

# Differential Validity for Cognitive Ability Tests in Employment and Educational Settings: Not Much More Than Range Restriction?

Philip L. Roth  
Clemson University

Huy Le  
University of Nevada–Las Vegas

In-Sue Oh  
Temple University

Chad H. Van Iddekinge  
Florida State University

Maury A. Buster  
Alabama State Personnel Department, Montgomery, Alabama

Steve B. Robbins  
Educational Testing Service, Princeton, New Jersey

Michael A. Campion  
Purdue University

The concept of differential validity suggests that cognitive ability tests are associated with varying levels of validity across ethnic groups, such that validity is lower in certain ethnic subgroups than in others. A recent meta-analysis has revived the viability of this concept. Unfortunately, data were not available in this meta-analysis to correct for range restriction within ethnic groups. We reviewed the differential validity literature and conducted 4 studies. In Study 1, we empirically demonstrated that using a cognitive ability test with a common cutoff decreases variance in test scores of Black subgroup samples more than in White samples. In Study 2, we developed a simulation that examined the effects of range restriction on estimates of differential validity. Results demonstrated that different levels of range restriction for subgroups can explain the apparent observed differential validity results in employment and educational settings (but not military settings) when no differential validity exists in the population. In Study 3, we conducted a simulation in which we examined how one corrects for range restriction affects the accuracy of these corrections. Results suggest that the correction approach using a common range restriction ratio for various subgroups may create or perpetuate the illusion of differential validity and that corrections are most accurate when done within each subgroup. Finally, in Study 4, we conducted a simulation in which we assumed differential validity in the population. We found that range restriction artificially increased the size of observed differential validity estimates when the validity of cognitive ability tests was assumed to be higher among Whites. Overall, we suggest that the concept of differential validity may be largely artifactual and current data are not definitive enough to suggest such effects exist.

*Keywords:* differential validity, adverse impact, personnel selection

Researchers have long been concerned with the predictive validity and adverse impact potential of selection tests (e.g., Aguinis & Smith, 2007; De Corte, 1999; Lawshe, 1987; Linn, 1978; Ployhart & Holtz, 2008; Reilly & Warech, 1994; Sackett, Schmitt, Ellingson, & Kabin, 2001). That is, how well do tests predict job performance and how might test scores differ for various groups? The issues are ones in which the “stakes are high” and the “underlying issues are extremely emotional” (Linn, 1978, p. 507). One manifestation of this debate relates to potential differential validity of cognitive ability tests. Dif-

ferential validity occurs when estimates of population criterion-related validity are different between subgroups, such as Whites and Blacks (Bobko & Bartlett, 1978; Linn, 1978). For example, are correlations between tests and job performance for a majority group (e.g., Whites) greater than the correlations for a minority group (e.g., Blacks)? There have long been competing schools of thought on this issue, with those arguing strongly for the existence of differential validity (e.g., Berry, Clark, & McClure, 2011; Fox & Lefkowitz, 1974; Lefkowitz & Fox, 1975) and those arguing strongly against the existence of differential

This article was published Online First September 30, 2013.

Philip L. Roth, Department of Management, Clemson University; Huy Le, Department of Management, Entrepreneurship and Technology, University of Nevada–Las Vegas; In-Sue Oh, Department of Human Resource Management, Fox School of Business, Temple University; Chad H. Van Iddekinge, Department of Management, Florida State University; Maury A. Buster, Alabama State Personnel Department, Montgomery, Alabama; Steve B. Rob-

bins, Educational Testing Service, Princeton, New Jersey; Michael A. Campion, Krannert Graduate School of Management, Purdue University.

We thank Frank Schmidt for his comments on an earlier version of this paper.

Correspondence concerning this article should be addressed to Philip L. Roth, Department of Management, Clemson University, 101 Sirmine Hall, Clemson, SC 29634-1305. E-mail: [rothp@clemson.edu](mailto:rothp@clemson.edu)

validity (e.g., Hunter, Schmidt, & Hunter, 1979; Kirchner, 1975; Lawshe, 1987).

There has recently been a resurgence of interest in differential validity after three decades of slumber. In particular, a recent meta-analysis in the *Journal of Applied Psychology* suggested that differential validity exists for Black and White subgroups in the prediction of both job performance and training performance (Berry et al., 2011). The use of meta-analysis is likely to increase salience of this issue, as meta-analyses are viewed as providing weighty scientific evidence in support of hypotheses or ideas (Aguinis, Dalton, Bosco, Pierce, & Dalton, 2011; Aguinis, Pierce, Bosco, Dalton, & Dalton, 2011; McDaniel, Rothstein, & Whetzel, 2006). The authors of the meta-analysis also suggested that the role of differential range restriction across subgroups deserves special research attention as a potential cause of observed differential validity (Berry et al., 2011). Unfortunately, they did not have the data to correct for range restriction or to examine this issue. This is likely due to the fact that good data in this area are extremely difficult to obtain.

Our purpose in this article is to examine the influence of range restriction on conclusions regarding differential validity for ethnic groups. We do not reexamine single group validity, as this idea is not logically possible (Bartlett, Bobko, & Pine, 1977; Bobko & Bartlett, 1978) or is best viewed as a relatively infrequent special case of differential validity (Linn, 1978). Nor do we focus on other predictors, such as interviews (Arvey, 1979), or on differential prediction (as per Bartlett, Bobko, Mosier, & Hannan, 1978, see also Aguinis, Culpepper, & Pierce, 2010; Mattern & Patterson, 2013). Instead, we focus on differential validity.

### Historical Debate

The existence of differential validity has been controversial for many years. We do not repeat a long literature review, as the literature has already been reviewed by others (e.g., Berry et al., 2011). Instead, we note there has been substantial discussion over the definition and existence of differential validity and related concepts (e.g., Boehm, 1977; Hunter et al., 1979).

### Education Literature

Historically, differential validity has been supported by a pattern of results found in the educational literature (e.g., Young, 1994, p. 1022). For example, Young (1994) reported results of a large-scale primary study of over 3,700 students (but only about 200 Blacks) and found that observed correlations between the SAT Verbal and Math scores and scholastic performance were higher for White students than for Black students. Interestingly, point estimates of the correlations were higher for Asian students than for White students. The issue of range restriction was not addressed by either Young's review of the literature or his empirical analyses.

### Employment Literature

Literature on employment uses of cognitive ability tests has tended to converge around the conclusion that differential validity generally is not a problem, though there has been dissent on this issue. For example, Katzell and Dyer (1977) "revived" the issue (p. 137) by suggesting that there was evidence of differential validity in many more of their samples than would be suggested by chance

(see also Katzell & Dyer, 1978). However, they warned that low power and methodological issues made the examination of the evidence difficult (see also Aguinis et al., 2010; Aguinis & Stone-Romero, 1997).

Other researchers questioned the existence of differential validity for cognitive tests (e.g., Boehm, 1978; Hunter & Schmidt, 1978; Hunter et al., 1979; Schmidt, Berner, & Hunter, 1973). For example, researchers challenged Katzell and Dyer's (1977) conclusions by noting that 26 instances in which the Black validity correlation coefficient was larger and 25 instances in which the White coefficient was larger (a more recent example also suggests Black validities may be higher in some instances; Gardner & Deadrick, 2012). Responses by researchers with a different view suggested that differential validity was observed only in roughly 8% of the 297 instances in studies in which some level of validity was observed, such that differential validity is rather uncommon (Boehm, 1977). Katzell and Dyer were also criticized for selecting only studies in which at least one validity coefficient was significant, which increased the chances of finding results supportive of differential validity (Hunter & Schmidt, 1978). In fact, Type I error rates were reported to be likely double the 5% level normally accepted (Schmidt & Hunter, 1980), and that differential validity happened at no more than chance rates once researchers considered issues such as differential range restriction and possible nonindependence of data (Hunter et al., 1979). Differential validity was declared on "intensive care" (Bartlett et al. 1978, p. 233), and the track record for differential validity was deemed "not impressive" (Fincher, 1975, p. 483).

### A Meta-Analysis

We understand why recent researchers have been intrigued by the use of meta-analysis to summarize the data on differential validity. Such an approach helps researchers to overcome the belief in the law of small numbers, which has been a stumbling block in differential validity research (Schmidt & Hunter, 1980), and to address vexing issues such as statistical power in such settings (e.g., Aguinis et al., 2010).

The results of the Berry et al. (2011) meta-analysis suggested evidence in support of differential validity. Their overall results for educational settings suggested a mean observed correlation (validity) between measures of overall cognitive ability and educational achievement (e.g., college grades) of .34 for Whites and .30 for Blacks (a difference of .04). For employment settings, the correlations between tests and measures of job performance were .19 for Whites and .16 for Blacks (a difference of .03). For military settings, the mean correlations (typically with training grades) were .34 for Whites and .17 for Blacks (a difference of .17). Overall Hispanic-White correlations were .34 for Whites and .30 for Hispanics (fewer moderator analyses were possible given the data available). Finally, there were only small overall validity differences between Whites (.34) and Asians (.33). On the basis of these results, the authors concluded that "enough evidence currently exists to conclude that it is likely observed test-criterion correlations differ for White and Black subgroups" and the next step is a "clear need" for examining underlying causes, particularly differential range restriction (Berry et al., 2011, pp. 892-893).

### The Critical Role of Range Restriction

Range restriction has received relatively little attention as a potential explanatory variable for differential validity. Only one study made more than passing mention of how range restriction might differ for various subgroups and how this could explain observed differences in validity coefficients (i.e., Hunter et al., 1979). Hunter et al. provided a brief illustration of differential range restriction by assuming a correlation between a cognitive ability test and a measure of job performance of .5, a White selection ratio of .4, and a standardized ethnic group difference of 1.0. They suggested that under these circumstances, the White observed validity coefficient would be expected to be .31 and the Black correlation would be expected to be .23. Hunter et al. noted that there were no studies available on the role of differential range restriction in observed validity differences. Likewise, Berry et al. (2011) lamented the lack of data on this issue, and given our own efforts to find quality estimates of range restriction in meta-analyses, we certainly understand their frustration with this state of affairs.

We suggest there is strong logical reason to suggest differential range restriction could account for a substantial portion of differences in observed validity coefficients. Figure 1 illustrates this reason (see also Aguinis & Smith, 2007, Figure 3). Cognitive ability tests have a strong record as valid predictors of job performance (e.g., Salgado et al., 2003; Schmidt, Shaffer, & Oh, 2008), and there are substantial mean differences between Blacks and Whites (and between Hispanics and Whites) on such tests (e.g., Sackett & Shen, 2010). Next, assume some common cutoff score (labeled as “cutoff point” in Figure 1). As illustrated in Figure 1, the cutoff score results in different selection ratios, and thus different range restriction, for minority and majority subgroup members. Range restriction should then affect the correlations between test scores and outcomes (e.g., job performance) differently for minority and majority subgroups, leading to the observed difference in validity coefficients. That is, a larger degree of range restriction is found on Black scores than on White scores (see the

right portion of each ethnic group distribution with selectees beyond the common cutoff score). As such, observed validity is likely lower (due to a larger degree of range restriction) for Blacks than for Whites.

A reviewer suggested that we also consider another scenario, one in which there is differential validity in the population (note that the logic above assumes a unitary validity value in the population). The reviewer wondered if the pattern of results might change under such circumstances because differential range restriction might have different effects under conditions that assume differential validity. In order to address this possibility, we conduct such a study below (see Study 4).

In sum, we describe four studies that examine differential validity related issues. Study 1 demonstrates how using a common cut score can result in more range restriction in Black samples than White samples in organizational and educational data. Study 2 demonstrates how range restriction can affect observed differential validity (when no differential validity exists in the population data). Study 3 addresses the issue that how one corrects for range restriction can affect differential validity conclusions. Finally, in Study 4, we assume differential validity exists in the population and examine how differential range restriction might obscure findings under these circumstances.

