PERSONNEL PSYCHOLOGY

THE STUDY OF PEOPLE AT WORK

PERSONNEL PSYCHOLOGY
2013, 66, 91–126

REVIEWING, CATEGORIZING, AND ANALYZING THE LITERATURE ON BLACK–WHITE MEAN DIFFERENCES FOR PREDICTORS OF JOB PERFORMANCE: VERIFYING SOME PERCEPTIONS AND UPDATING/CORRECTING OTHERS

PHILIP BOBKO
Gettysburg College

PHILIP L. ROTH
Clemson University

In both theoretical and applied literatures, there is confusion regarding accurate values for expected Black–White subgroup differences in personnel selection test scores. Much confusion arises because empirical estimates of standardized subgroup differences (*d*) are subject to many of the same biasing factors associated with validity coefficients (i.e., *d* is functionally related to a point-biserial *r*). To address such issues, we review/cumulate, categorize, and analyze a systematic set of many predictor-specific meta-analyses in the literature. We focus on confounds due to general use of concurrent, versus applicant, samples in the literature on Black–White *d*. We also focus on potential confusion due to different constructs being assessed within the same selection test method, as well as the influence of those constructs on *d*. It is shown that many types of predictors (such as biodata inventories or assessment centers) can have magnitudes of *d* that are much larger than previously thought. Indeed, some predictors (such as work samples) can have *d*s similar to that associated with paper-and-pencil tests of cognitive ability. We present more realistic values of *d* for both researcher and practitioner use. Implications for practice and future research are noted.

In the literature on personnel selection, as well as organizational application, much attention is given to both (a) the level of validity of individual predictors and (b) the level of adverse impact associated with such predictors. Organizations often focus on selection *validity* in order to increase the expected performance levels of their hires, increase organizational utility, and defend test use when adverse impact is present. Organizations

We thank the editor, two anonymous reviewers, Amy Hooper, and Barbara Bobko for helpful comments on earlier versions of this work.

Correspondence and requests for reprints should be addressed to Philip Bobko, Department of Management, Gettysburg College, Gettysburg, PA 17325; pbobko@gettysburg.edu.

 doi: 10.1111/peps.12007

also care about *adverse impact*. The reduction of adverse impact is an important legal issue (cf. Section 3 of the *Uniform Guidelines*, U.S. Equal Opportunity Employment Commission et al., 1978), and the notion of a diverse workforce has been conceptually linked to organizational outcomes (e.g, Joshi & Roh, 2009; McDaniel & Walls, 1997).

Across the decades, there have been periodic summaries of the point estimates of the validity of predictors of job performance (e.g., Ghiselli, 1966, 1973; Hunter & Hunter, 1984; Schmidt & Hunter, 1998; Schmitt, Gooding, Noe, & Kirsch, 1984). One ubiquitous finding is that tests of cognitive ability are some of the most empirically valid predictors of job performance (Schmidt & Hunter, 1998). However, as discussed in more detail later, cognitive ability tests are often associated with substantial mean score differences between Blacks and Whites. Thus, there has been a search for other predictors that might maintain validity while decreasing potential subgroup adverse impact. This search has also been motivated in practice by the statement in the *Uniform Guidelines* (1978, Section 3B) that employers consider feasible alternative selection systems that are "substantially equally valid" and have less adverse impact.

There are also some narrative reviews and meta-analytic studies that report values of Black–White *d*s for predictors of job performance. Although these efforts have added to the literature, many are plagued with some methodological problems (e.g., Ployhart & Holtz, 2008; Salgado, Viswesvaran, & Ones, 2001; see more detailed discussion later). We address these concerns by systematically reviewing, categorizing, and analyzing these previous reviews as well as recent predictor-specific meta-analyses. Thus, this study is not a single meta-analysis, per se, but a cumulation, investigation, and analysis of underlying patterns across sub-literatures on Black–White differences. Looking ahead, we categorize approximately 20 meta-analyses (e.g., by use of applicant vs. incumbent samples) as well as a few primary studies. That is, many of the meta-analyses about selection tests cumulated Black–White *d*s across studies, which were based on responses from incumbents. We compiled these analyses in our "incumbent" category. In contrast, several meta-analyses split their Black–White results into estimates of *d* from incumbent analyses and estimates of *d* from applicant studies. We placed the former meta-analytic estimates in our "incumbent" category and the latter estimates in a category called "applicant." We also found a few primary studies where subgroup *d*s were computed solely at the applicant level, and we also included them in our applicant category. We note that the median number of studies within each of the meta-analyses we categorized and complied is about 19 and the median sample size is about 17,500.

We focus on what we regard to be perhaps the two most important methodological problems in the existing assessments of subgroup differences: (a) computing *d* based on range-restricted incumbent samples

rather than applicant samples and (b) confounding a variety of constructs with specific methods of measurement. These methodological concerns have occasionally been mentioned in the domain of personnel selection (e.g., see Arthur & Villado, 2008, or Hunter & Hunter, 1984, for statements about construct-method confounds). However, much literature persistently ignores these issues. For example, many influential meta-analytic studies regarding levels of *d* aggregate primary studies that are subject to differing levels of range restriction (due to use of incumbent samples without correction) or confound methods and constructs.

In this paper, we first review some of the literature on each of the two above methodological concerns. We then revisit the literatures on predictors of job performance, including proposed alternatives to cognitive ability tests. For each predictor, we identify presumed levels of Black–White *d*—expectations often based on faulty estimates due to the above methodological issues. Using a variety of recent meta-analyses, we then summarize the more accurate levels of *d* that result when values are estimated at the applicant level (see Table 1). We aggregate findings by both methods (e.g., work samples) and constructs (e.g., cognitive ability). We note construct data and related issues for each method when such information is available. We then collate, in Table 2, the available construct-based estimates of *d* at the applicant level. It is important to again note that this compilation is primarily a collation of a variety of meta-analytic *d*s and not just a compilation of primary studies.

Effect Size Index

We use the statistic *d*; that is, the standardized subgroup difference associated with a given selection test. The value of *d* is computed by subtracting the minority group mean from the majority group mean and dividing this difference by the pooled within-group standard deviation of test scores (cf. Hunter & Schmidt, 2004). Values of *d* are independent of original scale choices, and they can be compared across tests and databases because they are standardized (cf. Sackett & Ellingson, 1997). In addition, values of *d* can be computed on job applicants without actually selecting individuals. In this regard, *d* can be considered as an index of "adverse impact potential."

We focus on the subgroups of Blacks and Whites. We chose to focus on these two subgroups because there is a relatively wider array of literature on Black–White differences in selection tests, yet scant available data on other racial and ethnic group minorities (e.g., Hispanics, as noted by Whetzel, McDaniel, & Nguyen, 2008).

Role of applicant vs. incumbent samples on estimates of d. We use the term "applicant sample" to mean that individuals are not prescreened on other tests/predictors such as cognitive ability tests, interviews, and so on.

TABLE 1
Updated, More Realistic Estimates of d *for Applicant Populations*

Test type	Potentially confounded *d*	Updated *d*	Underlying construct evidence
Cognitive ability	1.0[a]	.72 for moderate complexity jobs; .86 for low complexity jobs[b]	Already a "construct"; within job complexity level, overall *d* is lower; also varies by subfacet
Work samples	.38[c]; low	.73[d](varies at construct saturation level)	Large array of KSAs possible; .27 for oral communication and interpersonal, .27 for leadership and persuasion, .70 for written skills, .80 for cognitive processes and job knowledge[d]
Situational judgment tests	Range of .30 to .60; commonly .38;[e] moderate/ relatively low	Unclear, though likely larger; weighted avg. is .38, median is .67 (varies at job level from .09 to 1.04)[f]	Large array of KSAs potentially measured; cognitive saturation increases *d*; .19 for interpersonal, .65 for cognitive and job knowledge, 1.02 for leadership[f]
Biodata	Relatively low; small, commonly .33[g]	Larger; tentative avg. is .39,[h] median is .31; some variation by construct; one estimate is cross-job; one complex job is .73	Large array of KSAs potentially measured; expect cognitive saturation to influence *d* via academic factors
Assessment centers	Low; small (range .03 to .60)[i]	.56[j]	Large array of KSAs potentially measured; cognitive saturation increases *d*
Personality	Small and close to 0[k]	.04 for integrity[l] and −.09 for conscientiousness[m] from self-report measures (.30 for some incumbent interview ratings)[n]	Literature at construct level
Structured interviews	about .25[0]	Larger; tentative range from .31 to .46[p,q] (varies at construct level)	Large array of KSAs potentially measured; preliminary data suggests *d* ranges from .13 to .49 by category; however, these values are downwardly biased and computed across all structure levels

continued

TABLE 1 (continued)

Note. Numerical sources are footnoted; see text for details. Key sources are (k is the number of studies; n is the overall sample size):
(a) Hunter & Hunter (1984) (statistics not provided) and Ployhart & Holtz (2008) citing Roth, BeVier, Bobko, Switzer, & Tyler, (2001) ($k = 34$, $n = 464{,}201$).
(b) Roth, BeVier, et al. (2001) ($k = 18$, $n = 31{,}990$ and $k = 64$, $n = 125{,}654$, respectively).
(c) Schmitt, Clause, & Pulakos (1996)($k = 37$, $n = 15{,}738$).
(d) Roth, Bobko, et al. (2008) ($k = 21$, $n = 2476$ for overall value; k and n smaller for subcategories).
(e) Schmitt, Claus, & Pulakos (1996) ($k = 37$, $n = 15{,}738$) and Whetzel, et al. (2008) ($k = 62$, $n = 42{,}178$).
(f) Roth, Bobko, & Buster (2008) ($k = 6$, $n = 1337$; $k = 2$ and $n = 135$ for managerial jobs; $k = 4$ and $n = 1202$ for nonmanagerial jobs).
(g) DeCorte, Lievens, & Sackett (2007) and Ployhart & Holtz (2008), citing Bobko, Roth, Potosky (1999) ($k = 2$ cross-job samples, $n = 6{,}115$).
(h) Sample weighted value from Becton, Matthews, Hartley, & Whitaker (2009), Kriska (2001), and Dean (1999) ($k = 3$, $n = 18{,}776$).
(i) Hough & Oswald (2000) and Ployhart & Holtz (2008) (values from primary studies, statistics not provided).
(j) Dean, Roth, & Bobko (2008) ($k = 17$, $n = 8{,}210$).
(k) Gatewood, Feild & Barrick (2011) (textbook) and Hough, Oswald, & Ployhart (2001) (k not reported, $n = 50{,}199$ and 122,377).
(l) Ones and Viswesvaran (1998) ($k = 4$, $n = 590{,}394$).
(m) corrected d based on Foldes, Duehr, & Ones (2008) ($k = 67$, $n = 180{,}478$).
(n) Huffcutt, Conway, Roth, & Stone (2001) ($k = 15$, $n = 5{,}443$).
(o) Hough et al. (2001) and Ployhart & Holtz (2008), both citing Bobko et al. (1999) who use Huffcutt & Roth (1998) ($k = 21$, $n = 8{,}817$).
(p) Potosky, Bobko, & Roth (2005) (corrected value from Huffcutt & Roth, 1998, $k = 21$, $n = 8{,}817$).
(q) Roth, Van Iddekinge, Huffcutt, Eidson, & Bobko (2002) (primary study, $n = 358$).
(r) Huffcutt et al. (2001) (subcategory ks generally less than 7, ns generally 100–2000).

