences* (Educational Testing Service Research Report No. 00-15). Princeton, NJ: Educational Testing Service.
- \*Wynne, W. D. (2003). *An investigation of ethnic and gender intercept bias in the SAT's prediction of college freshman academic performance* (Unpublished doctoral dissertation). University of Texas at Austin.
- \*Young, J. W. (1994). Differential prediction of college grades by gender and by ethnicity: A replication study. *Educational and Psychological Measurement*, 54, 1022–1029.
- Young, J. W., & Kobrin, J. L. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis* (College Board Research Report No. 2001-6). New York, NY: College Board.



## Appendix

Table A1

*Information for Primary Studies Included in the Black–White Differential Validity Meta-Analysis*

Reference	Sample no.	N		Correlation		Test	Criterion	Criterion type	Domain	Year	Job complexity
		White	Black	White	Black						
AACTE (1992)	1	197	52	.39	.03	SAT	F GPA	Grades	Educ	1980s	n/a
Baggaley (1974)	1	464	65	.19	.10	SAT	C GPA	Grades	Educ	1970s	n/a
Berry & Sackett (2008b)	1	3,530	402	.22	.03	SAT	F GPA	Grades	Educ	1990s	n/a
	2	3,199	148	.28	.28	SAT	F GPA	Grades	Educ	1990s	n/a
	3	4,606	261	.43	.33	SAT	F GPA	Grades	Educ	1990s	n/a
	4	4,044	539	.24	.28	SAT	F GPA	Grades	Educ	1990s	n/a
	5	1,996	129	.18	.21	SAT	F GPA	Grades	Educ	1990s	n/a
	6	1,273	99	.28	.19	SAT	F GPA	Grades	Educ	1990s	n/a
	7	1,331	108	.49	.34	SAT	F GPA	Grades	Educ	1990s	n/a
	8	1,767	113	.23	.33	SAT	F GPA	Grades	Educ	1990s	n/a
	9	663	18	.25	.56	SAT	F GPA	Grades	Educ	1990s	n/a
	10	1,758	325	.26	.21	SAT	F GPA	Grades	Educ	1990s	n/a
	11	573	44	.49	.55	SAT	F GPA	Grades	Educ	1990s	n/a
	12	779	3	.49	-.34	SAT	F GPA	Grades	Educ	1990s	n/a
	13	4,356	405	.34	.41	SAT	F GPA	Grades	Educ	1990s	n/a
	14	519	21	.41	.09	SAT	F GPA	Grades	Educ	1990s	n/a
	15	3,313	131	.29	.36	SAT	F GPA	Grades	Educ	1990s	n/a
	16	866	12	.39	.48	SAT	F GPA	Grades	Educ	1990s	n/a
	17	332	100	.27	.39	SAT	F GPA	Grades	Educ	1990s	n/a
	18	2,641	26	.33	.08	SAT	F GPA	Grades	Educ	1990s	n/a
	19	2,251	1,048	.25	.47	SAT	F GPA	Grades	Educ	1990s	n/a
	20	2,513	227	.33	.39	SAT	F GPA	Grades	Educ	1990s	n/a
	21	2,572	51	.43	.53	SAT	F GPA	Grades	Educ	1990s	n/a
	22	1,049	19	.33	.59	SAT	F GPA	Grades	Educ	1990s	n/a
	23	1,112	113	.37	.53	SAT	F GPA	Grades	Educ	1990s	n/a
	24	3,394	439	.30	.26	SAT	F GPA	Grades	Educ	1990s	n/a
	25	3,045	139	.25	.26	SAT	F GPA	Grades	Educ	1990s	n/a
	26	3,092	134	.32	.04	SAT	F GPA	Grades	Educ	1990s	n/a
	27	494	34	.41	.35	SAT	F GPA	Grades	Educ	1990s	n/a
	28	255	4	.48	.80	SAT	F GPA	Grades	Educ	1990s	n/a
	29	254	26	.38	.56	SAT	F GPA	Grades	Educ	1990s	n/a
	30	1,933	138	.31	.18	SAT	F GPA	Grades	Educ	1990s	n/a
	31	8,314	437	.39	.36	SAT	F GPA	Grades	Educ	1990s	n/a
	32	2,221	89	.36	.34	SAT	F GPA	Grades	Educ	1990s	n/a
	33	823	57	.38	.36	SAT	F GPA	Grades	Educ	1990s	n/a
	34	4,361	366	.20	.24	SAT	F GPA	Grades	Educ	1990s	n/a
	35	955	218	.30	.27	SAT	F GPA	Grades	Educ	1990s	n/a
	36	7,184	1,072	.40	.38	SAT	F GPA	Grades	Educ	1990s	n/a
	37	4,332	60	.37	.50	SAT	F GPA	Grades	Educ	1990s	n/a
	38	11,202	692	.39	.33	SAT	F GPA	Grades	Educ	1990s	n/a
	39	3,707	96	.42	.41	SAT	F GPA	Grades	Educ	1990s	n/a
	40	6,439	246	.36	.30	SAT	F GPA	Grades	Educ	1990s	n/a
	41	1,811	196	.29	.10	SAT	F GPA	Grades	Educ	1990s	n/a
Breland & Griswold (1981)	1	5,236	583	.22	.42	SAT	EPT-Essay	Subjective	Educ	1980s	n/a
Bridgeman et al. (2000)	1	29,152	2,835	.33	.37	SAT	F GPA	Grades	Educ	1990s	n/a
	2	31,169	2,974	.34	.36	SAT	F GPA	Grades	Educ	1990s	n/a
Campbell et al. (1973)	1	285	168	.15	.19	Kit of Reference Tests	Multiple criteria	Composite	Employ	1970s	Medium
	2	236	99	.27	.28	Kit of Reference Tests	Multiple criteria	Composite	Employ	1970s	Medium
	3	50	38	.28	.30	Kit of Reference Tests	Supervisor rating	Subjective	Employ	1970s	Medium
	4	167	99	.23	.24	Kit of Reference Tests	Multiple criteria	Composite	Employ	1970s	Medium
Cleary (1968)	1	118	59	.50	.17	SAT	F GPA	Grades	Educ	1960s	n/a
	2	365	83	.41	.29	SAT	F GPA	Grades	Educ	1960s	n/a
	3	2,181	125	.48	.54	SAT	C GPA	Grades	Educ	1960s	n/a
Crawford et al. (1986)	1	945	176	.33	.39	ACT	C GPA	Grades	Educ	1980s	n/a
Davis & Kerner (1971)	1	n/a	200	n/a	.31	SAT	F GPA	Grades	Educ	1960s	n/a
	2	n/a	200	n/a	.22	SAT	F GPA	Grades	Educ	1960s	n/a

(Appendix continues)

Table A1 (continued)