### Study 1: Empirical Assessment of Differential Range Restriction

An anonymous reviewer suggested that it might be useful to examine how range restriction affects variance in cognitive ability test scores at various selection ratios in actual organizational and educational data. Although conceptual treatments of the effects of range restriction on validity are not hard to find (e.g., Arvey & Faley, 1988), organizational data assessing such issues within subgroups are more rare. Perhaps this is due to the difficulty of obtaining data on an entire applicant sample or the sensitive nature of such data and the willingness to share it (e.g., organizations are very protective of their job applicant data).

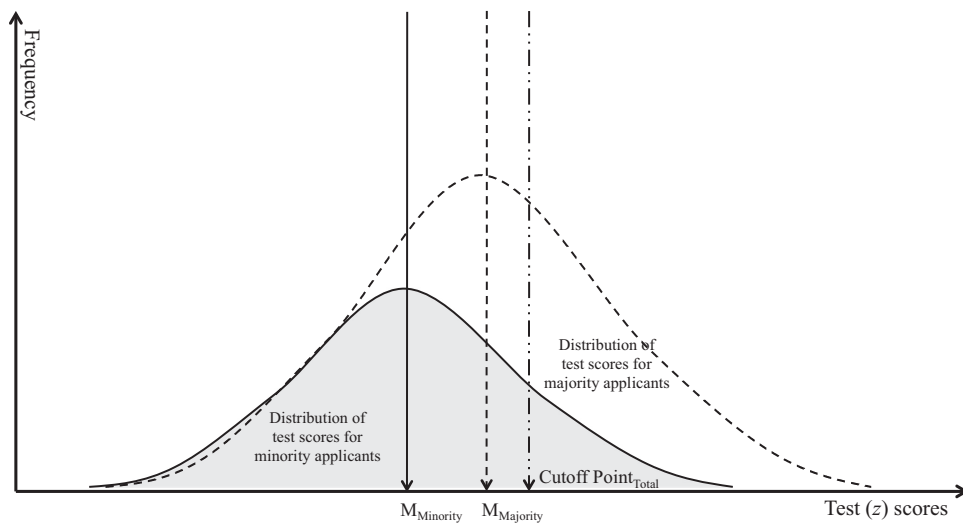


Figure 1. Illustration of differential range restriction on cognitive ability test scores. Test (z) scores represent standardized test scores. M = mean.

We briefly present data on how a common cutoff point would influence levels of range restriction in three organizational samples and one educational sample. For each sample, we report data for the entire applicant sample and then vary the selection ratios and examine variation (i.e., standard deviations) in the White and Black subgroups. We chose the selection ratios of .2, .4, .6, and .8 to maintain somewhat larger sample sizes and more stable results.

The first data set is for the job of administrative assistant. We collected data from a variety of state agencies in a southern state in which over 3,000 job applicants took a cognitive ability test. Administrative assistants help managers, typically in offices, complete their work in a variety of ways (e.g., compose and format documents, compile numbers). A test of mathematical and verbal ability was used as the first major hurdle in a selection system so that we have data for all applicants (i.e., we have “applicant” data as per Berry, Sackett, & Landers, 2007). Table 1 reports within-subgroup standard deviations (*SD*s), which tend to show a pattern where the Black subgroup *SD*s are more restricted than the White subgroup *SD*s. For example, the .80 selection ratio was associated with a *SD* for Blacks that was 64% of the *SD* for the entire/original Black applicant group, while the White subgroup *SD* was 75% of the entire White applicant group *SD*.

The second data set was for the job of engineering assistant from a single agency in a southern state. Engineering assistants help civil engineers build roadways, test materials used in roadway construction, and perform related functions (total *N* over 2,000 applicants). The sample of administrative assistants is independent of the sample of engineering assistants, and the nature of the work is markedly different. The engineering assistant test captures both mathematical and verbal ability. It was used as the first major hurdle in the selection system, so that we had access to all of the job applicants. The within-group *SD*s tend to show a pattern similar to the pattern above (see Table 1). For example, the selection ratio of .60 was associated with a Black *SD* that was 35% of the entire Black applicant *SD*, while the White *SD* was 44% of the *SD* of the entire White applicant *SD*.

The third data set was for a variety of professional jobs in a large federal agency. A test of verbal ability was used as the first hurdle

in the selection system, so that we had data for all job applicants (*N* over 100,000; see Table 2). The within-subgroup *SD*s tend to show a pattern where the Black subgroup *SD*s are more restricted than those of the White subgroup. For example, the .40 selection ratio was associated with an observed *SD* for Blacks that was 15% of the original Black applicant group *SD*, while the White subgroup *SD* was 26% of the original White applicant group *SD*.

Our fourth data set involved educational data from the ACT (see Table 3). These data were gathered from 1999 to 2006 on 277,262 students from 76 institutions. The sample was representative of the population of college applicants who took the test in terms of mean and *SD* of the ACT scores. Once again, the data suggested more range restriction for Black applicants than for White applicants (data were also available on Hispanic applicants in this sample). For example, the selection ratio of .60 was associated with a Black *SD* of 64% of the *SD* for the original Black subgroup (the figure was 67% for Hispanics). In contrast, the White *SD* was 74% of the *SD* of the original White applicant subgroup.

Overall, it appears that range restriction reduces variability in cognitive test scores more in Black applicant subgroups than in White applicant subgroups when a common cut point is used, in both employment and educational settings. This is not surprising theoretically, given the existence of standardized group differences between White and Black applicant subgroups on such tests (see Figure 1). Yet, analysis of multiple data sets of job applicants across three different jobs and a larger scale data set of college applicants provide empirical evidence for this pattern of results.

## Study 2: The Effect of Range Restriction on Differential Validity

In our second study, we sought to examine whether a differential amount of restriction in range manifests itself in terms of observed differential validities when there is no difference in validity between subgroups in the applicant/unrestricted population.

Table 1

*Restriction of Standard Deviations by Subgroup for Administrative Assistants and Engineering Assistants on a Cognitive Ability Test*

SR <sup>a,b</sup>	Race	Administrative assistants				Engineering assistants			
		<i>N</i>	<i>SD</i>	SR <sub>R</sub>	% <i>SD</i>	<i>N</i>	<i>SD</i>	SR <sub>R</sub>	% <i>SD</i>
100%	White	1,459	6.92	100%	100%	1,410	9.22	100%	100%
	Black	2,355	7.18	100%	100%	903	11.72	100%	100%
80%	White	1,293	5.19	89%	75%	1,251	6.08	89%	66%
	Black	1,767	4.6	75%	64%	607	6.23	67%	53%
60%	White	1,111	4.32	76%	62%	1,036	4.07	73%	44%
	Black	1,326	3.73	56%	52%	362	4.1	40%	35%
40%	White	810	3.37	56%	49%	760	2.47	54%	27%
	Black	805	2.85	34%	40%	192	2.5	21%	21%
20%	White	479	2.55	33%	37%	478	1.41	34%	15%
	Black	335	2.21	14%	31%	95	1.39	11%	12%

*Note.* SR = simulated selection ratio based on all subgroups; SR<sub>R</sub> = actual selection ratio for each subgroup; *SD* = standard deviation; %*SD* = percentage of the simulated hires as a percentage of the total subgroup's *SD* (this represents a within-group range restriction ratio).

<sup>a</sup> The actual SRs of 20%, 40%, and 60% for the engineering assistant sample slightly differ from the targeted SRs because of the number of tied scores by 1%, 2%, and 4%, respectively. <sup>b</sup> The actual SRs of 20% and 40% for the engineering assistant sample slightly differ from the targeted SRs because of the number of tied scores by 5% and 1%, respectively.

Table 2  
Restriction of Standard Deviations by Subgroup for Federal Professional Jobs on a Verbal Ability Test

SR	Race	<i>N</i>	<i>SD</i>	SR <sub>R</sub>	% <i>SD</i>
100%	White	100,884	7.87	100%	100%
	Black	16,902	12.88	100%	100%
80%	White	89,278	4.3	88%	55%
	Black	8,907	4.39	53%	34%
60%	White	69,316	2.82	69%	36%
	Black	4,810	2.69	28%	21%
40%	White	50,408	2.05	50%	26%
	Black	2,689	1.92	16%	15%
20%	White	25,585	1.31	25%	17%
	Black	1,018	1.19	6%	9%

Note. SR = simulated selection ratio based on all subgroups; SR<sub>R</sub> = actual selection ratio for each subgroup; *SD* = standard deviation; %*SD* = percentage of the simulated hires as a percentage of the total subgroup's *SD* (this represents a within-group range restriction ratio).

## Method

We based the simulation on meta-analytic values that estimate the non-range-restricted validities and standardized ethnic group differences for cognitive ability tests. That is, we started with correlations (and *ds*) based on applicant populations with no range restriction. We then “induced” range restriction based on various selection ratios. Again, we designed these simulations such that there is no differential validity in the applicant population before selection. However, when we select a certain percentage of applicants using a common cutoff score, differential range restriction ratios (selection ratios) for ethnic groups are induced due to the standardized ethnic group differences on predictor (test) scores (as noted above). Although we focused our simulations on employment tests, similar dynamics may occur in educational organizations.

**Values used in simulations.** We set up four different scenarios to simulate the influence of range restriction on observed differences in validities. We organized the scenarios around the issue of job complexity, which has been shown to moderate the size of ethnic group differences on cognitive ability tests (Roth, Bevier, Bobko, Switzer, & Tyler, 2001). In addition, a reviewer suggested we include an overall scenario as well.

**Scenario 1: Low complexity.** Our first scenario was meant to simulate selection for a job of low complexity. As such, the Black–White *d* value was set to .86 (Roth et al., 2001). We varied several factors within this simulation.

First, we varied the validity of the cognitive ability test to predict job performance. This required care, as Berry et al.'s (2011) estimates of differential validity were based on observed validities. So, we first took values from major meta-analyses (Hunter, 1986; Salgado et al., 2003; see also Salgado, Anderson, Moscoso, Bertua, & de Fruyt, 2003) and attenuated them for unreliability in the criterion (i.e., job performance). This ensured comparability to the Berry et al. meta-analysis. Across levels of job complexity, the range restriction (only) corrected correlations varied from .31 to .45. The validity corrected only for range restriction for low-complexity jobs was .31 (Hunter, 1986). Thus, we used .30 as our “focal validity” and then varied validity in increments of .05.

Second, we varied the selection ratio. We used selection ratios of .1, .3, .5, .7, and .9 (as per Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). In this way, we induced range restriction at various levels based on applicant level data. We selected applicants for our simulated jobs using top-down selection.

Third, we used sample sizes of 5,000 and 1 million applicants. We were careful to choose large numbers of applicants, considering the effect of *N* on estimation of *r*. That is, *r* is a negatively biased estimate of  $\rho$ , especially when sample size is small (Bobko & Schemmer, 1980; Hotelling, 1953). Given that we simulated selection with a smaller subgroup (e.g., 20%) at lower selection ratios (e.g., .1, .3), we were careful to maintain larger sample sizes to minimize the functional bias in our results.

Fourth, we varied the percentage of simulated Black applicants at 10% and 20% (as per Schmitt et al., 1997; see also Roth, Switzer, Van Iddekinge, & Oh, 2011).

We replicated each simulation condition 1,000 times in order to estimate the means and standard errors of the validity coefficients for the Black and White groups in the range-restricted samples (i.e., job incumbents).

**Scenario 2: Medium complexity.** We set the Black–White *d* value at .72 (Roth et al., 2001). In particular, validity estimates for medium-complexity jobs were .40 for Hunter (1986) and .38 for Salgado et al. (2003). We set the focal validity of .40 and varied the level of validity. We varied validity in two increments of .05, as above. We also used the same values for selection ratio, sample size, and percentage of simulated Black applicants.

**Scenario 3: High complexity.** We set the Black–White *d* value at .63 (Roth et al., 2001), though we note that this value is based on only two studies of job applicants (as suggested by a reviewer). The average high-complexity validity value was .45 (average of Hunter, 1986, and Salgado et al., 2003).