Minimum qualification or application blank screens typically do not cause samples to be excluded from categorization as an applicant sample (e.g., see Berry, Sackett, & Landers, 2007 or Roth, Bobko, McFarland, & Buster, 2008 for similar definitions).

Compared to applicant samples, range restriction in incumbent samples causes the magnitude of r, that is, the Pearson product moment correlation, to be biased downwards (cf. Ghiselli, 1964). Range restriction can be direct (prior selection on the same test) or indirect (selection on another test that is correlated with the test in question, or possibly restriction due to promoting good performers or firing bad ones). It is psychometrically well-known that use of incumbent samples can lead to substantial underestimates of r unless some type of range restriction correction to the applicant population is applied (cf. Schmidt, Hunter & Urry, 1976; Schmitt & Chan, 2006, p. 135).[1]

[1] In the relatively specific literature on cognitive ability, there has been a statement that predictive and concurrent observed validities are similar (Barrett, Phillips, & Alexander,

TABLE 2
Black–White, Applicant-Level d*s Presented Within Construct Category (by Method)*

Construct categories	Applicant-level *d*
Categories from Huffcutt et al. (2001)	
General cognitive ability	.72 or .86 (paper-and-pencil; by job complexity)[a]; .80 (work samples and job knowledge)[b]; .65 (SJTs)[c]; .60 (ACs)[**d]
Interpersonal & oral communication	.27 (work samples)[e]; .19 (SJTs)[f]; .28 (ACs)[**g]
Leadership	.27 (work samples; includes persuasion)[h]; 1.02 (SJTs)[i]
Job knowledge	.80 (work samples, includes cognitive ability)[j]
Conscientiousness	None for applicants
Emotional stability	None for applicants
Fit	None for applicants
Additional construct categories	
Integrity	.04 (self-report)[k]
Written skills	.70 (work samples)[l]

Note. Values of *d* based upon single, primary studies are denoted by **. All other reported values (except the SJT averages) are from meta-analytic cumulations.
Interpersonal and oral communication skills were separate in Huffcutt et al., but were combined here because applicant research did not distinguish between them.
Sources are (*k* is the number of studies; *n* is the overall sample size):
(a) Roth, BeVier, et al. (2001) ($k = 18$, $n = 31{,}990$ and $k = 64$, $n = 125{,}654$, respectively).
(b) Roth, Bobko, et al. (2008) ($k = 13$, $n = 785$).
(c) Roth, Bobko, & Buster (2008) ($k = 2$, $n = 305$).
(d) Goldstein, Riley, & Yusko (1999) as reported in Hough et al. (2001) (primary study, $n = 300$).
(e) Roth, Bobko, et al. (2008) ($k = 14$, $n = 1{,}275$).
(f) Roth, Bobko, & Buster (2008) ($k = 2$, $n = 897$).
(g) Goldstein, Reilly, & Yusko (1999) as reported in Hough et al. (2001) (primary study, $n = 300$).
(h) Roth, Bobko, et al. (2008) ($k = 14$, $n = 1{,}301$).
(i) Roth, Bobko, & Buster (2008) ($k = 2$, $n = 135$).
(j) Roth, Bobko et al. (2008) ($k = 13$, $n = 785$).
(k) Ones and Viswesvaran (1998) ($k = 4$, $n = 590{,}394$).
(l) Roth, Bobko, et al. (2008) ($k = 12$, $n = 675$).

A variety of additional factors might also influence the computed estimates of *r* in personnel selection. These factors (moderators) include complexity level of the job, faking, use of threat/warning, test coaching, test taker motivation, test response mode, and so forth. The Appendix provides specific instances of this literature, and we note that the first

1981; see also SIOP Principles, 2003, p. 15) as well as reactions to that statement (Guion & Cranny, 1982). However, findings of lower predictive validity for incumbent samples are discussed in the literature on biodata (e.g., Bliesener, 1996; Reilly & Warech, 1993, p. 145) and personality (Barrett, 2008; Hough, 1998; Morgeson et al., 2007).

part of this list also identifies methodological factors that could influence estimates of *d*. Researchers sometimes aggregate estimates of *r* across these factors—possibly because the amount of available research within each moderator is minimal, and aggregation becomes necessary or expedient.

The Link Between d *and* r

In spite of the fact that *d* and *r* are ordinal transformations of one another (cf. Bobko, Roth, & Bobko, 2001; Hunter & Schmidt, 2004), the literature on estimation of Black–White *d*s often ignores the influence of applicant versus incumbent samples. Indeed, as will be seen later, some meta-analytic estimates of *d* are based *solely* on studies of job incumbents. Thus, given that (a) *r*s are typically biased downwards and (b) *r* and *d* are related, then estimates of *d* are likely biased downwards. As also noted later, many recent statements in academic journals and reviews, practitioner publications, and legal testimony thus provide inaccurate estimates of *d* (e.g., underestimates of *d* for work sample exams, biodata, etc.), potentially misleading expectations about adverse impact, and questionable guidance to decision makers. That is, with a few exceptions (e.g., Hausknecht, Day, & Thomas, 2004; Weekley, Ployhart, & Harold, 2004), efforts in the field often ignore the fact that factors that influence estimates of validity (*r*) can also influence estimates of *d.*

To illustrate the dilemma, suppose a decision maker (not hypothetical in our experience) takes a value of $d = .38$ for work samples from the literature, notes that this is less than the literature values of 1.00 or .72 for cognitive ability tests, and then expends substantial resources developing a work sample test (or a set of such tests) to replace an existing cognitive ability test. Construct differences notwithstanding, we note that such a comparison mixes an estimate generally based on incumbent studies (i.e., .38) with estimates based on applicants (i.e., 1.00 or .72). When the practitioner's work sample is then used at the applicant level, the value of *d* is likely to be substantially larger and the practitioner will have spent additional resources without reducing adverse impact.

The concerns continue. The confounding of incumbent and applicant results continues to be of concern in the scholarly and applied literatures.[2] By way of example, the 2010 SIOP Myers award for best research practice

[2] There are times when a sample of incumbents of an organization can be appropriate for estimating *d*, that is, when incumbents apply for a promotional position and the test they are being given is for that position. For example, we consider such a possibility when discussing assessment centers, and the notion might apply to other types of tests. However, note that for data to be in our applicant category, individual respondents could be incumbents in the organization, but not incumbents in the job being tested/selected for.

involved an effort that developed computer adaptive personality measures; yet as noted in caution by those authors, all the *d*s provided were on incumbents (Kantrowitz, McClellan, Borman, Houston, & Schneider, 2009). Other recent examples in the academic literature include using incumbent *d*s in simulations for selection systems of applicants, directly comparing incumbent *d*s for one type of selection test (e.g., biodata) to applicant *d*s for another test (e.g., cognitive ability), or correcting *r*s for range restriction but not *d*s (both positive and negative; cf. DeCorte, Lievens, & Sackett, 2007; Finch, Edwards, & Wallace, 2009; Ployhart & Holtz, 2008; Van Iddekinge, Putka, & Campbell, 2011). Such examples are quite recent (others exist throughout prior years/decades), and they demonstrate the continuity of the concern, even by researchers we respect greatly.

Constructs Versus Methods

Some types of selection "tests" are more accurately considered as selection "methods." For example, it is well-known that interviews (or biodata inventories, situational judgment tests, etc.) can each be uniquely developed to target specific knowledges, skills, and abilities (KSAs). There are good general summaries indicating that confounds of methods and constructs exist in, and hinder, the literature on the estimation of *d* (e.g., Arthur & Villado, 2008; Chan & Schmitt, 1997; Sackett, Schmitt, Ellingson, & Kabin, 2001; Schmitt & Quinn, 2010). Other such statements appear in the specific contexts of work samples (Hough, Oswald, & Ployhart, 2001) and SJTs (Schmitt & Chan, 2006). However, it is common for researchers and professionals to ignore the method/construct distinction and to talk about "the *d*s" for work sample tests, SJTs, "noncognitive" tests, and so on. (e.g., Hornick, cited in Ricci et al. *v.* Destefano et al., 2006; Lievens, Buyse, & Sackett, 2005; Outtz, 1998; or Reilly & Warech, 1993).

Overall, regarding constructs and methods, the theoretical and applied literature is replete with investigations of the validity and adverse impact of "biodata," "SJTs," "trainability tests," and so on. Even a recent section on "testing and selection" of the *Journal of Applied Psychology* (2011, vol. 96, no. 6) includes analyses of "low fidelity simulations," "promotion retesting," "structured interviews," and "biodata." Thus, testing method is a standard unit of analysis, often without regard to underlying construct(s). As noted earlier, we consider both methods (because of their ubiquity of use) and constructs in the following analyses whenever possible.

Toward more accurate estimates of d. The choice of which predictors to consider in our review and analysis was difficult. We chose some predictors because they routinely appear in reviews of the literature on personnel selection (e.g., Schmidt & Hunter's, 1998, review of validity). In addition, work by Reilly et al. (e.g., Reilly, 1996; Reilly & Warech, 1993)

concluded that reasonable alternatives (i.e., substantially equal validity and less adverse impact) to paper-and-pencil tests of cognitive ability were work samples, biodata inventories, and assessment centers, and we chose those assessment tools. Although it was stated that, at the time, there were not enough data on situational judgment tests, we found recent additional work and included that method in our analyses. We also chose predictors for other reasons (e.g., personality for its popularity as a supplement to other tests).

Given the substantial dependency on use of incumbent samples in the literature, and consequent effects of range restriction, we expect that:

Expectation 1: More accurate, applicant-level estimates of *d* will be larger than usually denoted in reviews of the literature.[3]

Further, given the existing (but somewhat minimal) literature that looks at constructs within methods, we expect that:

Expectation 2: The types of constructs measured by predictor methods will influence the magnitude of *d* (e.g., more cognitive saturation, higher *d*).

The specific expectation (regarding general cognitive ability/g-saturation) is considered and supported later—that is, via analyses showing substantial positive correlations between g-saturation and levels of *d* (see also similar conclusions in Sackett et al.'s 2001 review of high stakes testing).

Overall, we anticipate that our applicant-level results when assessing Expectation 1 will be surprising to some individuals, in that alternative tests will often have substantially larger *d*s when considered at the applicant level. We also believe that Expectation 2 is important because, as noted above, the analysis of "methods" continues to permeate research and practice.