Reference	Sample no.	N		Correlation		Test	Criterion	Criterion type	Domain	Year	Job complexity
		White	Black	White	Black						
Distefano et al. (1976)	3	n/a	80	n/a	.30	SAT	F GPA	Grades	Educ	1960s	n/a
	4	n/a	118	n/a	.08	SAT	F GPA	Grades	Educ	1960s	n/a
	5	n/a	552	n/a	.32	SAT	F GPA	Grades	Educ	1960s	n/a
	6	n/a	187	n/a	.31	SAT	F GPA	Grades	Educ	1960s	n/a
	7	n/a	301	n/a	.23	SAT	F GPA	Grades	Educ	1960s	n/a
	8	201	97	.24	.23	SAT	F GPA	Grades	Educ	1960s	n/a
	9	200	67	.37	.26	SAT	F GPA	Grades	Educ	1960s	n/a
	10	175	32	.25	.24	SAT	F GPA	Grades	Educ	1960s	n/a
	11	193	41	.35	.18	SAT	F GPA	Grades	Educ	1960s	n/a
	12	200	43	.46	.48	SAT	F GPA	Grades	Educ	1960s	n/a
	13	187	38	.30	.64	SAT	F GPA	Grades	Educ	1960s	n/a
	1	34	36	.47	.43	State Civil Service and Otis Employment Test	Psychiatric Aide Test	Objective	Employ	1970s	Medium
Dittmar (1977)	1	233	115	.48	.49	SAT	F GPA	Grades	Educ	1970s	n/a
	2	270	155	.45	.35	SAT	F GPA	Grades	Educ	1970s	n/a
Duran (1983)	1	218	n/a	.49	n/a	SAT	GPA	Grades	Educ	1970s	n/a
	2	254	n/a	.44	n/a	SAT	GPA	Grades	Educ	1970s	n/a
Elliott & Strenta (1988)	1	521	66	.31	.14	SAT	C GPA	Grades	Educ	1980s	n/a
Farr et al. (1971)	1	178	126	0.39	.46	SAT	F GPA	Grades	Educ	1960s	n/a
	2	157	51	0.11	-.23	TMA	Supervisor rating	Subjective	Employ	1970s	Unknown
	3	99	84	.47	.05	AFQT	Final class standing	Grades	Military	1970s	Medium
Fox & Lefkowitz (1974)	1	46	54	.16	.21	SRA-P	Multiple criteria	Composite	Employ	1970s	Medium
Gael & Grant (1972)	1	193	106	.36	.27	BSQT 1	Multiple criteria	Composite	Employ	1960s	Medium
Gael et al. (1975a)	1	185	143	.50	.54	BSQT 1	Work sample	Objective	Employ	1970s	Low
Gael et al. (1975b)	1	464	501	.29	.41	BSQT 1	Work sample	Objective	Employ	1970s	Medium
Gardner & Deadrick (2008)	1	482	237	.07	.18	GATB	Piece rate	Objective	Employ	2000s	Low
Goldman & Hewitt (1975)	1	5,635	n/a	.29	n/a	SAT	C GPA	Grades	Educ	1970s	n/a
	2	5,500	n/a	.36	n/a	SAT	C GPA	Grades	Educ	1970s	n/a
	3	2,926	n/a	.31	n/a	SAT	C GPA	Grades	Educ	1970s	n/a
	4	3,127	n/a	.31	n/a	SAT	C GPA	Grades	Educ	1970s	n/a
Goldman & Hewitt (1976)	1	4259	272	.30	.25	SAT	C GPA	Grades	Educ	1970s	n/a
Goldman & Richards (1974)	1	210	n/a	.44	n/a	SAT	F GPA	Grades	Educ	1970s	n/a
	2	1,700	n/a	.32	n/a	SAT	F GPA	Grades	Educ	1970s	n/a
Grant & Bray (1970)	1	219	211	.36	.41	SCAT	Highest training level passed	Grades	Employ	1960s	Medium
Haney et al. (1976)	1	223	67	.29	.21	CAT	Course GPA	Grades	Educ	1970s	n/a
Houston & Novick (1987)	1	399	126	.36	.17	ASVAB Mech	Final training grade	Grades	Military	1980s	Medium
	2	512	101	.44	.01	ASVAB Mech	Final training grade	Grades	Military	1980s	Medium
	3	558	146	.45	0.24	ASVAB Mech	Final training grade	Grades	Military	1980s	Medium
	4	1,357	278	.43	.17	ASVAB Mech	Final training grade	Grades	Military	1980s	Medium
	5	1,656	210	.45	.35	ASVAB Mech	Final training grade	Grades	Military	1980s	Medium
	6	1,394	281	.41	.27	ASVAB Mech	Final training grade	Grades	Military	1980s	Medium
	7	1,176	177	.40	.16	ASVAB Mech	Final training grade	Grades	Military	1980s	Medium
	8	2,699	301	.44	.22	ASVAB Mech	Final training grade	Grades	Military	1980s	Medium
	9	2,695	361	.48	.21	ASVAB Mech	Final training grade	Grades	Military	1980s	Medium
Kallingal (1971)	1	511	225	.51	.47	Multiple tests	C GPA	Grades	Educ	1960s	n/a
Kirkpatrick et al. (1968)	1	100	26	.06	.12	SET	Supervisor rating	Subjective	Employ	1960s	Low
	2	39	33	.25	.04	Multiple tests	Supervisor rating	Subjective	Employ	1960s	Low
	3	23	39	-.05	.36	Multiple tests	Work sample	Objective	Employ	1960s	Low
	4	77	22	.55	-.02	PNG	Instructor rating	Subjective	Employ	1960s	Medium
	5	27	25	.69	.03	PNG	Instructor rating	Subjective	Employ	1960s	Medium
	6	437	98	.17	.30	General cognitive ability test	Supervisor rating	Subjective	Employ	1960s	Low
Lichtman (2008)	1	859	337	.31	.23	ACT	F GPA	Grades	Educ	2000s	n/a

(Appendix continues)

Table A1 (continued)