**Scenario 4: Across-complexity analyses.** In this case, we used meta-analytic figures across job complexity levels. We set *d* =

Table 3  
Restriction of Standard Deviations by Subgroup for Test Takers on the ACT (Between 1999/2000 and 2005/2006 Academic Years)

SR	Race	<i>N</i>	<i>SD</i>	SR <sub>R</sub>	% <i>SD</i>
100%	White	237,138	4.37	100%	100%
	Black	26,239	3.76	100%	100%
	Hispanic	13,885	4.29	100%	100%
80%	White	213,358	3.84	90%	88%
	Black	15,230	2.86	58%	76%
	Hispanic	11,411	3.53	82%	82%
60%	White	164,718	3.25	69%	74%
	Black	7,526	2.42	29%	64%
	Hispanic	7,755	2.89	56%	67%
40%	White	111,785	2.71	47%	62%
	Black	3,261	2.06	12%	55%
	Hispanic	4,953	2.42	36%	56%
20%	White	56,879	2.13	24%	49%
	Black	980	1.70	4%	45%
	Hispanic	2,140	1.89	15%	44%

Note. SR = simulated selection ratio based on all subgroups; SR<sub>R</sub> = actual selection ratio for each subgroup; *SD* = standard deviation; %*SD* = percentage of the simulated hires as a percentage of the total subgroup's *SD* (this represents a within-group range restriction ratio).

1.00 (Roth et al., 2001) and validity at .40 (averaged values from Hunter, 1986, and Salgado et al., 2003). We again varied the values for selection ratio, sample size, and percentage of Black simulated applicants.

**Procedure for generating data.** We describe our procedure for generating data in three parts, including assumptions, data generation, and accuracy checks.

**Assumptions.** We assumed that test scores were normally distributed within both the White subgroup and the Black subgroup job applicant populations. In addition, we assumed that the *SDs* of test scores in these two distributions were the same. The job applicant distribution is thus the joint distribution of the two distributions of the simulated White and Black job applicants.

**Generating data.** We generated the data using a SAS macro (available on request). We set the mean and *SD* of the test scores in the distribution of White subgroup at 0.00 and 1.00, respectively. We set the mean of the Black subgroup at  $-d_{WB}$  (with  $d_{WB}$  being the standardized difference of test scores between the two groups) and *SD* at 1.00. For example, we subtracted the value of .86 for our low-complexity simulation to simulate Black scores. Criterion scores were generated consistent with the simulation parameter in the condition. For example, the focal validity was .30 in the low-complexity condition. Next, we determined a cut score for the simulated applicants based on the condition in question. For example, we determined the cut score with which we would select the top 30% of all applicants. Finally, we calculated the observed correlation between the cognitive ability test and the performance measure for the selected applicants. For each condition (e.g., low complexity [ $d = .86$ ],  $N = 5,000$ , 20% simulated Black, and a selection ratio of .30), we repeated this procedure 1,000 times.

**Accuracy checks.** We performed a pair of key checks to examine the appropriateness of our simulation procedures. First, we checked to see if data generated corresponded with the parameters in our simulation. We checked each parameter (e.g., validity of .30) to the data we generated for a given condition (e.g., 20% simulated Blacks). For example, the validity parameter of .30 corresponded to correlations of .30 between our simulated test scores and our simulated criterion scores in the low-complexity conditions. Second, we checked to make sure our simulation procedure did not introduce any unexpected sources of bias. For this purpose, we simulated data with  $N = 1$  million for all the conditions. Results of these simulations are virtually free of sampling error, and they corresponded to the population parameters.

## Results

To be clear (and responsive to reviewer concerns), we reiterate the following results are based on an input matrix generated assuming a single value for validity (i.e., no differential validity) across subgroups in the population. In Study 4, we show results based on an input matrix generated assuming differential validity in the population.

**Scenario 1: Low complexity.** Table 4 contains the validity results for low-complexity jobs when sample size equals 5,000. For example, when 10% of the applicant pool was Black, when validity was .30, and when the selection ratio was .50, the observed White validity was .19 and the observed Black validity was .15 (even though we know that the “true” validity was .30). The difference in observed validities was .04. Similarly, the validity

difference across most selection ratios was about .03 to .05 when validity was set at .30. Observed differences in validity were at similar levels when validity was .25. Only when validity was set at .20 did the differences in observed validity drop to .02 to .03. Differences in validity were slightly larger as validity was set at .35 and .40. Differences in observed validities were similar in size and fairly stable when sample size increased from 5,000 to 1 million (see Table 5).

**Scenario 2: Medium complexity.** Results were fairly similar for our simulated medium-complexity scenario in Table 6 ( $N = 5,000$ ). For example, when validity was set at .40, there were 10% Blacks, and the selection ratio was .50; the White observed validity was .26 and the Black observed validity was .21. So, the observed differential validity was .05. Differences in observed validities were generally in the range of .04 to .06 for the selection ratios of .3 to .9. The differences in observed validities were similar when sample sizes were set at 1 million applicants to minimize sampling error (see Table 7).

**Scenarios 3 and 4: High-complexity and across-complexity analyses.** Tables 8 and 9 report results for high-complexity jobs. When the validity was set at .45, most differences in observed validity were in the range of .04 to .05. Results were again similar when we set sample size at 1 million. Tables 10 and 11 report results for our across-jobs scenario. When validity was set at .40, differences in observed validity were often .06 to .07.

Overall, many of the differences in observed validities were in the range of .04 to .05 (for various complexity scenarios). Such values are highly similar to the observed validity differences found by Berry et al. (2011) for organizational tests (.03) and educational tests (.04). They are not, however, similar to the differences for military tests (.17). Thus, the present data generally are consistent with the notion that differences in observed validity can largely be attributed to different amounts of range restriction when the overall population validity is the same value for both subgroups in employment and educational settings.

Another noteworthy point is that other artifacts mainly caused by a smaller number of Black applicants and hires vis-à-vis White counterparts, in addition to differential range restriction, can further create artifactual observed differential validity under some conditions (e.g., higher levels of focal validity, lower selection ratios, smaller percentages of Black applicants). For example, if focal validity is .40, selection ratio is .1, and portion of minority applicants is 10% (see Tables 4 and 5), differential validity is .08 when total number of applicants ( $N$ ) is 5,000, but it is substantially lower at .04 when  $N$  is 1 million. This noticeable change in differential validity moving from  $N = 5,000$  to  $N = 1$  million (.08 [ $W = .18$  vs.  $B = .09$ ] vs. .04 [ $W = .18$  vs.  $B = .14$ ]) is not due to the change in mean White validity but in mean Black validity. For example, small sample size and related artifacts negatively bias Black validity more than White validity in some cases (Bobko & Schemmer, 1980; Hunter & Schmidt, 2004). Given this, we caution researchers against drawing a conclusion regarding differential validity based on relatively small sample-based validation studies, where small sample size and related artifacts across subgroups can further create artifactual differential validity to a noticeable extent.

Table 4

*Differential Validity Results for Cognitive Ability Tests Based on Selection Ratio, Percentage of Black Applicants (10%, 20%) and Validity for Low-Complexity Jobs*

SR	% minority = 10%					% minority = 20%				
	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$
Focal validity = .30										
.1	.13	.043	.08	.392	.05	.13	.046	.10	.229	.03
.3	.16	.025	.12	.150	.04	.17	.025	.13	.096	.04
.5	.19	.020	.15	.094	.04	.19	.021	.15	.065	.05
.7	.22	.017	.17	.071	.05	.23	.017	.18	.047	.05
.9	.26	.014	.21	.051	.05	.26	.015	.22	.036	.05
Other validity = .20										
.1	.08	.045	.06	.377	.02	.09	.045	.07	.228	.02
.3	.11	.026	.08	.149	.02	.11	.026	.08	.098	.02
.5	.12	.020	.10	.096	.03	.13	.021	.10	.064	.03
.7	.15	.017	.11	.070	.03	.15	.018	.12	.047	.03
.9	.17	.015	.14	.054	.03	.17	.016	.14	.037	.03
Other validity = .25										
.1	.11	.045	.07	.380	.04	.11	.045	.08	.229	.03
.3	.13	.026	.10	.148	.03	.14	.026	.10	.097	.03
.5	.16	.020	.12	.096	.04	.16	.020	.13	.064	.04
.7	.18	.017	.14	.070	.04	.19	.017	.15	.047	.04
.9	.22	.015	.18	.053	.04	.22	.016	.18	.036	.04
Other validity = .35										
.1	.16	.044	.11	.379	.05	.16	.047	.12	.226	.04
.3	.19	.024	.15	.145	.04	.20	.026	.15	.094	.05
.5	.22	.019	.17	.097	.05	.23	.020	.18	.063	.05
.7	.26	.017	.20	.067	.06	.27	.017	.21	.045	.05
.9	.30	.014	.25	.052	.05	.31	.015	.26	.036	.05
Other validity = .40										
.1	.18	.044	.09	.390	.08	.18	.045	.14	.236	.04
.3	.22	.027	.17	.146	.06	.23	.026	.17	.096	.05
.5	.26	.019	.20	.095	.06	.27	.020	.21	.062	.06
.7	.30	.016	.24	.070	.06	.31	.017	.24	.045	.06
.9	.35	.014	.29	.049	.06	.35	.015	.29	.033	.06

Note.  $d = .86$ ,  $N = 5,000$ . SR = selection ratio;  $SE$  = standard error;  $r$  = mean validity across 1,000 replications;  $\Delta r_{W-B}$  = mean validity difference between the White and Black groups.

### Study 3: Correcting for Range Restriction in Differential Validity Studies

In the process of conducting our simulations, we were asked to contact researchers and organizations to request data we could use to assess differential range restriction and validity across subgroups. Most researchers had no data to share with us, and organizations tended to be very protective of their data. We did receive prompt and courteous responses from one very well known organization concerning how it corrects validity coefficients for range restriction when investigating differential validity. It used a multivariate correction for range restriction, but it corrected validity coefficients for Whites and other subgroups using an overall range restriction ratio approach rather than a within-group approach. This led us to wonder how the two approaches might influence validity calculations. Thus, we set out to compare how (a) a correction process that makes validity corrections within subgroups might compare to (b) a correction process that makes validity corrections with a common or unitary value.

### Method

We used the previous scenarios to compare the two approaches to corrections for range restriction. We used the focal validity estimate and associated standardized ethnic group difference estimates for each scenario. For example, we used the validity of .30 and  $d$  of .86 for low complexity and the focal validity of .40 and  $d$  of .72 for medium complexity. We varied sample sizes at 5,000 and 1 million and set the percentage of minorities at 20% in order to present a manageable set of simulations.