Thus, this paper contributes to the literature in many ways. After cumulating and categorizing a variety of meta-analyses, we show that many estimates of Black–White *d*s are faulty—because they reflect incumbent levels of *d* rather than the more realistic (and larger) values of *d* to be expected at the operational point of selection. Thus, researchers are using inaccurate values in their simulations, and practitioners may be expending resources on solutions that do not decrease adverse impact. Given that reviews of the literature and simulations continue to use values of *d* that are

[3] Our purpose is not to explain causal reasons behind any Black–White differences (see Ryan, 2001, or Outtz & Newman, 2010 for such efforts) but rather to make progress in the literature and help applied researchers by focusing on more accurate and realistic values of *d* for potential predictors of job performance.

not accurate (e.g., DeCorte, Lievens, & Sackett, 2007; Ployhart & Holtz, 2008), our contribution is also to raise awareness about these concerns, as well as to provide more accurate levels of differences for use in research and the practice of test development We further show, sometimes in a preliminary manner, patterns in the literature that relate to the influence of construct saturation on levels of *d*. This will help researchers and practitioners better design selection inventories with lower adverse impact. Because tests of cognitive ability are often the comparison predictor of interest, we begin our analysis there.

Overall Cognitive Ability Tests

Value in the literature: d = *1.00.* Prevailing statements in the academic literature and professional practice are that the Black–White *d* for cognitive ability is about 1.00. (Unless specified otherwise, our use of the phrase "cognitive ability" means tests of general cognitive ability and not tests that only focus on a specific facet of cognitive ability, such as spatial, numerical, etc.) A few lower values of *d* appear. Murphy, Cronin, and Tam (2003) indicate *d* is often between .80 and 1.00 (see p. 669), and Schmitt, Claus, and Pulakos (1996) report an overall *d* of .83 for cognitive ability (although their analysis may mix incumbent and applicant samples). However, the Black–White value of $d = 1.00$ is stated in the well-regarded review by Hunter and Hunter (1984) and has been repeated many times (e.g., Hough & Oswald's, 2000, review; Sackett, Borneman, & Connelly, 2008; Sackett & Wilk, 1994). In addition, Ployhart, and Holtz's (2008) review, citing Roth, BeVier, Bobko, Switzer, and Tyler (2001) states that $d = .99$. Thus, when *d*s for other predictor tests are considered, they are often compared to the referent value of 1.00. We also note that reported values of *d* are sometimes lower than 1.00 for specific subfacets of cognitive ability (cf. Hough et al., 2001, or Outtz & Newman, 2010).

More accurate d*s are .72 and .86.* Roth et al. (2001) provide a meta-analysis of the literature on Black–White *d*. Across all jobs and applicants, they find that *d* is about 1.00. However, those authors note that it is important to compute *d* using a *within*-job framework (i.e., across applicants applying to the same jobs or jobs of similar levels of complexity). Indeed, Sackett and Wilk (1994) provide an explanation—involving self-selection of applicants into particular career paths—regarding why *d* might be less when considered within job (see also Wilk, Desmarais, & Sackett, 1995).

Accounting for the within-job issue, Roth et al. report that the Black–White cognitive ability, applicant-level *d* is .72 for jobs of moderate complexity (e.g., first-level supervisor or skilled crafts job) and *d* is .86 for jobs of low complexity. Thus, if one is considering a predictor for a moderately

complex job that could be used in place of overall cognitive ability, then the comparison *d* value is appropriately .72 and not 1.00 because .72 is what the particular employer is likely to experience operationally when using a test of cognitive ability (see also Gatewood, Feild, & Barrick, 2011, p. 483).[4]

This is an important point. For example, Dean and Broach (2007) reported a concurrent $d = .78$ for a selection test for air traffic controllers. These authors indicated they were pleased that their value of *d* was less than the presumed value of 1.00. However, the value of .78 is as large (indeed, somewhat greater) than would be expected from a test of cognitive ability for moderately complex jobs (and we note further that the Dean and Broach value of .78 was computed on incumbents). Similarly faulty comparisons have been used in employment interview analyses (e.g., Huffcutt & Roth, 1998) or, as noted earlier, in simulations and overall reviews of the selection test literature.

In sum, in Table 1, which summarizes potentially confounded values of *d*, as well as our corrected/updated values, we note that there is often an expected operational *d* of 1.00 for cognitive ability tests yet a more realistic value is .72 (for moderate complexity jobs) or .86 (for low complexity jobs). Interestingly, Roth et al. (2001) also report a within-job meta-analytic value of $d = .38$ for studies based on incumbent samples, and this lower value is consistent with our expectation of underestimation of *d* in such types of samples. In sum, in Table 2, which provides applicant level *d*s by construct categories, we place the values of .72 and .86 for overall cognitive ability.

Work Samples

Potentially confounded value in the literature: d = *.38 and "low".* Most literature summarizing work sample tests implies that the Black–White *d* is low. The specific value of $d = .38$ permeates much of the current literature (e.g., reviews by Hough et al., 2001, or Salgado et al., 2001). The value of .38 comes from a study by Schmitt et al. (1996) who report their meta-analytic value from data across three journals. Indeed, their paper was motivated by an attempt to examine levels of *d* by type of construct (within method). Further, as those authors note (p. 119), it was often difficult to distinguish whether a test was a work sample or a job knowledge test. They explicitly noted that they combined data from

[4] Looking ahead, and as noted by a reviewer, the cognitive ability $d = 1.00$ is the only value that is reduced in Table 1. This is because the value was at the applicant level but was not within job (or within job type). Other values in Table 1 are generally within job (or within job type).

work sample tests, situational judgment tests (SJTs), and job knowledge tests. Thus, it is unclear if the value of .38 best represents the *d* for work samples. Further, as noted by others (e.g., Bobko, Roth, & Buster, 2005), the vast majority of studies available for Schmitt et al.'s analysis were incumbent studies or otherwise restricted studies because that is what was available in those journals.

Other reviews state that a work sample test "does reduce adverse impact" (Reilly & Warech, 1993, p. 141), that the Black–White *d* for work samples is smaller than the *d* for cognitive ability (Ployhart, 2006; Reilly, 1996), that the work sample *d* is about one-third of a standard deviation (i.e., about .33; Ployhart, Schneider, & Schmitt, 2006), and that the work sample *d* is "low" (Heneman & Judge, 2006; Outtz, 1998; Pulakos, 2005, p. 17 and 31, who defines low as between .10 and .30). In sum, in academic journal summaries, books, and practitioner manuals, the current expectation of a low Black–White *d* for work samples is quite common. We use the value of .38 and the word "low" in Table 1.

A somewhat updated value of *d* (of .52) has been noted by Ployhart and Holtz (2008). Although their value of *d* was taken from an analysis of job incumbents (i.e., Roth, Huffcutt, & Bobko, 2003), this value does separate work sample estimates from other methods (such as job knowledge tests). Regarding constructs, we note two review articles (in addition to work by Roth et al. noted later) that reported meta-analytic *d*s and mentioned specifically that work samples are a method and can thus target a variety of constructs (Hough et al., 2001; Ployhart & Holtz, 2008).

More accurate overall point estimate of d *is .73.* Bobko et al. (2005) partially addressed applicant level issues by reporting two applicant-level *d*s for work sample exams that were in the .70s. Roth et al. (2008) subsequently conducted an updated meta-analysis of the literature on work-sample exams used as predictors. Those authors included only work sample tests, and they also looked for studies that estimated Black–White *d*s at the applicant, versus incumbent, level. For concurrent/incumbent studies, they found that the *d* was .36 (based on $k = 19$ studies). However, for applicant studies, the *d* was .73 (based on $k = 21$ such studies), and the *d* was .67 when one influential applicant study was removed.

Thus, the overall expected level of adverse impact for work sample exams (i.e., $d = .73$) is almost twice as large as prior literature claims and expectations (literature based on range-restricted incumbents). Therefore, in Table 1, we note that although the field often expects an operational *d* of about .38, a more realistic overall value is .73 (constructs notwithstanding, as noted next). Note also that the value of $d = .73$ for work samples is highly similar to the value of $d = .72$ for tests of overall cognitive ability (estimated from moderate complexity jobs).

Regarding constructs, Roth et al. (2008) reported that *d*s for work sample exams vary, as expected, according to the central set of KSAs targeted by the work sample test. Their mean observed applicant Black–White *d* was .80 for work samples targeting cognitive processes and job knowledge, .27 for work samples targeting oral communication and interpersonal skills, .27 for work samples targeting leadership and persuasion, and .70 for work samples targeting written skills. We note these construct-specific *d*s in Table 2.

Also regarding constructs, some authors (e.g., Pulakos, Schmitt, & Chan, 1996; Schmitt & Mills, 2001) have suggested that *d*s for work sample exams might be lowered by using video-based presentations and simulations. However, as noted later, these lower *d*s can often be attributed to potential differences in the constructs being measured by these methods or to differential reliability (Sackett et al., 2001, or Schmitt & Quinn, 2010). In general, we note that more specific work on constructs and construct equivalence of alternative predictor methods (e.g., work sample exams) is needed.

Situational Judgment Tests (SJTs)

Potentially confounded values in the literature: Range of .30 to .60, commonly d = *.38; "moderate," or "relatively low."* The literature on SJTs is somewhat consistent in its claims and expectations. For example, SJT Black–White *d*s are expected to be small to moderate (Ployhart, 2006) and are stated to be "much smaller" than *d*s for cognitive ability (Hanson & Ramos, 1996, p. 123). It has been stated that SJTs will have less adverse impact than tests of cognitive ability (Clevenger, Pereira, Wiechman, Schmitt, & Harvey, 2001), are "promising ways to maintain validity and decrease adverse impact" (Lundquist, 2007, p. 1), and have "low or no adverse impact as compared to other types of written tests" (Joiner, 2007).

More specific point estimates of SJT Black–White *d*s include .38 from Whetzel, et al.'s (2008) meta-analysis, .43 for video-based tests and .61 for paper-and-pencil tests (Hough et al., 2001), .31 for video-based tests and .40 for written tests (Ployhart & Holtz, 2008), a range of values between .40 and .49 (Weekley, et al., 2004, who also cite a range of nonapplicant values between .52 and .85 from the prior literature), a value of .50 (Schmitt & Chan, 2006), and the value of .38 noted earlier (i.e., from Schmitt et al.'s, 1996 grouping of work samples, SJTs, and job knowledge tests). Thus, in Table 1 (and constructs notwithstanding), we enter the range of .30–.60 but also indicate that .38 is a commonly noted value.

The above literature is replete with data from incumbents rather than applicants, and several researchers have noted this issue. For example, although they state that SJTs have smaller *d*s, Clevenger et al. (2001),

Lievens et al. (2005), and Whetzel et al. (2008) all candidly note that most, if not all, data in their studies (including simulations and meta-analyses) were from incumbent samples.