Reference	Sample no.	N		Correlation		Test	Criterion	Criterion type	Domain	Year	Job complexity
		White	Black	White	Black						
Lopez (1966)	1	36	56	.07	.09	Mental ability test	Multiple criteria	Composite	Employ	1960s	Unknown
Mattern et al. (2008)	1	104,017	10,096	.30	.26	SAT	F GPA	Grades	Educ	2000s	n/a
McCornack (1983)	1	2,263	83	.22	.29	SAT	F GPA	Grades	Educ	1970s	n/a
	2	2,009	108	.24	.39	SAT	F GPA	Grades	Educ	1980s	n/a
McLaughlin et al. (1984)	1	4,780	6,985	.30	.13	ASVAB	Multiple criteria	Composite	Military	1980s	Unknown
	2	14,523	3,570	.30	.19	ASVAB	Multiple criteria	Composite	Military	1980s	Unknown
	3	4,527	3,111	.26	.10	ASVAB	Multiple criteria	Composite	Military	1980s	Unknown
	4	4,936	3,234	.36	.19	ASVAB	Multiple criteria	Composite	Military	1980s	Unknown
	5	474	624	.20	.11	ASVAB	Multiple criteria	Composite	Military	1980s	Unknown
	6	2,729	1,039	.25	.12	ASVAB	Multiple criteria	Composite	Military	1980s	Unknown
	7	6,941	3,316	.29	.14	ASVAB	Multiple criteria	Composite	Military	1980s	Unknown
	8	3,207	1,708	.25	.11	ASVAB	Multiple criteria	Composite	Military	1980s	Unknown
	9	6,682	956	.27	.14	ASVAB	Multiple criteria	Composite	Military	1980s	Unknown
Moffatt (1993)	1	456	31	.54	.16	SAT	C GPA	Grades	Educ	1980s	n/a
Morgan (1990)	1	89,013	5,162	.41	.35	SAT	F GPA	Grades	Educ	1970s	n/a
	2	89,524	4,086	.37	.29	SAT	F GPA	Grades	Educ	1980s	n/a
	3	74,586	5,095	.36	.30	SAT	F GPA	Grades	Educ	1980s	n/a
Noble et al. (1996)	1	275	275	.25	.27	ACT	Course grade	Grades	Educ	1990s	n/a
O'Leary et al. (1970)	1	83	33	-.04	-.16	Arithmetic reasoning test	Dollar and axle accuracy	Objective	Employ	1970s	Medium
	2	207	41	.08	-.01	CTMM	Supervisor rating	Subjective	Employ	1970s	High
	3	54	17	.02	.12	Otis Quick Score	Supervisor rating	Subjective	Employ	1970s	Medium
	4	273	30	.24	.21	TMA	Supervisor rating	Subjective	Employ	1970s	Low
	5	86	84	.36	.19	Verbal and arithmetic reasoning	Supervisor rating	Subjective	Employ	1970s	Low
	6	122	125	.14	.14	Verbal and arithmetic reasoning	Supervisor rating	Subjective	Employ	1970s	Low
	7	99	22	.06	.21	Verbal and arithmetic reasoning	Supervisor rating	Subjective	Employ	1970s	Low
	8	60	31	.03	-.04	Verbal and arithmetic reasoning	Supervisor rating	Subjective	Employ	1970s	Low
	9	106	24	.11	.13	Verbal and arithmetic reasoning	Supervisor rating	Subjective	Employ	1970s	Low
	10	74	17	.06	.13	Test of mental alertness	Multiple criteria	Composite	Employ	1970s	Low
Pandey (1971)	1	33	47	.32	.11	SCAT	C GPA	Grades	Educ	1970s	n/a
Patterson et al. (2009)	1	109,153	9,998	.30	.25	SAT	F GPA	Grades	Educ	2000s	n/a
Pearson (1993)	1	892	n/a	.29	n/a	SAT	C GPA	Grades	Educ	1990s	n/a
Pennock-Román (1990)	1	898	n/a	.36	n/a	SAT	F GPA	Grades	Educ	1980s	n/a
	2	1,304	n/a	.28	n/a	SAT	F GPA	Grades	Educ	1980s	n/a
	3	4,347	n/a	.36	n/a	SAT	F GPA	Grades	Educ	1980s	n/a
	4	2,565	n/a	.26	n/a	SAT	F GPA	Grades	Educ	1980s	n/a
	5	4,473	n/a	.35	n/a	SAT	F GPA	Grades	Educ	1980s	n/a
	6	2,033	n/a	.09	n/a	SAT	F GPA	Grades	Educ	1980s	n/a
Pfeifer & Sedlacek (1971)	1	178	126	.46	.56	SAT	F GPA	Grades	Educ	1960s	n/a
Roberts & Skinner (1996)	1	12,453	511	.35	.38	AFOQT-AA	Final training grade	Grades	Military	1980s	High
Scott (1976)	1	878	67	.21	-.04	ACT	J GPA	Grades	Educ	1970s	n/a
Sue & Abe (1988)	1	902	n/a	.23	n/a	SAT	F GPA	Grades	Educ	1980s	n/a
Synk & Swarthout (1987)	1	624	196	.05	-.01	GATB	Supervisor rating	Subjective	Employ	1980s	High
	2	81	44	.07	.11	GATB	Supervisor rating	Subjective	Employ	1980s	Low
	3	292	66	.27	.19	GATB	Supervisor rating	Subjective	Employ	1980s	High
	4	9,938	3,886	.19	.15	GATB	Supervisor rating	Subjective	Employ	1980s	Medium
	5	4,834	3,662	.20	.12	GATB	Supervisor rating	Subjective	Employ	1980s	Low

(Appendix continues)

Table A1 (continued)

Reference	Sample no.	N		Correlation		Test	Criterion	Criterion type	Domain	Year	Job complexity
		White	Black	White	Black						
Temp (1971)	1	100	100	.27	.26	SAT	F GPA	Grades	Educ	1960s	n/a
	2	99	98	.23	.15	SAT	F GPA	Grades	Educ	1960s	n/a
	3	104	104	.38	.30	SAT	F GPA	Grades	Educ	1960s	n/a
	4	93	92	.33	.18	SAT	F GPA	Grades	Educ	1960s	n/a
	5	140	140	.55	.51	SAT	F GPA	Grades	Educ	1960s	n/a
	6	92	102	.39	.25	SAT	F GPA	Grades	Educ	1960s	n/a
	7	100	99	.15	.07	SAT	F GPA	Grades	Educ	1960s	n/a
	8	97	100	.51	.27	SAT	F GPA	Grades	Educ	1960s	n/a
	9	100	100	.45	.06	SAT	F GPA	Grades	Educ	1960s	n/a
	10	95	100	.25	.39	SAT	F GPA	Grades	Educ	1960s	n/a
	11	69	68	.43	.08	SAT	F GPA	Grades	Educ	1960s	n/a
	12	100	39	.49	.41	SAT	F GPA	Grades	Educ	1960s	n/a
	13	109	104	.46	.15	SAT	F GPA	Grades	Educ	1960s	n/a
Tracey & Sedlacek (1985)	1	1,339	190	.40	.33	SAT	F GPA	Grades	Educ	1970s	n/a
	2	355	89	.41	.40	SAT	F GPA	Grades	Educ	1980s	n/a
U.S. Department of Labor (1985)	1	209	30	.36	.41	USESSATB	Supervisor rating	Subjective	Employ	1970s	Medium
Valentine (1977)	1	245	43	.39	.28	AFQT	Final school grade	Grades	Military	1970s	Medium
	2	171	43	.26	-.02	AFQT	Final school grade	Grades	Military	1970s	Medium
	3	317	55	.32	0.27	AFQT	Final school grade	Grades	Military	1970s	Medium
	4	664	230	.37	0.24	AFQT	Final school grade	Grades	Military	1970s	Low
	5	369	195	.32	0.16	AFQT	Final school grade	Grades	Military	1970s	Low
	6	1,849	181	.33	.30	AFQT	Final school grade	Grades	Military	1970s	Medium
	7	544	53	.35	.37	AFQT	Final school grade	Grades	Military	1970s	Medium
	8	2,163	244	.28	.21	AFQT	Final school grade	Grades	Military	1970s	Medium
	10	226	66	.27	-.05	AFQT	Final school grade	Grades	Military	1970s	Medium
	11	224	69	.27	.18	AFQT	Final school grade	Grades	Military	1970s	Medium
	12	75	24	.45	.10	AFQT	Final school grade	Grades	Military	1970s	Medium
	13	1,598	1,041	.32	.19	AFQT	Final school grade	Grades	Military	1970s	Medium
	15	4,559	1,073	.30	.14	AFQT	Final school grade	Grades	Military	1970s	Medium
	16	1,356	363	.41	.18	AFQT	Final school grade	Grades	Military	1970s	Medium
	17	241	52	.31	.23	AFQT	Final school grade	Grades	Military	1970s	Medium
	18	832	162	.35	.08	AFQT	Final school grade	Grades	Military	1970s	Medium
	19	912	154	.36	.18	AFQT	Final school grade	Grades	Military	1970s	Medium
	21	251	28	.38	.17	AFQT	Final school grade	Grades	Military	1970s	Medium
	23	653	160	.38	-.02	AFQT	Final school grade	Grades	Military	1970s	Low
	24	831	297	.34	.24	AFQT	Final school grade	Grades	Military	1970s	Medium
	25	505	75	.20	.19	AFQT	Final school grade	Grades	Military	1970s	Unknown
	26	215	36	.37	.03	AFQT	Final school grade	Grades	Military	1970s	Medium
	27	507	188	.22	.12	AFQT	Final school grade	Grades	Military	1970s	Unknown
	28	178	42	.25	.12	AFQT	Final school grade	Grades	Military	1970s	Low
	29	1,106	400	.44	.10	AFQT	Final school grade	Grades	Military	1970s	Unknown
	30	256	136	.13	.02	AFQT	Final school grade	Grades	Military	1970s	Medium
	31	367	265	.31	.05	AFQT	Final school grade	Grades	Military	1970s	Medium
	32	1,199	587	.32	.11	AFQT	Final school grade	Grades	Military	1970s	Medium
	33	481	360	.32	.13	AFQT	Final school grade	Grades	Military	1970s	Unknown
	34	439	100	.41	.24	AFQT	Final school grade	Grades	Military	1970s	Unknown
	35	1,503	1,078	.33	.20	AFQT	Final school grade	Grades	Military	1970s	High
	36	453	180	.50	.26	AFQT	Final school grade	Grades	Military	1970s	Medium
	37	2,172	1,222	.29	.13	AFQT	Final school grade	Grades	Military	1970s	Medium
	38	1,078	256	.35	.32	AFQT	Final school grade	Grades	Military	1970s	Unknown
	39	934	404	.39	.21	AFQT	Final school grade	Grades	Military	1970s	Unknown
	40	1,385	470	.39	.24	AFQT	Final school grade	Grades	Military	1970s	Unknown
	41	249	48	.37	.09	AFQT	Final school grade	Grades	Military	1970s	Unknown
	42	332	63	.24	.13	AFQT	Final school grade	Grades	Military	1970s	Unknown
	43	241	68	.40	.45	AFQT	Final school grade	Grades	Military	1970s	Unknown
Wightman (2000)	1	1,188	89	.25	.15	LSAT	CL GPA	Grades	Educ	1990s	n/a
	2	3,269	231	.24	.22	LSAT	CL GPA	Grades	Educ	1990s	n/a
	3	5,206	400	.27	.19	LSAT	CL GPA	Grades	Educ	1990s	n/a
	4	7,094	325	.29	.28	LSAT	CL GPA	Grades	Educ	1990s	n/a