### Results

The simulation results are summarized in Table 12. For example, the low-complexity scenario (validity of .30), a selection ratio of .50, and a sample size of 1 million was associated with an observed White validity of .20 and an observed Black validity of .15. Range restriction ratios (or  $u$  values; Hunter & Schmidt, 2004) are also reported. The  $u$  values index the ratio of the  $SD$  in the

Table 5

*Differential Validity Results for Cognitive Ability Tests Based on Selection Ratio, Percentage of Black Applicants (10%, 20%) and Validity for Low-Complexity Jobs*

SR	% minority = 10%					% minority = 20%				
	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$
Focal validity = .30										
.1	.13	.003	.10	.023	.03	.13	.003	.10	.015	.03
.3	.16	.002	.13	.010	.04	.17	.002	.13	.007	.04
.5	.19	.001	.15	.007	.04	.20	.001	.15	.005	.04
.7	.22	.001	.17	.005	.05	.23	.001	.18	.003	.05
.9	.26	.001	.21	.004	.05	.26	.001	.22	.003	.05
Other validity = .20										
.1	.09	.003	.07	.023	.02	.09	.003	.07	.016	.02
.3	.11	.002	.08	.010	.02	.11	.002	.08	.007	.02
.5	.13	.001	.10	.007	.03	.13	.001	.10	.005	.03
.7	.15	.001	.11	.005	.03	.15	.001	.12	.003	.03
.9	.17	.001	.14	.004	.03	.17	.001	.14	.003	.03
Other validity = .25										
.1	.11	.003	.09	.022	.02	.11	.003	.09	.015	.02
.3	.13	.002	.10	.010	.03	.14	.002	.11	.007	.03
.5	.16	.001	.12	.007	.04	.16	.001	.13	.005	.04
.7	.18	.001	.14	.005	.04	.19	.001	.15	.003	.04
.9	.22	.001	.18	.004	.04	.22	.001	.18	.003	.04
Other validity = .35										
.1	.15	.003	.12	.023	.03	.16	.003	.12	.016	.03
.3	.19	.002	.15	.010	.04	.20	.002	.15	.007	.04
.5	.22	.001	.18	.007	.05	.23	.001	.18	.005	.05
.7	.26	.001	.20	.005	.05	.27	.001	.21	.003	.06
.9	.30	.001	.25	.004	.05	.31	.001	.26	.003	.05
Other validity = .40										
.1	.18	.003	.14	.023	.04	.18	.003	.14	.016	.04
.3	.22	.002	.17	.010	.05	.23	.002	.18	.007	.05
.5	.26	.001	.20	.006	.06	.27	.001	.21	.004	.06
.7	.30	.001	.24	.005	.06	.31	.001	.24	.003	.06
.9	.35	.001	.29	.004	.06	.35	.001	.29	.002	.06

Note.  $d = .86$ ,  $N = 1$  million. SR = selection ratio;  $SE$  = standard error;  $r$  = mean validity across 1,000 replications;  $\Delta r_{W-B}$  = mean validity difference between the White and Black groups.

selected sample to the  $SD$  in the applicant population and reflect the amount of range restriction within each subgroup. The  $u$  value for the White subgroup was .63, and the  $u$  value for the Black subgroup was .49, suggesting more severe range restriction in the Black subgroup. The overall  $u$  value for the applicant population was .59. Both subgroup validities corrected back to .30 when within-subgroup  $u$  values were used for corrections. However, this was not the case when a unitary correction (i.e., based on the overall  $u$  values) was used. In this case, the validity for the White subgroup was .32 and that for the Black subgroup was .25. Thus, the correction for range restriction using the unitary  $u$  value appeared to overestimate the White validity and underestimate the Black validity. This appears to be a function of the way in which validity values were corrected, which fails to account for differential amounts of range restriction. Similar results were evident in medium- and high-complexity scenarios (see Table 12).<sup>1</sup>

Overall, it appears that it is quite important how corrections for range restriction are done when differential validity is investigated. Corrections that use within-group corrections ap-

pear to yield substantially more accurate results than those done with unitary, across-group estimates. As a reviewer suggested (and we concur), it is important to examine the number of Black individuals hired in simulation and organizational research on differential validity. The exact number of Black hires (and

<sup>1</sup> We note one other interesting finding. There were some indications of differential validity when sample sizes were 5,000 and the selection ratio was .10. We investigated such conditions and found that in many of these samples, less than 20 simulated Black individuals were hired (at times around 11 to 14). Thus, the validity estimates ( $r$ ) from these studies were based on small samples, and some results across the 1,000 replications were likely downwardly biased due to the combined effect of the negatively skewed sampling distribution of  $r$ s and the positively skewed sampling distribution of  $u$  values (Bobko & Schemmer, 1980; Hunter & Schmidt, 2004). Thus, we interpret this condition with caution. However, when sample sizes are 1 million, this negative bias does not appear to influence results when the selection ratio was set at .10.



Table 6

*Differential Validity Results for Cognitive Ability Tests Based on Selection Ratio, Percentage of Black Applicants (10%, 20%) and Validity for Medium-Complexity Jobs*

SR	% minority = 10%					% minority = 20%				
	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$
Focal validity = .40										
.1	.18	.044	.13	.289	.04	.18	.045	.13	.190	.04
.3	.22	.025	.18	.130	.04	.23	.027	.19	.085	.04
.5	.26	.019	.21	.079	.05	.26	.020	.22	.058	.05
.7	.30	.016	.24	.064	.06	.31	.017	.25	.044	.06
.9	.35	.014	.30	.048	.05	.35	.015	.30	.032	.05
Other validity = .30										
.1	.13	.044	.11	.306	.02	.13	.046	.11	.192	.02
.3	.16	.025	.13	.133	.03	.17	.027	.13	.088	.03
.5	.19	.019	.15	.087	.04	.19	.020	.16	.056	.04
.7	.22	.016	.18	.064	.04	.22	.018	.18	.044	.04
.9	.26	.015	.22	.052	.04	.26	.016	.23	.034	.04
Other validity = .35										
.1	.16	.046	.12	.305	.04	.16	.044	.12	.190	.04
.3	.19	.026	.15	.134	.04	.20	.027	.16	.088	.03
.5	.22	.020	.18	.084	.04	.23	.021	.19	.061	.04
.7	.26	.016	.21	.065	.05	.26	.017	.22	.042	.05
.9	.30	.014	.26	.049	.04	.31	.015	.27	.033	.04
Other validity = .45										
.1	.21	.045	.17	.292	.04	.21	.044	.18	.192	.03
.3	.26	.025	.21	.128	.05	.26	.026	.21	.088	.05
.5	.30	.019	.24	.088	.06	.30	.019	.25	.055	.06
.7	.34	.016	.28	.062	.06	.35	.016	.29	.043	.06
.9	.39	.013	.34	.046	.06	.40	.014	.35	.032	.05
Other validity = .50										
.1	.24	.042	.18	.309	.06	.24	.045	.20	.188	.04
.3	.29	.025	.23	.124	.06	.29	.025	.24	.085	.05
.5	.33	.019	.27	.083	.07	.34	.019	.28	.056	.06
.7	.38	.015	.32	.059	.06	.39	.015	.32	.043	.06
.9	.44	.013	.38	.045	.06	.45	.014	.39	.031	.06

Note.  $d = .72$ ,  $N = 5,000$ . SR = selection ratio; SE = standard error;  $r$  = mean validity across 1,000 replications;  $\Delta r_{W-B}$  = mean validity difference between the White and Black groups.

applicants) necessary in simulation and organizational research depends on selection conditions.<sup>2</sup>

#### Study 4: Investigating the Influence of Differential Validity

Our initial interest and research question involved “how much differential validity would we expect to observe if there is no differential validity in the population?” However, the reviewers strongly encouraged us to investigate another question. The reviewers urged us to also consider a situation in which differential validity is assumed to exist in the population and to examine its influence on observed differential validity.

Expectations for the outcomes of this study were more complex, given the nature of the question. We note two psychometric sources that could influence observed differential range restriction. First, there is differential range restriction. This should typically serve to restrict the range of Black scores more than White scores, as noted above. This source of variance is artifactual, because it is

<sup>2</sup> A reviewer asked us to provide guidance on the number of minority applicants needed in differential validity studies to allow accurate estimation of validity. We approached this empirically. We allowed only a 5% bias in estimation of range restriction corrected validity. We ran 300 conditions in which we varied selection ratio (.10, .30, .50, .70, .90); population validity (.30, .35, .40, .45, .50); standardized difference between groups (.63, .72, .86, 1.00); and proportion of minorities (.10, .20, .30). We ran 1,000 simulations within each condition. Using regression based on data across the conditions, we computed a formula for estimating the number of minority hires to be  $SN_b = 105 + 24.1 \times d_x - 42.2 \times pm - 121.7 \times SR$ , where  $SN_b$  is the number of minority hires,  $d_x$  = standardized mean group difference,  $pm$  = proportion of minorities, and  $SR$  = selection ratio. It is noted that the inclusion of population validity does not improve the prediction accuracy of the equation (adjusted  $R^2$ ), and thus it is excluded from the equation. For example, based on the formula, when selection ratio = .50, proportion of minority in the application pool = .20, and  $d = .72$ , researchers would find it helpful to make at least 53 minority hires. The total sample size needed for such a study would likely be close to 1,000 job applicants (including 200 minority applicants). Note that the regression weights for proportion minority and selection ratio might have a substantial influence on the estimate of the number of minority hires needed for accurate validity estimates.

Table 7

*Differential Validity Results for Cognitive Ability Tests Based on Selection Ratio, Percentage of Black Applicants (10%, 20%) and Validity for Medium-Complexity Jobs*

SR	% minority = 10%					% minority = 20%				
	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$
Focal validity = .40										
.1	.18	.003	.15	.020	.03	.18	.003	.15	.013	.03
.3	.22	.002	.18	.009	.04	.23	.002	.18	.006	.04
.5	.26	.001	.21	.006	.05	.26	.001	.22	.004	.05
.7	.30	.001	.25	.004	.05	.31	.001	.25	.003	.05
.9	.35	.001	.30	.003	.05	.35	.001	.30	.002	.05
Other validity = .30										
.1	.13	.003	.11	.020	.02	.13	.003	.11	.013	.02
.3	.16	.002	.13	.009	.03	.17	.002	.13	.006	.03
.5	.19	.001	.15	.006	.04	.19	.001	.16	.004	.04
.7	.22	.001	.18	.005	.04	.22	.001	.18	.003	.04
.9	.26	.001	.22	.003	.04	.26	.001	.22	.003	.04
Other validity = .35										
.1	.15	.003	.13	.020	.03	.16	.003	.13	.013	.03
.3	.19	.002	.16	.009	.04	.20	.002	.16	.006	.04
.5	.22	.001	.18	.006	.04	.23	.001	.19	.004	.04
.7	.26	.001	.21	.004	.05	.26	.001	.22	.003	.05
.9	.30	.001	.26	.004	.04	.31	.001	.26	.002	.04
Other validity = .45										
.1	.21	.003	.17	.019	.04	.21	.003	.17	.013	.04
.3	.26	.002	.21	.009	.05	.26	.002	.21	.006	.05
.5	.30	.001	.24	.006	.05	.30	.001	.25	.004	.06
.7	.34	.001	.28	.004	.06	.35	.001	.29	.003	.06
.9	.40	.001	.34	.003	.05	.40	.001	.35	.002	.05
Other validity = .50										
.1	.23	.003	.19	.019	.04	.24	.003	.20	.013	.04
.3	.29	.002	.24	.009	.05	.29	.002	.24	.006	.05
.5	.33	.001	.27	.006	.06	.34	.001	.28	.004	.06
.7	.38	.001	.32	.004	.06	.39	.001	.33	.003	.06
.9	.44	.001	.38	.003	.06	.45	.001	.39	.002	.06

*Note.*  $d = .72$ ,  $N = 1$  million. SR = selection ratio;  $SE$  = standard error;  $r$  = mean validity across 1,000 replications;  $\Delta r_{W-B}$  = mean validity difference between the White and Black groups.

due to differential range restriction. Second, there are now differential levels of variance in the criterion of job performance associated with test scores (before differential range restriction occurs). For example, there is less variance in the criterion of job performance that is related to cognitive ability in the Black distribution of scores by assuming a White validity of .40 for a cognitive ability test and a Black validity of .35 in the population. This difference in variances is due to “true” differences, because validities are set at different levels. Overall, it is possible that Black validity scores are doubly “attenuated” relative to White scores in many conditions that model higher White validity, as illustrated above.

## Method

We set two values in our simulation. First, we assumed a differential validity of .05. We thought this would be a plausible value, given the results of observed differential validity of .03 and .04 from Berry et al. (2011). We set  $d = .72$ , as per Scenario 2 above in a medium-complexity job, so as not to generate too many scenarios and tables.

We varied several other parameters. We varied the percentage of minority job applicants to be 10% and 20% of the sample and the sample size to be 5,000 and 1 million, as above. We also varied validity values and the direction of differential validity. We started by assuming White validity was higher, such that White validity was .40 versus Black validity of .35 and White validity was .45 versus Black validity of .40. We also modeled conditions in which Black validity was higher with the values of .40 and .45 (vs. White values of .35 and .45, respectively). We investigated differences in both directions, given previous research (see above), to provide a balanced approach. Finally, we varied the selection ratios at .1, .3, .5, .7, and .9, as in our other simulation studies (i.e., Studies 2 and 3).