More accurate d *is unclear, though likely larger than currently expected.* Because of dependence on incumbent samples, the Black–White *d* for SJTs is likely larger than the above values. A study by Weekley et al. (2004) is one of a few studies that does report both applicant and incumbent results.[5] Converting their point-biserial *r*s to *d*s results in an incumbent $d = .39$ and an applicant $d = .76$—nearly doubling the estimated effect size.

Although not a meta-analysis, Roth et al. (2008) reported SJT *d*s for six different jobs at the applicant level. For two managerial jobs (e.g., transportation administrator), the *d*s were 1.01 and 1.04. The other four jobs (e.g., accountant, department procurement officer) were associated with *d*s of .09, .25, .62, and .67. Computing the weighted average of these six *d*s results in a value of .38, and the median is .67. These applicant-level values, which depend on the type of job (and presumably constructs), are noted in Table 1.

Regarding constructs, the breadth of KSAs targeted by SJTs is illustrated by noting that 16 constructs are assessed in the previously mentioned Clevenger et al. (2001) article. More generally, as McDaniel, Morgeson, et al. (2001) note, SJTs are "measurement methods that may assess a variety of constructs" (p. 732). And, as noted by a reviewer, there is also potential variation in the content of the stimuli and response formats (cf. McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006).

Additional recent work on the role of constructs is encouraging. Although exclusively focused on validity, Christian, Edwards, and Bradley (2010) recently analyzed SJTs in regard to construct saturation. Using a taxonomy of constructs from Huffcutt et al.'s (2001) analysis of interviews, Christian et al. found that many SJTs were linked to leadership skills, interpersonal skills, or personality; they also found that SJTs often involved a heterogeneous composite of KSAs.

In regard to constructs and levels of adverse impact, McDaniel and colleagues (McDaniel, Hartman, Whetzel, & Grubb, 2007; Whetzel et al., 2008) investigated the link between SJT scores, cognitive ability, and personality constructs. Their work demonstrates that cognitive saturation of the SJT influences the value of Black–White *d* ($r = .77$, across 60+ samples and 40,000+ participants).

[5] We urge caution in interpreting the Weekley et al. study. First, their database is both cross-job and cross-organization in nature. Second, their two ethnic groups are White and non-White, rather than White and Black. Third, specific *d*s in a subsequent table in their study appear to contradict earlier values of *d*.

The Roth et al. (2008) study also investigated sources of construct variation in their applicant-level *d*s. Using Huffcutt et al.'s taxonomy, they reported that Black–White *d*s were 1.01 and 1.04 for the two SJTs focusing on leadership constructs, .62 and .67 for the two SJTs focusing on cognitive ability/job knowledge and .09 and .25 for the two SJTs focusing on interpersonal skills. Regarding their two largest *d*s, they suggested that upper-level leadership skills might involve metacognitive abilities. Within their three categories, we averaged their *d*s by sample sizes and placed these applicant-level, construct values in Table 2, although we note that future research might be useful to confirm these preliminary conference paper results.

Thus, in Table 1, we note that a singular estimate of applicant-level *d* for SJTs is unclear, although likely larger than the commonly reported value of .38. Regarding constructs, we note that SJTs can potentially assess a large array of KSAs, that different constructs will likely be associated with different levels of *d* (values ranging from .19 to 1.02 are reported in both Table 1 and Table 2), and that cognitive saturation increases SJT *d*.)

Biodata

Potentially confounded value in the literature: d *is relatively low, if not small.* Reilly and Warech (1993) stated that biodata may have less adverse impact than cognitive tests, Schmitt and Pulakos (1998) noted that biodata may be popular because of "evidence that biodata exhibit little or no subgroup differences" (p. 167), and Hough et al. (2001) stated that *d*s for biodata involve "small differences" (p. 167).

Recent point estimates of *d* from the literature note a meta-analytic value of .33 (DeCorte, Lievens, & Sackett, 2007, Outtz, 2002; Ployhart & Holtz, 2008, and Salgado et al., 2001). All of these studies cite Bobko, Roth, and Potosky (1999), who pointed out that their values were not corrected for range restriction and that estimates in their article would likely be negatively biased.

One primary study deserves mention because it is highly cited and is influential in the above numerical values—Gandy, Dye, and MacLane's (1994) development and analysis of a federal government biodata form (Individual Achievement Record). Their well-known form and large-scale analysis ($n = 5{,}868$ across jobs), reports a Black–White *d* of .35 (or .27 after focused item deletion). However, all participants were job incumbents. Indeed, most of the analyses referenced above are based on incumbent studies (as noted by some of those authors). For example, Reilly and Warech (1993, p. 145) noted that the *d*s in their review were concurrent and therefore difficult to generalize to an applicant population. (Reilly,

1996, restated the idea that biodata has less adverse impact than cognitive tests but did not restate that most samples are concurrent.)

Regarding constructs, biodata items/scales are a method of assessing a variety of constructs. For example, Gandy et al.'s (1994) form assesses the four underlying factors of (a) work competency, (b) high school achievement, (c) college achievement, and (d) leadership skills. Unfortunately, Black–White subgroup *d*s are not provided separately for these factor scores, although consistent with the work of Goldstein or Whetzel and their colleagues, it might be expected that *d*s are larger for biodata components that are more cognitively saturated. We encourage such work and note that the U.S. military has been interested in the construct saturation of biodata forms (Kilcullen, White, Mumford, & Mack, 1995).

More accurate d *is unclear, though likely larger than currently expected.* There is little published literature on Black–White biodata *d*s at the applicant level. It is also difficult to find such data or find organizations that will share such data (see Potosky, Bobko, & Roth, 2005, for a similar sentiment), and articles reporting biodata *d*s often use incumbent samples (e.g., Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004).

Potosky et al. (2005) note two biodata studies for which applicant level values exist. In one study, Dean (1999) administered a biodata form (based on Gandy et al.'s, 1994, work) to a large sample of air traffic controller trainees ($n = 3{,}599$). When d was computed on the controller trainees (i.e., individuals previously selected using a test saturated by cognitive ability), the value was .34—which is reminiscent of the above concurrent values. When d was corrected for known levels of indirect range restriction, the estimated applicant value of d was .73. Note that this larger value of d is similar in magnitude to the Black–White d for cognitive ability tests for medium complexity jobs.[6]

The other study noted by Potosky et al. (2005) is from Kriska (2001), who reports a Black–White biodata $d = .27$ ($n = 1876$) when assessing police officer candidates. Potosky et al.'s sample-weighted average of these two *d*s was .57. More recently, Becton, Matthews, Hartley, and Whitaker (2009) provided an overall, Black–White applicant-level $d = .31$ ($n = 13{,}301$) for a biodata measure used for several different types of jobs in hospitals (e.g., laundry worker, nurse, manager).[7]

[6] Again, although it is outside the scope of this article to explain causal reasons behind ethnic group differences, we refer the reader to Schmitt et al. (2003) who review that topic and note the potential for nonzero d due to cultural differences in approaches to life experiences.

[7] Becton et al. also report cross-job biodata *d*s for each of three subscales/constructs. In particular, *d*s are .27, .26, and .36 for commitment, performance, and retention prediction subscales, respectively. Because these subscales were criterion focused, and not predictor construct focused, we do not incorporate them in Table 2.

In sum, in Table 1 we note that the field's expectations for Black–White *d*s are relatively low (.20 to .35). Given general reliance on incumbent samples, we expect that operational values of *d* could be larger. In fact, the values in the three applicant-level studies noted above (.73, .27, .31) are sometimes inside, and sometimes outside of, that range (particularly in the cognitively complex job of air traffic controller.) A sample-weighted combination of the three applicant-level values results in an overall $d = .39$. We record a potentially more realistic value of .39 in Table 1, although we caution the reader that this value is based on only three studies, and one of them is cross-job. We also call for more research that reports applicant-level biodata *d*s by targeted constructs.

Assessment Centers

Potentially confounded value in the literature: d *is relatively low, if not small.* The literature on assessment centers (ACs) has been optimistic about levels of Black–White *d*s. For example, Ployhart's (2006) review states that *d*s are expected by many in the field to be "generally lower" (p. 881). Earlier summaries state that ACs have "less adverse impact than tests" (Reilly & Warech, 1993, p. 156) and that AC *d*s are "low" (Outtz, 1998, p. 55). In addition, Hoffman and Thornton (1997, p. 166) state that half of the AC studies they reviewed had nonsignificant levels of Black–White *d*.

Point estimates of *d* from literature summaries include the range .03 to .40 (Hough & Oswald, 2000; citing Goldstein, Reilly, & Yusko, 1999), as well as values of .60 or less (Ployhart & Holtz, 2008). Once again, the point estimates of *d* are based on analyses that include incumbent, range-restricted samples.

More accurate d *is about .56.* Dean, Roth, and Bobko (2008) conducted a meta-analysis in which they added new data to the literature. They also coded available AC studies as being based on incumbents versus applicants (applicants were potentially incumbents in the organization, but they were applicants for the upper level positions associated with the ACs). They found an incumbent $d = .32$ (based on $k = 6$ studies) and an applicant $d = .56$ (based on $k = 10$ studies). The increase in *d* for applicant studies is consistent with our expectations. In Table 1, we thus note that the field's assumptions that Black–White *d*s are relatively low, and we record a potentially more realistic, operational (applicant level) value of .56, which interestingly is close to the maximum value noted in prior reviews.

Construct issues also clearly need to be investigated because assessment centers can simultaneously focus on a variety of KSAs (cognitive, oral communication, etc.). However, assessment center articles do not

generally report results for ethnic group *d*s by construct (Dean et al., 2008). The literature on AC *validity* might provide useful frameworks for the types of categories that are most theoretically relevant (e.g., Arthur, Day, McNelly, & Edens, 2003; Hoffman & Woehr, 2009). Goldstein, Yusko, Braverman, & Smith, (1998) work—showing that cognitive saturation influences assessment centers *d*s—is a good start in this area. In support of this notion, a 1999 conference presentation by Goldstein, Riley, and Yusko (as reported by Hough et al., 2001) indicates that $d = .28$ for AC exercises requiring interpersonal skills and $d = .60$ for exercises requiring cognitive abilities. This primary study used a sample of about 300 job applicants for promotion within a public service position (Goldstein, personal communication, February 2011), so we incorporate these two values in Table 2, although we note they are the only values in Table 2 from single samples.

Personality

Personality tests are often used as examples of possible ways to reduce levels of adverse impact—by incorporating such "noncognitive" tests in selection composites. On the other hand, recent summaries question the utility of using personality tests in operational settings (cf. Morgeson, et al., 2007). Concerns include faking,[8] low validity, and the self-report nature of most instruments. In regard to the current review, a useful aspect to the personality literature is that most of the work is, by definition, focused on constructs. We consider two dimensions of personality that are often presumed to have among the highest levels of selection validity within the personality domain: measures of "conscientiousness" and "integrity."