(Appendix continues)



Table A1 (*continued*)

Reference	Sample no.	N		Correlation		Test	Criterion	Criterion type	Domain	Year	Job complexity
		White	Black	White	Black						
Wightman & Muller (1990)	5	1,649	64	.29	.31	LSAT	CL GPA	Grades	Educ	1990s	n/a
	6	194	237	.33	.32	LSAT	CL GPA	Grades	Educ	1990s	n/a
	1	31	165	.42	.12	LSAT	FL GPA	Grades	Educ	1980s	n/a
	2	114	202	.34	.29	LSAT	FL GPA	Grades	Educ	1980s	n/a
	3	390	65	.28	.18	LSAT	FL GPA	Grades	Educ	1980s	n/a
	4	643	97	.38	.37	LSAT	FL GPA	Grades	Educ	1980s	n/a
	5	419	60	.26	.34	LSAT	FL GPA	Grades	Educ	1980s	n/a
	6	453	69	.40	.35	LSAT	FL GPA	Grades	Educ	1980s	n/a
	7	1,034	130	.39	.42	LSAT	FL GPA	Grades	Educ	1980s	n/a
	8	434	52	.33	.36	LSAT	FL GPA	Grades	Educ	1980s	n/a
	9	294	33	.36	.57	LSAT	FL GPA	Grades	Educ	1980s	n/a
	10	284	30	.48	.17	LSAT	FL GPA	Grades	Educ	1980s	n/a
	11	579	58	.29	.30	LSAT	FL GPA	Grades	Educ	1980s	n/a
	12	1,002	98	.34	.27	LSAT	FL GPA	Grades	Educ	1980s	n/a
	13	402	41	.29	.51	LSAT	FL GPA	Grades	Educ	1980s	n/a
	14	513	47	.25	.36	LSAT	FL GPA	Grades	Educ	1980s	n/a
	15	477	40	.15	.06	LSAT	FL GPA	Grades	Educ	1980s	n/a
	16	467	38	.41	.32	LSAT	FL GPA	Grades	Educ	1980s	n/a
	17	445	36	.33	.12	LSAT	FL GPA	Grades	Educ	1980s	n/a
	18	401	32	.27	.71	LSAT	FL GPA	Grades	Educ	1980s	n/a
	19	961	76	.36	.43	LSAT	FL GPA	Grades	Educ	1980s	n/a
	20	534	42	.12	.35	LSAT	FL GPA	Grades	Educ	1980s	n/a
	21	980	79	.28	.52	LSAT	FL GPA	Grades	Educ	1980s	n/a
	22	450	35	.26	.10	LSAT	FL GPA	Grades	Educ	1980s	n/a
	23	489	38	.25	.32	LSAT	FL GPA	Grades	Educ	1980s	n/a
	24	557	45	.15	.35	LSAT	FL GPA	Grades	Educ	1980s	n/a
	25	609	46	.46	.26	LSAT	FL GPA	Grades	Educ	1980s	n/a
	26	813	62	.33	.07	LSAT	FL GPA	Grades	Educ	1980s	n/a
	27	560	41	.21	.27	LSAT	FL GPA	Grades	Educ	1980s	n/a
	28	703	51	.24	.13	LSAT	FL GPA	Grades	Educ	1980s	n/a
	29	610	42	.36	.33	LSAT	FL GPA	Grades	Educ	1980s	n/a
	30	769	52	.28	.29	LSAT	FL GPA	Grades	Educ	1980s	n/a
	31	656	43	.40	.25	LSAT	FL GPA	Grades	Educ	1980s	n/a
	32	500	32	.21	.58	LSAT	FL GPA	Grades	Educ	1980s	n/a
	33	643	40	.26	.38	LSAT	FL GPA	Grades	Educ	1980s	n/a
	34	549	34	.31	.23	LSAT	FL GPA	Grades	Educ	1980s	n/a
	35	623	38	.37	.56	LSAT	FL GPA	Grades	Educ	1980s	n/a
	36	756	46	.41	.30	LSAT	FL GPA	Grades	Educ	1980s	n/a
	37	671	38	.32	.65	LSAT	FL GPA	Grades	Educ	1980s	n/a
	38	684	39	.26	.47	LSAT	FL GPA	Grades	Educ	1980s	n/a
	39	710	37	.19	-.07	LSAT	FL GPA	Grades	Educ	1980s	n/a
	40	886	46	.40	.40	LSAT	FL GPA	Grades	Educ	1980s	n/a
	41	701	35	.38	.43	LSAT	FL GPA	Grades	Educ	1980s	n/a
	42	1,040	50	.33	.10	LSAT	FL GPA	Grades	Educ	1980s	n/a
	43	961	55	.38	.61	LSAT	FL GPA	Grades	Educ	1980s	n/a
	44	604	31	.27	.50	LSAT	FL GPA	Grades	Educ	1980s	n/a
	45	829	36	.16	.20	LSAT	FL GPA	Grades	Educ	1980s	n/a
	46	954	43	.41	.15	LSAT	FL GPA	Grades	Educ	1980s	n/a
	47	1,079	47	.27	.52	LSAT	FL GPA	Grades	Educ	1980s	n/a
	48	1,235	53	.29	.48	LSAT	FL GPA	Grades	Educ	1980s	n/a
	49	1,262	59	.24	.25	LSAT	FL GPA	Grades	Educ	1980s	n/a
	50	1,053	39	.37	.18	LSAT	FL GPA	Grades	Educ	1980s	n/a
	51	1,253	41	.26	.46	LSAT	FL GPA	Grades	Educ	1980s	n/a
	52	255	n/a	.51	n/a	LSAT	FL GPA	Grades	Educ	1980s	n/a
	53	631	n/a	.23	n/a	LSAT	FL GPA	Grades	Educ	1980s	n/a
	54	1,108	n/a	.25	n/a	LSAT	FL GPA	Grades	Educ	1980s	n/a
Wilbourn et al. (1984)	1	107	30	.57	.72	AFQT	Training grade	Grades	Military	1980s	High
	2	192	53	.40	.47	AFQT	Training grade	Grades	Military	1980s	High