## Results

The simulation results are found in Tables 13, 14, 15, and 16. The pattern of results is fairly stable across conditions. When there is (assumed) differential validity in the population in favor of Whites, the observed differential validity is more pronounced than true differential validity. For example, Table 13 shows that when

Table 8

*Differential Validity Results for Cognitive Ability Tests Based on Selection Ratio, Percentage of Black Applicants (10%, 20%) and Validity for High-Complexity Jobs*

SR	% minority = 10%					% minority = 20%				
	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$
Focal validity = .45										
.1	.21	.044	.16	.266	.04	.21	.043	.17	.178	.04
.3	.25	.024	.21	.121	.05	.26	.026	.21	.081	.05
.5	.30	.019	.24	.080	.05	.30	.020	.25	.054	.05
.7	.34	.015	.29	.061	.05	.35	.017	.29	.041	.05
.9	.40	.013	.35	.045	.05	.40	.014	.35	.032	.05
Other validity = .35										
.1	.15	.045	.12	.269	.04	.15	.044	.12	.177	.03
.3	.19	.026	.16	.121	.03	.20	.026	.16	.079	.04
.5	.22	.019	.19	.083	.04	.23	.020	.19	.056	.04
.7	.26	.016	.22	.060	.04	.26	.017	.22	.042	.04
.9	.30	.014	.26	.049	.04	.31	.014	.27	.033	.04
Other validity = .40										
.1	.18	.045	.13	.267	.05	.18	.045	.15	.171	.03
.3	.22	.025	.18	.120	.04	.22	.025	.19	.082	.04
.5	.26	.019	.21	.084	.05	.26	.020	.22	.058	.04
.7	.30	.016	.25	.062	.05	.30	.016	.26	.042	.05
.9	.35	.013	.31	.047	.04	.35	.014	.31	.031	.04
Other validity = .50										
.1	.23	.044	.18	.280	.05	.24	.044	.20	.170	.04
.3	.29	.025	.24	.118	.04	.29	.025	.24	.079	.05
.5	.33	.018	.28	.082	.05	.34	.019	.29	.053	.05
.7	.38	.016	.32	.060	.06	.39	.015	.33	.039	.06
.9	.44	.012	.39	.044	.05	.45	.013	.40	.031	.05
Other validity = .55										
.1	.26	.044	.21	.266	.06	.27	.044	.22	.167	.05
.3	.33	.023	.28	.113	.05	.33	.025	.28	.074	.05
.5	.37	.018	.32	.079	.06	.38	.018	.32	.053	.06
.7	.43	.015	.37	.055	.06	.43	.015	.37	.037	.06
.9	.49	.012	.44	.040	.05	.49	.012	.44	.029	.05

Note.  $d = .63$ ,  $N = 5,000$ . SR = selection ratio;  $SE$  = standard error;  $r$  = mean validity across 1,000 replications;  $\Delta r_{W-B}$  = mean validity difference between the White and Black groups.

the simulated White validity is .40, the Black validity is .35, and the selection ratio is .5, the observed differential validity is .08 (which is greater than the population difference of .05). A similar pattern is observed in Table 13 when the White validity is .45, the Black validity is .40, and the selection ratio is .5. This general pattern is also apparent in Tables 14–16 with varying sample sizes and proportions of simulated Black applicants. Thus, differential range restriction increases the observed differential validity when White validities are assumed to be higher in the population.

The pattern of results is different when simulated Black validities are higher than White validities in the population. For example, when the Black validity is set at .45, the White validity is .40, and the selection ratio is .5, the observed differential validity is .02 (see, e.g., Table 13). In this case, differential validity is smaller than the true differential validity set at .05. It appears that differential range restriction masks the amount of true differential validity. In this case, the two attenuating forces (i.e., less error variance yet more range restriction in the Black score distribution) appear to act in opposite directions and push observed validity

estimates for subgroups closer together than they are in the population.

## Discussion

The results of Study 4 deserve some comment. First, the results contribute to our understanding of what might happen in true differential validity situations. When White validities are higher, it appears that true differential validity is amplified by differential range restriction. That is, a true differential validity (e.g., .05) is associated with larger observed differences (e.g., .08). In contrast, when Black validities are higher (e.g., .05), observed differences in validity were masked/lessened by differential range restriction (e.g., .02).

The modeling of true differential validity in our simulation might also help us understand what could have happened in previous analyses of differential validity. One might have observed a validity difference of .03 or .04 in favor of Whites (reported in Berry et al., 2011). Given the existence of (a) differential range

Table 9

*Differential Validity Results for Cognitive Ability Tests Based on Selection Ratio, Percentage of Black Applicants (10%, 20%) and Validity for High-Complexity Jobs*

SR	% minority = 10%					% minority = 20%				
	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$
Focal validity = .45										
.1	.21	.003	.17	.018	.03	.21	.003	.18	.012	.03
.3	.25	.002	.21	.009	.04	.26	.002	.22	.006	.04
.5	.30	.001	.25	.005	.05	.30	.001	.25	.004	.05
.7	.34	.001	.29	.004	.05	.35	.001	.29	.003	.05
.9	.40	.001	.35	.003	.05	.40	.001	.35	.002	.05
Other validity = .35										
.1	.15	.003	.13	.018	.02	.16	.003	.13	.012	.02
.3	.19	.002	.16	.008	.03	.19	.002	.16	.006	.03
.5	.22	.001	.19	.006	.04	.23	.001	.19	.004	.04
.7	.26	.001	.22	.004	.04	.26	.001	.22	.003	.04
.9	.30	.001	.27	.004	.04	.31	.001	.27	.002	.04
Other validity = .40										
.1	.18	.003	.15	.018	.03	.18	.003	.15	.012	.03
.3	.22	.002	.19	.008	.04	.23	.002	.19	.006	.04
.5	.26	.001	.22	.006	.04	.26	.001	.22	.004	.04
.7	.30	.001	.25	.004	.05	.30	.001	.26	.003	.05
.9	.35	.001	.31	.003	.04	.35	.001	.31	.002	.04
Other validity = .50										
.1	.23	.003	.20	.017	.04	.24	.003	.20	.012	.04
.3	.29	.002	.24	.008	.05	.29	.002	.25	.006	.05
.5	.33	.001	.28	.006	.05	.34	.001	.29	.004	.05
.7	.38	.001	.33	.004	.06	.39	.001	.33	.003	.06
.9	.44	.001	.39	.003	.05	.45	.001	.40	.002	.05
Other validity = .55										
.1	.26	.003	.22	.017	.04	.27	.003	.23	.012	.04
.3	.33	.002	.27	.008	.05	.33	.002	.28	.006	.05
.5	.37	.001	.32	.005	.06	.38	.001	.32	.004	.06
.7	.43	.001	.37	.004	.06	.43	.001	.37	.003	.06
.9	.49	.001	.44	.003	.05	.49	.001	.44	.002	.05

*Note.*  $d = .63$ ,  $N = 1$  million. SR = selection ratio;  $SE$  = standard error;  $r$  = mean validity across 1,000 replications;  $\Delta r_{W-B}$  = mean validity difference between the White and Black groups.

restriction and (b) true differential validity, it is possible that true levels of differential validity were less than .03 or .04, given the above results. That is, going backwards from the observed differential validity to true differential validity, values of .03 or .04 could be overestimates of differential validity.

Second, we suggest considerable caution in the interpretation of Study 4. Given the results of Study 2, all the differences in validity in organizational and educational samples could be due to artifactual variance. We respect the obfuscating power of range restriction in this situation and suggest that modeling differential validity requires a number of key assumptions, as discussed. Further, we note that there are a number of “moving parts” in Study 4, due to both differential range restriction and differential amounts of variance associated with test scores. Given both dynamics, where we set the values may be important, and there is little guidance on how to do this with certainty. Thus, we are careful not to overinterpret these values at the present time.

## General Discussion

We set out to examine the effect of range restriction on evidence of Black–White differential validity for cognitive ability tests. We thought it important to revisit the issue of differential validity, given the conclusions in a recent meta-analysis. We first examined empirical data to see if our thinking on differential range restriction within subgroups was warranted. Our examination of applicant data from several jobs and organizations (e.g., administrative assistant, engineering assistant) suggests that a common cut point on cognitive tests results in more range restriction for Black applicants than for White applicants.

Second, we used simulated data to examine how differential range restriction might influence observed differences in validity. We based our work on unrestricted estimates (i.e., validity and standardized ethnic group differences) from meta-analyses (e.g., Hunter, 1986; Roth et al., 2011). As suggested by a reviewer, we note there was no differential validity in the population values for

Table 10

*Differential Validity Results for Cognitive Ability Tests Based on Selection Ratio, Percentage of Black Applicants (10%, 20%) and Validity Across All Jobs*

SR	% minority = 10%					% minority = 20%				
	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$
Focal validity = .40										
.1	.18	.043	.11	.478	.07	.18	.045	.13	.275	.05
.3	.22	.025	.16	.164	.06	.23	.025	.17	.109	.06
.5	.26	.019	.19	.103	.07	.27	.019	.20	.069	.07
.7	.30	.016	.23	.073	.07	.31	.017	.23	.049	.07
.9	.35	.014	.28	.052	.07	.35	.014	.28	.036	.07
Other validity = .30										
.1	.13	.044	.08	.486	.05	.13	.045	.09	.272	.04
.3	.16	.026	.12	.169	.04	.17	.026	.12	.109	.04
.5	.19	.020	.14	.105	.05	.20	.020	.15	.069	.05
.7	.22	.017	.17	.074	.05	.23	.017	.17	.050	.06
.9	.26	.015	.20	.054	.06	.26	.015	.21	.038	.05
Other validity = .35										
.1	.15	.044	.09	.483	.06	.16	.045	.11	.277	.05
.3	.19	.025	.14	.167	.05	.20	.026	.14	.108	.05
.5	.23	.020	.17	.103	.06	.23	.020	.17	.068	.06
.7	.26	.016	.20	.074	.06	.27	.017	.20	.049	.06
.9	.30	.014	.24	.054	.06	.31	.015	.25	.036	.06
Other validity = .45										
.1	.21	.043	.13	.488	.08	.21	.044	.15	.271	.06
.3	.26	.025	.19	.166	.07	.26	.025	.20	.107	.07
.5	.30	.019	.22	.102	.08	.31	.019	.23	.067	.08
.7	.34	.016	.26	.071	.08	.35	.016	.27	.048	.08
.9	.40	.013	.32	.051	.08	.40	.014	.32	.035	.08
Other validity = .50										
.1	.23	.044	.15	.482	.08	.24	.043	.17	.275	.07
.3	.29	.024	.21	.163	.08	.30	.024	.23	.106	.07
.5	.34	.018	.25	.101	.08	.35	.019	.26	.065	.08
.7	.38	.015	.29	.070	.09	.39	.016	.30	.047	.09
.9	.44	.013	.36	.049	.08	.45	.013	.36	.034	.08

Note.  $d = 1.00$ ,  $N = 5,000$ . SR = selection ratio;  $SE$  = standard error;  $r$  = mean validity across 1,000 replications;  $\Delta r_{W-B}$  = mean validity difference between the White and Black groups.

these analyses (from Studies 2 and 3). This allowed us to induce various levels of range restriction that might occur in selection settings and to examine observed differential validity under various conditions. We generally found that different levels of range restriction across subgroups could result in observed differential validity of roughly .05 for both medium- and lower complexity jobs. In addition, we suggest that researchers should attend to the small number of Blacks hired in some situations, given the negative bias involved in calculating and aggregating  $r$ s with small sample sizes.

Third, we examined the implications for how one corrects for range restriction in differential validity analyses. We compared within-subgroup corrections to an across-subgroups correction approach and found that within-group corrections generally are more accurate than unitary corrections. In fact, a unitary approach tends to overestimate White subgroup validity and underestimate Black subgroup validity.

Fourth, we made efforts to model the effects of differential range restriction, assuming that differential validity actually does exist in the

population. We found that observed differential validity was higher than (assumed) true differential validity when White validities were higher than those for Blacks in the population. For example, simulated true differential validity was .05 and observed differential validity was .08. Thus, it is possible that previous analyses of differential validity might have overestimated the magnitude of differential validity. Yet, we are cautious not to overinterpret such results, given the important role of artifacts found in Study 2 and the lack of data to inform decisions about certain parameters to simulate a situation in which differential validity is assumed to exist in the population.