Potentially confounded value in the literature: d *is small and relatively close to zero for both conscientiousness and integrity*. Regarding the general category of personality tests, Reilly and Warech (1993) state that personality researchers "consistently find no adverse impact for minorities" (p. 186). Schmitt et al.'s (1996) meta-analysis reports a Black–White *d* for "personality" to be .09. Hough et al.'s (2001) meta-analysis finds "minimal differences" (p. 161) between Blacks and Whites on an array of personality factors and their subfacets (the *d*s range from .12 to −.31). Textbooks also note that *d*s for personality tests are low or near zero (e.g., Gatewood et al., 2011; Heneman & Judge, 2006).

For the more specific dimension of conscientiousness, Hough et al. (2001) report $d = .06$. Foldes, Duehr, and Ones (2008) conduct what they

[8] Indeed, a narrative review of personality testing (Hough & Oswald, 2008) "makes no attempt to resolve the insidious problem that job applicants might distort their responses—even lie" (p. 272).

regard to be a more extensive meta-analysis than Hough et al. and report $d = -.07$ when combining results for measures of conscientiousness and its subfacets. That is, the value of d is relatively small and, if anything, Blacks score higher on average than Whites.

For the specific dimension of integrity, Ones and Viswesvaran (1998) conducted a meta-analysis across four large-scale data sets and reported $d = .04$.

More accurate d *may be similar for integrity, but larger in magnitude for conscientiousness.* Weekley et al. (2004) and Oh, Wang, and Mount (2011) note that many of the primary studies in the personality literature are concurrent, rather than predictive, in nature. Indeed, in regard to the Foldes et al. (2008) meta-analysis discussed above, those authors indicate that over 70% of their organizational samples were for incumbents, so the estimate of the conscientiousness d may be influenced by range restriction. On the other hand, Ones and Viswesvaran's (1998) specific analysis of integrity tests uses "applicants for a broad range of jobs" (p. 37), so the point estimate of $d = .04$ for integrity tests seems plausibly useful. We report that value in Tables 1 and 2, although with the caveat that cross-job variance might have influenced this value. We also note that the quality of studies that are used in meta-analyses of the validity of integrity tests has been questioned (cf. Morgeson et al., 2007, or Van Iddekinge, Roth, Raymark, & Odle-Dusseau, 2012). We encourage the field to resolve this issue and consider how this resolution might, in turn, influence estimates of d.

Regarding a more accurate d for conscientiousness tests, we note that Hough (1998) reports a .07 change in *validity* when comparing concurrent studies to predictive studies yet provides no such methodological comparison for ds. If one assumes that the (corrected) validity for conscientiousness measures is about .20 (Hurtz & Donovan, 2000), then a change of .07 represents about a 33% change in validity estimates. In turn, applicant level ds for conscientiousness might also be somewhat higher than usually reported. For example, assuming 80% majority, a d of $-.07$ leads to a point-biserial r of $-.028$ between test scores and race. Increasing the magnitude of this correlation by 33% and then back-translating the resulting value of r gives $d = -.093$. Thus, in Table 1, we report $d = -.09$ for self-report measures of conscientiousness. This value is still relatively small in magnitude (and negative) compared to values of d for other selection devices.

We note that many, if not most, of the assessments of personality constructs are self-report. There has been much written about how self-report assessments may (or may not) influence levels of validity (e.g., Morgeson et al., 2007; Ones, Viswesvaran, & Schmidt, 1993), but we are not aware of systematic research about how self-report may influence

subgroup *d*s. In this regard, we later discuss work by Huffcutt et al. (2001) in the interviewing literature. Those authors report that interview subscores that target conscientiousness are associated with $d = .30$ (and even this value is generally based upon incumbent/restricted samples). Thus, in Table 1, we also note such a possible value of $d = .30$ for conscientiousness (i.e., in assessments from employment interviews).

Using the concept of conditional reasoning tests, James and his colleagues provide another notable exception to the use of self-report assessments in personality. For the construct of "aggression," James et al. (2005) report Black–White differences for a sample of undergraduates and a sample of temporary employees (i.e., none of their samples are for job applicants). Converting their reported *r*s to *d*s results in values of $d = .20$ and $d = .18$.

Structured Interviews

Potentially confounded value in the literature: d *of about .25.* For interviews, literature up to the late 1990s generally suggested fairly low levels of *d*. For example, Schmitt et al. (1996) reported $d = .15$ (based on six studies) and Reilly and Warech (1993, p. 161) stated that the "limited data" indicated that interviews have "probably less adverse impact than cognitive ability tests" (see also Reilly, 1996). Other authors also suggest that adverse impact potential for structured interviews is low (Outtz, 1998).

Regarding prior specific values, Huffcutt and Roth's (1998) meta-analysis reported that structured interviews were associated with a mean observed $d = .23$ (from 21 studies) whereas unstructured interviews had a $d = .32$ (10 studies). Structured interviews are also stated to have *d*s of one-third or one-fifth of Black–White differences on cognitive ability tests (Outtz, 2002), and interviews in general are stated to have a *d* of about .25 (Hough et al., 2001 citing Huffcutt and Roth, 1998; see also Ployhart & Holtz, 2008).

Regarding constructs, Huffcutt et al. (2001) investigated the types of constructs assessed in interviews. Although based on small numbers of samples (sometimes only 2 or 3 studies), they reported that mean levels of *d* varied as a function of the construct being assessed in the interview.[9] For example, across both low and high structure interviews, they report $d = .13$ for applied mental skills (problem solving), $d = .18$ for Extraversion, $d = .25$ for leadership skills, $d = .30$ for conscientiousness, $d = .39$ for communication skills, and $d = .49$ for cognitive ability.

[9] We note that Huffcutt et al. (2001) computed mean *d*s "by giving each study coefficient equal weight" (p. 902) rather than weighting by sample size.

More accurate overall d *is likely larger.* Many of the primary studies in the Huffcutt and Roth (1998) meta-analysis were based on incumbents, and/or the results were range-restricted by previous stages of selection (as noted also by Roth et al., 2002). Thus, Potosky et al. (2005) took the value of $d = .23$ from Huffcutt and Roth and corrected its magnitude for range restriction (based on a range restriction estimate from Huffcutt & Arthur, 1994). The corrected value of the meta-analytic d was .31. Roth et al. (2002) also reported that a behavior description interview was estimated to have a d of .46 at the applicant level. Thus, in Table 1, we enter a range of values between $d = .31$ and $d = .46$. More primary research is needed that looks at subgroup differences in interviews for applicants who have not been prescreened on other selection tests.

Regarding constructs, Huffcutt et al.'s (2001) findings suggest that, once again, constructs matter, and we indicate the potential influence of interview constructs in Table 1. We also note that the magnitudes of the estimated construct-level *d*s are likely influenced (in a downward manner) by range restriction due to use of incumbent samples and/or preselection on other tests, and they confound structured and unstructured interviews, so we do not incorporate them in Table 2.

Other Predictors

There are some commonly used predictors for which little is known regarding accurate levels of Black–White d. These predictors include academic grades, accomplishment records, trainability tests, assessments of job knowledge, minimum qualifications, and training and experience records. For space considerations we do not discuss them here, although the sparse evidence indicates *d*s are not negligible (e.g., .78 for grades, 1.07 for trainability tests, .33 for accomplishment records, and .50 for job knowledge—the latter two *d*s being on incumbents; Roth & Bobko, 2000, Roth, Buster, & Bobko, 2011, Hough, 1984, and Salgado et al., 2001, respectively.)

Discussion

This paper addresses issues that are fundamental to personnel selection. We focused on the use of more accurate estimates for Black–White d in selection tests—in order to maximally inform researchers and help guide organizational decision makers. We first considered issues based on the analysis of applicant versus incumbent samples. Although more studies using applicant-level values of Black–White d are needed in the literature (see such a call later), the pattern of results for available applicant-level datasets is generally clear. That is, from a synthesis of major meta-analyses

and prior reviews across many types of tests, it was concluded that many in the field may be underestimating the extent of potential for adverse impact. We also considered issues related to targeted constructs, such as cognitive ability or personality. We noted a pattern suggesting that tests targeting cognitive abilities increase subgroup differences, whereas tests targeting social skills might reduce subgroup differences (although more research is needed). We summarize these issues, and their implications, in more detail later.

Values of Applicant Versus Incumbent d

The applicant-level values presented in Table 1 indicate that Black–White *d*s (and the potential for adverse impact) are often substantially larger than currently presumed by researchers and/or practitioners due to the confounding artifact of range restriction. Thus, our first expectation was generally supported.

For applicant samples, some types of predictors have levels of *d* that are similar to the *d* associated with the usual comparison predictor in the literature (i.e., paper-and-pencil tests of cognitive ability). For example, for jobs that are of moderate complexity, the *d* for cognitive ability tests has a mean estimated value of about .72 (see Table 1). By comparison, overall work sample test scores are expected to have somewhat similar, or larger, values of *d*. Assessment center scores, SJTs, and possibly some biodata inventories are also associated with levels of *d* that may approach the cognitive ability comparison values.

Practical Implications (Including Adverse Impact)

More realistic expectations. As noted, the findings in Table 1 indicate that, in operational practice, many types of predictors might be associated with substantial levels of *d* (and substantial adverse impact). This stands in contrast to many summary claims in the literature. The applicant-level values in Table 1 are helpful to practitioners and organizational decision makers because they are assembled at a relatively common point in selection systems. These estimates thus provide more realistic expectations about the potential for adverse impact when considering/comparing alternative predictors (e.g., alternatives as a replacement for, or in combination with, existing predictors).

In particular, we have seen organizations spend tens (and hundreds) of thousands of dollars on "alternative predictor" development. The resources were likely spent with good intentions because the values of *d* in the literature were low (as they were based on incumbent samples). After implementing the new selection system, the level of *d* (and adverse

impact) was quite large—likely because the new system was operationally being administered to applicants not incumbents. The values of *d* in the third column of Table 1 may help avoid such nonpositive "selection system surprises." In addition, values of *d* are often used in research simulations and analyses of predictor matrices (e.g., DeCorte et al., 2007; Finch et al., 2009). It would be important to use appropriate values in these simulations so that conclusions are not biased towards solutions associated with incumbent level statistics.

More unanticipated adverse impact. We further note that even moderate levels of *d* can often cause adverse impact to occur in practice (cf. Sackett & Ellingson, 1997, table 2). For example, Potosky et al. (2005) demonstrated that the four-fifths rule will be violated when comparing Blacks to Whites for almost any composite set of selection devices they analyzed, unless selection rates were over 90%. In their focus on SJTs, and using the value of $d = .38$, Whetzel et al. (2008) found similar outcomes. Also, Foldes et al. (2008) conclude that "about half" (p. 607) of their racial group comparisons across several personality constructs could lead to findings of adverse impact. Incorporation of the appropriately larger levels of Black–White *d* at the applicant level for many predictors (see Table 1) could add to the pessimistic nature of these demonstrations.