(Appendix continues)

Table A1 (continued)

Reference	Sample no.	N		Correlation		Test	Criterion	Criterion type	Domain	Year	Job complexity
		White	Black	White	Black						
	3	242	106	.39	.17	AFQT	Training grade	Grades	Military	1980s	Medium
	4	89	39	.28	.22	AFQT	Training grade	Grades	Military	1980s	High
	5	208	27	.29	.06	AFQT	Training grade	Grades	Military	1980s	High
	6	296	28	.23	.26	AFQT	Training grade	Grades	Military	1980s	High
	7	211	27	.25	.50	AFQT	Training grade	Grades	Military	1980s	High
	8	310	25	.36	.23	AFQT	Training grade	Grades	Military	1980s	Medium
	9	324	25	.35	.33	AFQT	Training grade	Grades	Military	1980s	Unknown
	10	454	85	.46	.41	AFQT	Training grade	Grades	Military	1980s	Medium
	11	267	81	.27	.23	AFQT	Training grade	Grades	Military	1980s	Medium
	12	353	66	.28	.04	AFQT	Training grade	Grades	Military	1980s	Medium
	13	1,080	126	.40	.38	AFQT	Training grade	Grades	Military	1980s	Medium
	14	133	28	.42	.13	AFQT	Training grade	Grades	Military	1980s	Medium
	15	268	47	.42	.13	AFQT	Training grade	Grades	Military	1980s	High
	16	129	27	.36	.15	AFQT	Training grade	Grades	Military	1980s	High
	17	492	49	.41	.16	AFQT	Training grade	Grades	Military	1980s	Medium
	18	1,930	207	.45	.41	AFQT	Training grade	Grades	Military	1980s	Medium
	19	1913	250	.45	.27	AFQT	Training grade	Grades	Military	1980s	Medium
	20	128	36	.45	.31	AFQT	Training grade	Grades	Military	1980s	High
	21	668	120	.40	.16	AFQT	Training grade	Grades	Military	1980s	High
	22	66	37	.49	.50	AFQT	Training grade	Grades	Military	1980s	Unknown
	23	72	31	.53	.47	AFQT	Training grade	Grades	Military	1980s	Medium
	24	124	36	.47	.59	AFQT	Training grade	Grades	Military	1980s	Medium
	25	309	117	.33	.37	AFQT	Training grade	Grades	Military	1980s	Medium
	26	359	155	.38	.33	AFQT	Training grade	Grades	Military	1980s	Unknown
	27	1,135	630	.42	.21	AFQT	Training grade	Grades	Military	1980s	High
	28	483	174	.54	.38	AFQT	Training grade	Grades	Military	1980s	Medium
	29	3,809	772	.41	.29	AFQT	Training grade	Grades	Military	1980s	Medium
	30	1,422	400	.45	.31	AFQT	Training grade	Grades	Military	1980s	Medium
	31	455	129	.50	.35	AFQT	Training grade	Grades	Military	1980s	Medium
	32	89	26	.43	.02	AFQT	Training grade	Grades	Military	1980s	Medium
	33	167	69	.49	.19	AFQT	Training grade	Grades	Military	1980s	Medium
	34	157	60	.39	.03	AFQT	Training grade	Grades	Military	1980s	Medium
Wynne (2003)	1	379	n/a	.33	n/a	SAT	F GPA	Grades	Educ	2000s	n/a
Young (1994)	1	3,166	211	.27	.16	SAT	C GPA	Grades	Educ	1980s	n/a

*Note.* Multiple tests indicates that the original study used multiple cognitive ability tests. Multiple criteria indicates that the original study used multiple criteria. For the overall analysis the multiple predictors and/or criteria were averaged (with composite formulas where possible); the calculated correlations are presented above. For the objective/subjective moderator analysis, when the original study presented data separately for each type of criteria we used this data in the analysis. (If the original study did not present data separately by criterion type, and there were both subjective and objective criteria in this composite, we did not include the study in the moderator analysis.) Description of multiple tests and multiple criteria: Campbell et al. (1973) used supervisor ratings and a job knowledge test (Study 1); supervisor ratings, a job knowledge test, and a work sample (Study 2); and supervisor ratings and a work sample (Study 4). Lopez (1966) used supervisor's ranking and tolls accuracy rate. Fox and Lefkowitz (1974) used objective performance data, supervisor ratings, and supervisor work group rankings. Gael and Grant (1972) used a job knowledge test and a work sample. O'Leary et al. (1970) used supervisor ratings and two objective measures. In the military domain, McLaughlin et al. (1984) used training grades and work sample performance. Kallingal (1971) used Michigan State University (MSU) English, MSU reading, and College Qualifications Test (CQT) Verbal, Informational, and Numerical. Kirkpatrick et al. (1968) used the Science Research Associates Non-Verbal Form, Differential Aptitude Tests (DAT) Abstract Reasoning, and Total score (the total number correct on four subtests: vocabulary, numerical, checking, and coding) in Study 2, and the Numerical Ability Test, Form B, of the DAT (differential aptitude tests) and the Gates Reading Survey in Study 3. Roberts and Skinner (1996) used both a final training course grade and a subjective rating criterion; however, in this instance, we chose to use only the final course grade, as the subjective rating criteria had especially low reliability. AACTE = American Association of Colleges for Teacher Education; ACT = American College Testing; AFOQT-AA = Air Force Officer Qualification Test—Academic Aptitude; AFQT = Armed Forces Qualification Test; ASVAB = Armed Services Vocational Aptitude Battery; ASVAB Mech = Armed Services Vocational Aptitude Battery Mechanical composite; Axle accuracy = for a given toll collector, this was measured by the ratio of the total number of transactions in a month to the number of errors in axle count in that month; BSQT 1 = Bell Systems Qualification Test 1 (verbal and quantitative); CAT = California Achievement Test; C GPA = cumulative grade point average; CL GPA = cumulative law school grade point average; CTMM = California Test of Mental Maturity; Dollar accuracy = for a given toll collector, this was measured in terms of the ratio of the total number of transactions in a month that the toll collector completed to the amount of error (in dollars) in the toll receipts turned in during that month; Educ = educational domain; Employ = employment domain; EPT-Essay = 45-min essay in response to a specific topic. All essays were scored on a 6-point scale by two independent raters (third rater if disagreement existed), and scores were combined for a total between 2 and 12; F GPA = freshman grade point average; FL GPA = freshman law school grade point average; GPA = grade point average; Kit of Reference Tests = Kit of Reference Tests for Cognitive Factors included many subtests (i.e., coordination, hidden figures, vocabulary, object-number, card rotations, CS arithmetic, map planning, surface development, maze tracing speed, following oral directions, identical pictures, extended range vocabulary, necessary arithmetic operations), which were averaged; GATB = General Aptitude Test Battery; J GPA = junior grade point average; LSAT = Law School Admission Test; n/a = not applicable; Otis Quick Score = Otis Quick Score Mental Ability Test; PNG = Pre-Nursing and Guidance Examination; Psychiatric Aide Tests = job knowledge test for psychiatric attendants; SAT = Scholastic Aptitude Test; SCAT = School and College Ability Test; SET = short employment test, which contained verbal, quantitative, and clerical subscales (we used only the verbal and quantitative scores); SRA-P = Science Research Associates Pictorial Reasoning Test; State Civil Service and Otis Employment Tests = these are both general ability tests; TMA = Thurstone Test of Mental Alertness; USES SATB = United States Employment Service Specific Aptitude Test Battery.