### Implications for Recent Thinking About Differential Validity

The results of our analyses are important because the current levels of artifactually induced differential validity are as large as (or slightly larger than) the level of observed differential validity reported in the recent meta-analysis of employment and educational samples (Berry et al., 2011). Berry et al.'s results were often

Table 11

*Differential Validity Results for Cognitive Ability Tests Based on Selection Ratio, Percentage of Black Applicants (10%, 20%) and Validity Across All Jobs*

SR	% minority = 10%					% minority = 20%				
	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$
Focal validity = .40										
.1	.18	.003	.14	.027	.04	.18	.003	.14	.018	.04
.3	.22	.002	.17	.012	.06	.23	.002	.17	.008	.06
.5	.26	.001	.20	.007	.06	.27	.001	.20	.005	.07
.7	.30	.001	.23	.005	.07	.31	.001	.23	.003	.07
.9	.35	.001	.28	.004	.07	.35	.001	.28	.003	.07
Other validity = .30										
.1	.13	.003	.10	.028	.03	.13	.003	.10	.018	.03
.3	.16	.002	.12	.012	.04	.17	.002	.12	.008	.04
.5	.19	.001	.14	.008	.05	.20	.001	.15	.005	.05
.7	.22	.001	.17	.005	.05	.23	.001	.17	.004	.05
.9	.26	.001	.20	.004	.05	.26	.001	.21	.003	.05
Other validity = .35										
.1	.15	.003	.12	.028	.04	.16	.003	.12	.018	.04
.3	.19	.002	.15	.012	.05	.20	.002	.15	.008	.05
.5	.23	.001	.17	.007	.06	.23	.001	.17	.005	.06
.7	.26	.001	.20	.005	.06	.27	.001	.20	.003	.06
.9	.30	.001	.24	.004	.06	.31	.001	.25	.003	.06
Other validity = .45										
.1	.21	.003	.16	.027	.05	.21	.003	.16	.018	.05
.3	.26	.002	.19	.011	.06	.26	.002	.20	.008	.06
.5	.30	.001	.22	.008	.07	.31	.001	.23	.005	.07
.7	.34	.001	.26	.005	.08	.35	.001	.27	.004	.08
.9	.40	.001	.32	.004	.08	.40	.001	.32	.003	.08
Other validity = .50										
.1	.23	.003	.18	.026	.05	.24	.003	.18	.018	.06
.3	.29	.002	.22	.011	.07	.30	.002	.23	.007	.07
.5	.34	.001	.26	.007	.08	.35	.001	.26	.005	.08
.7	.38	.001	.30	.005	.09	.39	.001	.30	.003	.09
.9	.44	.001	.36	.003	.08	.45	.001	.36	.002	.08

*Note.*  $d = 1.00$ ,  $N = 1$  million. SR = selection ratio;  $SE$  = standard error;  $r$  = mean validity across 1,000 replications;  $\Delta r_{W-B}$  = mean validity difference between the White and Black groups.

in the range of .03 to .04. Our results were often in the range of .03 to .07. This is important because our results in Study 2 are solely attributable to range restriction, and thus the existence of differential validity was artifactual. Thus, our analysis suggests that observed differential validity in employment and educational tests may largely reflect an artifact of range restriction when there are no true population-level differences in validity across subgroups, as in the population parameters that we designed.

Importantly, applying our results to current tests of cognitive ability in employment settings does not require strong assumptions. One must merely assume that organizations use such tests. For example, [Berry et al. \(2011\)](#) gathered a substantial amount of their data from the previous meta-analysis of [Synk and Swarthout \(1987\)](#) in which the General Aptitude Test Battery was used in hiring. Second, one does not need to assume a stringent selection ratio. In our analyses, selection ratios had to be .9 to induce observed differential validity. Finally, one only needs to assume moderate validities (e.g., .40 or .30 for medium-complexity jobs) in the population. Following these conditions, one would expect

observed differential validity to be in the range of .03 to .07 in many cases. Thus, our data suggest that different levels of range restriction for ethnic groups can generally explain the magnitude of [Berry et al.](#)'s results for cognitive tests used in employment settings and perhaps tests used in educational settings.

However, in fairness to [Berry et al. \(2011\)](#), data necessary for corrections for range restriction were not available to them (see their discussion on p. 887). We found similar difficulty in obtaining any such estimates. For example, many organizations did not have such data, and other organizations appeared reticent about sharing them. Thus, we truly understood previous authors' likely levels of frustration at finding good data.

It may also be useful to consider two related issues to finding (any) range restriction data. First, researchers would need to be careful to understand the selection systems in operation (e.g., presence of diversity-related efforts) and how they influenced cut scores,  $r$ , and  $u$  within the various subgroups. Second, researchers would need to be careful in their choice of standard deviations. It is possible to find suboptimal estimates of  $u$  (e.g.,

Table 12  
*The Influence of Using Within-Group Range Restriction Corrections (% of Minority = 20%)*

SR	$r_W$	$r_B$	$u_W$	$u_B$	$u_T$	Correcting for RR using $u_W$ and $u_B$			Correcting for RR using unitary $u_T$		
						$\rho_W$	$\rho_B$	$\Delta\rho_{W-B}$	$\rho_W$	$\rho_B$	$\Delta\rho_{W-B}$
Low-complexity jobs											
Focal validity = .30, $d = .86$ , $N = 5,000$											
.1	.13	.10	.42	.32	.40	.29	.22	.08	.31	.19	.12
.3	.17	.13	.54	.41	.50	.30	.29	.01	.32	.24	.07
.5	.19	.15	.63	.49	.59	.30	.29	.01	.32	.24	.08
.7	.23	.18	.74	.57	.69	.30	.30	.00	.32	.25	.07
.9	.26	.22	.86	.70	.82	.30	.30	.00	.32	.26	.05
Focal validity = .30, $d = .86$ , $N = 1$ million											
.1	.13	.10	.42	.34	.40	.30	.30	.00	.32	.25	.06
.3	.17	.13	.54	.42	.50	.30	.30	.00	.32	.25	.07
.5	.20	.15	.63	.49	.59	.30	.30	.00	.32	.25	.07
.7	.23	.18	.74	.58	.69	.30	.30	.00	.32	.25	.06
.9	.26	.22	.86	.71	.82	.30	.30	.00	.32	.26	.05
Medium-complexity jobs											
Focal validity = .40, $d = .72$ , $N = 5,000$											
.1	.18	.13	.42	.33	.40	.39	.30	.09	.40	.27	.14
.3	.23	.19	.53	.43	.51	.40	.39	.01	.42	.34	.07
.5	.26	.22	.63	.50	.60	.40	.40	.00	.42	.35	.07
.7	.31	.25	.73	.60	.70	.40	.40	.00	.42	.35	.07
.9	.35	.30	.86	.73	.83	.40	.40	.00	.42	.36	.06
Focal validity = .40, $d = .72$ , $N = 1$ million											
.1	.18	.15	.42	.35	.40	.40	.40	.00	.42	.35	.06
.3	.23	.18	.53	.43	.51	.40	.40	.00	.42	.35	.07
.5	.26	.22	.63	.51	.60	.40	.40	.00	.42	.35	.07
.7	.31	.25	.73	.60	.70	.40	.40	.00	.42	.35	.07
.9	.35	.30	.86	.73	.83	.40	.40	.00	.41	.36	.05
High-complexity jobs											
Focal validity = .45, $d = .63$ , $N = 5,000$											
.1	.21	.17	.42	.34	.41	.45	.37	.08	.46	.34	.12
.3	.26	.21	.53	.44	.51	.45	.43	.01	.47	.39	.08
.5	.30	.25	.63	.52	.60	.45	.45	.01	.47	.39	.07
.7	.35	.29	.73	.61	.70	.45	.45	.00	.47	.40	.06
.9	.40	.35	.86	.75	.83	.45	.45	.00	.46	.41	.05
Focal validity = .45, $d = .63$ , $N = 1$ million											
.1	.21	.18	.42	.35	.41	.45	.45	.00	.46	.40	.06
.3	.26	.22	.53	.44	.51	.45	.45	.00	.47	.40	.07
.5	.30	.25	.63	.52	.60	.45	.45	.00	.47	.40	.07
.7	.35	.29	.73	.61	.70	.45	.45	.00	.47	.40	.06
.9	.40	.35	.86	.75	.83	.45	.45	.00	.46	.41	.05
All jobs											
Focal validity = .40, $d = 1.00$ , $N = 5,000$											
.1	.18	.13	.42	.30	.39	.40	.27	.13	.42	.24	.18
.3	.23	.17	.54	.40	.50	.40	.38	.02	.43	.32	.11
.5	.27	.20	.64	.47	.59	.40	.40	.00	.43	.33	.10
.7	.31	.23	.74	.55	.68	.40	.40	.00	.43	.33	.10
.9	.35	.28	.86	.68	.80	.40	.40	.00	.42	.35	.08
Focal validity = .40, $d = 1.00$ , $N = 1$ million											
.1	.18	.14	.43	.32	.39	.40	.40	.00	.43	.34	.09
.3	.23	.17	.54	.40	.50	.40	.40	.00	.43	.33	.10
.5	.27	.20	.64	.47	.59	.40	.40	.00	.43	.33	.10
.7	.31	.23	.74	.55	.68	.40	.40	.00	.43	.33	.09
.9	.35	.28	.86	.68	.80	.40	.40	.00	.42	.35	.08

*Note.* Some values for the change in the rho estimate between the Black and White subgroups ( $\Delta\rho_{W-B}$ ) may round to slightly different values. For example, the value of .07 in the low-complexity condition with  $N = 5,000$  and a SR of .30 is a function of rounding the relevant values to two digits. SR = selection ratio;  $r$  = mean validity across 1,000 replications;  $u$  = range restriction ratio ( $_W$  = White,  $_B$  = Black,  $_T$  = Total);  $\rho$  = mean corrected  $r$  for range restriction (RR) across 1,000 replications;  $\Delta\rho_{W-B}$  = mean corrected validity difference between the White and Black groups.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 13  
*The Influence of Differential Validity (.05) in the Population (% of Minority = 10%, N = 5,000)*

Unrestricted focal validity			Restricted (observed) validity						Correcting for RR using $u_W$ and $u_B$			Correcting for RR using unitary $u_T$		
W	B	W-B	SR	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$	$\rho_W$	$\rho_B$	$\Delta \rho_{W-B}$	$\rho_W$	$\rho_B$	$\Delta \rho_{W-B}$
.40	.35	.05	.1	.18	.044	.12	.304	.06	.40	.22	.18	.40	.20	.20
.40	.35	.05	.3	.22	.025	.16	.130	.07	.40	.32	.08	.41	.28	.13
.40	.35	.05	.5	.26	.019	.18	.086	.08	.40	.34	.06	.41	.29	.12
.40	.35	.05	.7	.30	.016	.21	.063	.09	.40	.35	.05	.41	.29	.12
.40	.35	.05	.9	.35	.014	.26	.049	.09	.40	.35	.05	.41	.31	.10
.45	.40	.05	.1	.21	.044	.14	.302	.07	.45	.26	.19	.46	.24	.22
.45	.40	.05	.3	.26	.025	.18	.129	.08	.45	.37	.08	.46	.32	.14
.45	.40	.05	.5	.30	.019	.21	.085	.09	.45	.39	.06	.46	.33	.13
.45	.40	.05	.7	.34	.016	.25	.063	.09	.45	.40	.05	.46	.34	.12
.45	.40	.05	.9	.40	.013	.30	.048	.10	.45	.40	.05	.46	.35	.11
.35	.40	-.05	.1	.15	.045	.14	.304	.02	.34	.26	.09	.35	.24	.11
.35	.40	-.05	.3	.19	.026	.18	.129	.01	.35	.37	-.02	.36	.32	.04
.35	.40	-.05	.5	.22	.020	.21	.084	.01	.35	.39	-.04	.36	.33	.03
.35	.40	-.05	.7	.26	.016	.25	.063	.01	.35	.40	-.05	.36	.34	.02
.35	.40	-.05	.9	.30	.014	.30	.048	.00	.35	.40	-.05	.36	.35	.01
.40	.45	-.05	.1	.18	.044	.16	.302	.02	.40	.30	.09	.40	.28	.13
.40	.45	-.05	.3	.22	.025	.20	.129	.02	.40	.41	-.02	.41	.36	.05
.40	.45	-.05	.5	.26	.019	.24	.083	.02	.40	.44	-.04	.41	.38	.03
.40	.45	-.05	.7	.30	.016	.28	.062	.02	.40	.45	-.05	.41	.38	.02
.40	.45	-.05	.9	.35	.014	.34	.047	.01	.40	.45	-.05	.41	.40	.01

Note. SR = selection ratio;  $r$  = mean validity across 1,000 replications;  $SE$  = standard error;  $u$  = range restriction ratio ( $_W$  = White,  $_B$  = Black,  $_T$  = Total);  $\rho$  = estimated mean corrected  $r$  for range restriction (RR) across 1,000 replications;  $\Delta \rho_{W-B}$  = mean corrected validity difference between the White and Black groups.

using national norm or pooled incumbent  $SDs$  instead of the unrestricted/applicant  $SD$ ). For example, national norm-based estimates might lead to underestimation of  $u$  (and overcorrection for range restriction), and pooled estimates across jobs

based on incumbent data could, at times, lead to overestimation of  $u$  (and undercorrection for range restriction). Again, understanding the dynamics of selection within subgroups would be important under these circumstances.