Feasibility. The feasibility of predictors in Table 1 should also be considered. For example, although interviews are generally associated with somewhat lower Black–White *d*s (constructs assessed notwithstanding), it might be costly and impractical to use interviews as an initial selection hurdle if large numbers of applicants are expected. Similar comments might apply to assessment centers and work samples. (We also note that the applicant values of *d* presented in Table 1 are appropriate for their intended use; for example, the assessment center values were based on applicants for, not incumbents of, the higher level job even if the applicants were incumbents of the organization.)

Predictor composites. Our analysis was focused on predictors that are discussed in the literature as alternatives to paper-and-pencil tests of cognitive ability. However, rather than replacing a cognitive ability test with a different test, an organization might add several alternative predictors to their selection process—either sequentially or in combination (cf. DeCorte et al., 2007; Sackett & Roth, 1996).

It is often suggested that adding "low-*d*" predictors, such as biodata scores or personality construct scores, to more "traditional" test scores (e.g., paper-and-pencil cognitive ability) is a useful notion, and that diversity–validity tradeoffs will be lessened (e.g., Ployhart & Holtz, 2008, p. 168). The suggestion to add such predictors may help content validity because a greater number of important KSAs are potentially being covered. However, such a suggestion is also made because of the belief

that overall levels of *d* will go down—yet there are concerns with this assumption.

First, even if the composite *d* is reduced, the reduction might not be as much as expected. For example, if two uncorrelated predictors have $d = 1$ and $d = 0$, the *d* for an equally weighted composite is about .70, rather than the .50 one might intuitively expect (Sackett & Ellingson, 1997). In addition, Potosky et al. (2005) simulate a variety of selection composite scenarios and conclude that any reductions in adverse impact are "more modest" (p. 304) than might be thought. Use of the (generally larger) applicant level values in our Table 1, third column, could add substantial further caution to expectations of sizeable reductions in *d*.

Second, as also noted by Sackett and Ellingson (1997), some composites might unintentionally increase levels of *d* in operational settings. For example, using the applicant-level values in Table 1, and assuming a .30 correlation between work sample scores and a cognitive ability test, the *d* for the composite of these two tests will be larger than the *d* for either predictor alone (equations available upon request). Potosky et al. (2005) reported similar findings. That is, the level of *d* for their regression-weighted composite of cognitive ability and biodata was larger than the *d* for either predictor alone (and their two *d* values were somewhat similar to those in our Table 1).

Constructs Matter

This phrase rings true and has been noted eloquently by some in the field. However, the moderating role of constructs remains largely ignored in research on subgroup *d*s for selection tests (with some empirical exceptions such as Goldstein et al., 2007, and McDaniel et al., 2007, or theoretical exceptions such as Arthur & Villado, 2008, as noted above).

When compiling this paper, we envisioned (and attempted to assemble) several types of multitrait (KSA)-multimethod tables/syntheses. However, the extant literature on construct-level *d*s at the common, applicant level of analysis is sparse. That is, the majority of cells in such method-by-construct tables contain no data. Despite this limitation, we used Huffcutt et al.'s (2001) construct categories and assembled Table 2, which begins to address this issue and could aid/alert decision makers when designing various selection systems. Because some constructs might have the same label, yet not be equivalent (in a measurement sense), we thus denoted the entries as "construct categories" rather than "constructs." Nonetheless, from the pattern of results in Table 2, it is apparent that cognitive ability is associated with relatively high *d*s across method (as hypothesized by us and others). Interpersonal and oral communication skills appear to be associated with relatively less potential adverse impact. It is also noted

that personality constructs are associated with relatively low *d*s (at least for self-report measures). Further, we note that leadership constructs are associated with a wide variation in Black–White *d*s (again, perhaps due to differences in the definition of "leadership"). These construct-specific findings are suggestive but tentative, and once again we call for future research on these construct-level patterns (see future research section later).

Thus, the applicant-level results support our second expectation; that is, the pattern supports the notion that magnitudes of *d* vary depending on what KSA (or set of KSAs) is being measured. As discussed by Sackett et al. (2001) or Schmitt and Quinn (2010), differences between test types (i.e., differences in methods such as video vs. paper-and-pencil tests) can often be attributed to differences in underlying constructs.

Cognitive Saturation

We have specifically noted that saturation of tests with cognitive ability generally increases levels of Black–White *d*. This triangulates with empirical evidence in SJT research (McDaniel et al., 2007; Whetzel et al., 2008) and in AC research (Goldstein et al., 1998, 2001; Hough & Oswald, 2000), as well as narrative conjectures in earlier reviews of high stakes testing (Sackett et al., 2001). Indeed, although not reported above, Roth et al. (2011) recently noted a corrected *r* of .80 between g-saturation and *d* for trainability tests.

"Method" Revisited

Coming full circle back to the concept of "method," we note that in many personnel selection situations it is the naturally occurring interaction of multiple constructs that is intentionally targeted and assessed by some methods rather than a singular construct. That is, a focus on a particular "method" (as in Table 1) might be (a) theoretically appropriate or (b) useful to practitioners. For example, many work sample tests involve an interplay between a variety of skills and abilities; and the interaction between these KSAs is what is being targeted/measured (e.g., as they occur on the job) not each KSA separately. Or, biodata items sometimes specifically measure past behavior (i.e., past behavior predicts future behavior). Given that behavior is multifaceted and determined, such biodata items are theoretically more "molar" than assessments of singular KSAs, and it may not be logical to attempt to disentangle each trait/interest/motive separately. Thus, the entries in Table 1 that are associated with testing "methods" have utility beyond the construct-specific values.

Future Research

A variety of research suggestions and needs result from our analysis and synthesis. For example, and perhaps most central to the current review, researchers who are interested in subgroup differences in selection tests should collect and analyze applicant-level data when feasible. Reported statistics would include means and standard deviations—for the overall sample and for each subgroup. These statistics could also be reported by construct (when feasible) within subgroup, and empirical correlations with any available marker tests could be noted. If incumbent-level data are gathered, then attempts should be made to determine the nature of any range restriction (direct/indirect), the degree of restriction (e.g., ratio of restricted to unrestricted standard deviations), and any other associated factors (e.g., the type of variable that caused the restriction, whether or not strict top-down selection was used, etc.). Similarly, if applicants in a study on a predictor of interest (e.g., structured interview) have been screened on another predictor (e.g., a cognitive ability test), the researchers should attempt to correct for subsequent range restriction that might have occurred on the predictor of interest.

Regarding other future research needs, we noted almost no applicant-level research for other selection devices not included in Table 1, such as training and experience evaluations, minimum qualifications, grades, social media information, and so forth. Given the ubiquity of such devices in selection screens, it is clear that research is needed in any of these domains.

As another example, leadership assessments were associated with substantial variation in *d*s. Research is needed to unpack these findings and to investigate if some of the variation is due to differences in the definitions of leadership or to other factors.

Research is also needed that investigates to what extent the lower levels of *d* for personality tests are associated with the usual self-report nature of such tests. For example, we earlier noted that larger personality *d*s were obtained from interviews. As another possibility, Oh et al. (2011) consider personality assessments based on reports of others, although there are no reported values of *d.* There is also research on the use of forced-choice responding to personality inventories (cf. McCloy, Heggestad, & Reeve, 2005), including the notion that such responding might decrease faking without decreasing validity. Unfortunately, this published literature does not report applicant level ethnic-group differences.[10] However, the literature does show that using forced-choice responding

[10] One military technical report (Knapp, Heggestad, & Young, 2004) did report differences, but they were based on incumbents.

increases the correlation of personality factors with cognitive ability (e.g., Christiansen, Burns, & Montgomery, 2005; Vasilopoulos, Cucina, Dyomina, Morewitz, & Reilly, 2006). Thus, forced choice methods might increase the magnitudes of *d* for personality scales. Interestingly, a reviewer also commented that another research domain might involve the extent to which faking by applicants restricted the range of scores and, hence, potentially reduced observed applicant-level values of *d*.

In addition, the role of differential unreliability across constructs (i.e., not just differential unreliability across methods) might matter. For example, Bobko et al. (2005) and Roth et al. (2008) noted that the low-*d* work sample exercises (e.g., the "human relations" score and other role-play scores) in their analyses were generally associated with lower interrater reliabilities (see also Sackett et al., 2001). Future applicant-level research might also account for differential reliability so as to ameliorate this potential confound.

As yet another need, note that we chose to look at Black–White differences because data were almost nonexistent for other ethnic groups (Foldes et al. is an exception in the personality literature). Given that Hispanics are now the largest minority group in the workforce (cf. Whetzel et al., 2008), we call for applicant level Hispanic–White research in just about all test domains. Such a call could readily be generalized to other ethnic groups (e.g., Asians) or to other national contexts (e.g., as noted by a reviewer, the Black–White parallel in France might be native French and North Africans).

Applicant-level research is needed on methods such as biodata inventories or structured interviews. These latter two methods are ubiquitous in selection (biodata as an early screen; interviews as a later hurdle) yet we know little about what constructs underlie any applicant-level adverse impact.

As noted earlier and above, research is generally needed that reports on applicant level *d*s for constructs within methods. Sometimes this parsing of constructs will be difficult. For example, a single work sample exercise might require technical expertise, creativity in coming up with a solution, and then written/oral skill in communicating that solution to the person scoring the test. Some researchers (e.g., McDaniel et al., 2007) have dealt with this issue by also giving construct-level marker tests to research participants, computing correlations between the marker tests and the selection device, and reporting levels of construct "saturation."

The potential confound between method and construct also leads us to suggest that future research efforts be mindful of the phrase "noncognitive predictors." That is, the phrase "noncognitive" has been applied to most selection tests that are not multiple choice, cognitive ability tests—yet such a phrase is a potential misnomer for several reasons. First, some

"noncognitive" tests incorporate multiple choice/option items (e.g., personality tests or SJTs). Second, many so-called alternative selection tests involve cognition. For example, SJTs have been labeled as noncognitive alternatives, yet it is difficult to imagine that "judgment" (i.e., the J in SJT) does not involve cognitive processing. Some work sample tests, assessment centers, or biodata experiences (e.g., academic achievement factors) are likely cognitively saturated. We suggest that selection devices be labeled with more clarity and that the phrase "noncognitive" be avoided.

Further, we reported on several methods in this paper (e.g., work samples, assessment centers, biodata), but the methods could be parsed further by presentation mode (e.g., paper, video, or oral stimuli). Other dimensions to "method" include scoring, timing, response mode, and so forth (for examples of studies on open-ended response mode, see Arthur, Edwards, & Barrett, 2002, or Edwards & Arthur, 2007).