(Appendix continues)

Table A2

*Information for Primary Studies Included in the Hispanic–White Differential Validity Meta-Analysis*

Reference	Sample no.	N		Correlation		Test	Criterion	Criterion type	Domain	Year	Job complexity
		White	Hispanic	White	Hispanic						
Berry & Sackett (2008b)	1	3,530	336	.22	.10	SAT	F GPA	Grades	Educ	1990s	n/a
	2	3,199	49	.28	.06	SAT	F GPA	Grades	Educ	1990s	n/a
	3	4,606	479	.43	.38	SAT	F GPA	Grades	Educ	1990s	n/a
	4	4,044	352	.24	.29	SAT	F GPA	Grades	Educ	1990s	n/a
	5	1,996	145	.18	.34	SAT	F GPA	Grades	Educ	1990s	n/a
	6	1,273	80	.28	.14	SAT	F GPA	Grades	Educ	1990s	n/a
	7	1,331	50	.49	.38	SAT	F GPA	Grades	Educ	1990s	n/a
	8	1,767	180	.23	.36	SAT	F GPA	Grades	Educ	1990s	n/a
	9	663	13	.25	-.01	SAT	F GPA	Grades	Educ	1990s	n/a
	10	1,758	350	.26	.28	SAT	F GPA	Grades	Educ	1990s	n/a
	11	573	30	.49	.55	SAT	F GPA	Grades	Educ	1990s	n/a
	12	779	19	.49	.32	SAT	F GPA	Grades	Educ	1990s	n/a
	13	4,356	180	.34	.19	SAT	F GPA	Grades	Educ	1990s	n/a
	14	519	7	.41	.55	SAT	F GPA	Grades	Educ	1990s	n/a
	15	3,313	65	.29	.27	SAT	F GPA	Grades	Educ	1990s	n/a
	16	866	34	.39	.41	SAT	F GPA	Grades	Educ	1990s	n/a
	17	332	35	.27	.28	SAT	F GPA	Grades	Educ	1990s	n/a
	18	2,641	20	.33	.02	SAT	F GPA	Grades	Educ	1990s	n/a
	19	2,251	486	.25	.37	SAT	F GPA	Grades	Educ	1990s	n/a
	20	2,513	64	.33	.42	SAT	F GPA	Grades	Educ	1990s	n/a
	21	2,572	86	.43	.33	SAT	F GPA	Grades	Educ	1990s	n/a
	22	1,049	57	.33	.47	SAT	F GPA	Grades	Educ	1990s	n/a
	23	1,112	188	.37	.26	SAT	F GPA	Grades	Educ	1990s	n/a
	24	3,394	247	.30	.17	SAT	F GPA	Grades	Educ	1990s	n/a
	25	3,045	45	.25	.28	SAT	F GPA	Grades	Educ	1990s	n/a
	26	3,092	30	.32	.11	SAT	F GPA	Grades	Educ	1990s	n/a
	27	494	122	.41	.50	SAT	F GPA	Grades	Educ	1990s	n/a
	28	255	12	.48	.39	SAT	F GPA	Grades	Educ	1990s	n/a
	29	254	146	.38	.35	SAT	F GPA	Grades	Educ	1990s	n/a
	30	1,933	184	.31	.20	SAT	F GPA	Grades	Educ	1990s	n/a
	31	8,314	1,281	.39	.38	SAT	F GPA	Grades	Educ	1990s	n/a
	32	2,221	266	.36	.38	SAT	F GPA	Grades	Educ	1990s	n/a
	33	823	55	.38	.19	SAT	F GPA	Grades	Educ	1990s	n/a
	34	4,361	604	.20	.19	SAT	F GPA	Grades	Educ	1990s	n/a
	35	955	650	.30	.29	SAT	F GPA	Grades	Educ	1990s	n/a
	36	7,184	95	.40	.28	SAT	F GPA	Grades	Educ	1990s	n/a
	37	4,332	156	.37	.24	SAT	F GPA	Grades	Educ	1990s	n/a
	38	11,202	2,434	.39	.37	SAT	F GPA	Grades	Educ	1990s	n/a
	39	3,707	116	.42	.27	SAT	F GPA	Grades	Educ	1990s	n/a
	40	6,439	317	.36	.31	SAT	F GPA	Grades	Educ	1990s	n/a
	41	1,811	41	.29	.51	SAT	F GPA	Grades	Educ	1990s	n/a
Breland & Griswold (1981)	1	5,236	445	.22	.39	SAT	EPT-Essay	Subjective	Educ	1980s	n/a
Bridgeman et al. (2000)	1	29,152	3,225	.33	.32	SAT	F GPA	Grades	Educ	1990s	n/a
	2	31,169	3,451	.34	.31	SAT	F GPA	Grades	Educ	1990s	n/a
Dittmar (1977)	1	233	209	.48	.41	SAT	F GPA	Grades	Educ	1970s	n/a
	2	270	292	.45	.25	SAT	F GPA	Grades	Educ	1970s	n/a
Duran (1983)	1	218	187	.49	.42	SAT	GPA	Grades	Educ	1970s	n/a
	2	254	266	.44	.21	SAT	GPA	Grades	Educ	1970s	n/a
Goldman & Hewitt (1975)	1	5,635	261	.29	.16	SAT	C GPA	Grades	Educ	1970s	n/a
	2	5,500	84	.36	.43	SAT	C GPA	Grades	Educ	1970s	n/a
	3	2,926	180	.31	.22	SAT	C GPA	Grades	Educ	1970s	n/a
	4	3,127	131	.31	.23	SAT	C GPA	Grades	Educ	1970s	n/a
Goldman & Hewitt (1976)	1	4,259	188	.30	.18	SAT	C GPA	Grades	Educ	1970s	n/a
Goldman & Richards (1974)	1	210	42	.44	.25	SAT	F GPA	Grades	Educ	1970s	n/a
	2	1,700	110	.32	.19	SAT	F GPA	Grades	Educ	1970s	n/a
Haney et al. (1976)	1	223	73	.29	.21	CAT	Course GPA	Grades	Educ	1970s	n/a
Lichtman (2008)	1	859	48	.31	.59	ACT	F GPA	Grades	Educ	2000s	n/a
Mattern et al. (2008)	1	104,017	10,486	.30	.29	SAT	F GPA	Grades	Educ	2000s	n/a