Table 14  
*The Influence of Differential Validity (.05) in the Population (% of Minority = 10%, N = 1 Million)*

Unrestricted focal validity			Restricted (observed) validity						Correcting for RR using $u_W$ and $u_B$			Correcting for RR using unitary $u_T$		
W	B	W-B	SR	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$	$\rho_W$	$\rho_B$	$\Delta \rho_{W-B}$	$\rho_W$	$\rho_B$	$\Delta \rho_{W-B}$
.40	.35	.05	.1	.18	.003	.13	.020	.05	.40	.35	.05	.41	.30	.11
.40	.35	.05	.3	.22	.002	.16	.009	.07	.40	.35	.05	.41	.30	.11
.40	.35	.05	.5	.26	.001	.18	.006	.08	.40	.35	.05	.41	.30	.11
.40	.35	.05	.7	.30	.001	.21	.005	.09	.40	.35	.05	.41	.30	.11
.40	.35	.05	.9	.35	.001	.26	.004	.09	.40	.35	.05	.41	.31	.10
.45	.40	.05	.1	.21	.003	.15	.020	.06	.45	.40	.05	.46	.35	.11
.45	.40	.05	.3	.26	.002	.18	.009	.07	.45	.40	.05	.46	.34	.12
.45	.40	.05	.5	.30	.001	.21	.006	.08	.45	.40	.05	.46	.34	.12
.45	.40	.05	.7	.34	.001	.25	.004	.09	.45	.40	.05	.46	.34	.12
.45	.40	.05	.9	.40	.001	.30	.003	.10	.45	.40	.05	.46	.35	.11
.35	.40	-.05	.1	.15	.003	.15	.019	.01	.35	.40	-.05	.36	.34	.01
.35	.40	-.05	.3	.19	.002	.18	.009	.01	.35	.40	-.05	.36	.34	.02
.35	.40	-.05	.5	.22	.001	.21	.006	.01	.35	.40	-.05	.36	.34	.02
.35	.40	-.05	.7	.26	.001	.25	.004	.01	.35	.40	-.05	.36	.34	.02
.35	.40	-.05	.9	.30	.001	.30	.003	.00	.35	.40	-.05	.36	.35	.00
.40	.45	-.05	.1	.18	.003	.17	.020	.01	.40	.45	-.05	.41	.39	.02
.40	.45	-.05	.3	.22	.002	.21	.009	.01	.40	.45	-.05	.41	.39	.02
.40	.45	-.05	.5	.26	.001	.24	.006	.02	.40	.45	-.05	.41	.39	.02
.40	.45	-.05	.7	.30	.001	.28	.004	.02	.40	.45	-.05	.41	.39	.02
.40	.45	-.05	.9	.35	.001	.34	.003	.01	.40	.45	-.05	.41	.40	.01

Note. SR = selection ratio;  $r$  = mean validity across 1,000 replications;  $SE$  = standard error;  $u$  = range restriction ratio ( $_W$  = White,  $_B$  = Black,  $_T$  = Total);  $\rho$  = estimated mean corrected  $r$  for range restriction (RR) across 1,000 replications;  $\Delta \rho_{W-B}$  = mean corrected validity difference between the White and Black groups.



Table 15  
*The Influence of Differential Validity (.05) in the Population (% of Minority = 20%, N = 5,000)*

Unrestricted focal validity			Restricted (observed) validity						Correcting for RR using $u_W$ and $u_B$			Correcting for RR using unitary $u_T$		
W	B	W-B	SR	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$	$\rho_W$	$\rho_B$	$\Delta\rho_{W-B}$	$\rho_W$	$\rho_B$	$\Delta\rho_{W-B}$
.40	.35	.05	.1	.18	.045	.13	.192	.05	.40	.28	.11	.41	.26	.16
.40	.35	.05	.3	.23	.026	.16	.086	.07	.40	.34	.06	.42	.29	.12
.40	.35	.05	.5	.26	.020	.18	.058	.08	.40	.35	.05	.42	.30	.12
.40	.35	.05	.7	.30	.016	.22	.044	.09	.40	.35	.05	.42	.30	.11
.40	.35	.05	.9	.35	.014	.26	.034	.09	.40	.35	.05	.41	.31	.10
.45	.40	.05	.1	.21	.045	.14	.191	.07	.45	.32	.13	.46	.29	.18
.45	.40	.05	.3	.26	.025	.18	.086	.08	.45	.39	.06	.47	.34	.13
.45	.40	.05	.5	.30	.019	.21	.058	.09	.45	.40	.05	.47	.34	.13
.45	.40	.05	.7	.35	.016	.25	.044	.10	.45	.40	.05	.47	.35	.12
.45	.40	.05	.9	.40	.014	.30	.033	.09	.45	.40	.05	.47	.36	.11
.35	.40	-.05	.1	.16	.045	.14	.191	.01	.35	.32	.03	.36	.29	.07
.35	.40	-.05	.3	.20	.026	.18	.087	.01	.35	.39	-.04	.37	.34	.03
.35	.40	-.05	.5	.23	.020	.22	.058	.01	.35	.40	-.05	.37	.34	.02
.35	.40	-.05	.7	.26	.017	.25	.043	.01	.35	.40	-.05	.37	.35	.02
.35	.40	-.05	.9	.31	.015	.30	.033	.00	.35	.40	-.05	.36	.36	.00
.40	.45	-.05	.1	.18	.045	.17	.191	.02	.40	.37	.03	.41	.33	.08
.40	.45	-.05	.3	.23	.026	.21	.085	.02	.40	.44	-.04	.42	.38	.03
.40	.45	-.05	.5	.26	.020	.25	.057	.02	.40	.44	-.05	.42	.39	.03
.40	.45	-.05	.7	.30	.017	.29	.043	.02	.40	.45	-.05	.42	.40	.02
.40	.45	-.05	.9	.35	.015	.35	.032	.01	.40	.45	-.05	.41	.41	.01

Note. SR = selection ratio;  $r$  = mean validity across 1,000 replications;  $SE$  = standard error;  $u$  = range restriction ratio ( $_W$  = White,  $_B$  = Black,  $_T$  = Total);  $\rho$  = estimated mean corrected  $r$  for range restriction (RR) across 1,000 replications;  $\Delta\rho_{W-B}$  = mean corrected validity difference between the White and Black groups.

**Implications for Theory and Practice**

Our results also have implications for interpreting the wider differential validity literature. [Katzell and Dyer \(1977\)](#) noted in the

title to their article that differential validity was being “revived.” Commenting on this debate, [Linn \(1978\)](#) suggested that there was a large amount of evidence that differential validity differences were so small as not to be of practical significance or to require

Table 16  
*The Influence of Differential Validity (.05) in the Population (% of Minority = 20%, N = 1 Million)*

Unrestricted focal validity			Restricted (observed) validity						Correcting for RR using $u_W$ and $u_B$			Correcting for RR using unitary $u_T$		
W	B	W-B	SR	$r_W$	$SE_W$	$r_B$	$SE_B$	$\Delta r_{W-B}$	$\rho_W$	$\rho_B$	$\Delta\rho_{W-B}$	$\rho_W$	$\rho_B$	$\Delta\rho_{W-B}$
.40	.35	.05	.1	.18	.003	.13	.013	.05	.40	.35	.05	.42	.31	.11
.40	.35	.05	.3	.23	.002	.16	.006	.07	.40	.35	.05	.42	.30	.11
.40	.35	.05	.5	.26	.001	.19	.004	.08	.40	.35	.05	.42	.30	.12
.40	.35	.05	.7	.30	.001	.22	.003	.09	.40	.35	.05	.42	.31	.11
.40	.35	.05	.9	.35	.001	.26	.002	.09	.40	.35	.05	.41	.31	.10
.45	.40	.05	.1	.21	.003	.15	.013	.06	.45	.40	.05	.47	.35	.12
.45	.40	.05	.3	.26	.002	.18	.006	.07	.45	.40	.05	.47	.35	.12
.45	.40	.05	.5	.30	.001	.22	.004	.09	.45	.40	.05	.47	.35	.12
.45	.40	.05	.7	.35	.001	.25	.003	.09	.45	.40	.05	.47	.35	.12
.45	.40	.05	.9	.40	.001	.30	.003	.09	.45	.40	.05	.47	.36	.10
.35	.40	-.05	.1	.16	.003	.15	.013	.01	.35	.40	-.05	.36	.35	.01
.35	.40	-.05	.3	.20	.002	.18	.006	.01	.35	.40	-.05	.37	.35	.02
.35	.40	-.05	.5	.23	.001	.22	.004	.01	.35	.40	-.05	.37	.35	.02
.35	.40	-.05	.7	.26	.001	.25	.003	.01	.35	.40	-.05	.37	.35	.02
.35	.40	-.05	.9	.31	.001	.30	.002	.00	.35	.40	-.05	.36	.36	.00
.40	.45	-.05	.1	.18	.003	.17	.013	.01	.40	.45	-.05	.42	.40	.02
.40	.45	-.05	.3	.23	.002	.21	.006	.01	.40	.45	-.05	.42	.39	.02
.40	.45	-.05	.5	.26	.001	.25	.004	.02	.40	.45	-.05	.42	.39	.03
.40	.45	-.05	.7	.31	.001	.29	.003	.02	.40	.45	-.05	.42	.40	.02
.40	.45	-.05	.9	.35	.001	.35	.002	.01	.40	.45	-.05	.41	.41	.01

Note. SR = selection ratio;  $r$  = mean validity across 1,000 replications;  $SE$  = standard error;  $u$  = range restriction ratio ( $_W$  = White,  $_B$  = Black,  $_T$  = Total);  $\rho$  = estimated mean corrected  $r$  for range restriction (RR) across 1,000 replications;  $\Delta\rho_{W-B}$  = mean corrected validity difference between the White and Black groups.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

separate analyses (see p. 510). Other researchers have suggested differential validity has been largely discredited and is a distraction (Bartlett et al., 1978; Bobko & Bartlett, 1978; Hunter et al., 1979; Schmidt et al., 1973; Schmidt & Hunter, 1981). Our results appear to be consistent with the latter position. We were fortunate to bring some empirical evidence and several simulations to bear on this issue based on recent meta-analytic evidence, which was used to design population matrices for analysis (Roth et al., 2011). This way, we could be reasonably sure of the starting values used in Studies 2 and 3, and then we could systematically demonstrate how range restriction would influence differential validity results (see also Study 4).