More research might be conducted on some of the factors noted in the Appendix (factors which might influence both *d* and *r*). We hypothesized, and confirmed, that the influence of range restriction (and incumbent-based estimates) on *d* was potentially substantial. Regarding other factors in the Appendix, both Ployhart and Holtz (2008) and Schmitt and Quinn (2010) provide summaries of current knowledge about several of these factors. They note that the influence of some factors is likely small (e.g., coaching, availability of retesting) although other possible factors might be more promising ways to reduce adverse impact (e.g., alternative modes of presenting test stimuli). In this latter instance, we note that constructs might then change.

In sum, we suggested that the *d* values for many selection tests may be larger than researchers and practitioners have generally thought. Although applicant-level statistics are more difficult to find than incumbent-level statistics, we note the pattern is strong across the available data. We further noted that consideration of values of *d* at a common, applicant level will better inform researchers and decision makers about realistic reductions in levels of adverse impact for selection systems. We also reviewed the role of constructs (e.g., cognitive ability, social skills, etc.) on levels of *d*, confirmed again in the limited available research that constructs might matter, and delineated research that can increase our applied and theoretical understanding of subgroup differences in personnel selection. We hope that use of the more accurate values in Tables 1 and 2 in simulations, analyses, and organizational decisions assist in that process.

REFERENCES

Arthur W, Day E, McNelly T, Edens P. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. PERSONNEL PSYCHOLOGY, *56*, 125–154.

Arthur W, Edwards B, Barrett G. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil formats. PERSONNEL PSYCHOLOGY, *55*, 985–1008.

Arthur W, Villado A. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*, 435–442.

Barrett G. (2008). Practitioner's view of personality testing and industrial-organizational psychology: Practical and legal issues. *Industrial and Organizational Psychology*, *1*, 299–302.

Barrett G, Phillips J, Alexander R. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology*, *66*, 1–6.

Becton B, Matthews M, Hartley D, Whitaker D. (2009). Using biodata to predict turnover, organizational commitment, and job performance in healthcare. *International Journal of Selection and Assessment*, *17*, 189–202.

Berry C, Sackett P, Landers R. (2007). Revisiting interview–cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. PERSONNEL PSYCHOLOGY, *60*, 837–874.

Bliesener T. (1996). Methodological moderators in validating biographical data in personnel selection. *Journal of Occupational and Organizational Psychology*, *69*, 107–120.

Bobko P. (2001). *Correlation and regression: Principles and applications for industrial/organizational psychology and management* (2nd ed.). London: Sage.

Bobko P, Roth P, Bobko C. (2001). Correcting the effect size of *d* for range restriction and unreliability. *Organizational Research Methods*, *4*, 46–61.

Bobko P, Roth P, Buster M. (2005). Work sample selection tests and expected reduction in adverse impact: A cautionary note. *International Journal of Selection and Assessment*, *13*, 1–10.

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

Callinan M, Robertson I. (2000). Work sample testing. *International Journal of Selection and Assessment, 8,* 248–260.

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

Christian M, Edwards B, Bradley J. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. PERSONNEL PSYCHOLOGY, *63*, 83–117.

Christiansen N, Burns G, Montgomery G. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, *18*, 267–307.

Clevenger J, Pereira G, Wiechman E, Schmitt N, Harvey V. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology*, *86*, 410–417.

Dean M. (1999). *On biodata construct validity, criterion validity and adverse impact*. Unpublished doctoral dissertation, Louisiana State University.

Dean M, Broach D. (2007). *Multi-sample investigation of biodata validity and demographic group differences*. Paper presented at the 22nd Annual Conference of the Society for Industrial and Organizational Psychology, New York.

Dean M, Roth P, Bobko P. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology*, *93*, 685–691.

DeCorte W. (1999). Weighting job performance predictors to both maximize the quality of the selected workforce and control the level of adverse impact. *Journal of Applied Psychology*, *84*, 695–702.

DeCorte W, Lievens F, Sackett P. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology*, *92*, 1380–1393.

Edwards B, Arthur W. (2007). An examination of factors contributing to a reduction in subgroup differences on a construct-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology*, *92*, 794–801.

Finch D, Edwards B, Wallace C. (2009). Multistage selection strategies: Simulating the effects on adverse impact and expected performance for various predictor combinations. *Journal of Applied Psychology*, *94*, 318–340.

Foldes H, Duehr E, Ones D. (2008). Group differences in personality: Meta-analyses comparing five U.S. racial groups. PERSONNEL PSYCHOLOGY, *61*, 579–616.

Gandy J, Dye D, MacLane C. (1994). Federal government selection: The individual achievement record. In Stokes G, Mumford M (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 275–309). Palo Alto, CA: Consulting Psychologists Press.

Gatewood R, Feild H, Barrick M. (2011). *Human resource selection* (*7*th ed.). Mason, OH: South-Western.

Ghiselli E. (1964). *Theory of psychological measurement*. New York, NY: McGraw-Hill.

Ghiselli E. (1966). *The validity of occupational aptitude tests*. New York, NY: Wiley.

Ghiselli E. (1973). The validity of aptitude tests in personnel selection. PERSONNEL PSYCHOLOGY, *26*, 461–477.

Goldstein H, Riley Y, Yusko K. (1999). *Exploration of Black–White subgroup differences on interpersonal constructs.* Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Goldstein H, Yusko K, Braverman E, Smith D, Chung B. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. PERSONNEL PSYCHOLOGY, *51*, 357–374.

Goldstein H, Yusko K, Nicolopoulos V. (2001). Exploring Black–White subgroup differences of managerial competencies. PERSONNEL PSYCHOLOGY, *54*, 783–807.

Guion R, Cranny C. (1982). A note on concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology*, *67*, 239–244.

Gutenberg R, Arvey R, Osburn H, Jeanneret P. (1983). Moderating effects of decision-making/information-processing job dimensions on test validities. *Journal of Applied Psychology*, *68*, 602–608.

Hanson M, Ramos R. (1996). Situational judgment tests. In Barrett R (Ed.), *Fair employment strategies in human resource management* (pp. 119–124). Westport, CT: Quorum/Greenwood

Hattrup K, Rock J, Scalia C. (1997). The effects of varying conceptualizations of job performance on adverse impact, minority hiring, and predicted performance. *Journal of Applied Psychology*, *82*, 656–664.

Hausknecht J, Day P, Thomas S. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. PERSONNEL PSYCHOLOGY, *57*, 639–683.

Heneman H, Judge T. (2006). *Staffing organizations* (5th ed.). Boston, MA: Irwin/McGraw-Hill.

Hoffman B, Woehr D. (2009). Assessment center construct-related validity: Stepping beyond the MTMM matrix. *Journal of Vocational Behavior*, *75*, 173–182.

Hoffman C, Thornton G. (1997). Examining selection utility where competing predictors differ in adverse impact. PERSONNEL PSYCHOLOGY, *50*, 455–470.

Hough L. (1984). Development and evaluation of the accomplishment record method of selecting and promoting professionals. *Journal of Applied Psychology*, *69*, 135–146.

Hough L. (1998). Personality at work: Issues and evidence. In Hakel M (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 131–159). Mahwah, NJ: Erlbaum.

Hough L, Furnham A. (2003). Importance and use of personality variables in work settings. In Weiner I, Borman W, Ilgen D, Klimoski R (Eds.), *Handbook of Psychology*: Vol. 12. *Industrial and Organizational Psychology* (pp. 131–169). New York, NY: Wiley.

Hough L, Ones D. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In Anderson N, Ones D, Sinangil H, Viswesvaran C (Eds.), *Handbook of Industrial, Work, and Organizational Psychology*, (Vol. 1, pp. 233–277). London/New York, NY: Sage.

Hough L, Oswald F. (2000). Personnel selection: Looking toward the future—Remembering the past. *Annual Review of Psychology*, *51*, 631–664.

Hough L, Oswald F. (2008). Personality testing and industrial-organizational psychology: Reflections, progress, and prospects. *Industrial and Organizational Psychology*, *1*, 272–290.

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

Huffcutt A, Arthur W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, *79*, 184–190.

Huffcutt A, Conway J, Roth P, Klehe U. (2004). The impact of job complexity and study design on situational and behavior description interview validity. *International Journal of Selection and Assessment*, *12*, 262–273.

Huffcutt A, Conway J, Roth P, Stone N. (2001). Identification and meta-analysis of constructs measured in employment interviews. *Journal of Applied Psychology*, *86*, 897–913.

Huffcutt A, Roth P. (1998). Racial group differences in interview evaluations. *Journal of Applied Psychology*, *83*, 288–297.

Hunter J. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization to the General Aptitude Test Battery (GATB)*. USES Test Research Report No. 45. Washington, DC: US Department of Labor.

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

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

Hurtz G, Donovan J. (2000). Personality and job performance: The big five revisited. *Journal of Applied Psychology*, *85*, 869–879.

James L, McIntyre M, Glisson C, Green P, Patton T, LeBreton J, . . . Williams L. (2005). A conditional reasoning measure for aggression. *Organizational Research Methods*, *8*, 69–99.

Joiner DA. (2007). Why situational judgment tests have become so popular. *Assessment Council News*, p. 14.

Joshi A, Roh H. (2009). The role of context in work team diversity research: A meta-analytic review. *Academy of Management Journal*, *52*, 599–627.

Kantrowitz T, McClellan R, Borman W, Houston J, Schneider R. (2009). *Validation of computer adaptive personality scales for commercial use*. Paper presented at the 24th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Kilcullen R, White L, Mumford M, Mack H. (1995). Assessing the construct validity of rational biodata scales. *Military Psychology*, *7*, 17–28.

Knapp D, Heggestad E, Young M. (2004). *Understanding and improving the Assessment of Individual Motivation (AIM) in the Army's GED plus program*. Technical Report

2004-03, Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

Kriska S. (2001). *The validity-adverse impact trade-off: Real data and mathematical model estimates*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Lievens F, Buyse T, Sackett P. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology*, *90*, 442–452.

Lundquist K. (2007). *Statement of Kathleen K. Lundquist*. Employment and screening meeting held by the Equal Employment Opportunity Commission. Washington, DC, May 16, 2007.

McCloy R, Heggestad E, Reeve C. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods*, *8*, 222–248.

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

McDaniel M, Morgeson F, Finnegan E, Campion M, Braverman E. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, *86*, 730–740.

McDaniel R, Walls M. (1997). Diversity as a management strategy for organizations: A view through the lenses of chaos and quantum theories. *Journal of Management Inquiry*, *6*, 363–375.

McDaniel M, Whetzel D, Hartman N, Nguyen N, Grubb W. (2006). Situational judgment tests: Validity and an integrative model. In Ployhart R,Weekley J (Eds.), *Situational judgment tests: Theory, measurement, and application*, pp. 183–204, Mahwah, NJ: Jossey Bass.