*(Appendix continues)*

Table A2 (*continued*)

Reference	Sample no.	<i>N</i>		Correlation		Test	Criterion	Criterion type	Domain	Year	Job complexity
		White	Hispanic	White	Hispanic						
McCornack (1983)	1	2,263	94	.22	.16	SAT	F GPA	Grades	Educ	1970s	n/a
	2	2,009	115	.24	.44	SAT	F GPA	Grades	Educ	1980s	n/a
Morgan (1990)	1	89,013	1,575	.41	.30	SAT	F GPA	Grades	Educ	1970s	n/a
	2	89,524	1,354	.37	.31	SAT	F GPA	Grades	Educ	1980s	n/a
	3	74,586	2,192	.36	.27	SAT	F GPA	Grades	Educ	1980s	n/a
Patterson et al. (2009)	1	109,153	12,717	.30	.28	SAT	F GPA	Grades	Educ	2000s	n/a
Pearson (1993)	1	892	220	.29	.28	SAT	C GPA	Grades	Educ	1990s	n/a
Pennock-Román (1990)	1	898	110	.36	.25	SAT	F GPA	Grades	Educ	1980s	n/a
	2	1,304	70	.28	.39	SAT	F GPA	Grades	Educ	1980s	n/a
	3	4,347	637	.36	.27	SAT	F GPA	Grades	Educ	1980s	n/a
	4	2,565	135	.26	.27	SAT	F GPA	Grades	Educ	1980s	n/a
	5	4,473	129	.35	.09	SAT	F GPA	Grades	Educ	1980s	n/a
	6	2,033	177	.09	.09	SAT	F GPA	Grades	Educ	1980s	n/a
Scott (1976)	1	878	66	.21	.20	ACT	J GPA	Grades	Educ	1970s	n/a
Wightman (2000)	1	1,188	41	.25	.22	LSAT	CL GPA	Grades	Educ	1990s	n/a
	2	3,269	109	.24	.38	LSAT	CL GPA	Grades	Educ	1990s	n/a
	3	5,206	114	.27	.47	LSAT	CL GPA	Grades	Educ	1990s	n/a
	4	7,094	182	.29	.36	LSAT	CL GPA	Grades	Educ	1990s	n/a
	5	1,649	40	.29	.37	LSAT	CL GPA	Grades	Educ	1990s	n/a
Wightman & Muller (1990)	2	114	59	.34	.11	LSAT	FL GPA	Grades	Educ	1980s	n/a
	6	453	41	.40	.53	LSAT	FL GPA	Grades	Educ	1980s	n/a
	7	1,034	41	.39	.31	LSAT	FL GPA	Grades	Educ	1980s	n/a
	13	402	36	.29	.15	LSAT	FL GPA	Grades	Educ	1980s	n/a
	21	980	31	.28	.57	LSAT	FL GPA	Grades	Educ	1980s	n/a
	24	557	31	.15	.63	LSAT	FL GPA	Grades	Educ	1980s	n/a
	26	813	32	.33	.36	LSAT	FL GPA	Grades	Educ	1980s	n/a
	38	684	31	.26	.49	LSAT	FL GPA	Grades	Educ	1980s	n/a
	43	961	210	.38	.43	LSAT	FL GPA	Grades	Educ	1980s	n/a
	44	604	58	.27	.14	LSAT	FL GPA	Grades	Educ	1980s	n/a
	46	954	66	.41	.30	LSAT	FL GPA	Grades	Educ	1980s	n/a
	47	1,079	36	.27	.32	LSAT	FL GPA	Grades	Educ	1980s	n/a
	48	1,235	47	.29	.40	LSAT	FL GPA	Grades	Educ	1980s	n/a
	49	1,262	179	.24	.38	LSAT	FL GPA	Grades	Educ	1980s	n/a
	50	1,053	37	.37	.53	LSAT	FL GPA	Grades	Educ	1980s	n/a
Wynne (2003)	52	255	47	.51	.51	LSAT	FL GPA	Grades	Educ	1980s	n/a
	53	631	58	.23	.12	LSAT	FL GPA	Grades	Educ	1980s	n/a
	54	1,108	32	.25	.23	LSAT	FL GPA	Grades	Educ	1980s	n/a
	1	379	132	.33	.39	SAT	F GPA	Grades	Educ	2000s	n/a
	1	3,166	140	.27	.22	SAT	C GPA	Grades	Educ	1980s	n/a

*Note.* ACT = American College Testing; CAT = California Achievement Test; CL GPA = cumulative law school grade point average; C GPA = cumulative grade point average; Educ = educational domain; EPT-Essay = 45-min essay in response to a specific topic. All essays were scored on a 6-point scale by two independent raters (third rater if disagreement existed), and scores were combined for a total between 2 and 12; F GPA = freshman grade point average; FL GPA = freshman law school grade point average; GPA = grade point average; J GPA = junior grade point average; LSAT = Law School Admission Test; n/a = not applicable; SAT = Scholastic Aptitude Test.

(*Appendix continues*)



Table A3

*Information for Primary Studies Included in the Asian–White Differential Validity Meta-Analysis*

Reference	Sample no.	N		Correlation		Test	Criterion	Criterion type	Domain	Year	Job complexity
		White	Asian	White	Asian						
Berry & Sackett (2008b)	1	3,530	509	.22	–.02	SAT	F GPA	Grades	Educ	1990s	n/a
	2	3,199	63	.28	.33	SAT	F GPA	Grades	Educ	1990s	n/a
	3	4,606	1,170	.43	.40	SAT	F GPA	Grades	Educ	1990s	n/a
	4	4,044	1,132	.24	.20	SAT	F GPA	Grades	Educ	1990s	n/a
	5	1,996	772	.18	.17	SAT	F GPA	Grades	Educ	1990s	n/a
	6	1,273	192	.28	.12	SAT	F GPA	Grades	Educ	1990s	n/a
	7	1,331	35	.49	.23	SAT	F GPA	Grades	Educ	1990s	n/a
	8	1,767	161	.23	.17	SAT	F GPA	Grades	Educ	1990s	n/a
	9	663	6	.25	.52	SAT	F GPA	Grades	Educ	1990s	n/a
	10	1,758	147	.26	.26	SAT	F GPA	Grades	Educ	1990s	n/a
	11	573	19	.49	.50	SAT	F GPA	Grades	Educ	1990s	n/a
	12	779	24	.49	.56	SAT	F GPA	Grades	Educ	1990s	n/a
	13	4,356	680	.34	.29	SAT	F GPA	Grades	Educ	1990s	n/a
	14	519	37	.41	.15	SAT	F GPA	Grades	Educ	1990s	n/a
	15	3,313	46	.29	.16	SAT	F GPA	Grades	Educ	1990s	n/a
	16	866	86	.39	.33	SAT	F GPA	Grades	Educ	1990s	n/a
	17	332	174	.27	.34	SAT	F GPA	Grades	Educ	1990s	n/a
	18	2,641	41	.33	.30	SAT	F GPA	Grades	Educ	1990s	n/a
	19	2,251	3,527	.25	.32	SAT	F GPA	Grades	Educ	1990s	n/a
	20	2,513	49	.33	.00	SAT	F GPA	Grades	Educ	1990s	n/a
	21	2,572	272	.43	.43	SAT	F GPA	Grades	Educ	1990s	n/a
	22	1,049	123	.33	.37	SAT	F GPA	Grades	Educ	1990s	n/a
	23	1,112	327	.37	.23	SAT	F GPA	Grades	Educ	1990s	n/a
	24	3,394	52	.30	.39	SAT	F GPA	Grades	Educ	1990s	n/a
	25	3,045	32	.25	.25	SAT	F GPA	Grades	Educ	1990s	n/a
	26	3,092	30	.32	.44	SAT	F GPA	Grades	Educ	1990s	n/a
	27	494	82	.41	.49	SAT	F GPA	Grades	Educ	1990s	n/a
	28	255	3	.48	–.87	SAT	F GPA	Grades	Educ	1990s	n/a
	29	254	103	.38	.53	SAT	F GPA	Grades	Educ	1990s	n/a
	30	1,933	462	.31	.22	SAT	F GPA	Grades	Educ	1990s	n/a
	31	8,314	324	.39	.37	SAT	F GPA	Grades	Educ	1990s	n/a
	32	2,221	58	.36	.23	SAT	F GPA	Grades	Educ	1990s	n/a
	33	823	77	.38	.43	SAT	F GPA	Grades	Educ	1990s	n/a
	34	4,361	265	.20	.23	SAT	F GPA	Grades	Educ	1990s	n/a
	35	955	1,287	.30	.30	SAT	F GPA	Grades	Educ	1990s	n/a
	36	7,184	475	.40	.43	SAT	F GPA	Grades	Educ	1990s	n/a
	37	4,332	473	.37	.32	SAT	F GPA	Grades	Educ	1990s	n/a
	38	11,202	2,836	.39	.44	SAT	F GPA	Grades	Educ	1990s	n/a
	39	3,707	133	.42	.53	SAT	F GPA	Grades	Educ	1990s	n/a
	40	6,439	2,674	.36	.40	SAT	F GPA	Grades	Educ	1990s	n/a
	41	1,811	35	.29	.44	SAT	F GPA	Grades	Educ	1990s	n/a
Breland & Griswold (1981)	1	5,236	606	.22	.34	SAT	EPT-Essay	Subjective	Educ	1980s	n/a
Bridgeman et al. (2000)	1	29,152	7,814	.33	.35	SAT	F GPA	Grades	Educ	1990s	n/a
	2	31,169	7,865	.34	.37	SAT	F GPA	Grades	Educ	1990s	n/a
Goldman & Hewitt (1976)	1	4,259	852	.30	.32	SAT	C GPA	Grades	Educ	1970s	n/a
Lichtman (2008)	1	859	122	.31	.44	ACT	F GPA	Grades	Educ	2000s	n/a
Mattern et al. (2008)	1	104,017	14,109	.30	.30	SAT	F GPA	Grades	Educ	2000s	n/a
McCornack (1983)	1	2,263	82	.22	.26	SAT	F GPA	Grades	Educ	1970s	n/a
	2	2,009	148	.24	.53	SAT	F GPA	Grades	Educ	1980s	n/a
Morgan (1990)	1	89,013	2,535	.41	.42	SAT	F GPA	Grades	Educ	1970s	n/a
	2	89,524	3,585	.37	.36	SAT	F GPA	Grades	Educ	1980s	n/a
	3	74,586	4,375	.36	.37	SAT	F GPA	Grades	Educ	1980s	n/a
Patterson et al. (2009)	1	109,153	14,363	.30	.30	SAT	F GPA	Grades	Educ	2000s	n/a
Sue & Abe (1988)	1	902	3,922	.23	.30	SAT	F GPA	Grades	Educ	1980s	n/a
Wightman (2000)	1	1,188	90	.25	.13	LSAT	CL GPA	Grades	Educ	1990s	n/a
	2	3,269	275	.24	.33	LSAT	CL GPA	Grades	Educ	1990s	n/a
	3	5,206	202	.27	.30	LSAT	CL GPA	Grades	Educ	1990s	n/a
	4	7,094	256	.29	.33	LSAT	CL GPA	Grades	Educ	1990s	n/a
Wynne (2003)	1	379	325	.33	.39	SAT	F GPA	Grades	Educ	2000s	n/a
Young (1994)	1	3,166	186	.27	.35	SAT	C GPA	Grades	Educ	1980s	n/a

*Note.* C GPA = cumulative grade point average; CL GPA = cumulative law school grade point average; Educ = educational domain; EPT-Essay = 45-min essay in response to a specific topic. All essays were scored on a 6-point scale by two independent raters (third rater if disagreement existed), and scores were combined for a total between 2 and 12; F GPA = freshman grade point average; LSAT = Law School Admission Test; n/a = not applicable; SAT = Scholastic Aptitude Test.

(Appendix continues)

Table A4

*Meta-Analytic Results Both for All Studies and for Only Employment Studies When Including vs. Excluding the GATB Validity Studies*

Variable	<i>N</i>		<i>k</i>		$\bar{r}$		$SD_r$		% var		95% CI	
	White	Black	White	Black	White	Black	White	Black	White	Black	White	Black
All studies, GATB included	903,779	112,194	405	392	.33	.24	0.06	0.11	6.89	21.16	[.32, .34]	[.23, .25]
All studies, GATB excluded	888,010	104,340	292	279	.34	.25	0.06	0.10	8.48	23.69	[.33, .34]	[.23, .26]
Employment samples, GATB included	20,399	10,350	143	143	.19	.16	0.07	0.10	30.53	31.58	[.18, .20]	[.14, .18]
Employment samples, GATB excluded	4,630	2,707	30	30	.21	.26	0.14	0.16	32.47	42.97	[.16, .26]	[.20, .31]

*Note.* *N* = total sample sizes; *k* = number of correlations;  $\bar{r}$  = mean sample-size-weighted observed correlation;  $SD_r$  = sample-size-weighted observed standard deviation of correlations; % var = percentage of variance attributable to sampling error; CI = confidence interval; GATB = General Aptitude Test Battery.

Received September 17, 2009

Revision received December 13, 2010

Accepted January 17, 2011 ■

### E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <http://notify.apa.org/> and you will be notified by e-mail when issues of interest to you become available!