Morgeson F, Campion M, Dipboye R, Hollenbeck J, Murphy K, Schmitt N. (2007). Reconsidering the use of personality tests in personnel selection contexts. PERSONNEL PSYCHOLOGY, *60*, 683–729.

Motowidlo S, Van Scotter J. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology*, *79*, 475–480.

Muchinsky P. (1996). The correction for attenuation. *Educational and Psychological Measurement*, *56*, 63–75.

Murphy K, Cronin B, Tam A. (2003). Controversy and consensus regarding the use of cognitive ability testing in organizations. *Journal of Applied Psychology*, *88*, 660–671.

Oh I, Wang G, Mount M. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology*, *96*, 762–773.

Ones D, Dilchert S, Viswesvaran C, Judge T. (2007). In support of personality assessment in organizational settings. PERSONNEL PSYCHOLOGY, *60*, 995–1027.

Ones D, Viswesvaran C (1998). Gender, age, and race differences on overt integrity tests: Results across four large-scale applicant data sets. *Journal of Applied Psychology*, *83*, 35–42.

Ones D, Viswesvaran C, Schmidt F. (1993). Comprehensive meta-analysis of integrity test validation: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*, 679–703.

Oswald F, Schmitt N, Kim B, Ramsay L, Gillespie M. (2004). Developing a biodata measure and situational judgment inventory as a predictors of college student performance. *Journal of Applied Psychology*, *89*, 187–207.

Outtz J. (1998). Testing medium, validity, and test performance. In Hakel M (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection*, pp. 41–57. Mahwah, NJ: Erlbaum.

Outtz J. (2002). The role of cognitive ability tests in employment selection. *Human Performance*, *15*, 161–171.

Outtz J,Newman D. (2010). A theory of adverse impact. In Outtz J (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection*. (pp. 53–94). New York, NY: Routledge/Taylor & Francis.

Ployhart R. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management*, *32*, 868–897.

Ployhart R, Holtz B. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. PERSONNEL PSYCHOLOGY, *61*, 153–172.

Ployhart R, Schneider B, Schmitt N. (2006). *Staffing organizations: Contemporary practice and research*. Mahwah, NJ: Erlbaum.

Potosky D, Bobko P, Roth P. (2005). Forming composites of cognitive ability and alternative measures to predict job performance and reduce adverse impact: Corrected estimates and realistic expectations. *International Journal of Selection and Assessment*, *13*, 304–315.

Pulakos E. (2005). *Selection assessment methods*. Alexandria, VA: SHRM Foundation.

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

Reilly R. (1996). Alternative selection procedures. In Barrett R (Ed.), *Fair employment strategies in human resource management*. Westport, CT: Quorum.

Reilly R, Warech M (1993). The validity and fairness of alternatives to cognitive ability tests. In Wing L, Gifford B (Eds.), *Policy issues in employment testing*. Boston, MA: Kluwer.

Ricci et al v. Destefano et al.; civil no. 3:04cv1109 (JBA, 2006).

Roth P, Bobko P. (2000). College grade point average as a personnel selection device: Ethnic group differences and potential adverse impact. *Journal of Applied Psychology*, *85*, 399–406.

Roth P, BeVier C, Bobko P, Switzer F 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, Bobko P, Buster M. (August, 2008). *Applicant Black–White differences on situational judgment tests: They may be surprising*. presented at the Annual Meeting of the Academy of Management, Anaheim, CA.

Roth P, Bobko P, McFarland L, Buster M. (2008). Work sample tests in personnel selection: A meta-analysis of Black–White differences in overall and exercise scores. PERSONNEL PSYCHOLOGY, *61*, 637–662.

Roth P, Bobko P, Switzer F. II, Dean M. (2001). Prior selection causes biased estimates of standardized ethnic group differences: Simulation and analysis. PERSONNEL PSYCHOLOGY, *54*, 591–617.

Roth P, Buster M, Bobko P. (2011). Updating the trainability tests literature on Black–White subgroup differences and reconsidering criterion-related validity. *Journal of Applied Psychology*, *96*, 34–45.

Roth P, Huffcutt A, Bobko, P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88,* 694–706.

Roth P, Van Iddekinge C, Huffcutt A, Eidson C, Bobko P. (2002). Correcting for range restriction in structured interview ethnic group differences: The values may be larger than we thought. *Journal of Applied Psychology*, *87*, 369–376.

Russell C, Settoon R, McGrath R, Blanton A, Kidwell R, Lohrke F, . . . Danforth G. (1994). Investigator characteristics as moderators of personnel selection research: A meta-analysis. *Journal of Applied Psychology*, *79*, 163–170.

Ryan A. (2001). Explaining the Black–White test score gap: The role of test perceptions. *Human Performance*, *14*, 45–75.

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

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

Sackett P, Lievens F. (2008). Personnel selection. *Annual Review of Psychology*, *59*, 419–450.

Sackett P, Roth L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. PERSONNEL PSYCHOLOGY, *49*, 1–18.

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

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

Sackett P, Yang H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*, 112–118.

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

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

Schmidt F, Hunter J, Urry V. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology*, *61*, 473–485.

Schmitt N, Chan D. (2006). Situational judgment tests: Method or construct. In Weekley J, Ployhart R (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 135–155). Mahwah, NJ: Erlbaum.

Schmitt N, Clause C, Pulakos E. (1996). Subgroup differences associated with different measures of some common job relevant constructs. In Cooper C, Robertson I (Eds.), *International review of industrial and organizational psychology*, (Vol. 11, pp. 115–139). New York: Wiley.

Schmitt N, Cortina J, Ingerick M, Wiechmann D. (2003). Personnel selection and employee performance. In Borman W, Ilgen D, Klimoski R (Eds.), *Handbook of psychology*: Vol. 12: *Industrial and organizational psychology* (pp. 77–105). Hoboken, NJ: Wiley.

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

Schmitt N, Mills A. (2001). Traditional tests and job simulations: Minority and majority performance and test validities. *Journal of Applied Psychology*, *86*, 451–458.

Schmitt N, Pulakos E. (1998). Biodata and differential prediction: Some reservations. In Hakel M (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection.* (pp. 167–182). Mahwah, NJ: Erlbaum.

Schmitt N, Quinn A. (2010). Reductions in measured subgroup mean differences: What is possible? In Outtz J (Ed.), *Adverse impact: Implications for organizational staffing*

and high stakes selection. (pp. 425–451). New York, NY: Routledge/Taylor & Francis.

Society for Industrial and Organizational Psychology, Inc. (2003), Principles for the validation and use of personnel selection procedures (4th Ed.). Bowling Green, OH: Author.

Stokes G, Hogan J, Snell A. (1993). Comparability of incumbent and applicant samples for the development of biodata keys: The influence of social desirability. PERSONNEL PSYCHOLOGY, *46*, 739–762.

Thorndike R. (1949). *Personnel selection: Test and measurement techniques*. New York, NY: Wiley.

U.S. Equal Opportunity Employment Commission, U.S. Civil Service Commission, U.S. Department of Labor, U.S. Department of Justice (1978). Uniform Guidelines on employee selection procedures. *Federal Register*, *43*, 38295–38309.

Van Iddekinge C, Morgeson F, Schleicher D, Campion M. (2011). Can I retake it? Exploring subgroup differences and criterion-related validity in promotion retesting. *Journal of Applied Psychology*, *96*, 941–955.

Van Iddekinge C, Putka D, Campbell J. (2011). Reconsidering vocational interests for personnel selection: The validity of an interest-based selection test in relation to job knowledge, job performance, and continuance intentions. *Journal of Applied Psychology*, *96*, 13–33.

Van Iddekinge C, Roth P, Raymark P, Odle-Dusseau H. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology*, *97*, 499–530.

Vasilopoulos N, Cucina J, Dyomina N, Morewitz C, Reilly R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance*, *19*, 175–199.

Vasilopoulos N, Cucina J, McElreath J. (2005). Do warnings of response verification moderate the relationship between personality and cognitive ability? *Journal of Applied Psychology*, *90*, 306–322.

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

Whetzel D, McDaniel M, Nguyen N. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*, 291–309.

Wilk S, Desmarais L, Sackett P. (1995). Gravitation to jobs commensurate with ability: Longitudinal and cross-sectional tests. *Journal of Applied Psychology*, *80*, 79–85.

APPENDIX
Moderators That Can Potentially Influence the Magnitude of r

Moderators that might also influence *d*

a. *Range restriction* (e.g., Bobko, 2001; Ghiselli, 1964; Sackett & Yang, 2000; Thorndike, 1949)
b. *Measurement unreliability* (see Muchinsky, 1996 for a review, or Bobko et al., 2001, for a discussion of correcting *d* for unreliability)
c. *Complexity* level of the job (e.g., Gutenberg, Arvey, Osburn, & Jeanneret, 1983; Huffcutt, Conway, Roth, & Klehe, 2004; Hunter, 1983; Schmitt, Cortina, Ingerick, & Wiechmann, 2003)
d. *Faking* (e.g., Hough & Ones, 2001; McDaniel, Hartman, & Whetzel, unpublished data; Stokes, Hogan, & Snell, 1993)
e. *Threat/warning* (e.g., Vasilopoulos, Cucina, & McElreath, 2005)
f. Test *coaching* (cf. Hough et al., 2001)
g. Test-taker *motivation* (cf. Stokes et al., 1993)
h. Sample not representative because *oversample* range of minorities (e.g., Hough et al., 2001)
i. *Perceptions* of applicants (cf. Hough et al., 2001, although see Sackett & Lievens, 2008, for a statement that effects are meager for several factors, such as coaching, motivation, and perceptions)
j. The type of test *response mode* (Arthur, Edwards, & Barrett, 2002; Edwards & Arthur, 2007)
k. Opportunity for applicants to *retake the test* (VanIddekinge, Morgeson, Schleicher, & Campion, 2011)
l. *Characteristics of the article's authors*, such as experience, employment, etc. (e.g., Russell et al., 1994; VanIddekinge et al., 2011).

Moderators that might influence *r* but likely not *d*

m. The possibility that in concurrent designs the same behavior influences both the predictor score and the criterion measure and hence the *validity is "built-in"* (e.g., Callinan & Robertson, 2000; Huffcutt et al., 2004; Oswald et al., 2004)
n. *Criterion type* (e.g., task, OCB, maximal/typical, counterproductive behavior; see, e.g., Hough & Furnham, 2003; Motowidlo & Van Scotter, 1994; Ones, Dilchert, Viswesvaran, & Judge, 2007; Schmitt et al., 2003)
o. Differential *weighting of criteria* (e.g., Hattrup, Rock, & Scalia, 1997; DeCorte, 1999; DeCorte et al., 2007)