We also note that our results do not explain the level of differential validity in military settings. Berry et al. (2011) found larger levels of differential validity of .34 for Whites and .17 for Blacks for predicting primarily training performance. It is possible that the military is a special type of organization, especially when one is studying selection-related issues. This may be due to the use of cognitive ability measures in both selection and job classification. That is, many military occupational specialties are allotted a higher number of high scorers on cognitive ability tests (e.g., electronic repair shops) than are other specialties. This two-step or double range restriction process could be partially responsible for the apparent differential validity in military studies. Alternatively, there may be a moderator effect for measures of job performance (often used in employment settings) versus training performance (often used in Berry et al.'s analysis of military settings).

Our results also affirm certain common practices in simulations research. It is relatively common in such research to start with a matrix of validities and standardized ethnic group differences (e.g., De Corte, Lievens, & Sackett, 2006, 2007). Such analyses can then continue into various selection procedures, such as multistage selection (e.g., Finch, Edwards, & Wallace, 2009; Sackett & Roth, 1996), composite formation (e.g., Potosky, Bobko, & Roth, 2005; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997), the role of range restriction (e.g., Roth et al., 2011), or the planning of selection systems (Ployhart & Holtz, 2008). Our findings suggest that there is no need for such simulations to consider the issue of differential validity in the population.

Our results also have implications for practitioners. Although there are commentaries consistent with the existence of differential validity (e.g., Helms, 2011; Tonowski, 2011), our results are consistent with the SIOP *Principles* (Society for Industrial and Organizational Psychology, 2003) that urge a focus, if needed, on differential prediction (e.g., moderated multiple regression analyses; Bartlett et al., 1978) rather than differential validity (e.g., the analysis of correlation coefficients for subgroups). In fact, it appears that differential validity was not emphasized in the *Principles*. Our results suggest this is appropriate, given that observed differential validity may be largely due to the influence of the research artifact of range restriction. In particular, when practitioners are concerned with the effects of ethnic group differences (and fairness) in a particular sample, we recommend the use of moderated multiple regression (Bartlett et al., 1978).

Furthermore, if practitioners feel they must investigate the issue of differential validity, such investigation will likely require fairly large samples (see our footnote above for one approach to estimation) and the calculation of individual validity coefficients and within-subgroup *SDs*. Range restriction correction formulae are

available in a number of places and are important aspects of such investigations (e.g., Sackett & Yang, 2000; Schmidt, Oh, & Le, 2006; see also Bobko, Roth, & Bobko, 2001), although we are aware of the difficulty in obtaining appropriate unrestricted *SDs*. We strongly suggest that practitioners correct validities using a within-groups approach. This approach takes into account differential range restriction with the various subgroups and allows this restriction to be accurately modeled. In contrast, we caution against across-groups corrections, which do not appear to correctly model the dynamics of range restriction in many samples.

## Limitations

We note several potential limitations of our research. For one, some of our conclusions are based on simulations, and simulation results are only as good as the input values that underlie the generation of the data. We made efforts to tie our population estimates (*r* and *d*) to the meta-analytic literature and to studies that did not suffer from range restriction. We also simulated a range of situations with respect to validity and selection ratio. However, it is possible that our findings might not apply to certain atypical selection situations we could not anticipate.

Furthermore, we examined the differential validity of cognitive ability tests only with respect to ethnicity. The findings of this and previous research might not generalize to other predictors (e.g., structured interviews, work sample tests) or to other subgroup comparisons (e.g., sex, age). Our findings do not explain observed validity differences in military settings.

We modeled only direct range restriction. That is, we modeled how using cognitive ability tests alone within selection might result in differential validity. Yet, there are also likely sources of indirect range restriction in many data sets. For example, organizations might use a multiple hurdle selection system in which a battery of cognitively related tests was administered first and an interview was administered later. In this case, the interview score could indirectly restrict the range of cognitive ability scores in employment settings. If this were the case, the restriction due to both direct and indirect restriction would likely be greater than what we modeled. We did not examine the issue of differential reliability, but, psychometrically speaking, test score reliability could be lower for Blacks due to more restricted variance in test scores.

Finally, we were also asked by reviewers to reiterate that our simulations in Study 2 assumed a unitary estimate of validity in the population. Our simulations in Study 4 did simulate differential validity in the population. However, we again urge caution in interpreting Study 4 results, given that all the variance in organizational and educational settings could be attributed to artifacts and that quality data to inform our starting values were lacking.

## Future Research Directions

There are some avenues of future research related to differential validity. One line of research might involve high-quality empirical studies. We were struck by the lack of high-quality studies in this area. We suspect that Berry et al. (2011) and many others feel the same way (and we truly empathize with their efforts to search for such data). Thus, empirical investigations could be useful, and we urge care in conducting such studies. Some data might be found where range restriction is minimized. In other cases, studies need to be carefully designed to facilitate accurate corrections for range restriction.

Future research also could examine more complex selection scenarios, such as multiple hurdle selection systems. For example, researchers might examine the use of a measure of cognitive ability and then a structured interview. Similarly, selection composites that might include cognitive ability tests and grade point average information might be used in educational settings. These selection systems could be important, as both direct and indirect range restriction is modeled. Values for such simulations could be drawn from other simulation studies (e.g., Roth et al., 2011; Sackett, De Corte, & Lievens, 2010). Further, we urge that researchers simulate large numbers of applicants in order not to confound differential range restriction (or other artifacts) with small sample issues, as the number of minority hires can be very small under some selection conditions (see Footnote 2 for more details).

Future research also might involve taking effort that might be addressed toward examining differential validity and applying it elsewhere in staffing/selection research. For example, we still know comparatively little about construct-based, unrestricted gender and ethnic differences in validity and standardized group validities for employment interviews and situational judgment tests (Whetzel, McDaniel, & Nguyen, 2008).

Finally, our simulation results did not account for the amount of observed differential validity in military settings, and thus this topic may deserve further research. One line of such research would be to examine the use of cognitive ability tests in both selection and classification processes. It is also possible that military studies used training grades as a criterion, and this may have implications for validity studies.

## Conclusion

Belief in the existence of differential validity for cognitive tests was revived in the 1970s, and it appears to have been “re-revived” by a recent meta-analysis. Our results suggest that the lack of correction for range restriction in the recent meta-analysis is problematic and that researchers can expect to observe differential validity in the neighborhood of .03 to .07 for cognitive ability tests largely on the basis of range restriction (even when there is no such validity difference in the population). At this time, it appears differential range restriction can explain the observed differential validity seen in previous analyses of organizational and educational data.

## References

- Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in pre-employment testing. *Journal of Applied Psychology, 95*, 648–680. doi:10.1037/a0018714
- Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2011). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management, 37*, 5–38. doi:10.1177/0149206310377113
- Aguinis, H., Pierce, C. A., Bosco, F. A., Dalton, D. R., & Dalton, C. M. (2011). Debunking myths and urban legends about meta-analysis. *Organizational Research Methods, 14*, 306–331. doi:10.1177/1094428110375720
- Aguinis, H., & Smith, M. A. (2007). Understanding the impact of test validity and bias on selection errors and adverse impact. *Personnel Psychology, 60*, 165–199. doi:10.1111/j.1744-6570.2007.00069.x
- Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82*, 192–206. doi:10.1037/0021-9010.82.1.192
- Arvey, R. D. (1979). Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin, 86*, 736–765. doi:10.1037/0033-2909.86.4.736
- Arvey, R. D., & Faley, R. H. (1988). *Fairness in selecting employees* (2nd ed.). Reading, MA: Addison Wesley.
- Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. L. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31*, 233–241. doi:10.1111/j.1744-6570.1978.tb00442.x
- Bartlett, C. J., Bobko, P., & Pine, S. M. (1977). Single-group validity: Fallacy of the facts? *Journal of Applied Psychology, 62*, 155–157. doi:10.1037/0021-9010.62.2.155
- Berry, C. M., Clark, M. A., & McClure, T. K. (2011). Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology, 96*, 881–906. doi:10.1037/a0023222
- Berry, C. M., Sackett, P. R., & Landers, R. N. (2007). Revisiting interview–cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology, 60*, 837–874. doi:10.1111/j.1744-6570.2007.00093.x
- Bobko, P., & Bartlett, C. J. (1978). Subgroup validities: Differential definitions and differential prediction. *Journal of Applied Psychology, 63*, 12–14. doi:10.1037/0021-9010.63.1.12
- Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of *d* for range restriction and unreliability. *Organizational Research Methods, 4*, 46–61. doi:10.1177/109442810141003
- Bobko, P., & Schemmer, M. (1980). Note on standardized regression estimates. *Psychological Bulletin, 88*, 233–236. doi:10.1037/0033-2909.88.1.233
- Boehm, V. R. (1977). Differential prediction: A methodological artifact? *Journal of Applied Psychology, 62*, 146–154. doi:10.1037/0021-9010.62.2.146
- Boehm, V. R. (1978). Populations, preselection, and practicalities: A reply to Hunter and Schmidt. *Journal of Applied Psychology, 63*, 15–18. doi:10.1037/0021-9010.63.1.15
- De Corte, W. (1999). Weighing 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. doi:10.1037/0021-9010.84.5.695
- De Corte, W., Lievens, F., & Sackett, P. R. (2006). Predicting adverse impact and mean criterion performance in multistage selection. *Journal of Applied Psychology, 91*, 523–537. doi:10.1037/0021-9010.91.3.523
- De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92*, 1380–1393. doi:10.1037/0021-9010.92.5.1380
- Finch, D. M., Edwards, B. D., & Wallace, C. J. (2009). Multistage selection strategies: Simulating the effects on adverse impact and expected performance for various predictor combinations. *Journal of Applied Psychology, 94*, 318–340.
- Fincher, C. (1975). Differential validity and test bias. *Personnel Psychology, 28*, 481–500. doi:10.1111/j.1744-6570.1975.tb01387.x
- Fox, H., & Lefkowitz, J. (1974). Differential validity: Ethnic group as a moderator in predicting job performance. *Personnel Psychology, 27*, 209–223. doi:10.1111/j.1744-6570.1974.tb01529.x
- Gardner, D. G., & Deadrick, D. L. (2012). Moderation of selection procedure validity by employee race. *Journal of Managerial Psychology, 27*, 365–382. doi:10.1108/02683941211220180
- Helms, J. E. (2012). A legacy of eugenics underlies racial-group comparisons in intelligence testing. *Industrial and Organizational Psychology, 5*, 176–179. doi:10.1111/j.1754-9434.2012.01426.x

- Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society, Series B*, 15, 193–225.
- Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, 29, 340–362. doi:10.1016/0001-8791(86)90013-8
- Hunter, J. E., & Schmidt, F. L. (1978). Differential and single-group validity of employment tests by race: A critical analysis of three recent studies. *Journal of Applied Psychology*, 63, 1–11. doi:10.1037/0021-9010.63.1.1
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting for error and bias in research findings* (2nd ed). Newbury Park, NJ: Sage.
- Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, 86, 721–735. doi:10.1037/0033-2909.86.4.721
- Katzell, R. A., & Dyer, F. J. (1977). Differential validity revived. *Journal of Applied Psychology*, 62, 137–145. doi:10.1037/0021-9010.62.2.137
- Katzell, R. A., & Dyer, F. J. (1978). On differential validity and bias. *Journal of Applied Psychology*, 63, 19–21. doi:10.1037/0021-9010.63.1.19
- Kirchner, W. K. (1975). Some questions about “Differential validity: Ethnic group as a moderator in predicting job performance”. *Personnel Psychology*, 28, 341–343. doi:10.1111/j.1744-6570.1975.tb01541.x
- Lawshe, C. H. (1987). Adverse impact: Is it a viable concept? *Professional Psychology: Research and Practice*, 18, 492–497. doi:10.1037/0735-7028.18.5.492
- Lefkowitz, J., & Fox, H. (1975). Some answers to “Some questions about ‘Differential validity: Ethnic group as a moderator in predicting job performance.’” *Personnel Psychology*, 28, 345–349. doi:10.1111/j.1744-6570.1975.tb01542.x
- Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology*, 63, 507–512. doi:10.1037/0021-9010.63.4.507
- Mattern, K. D., & Patterson, B. F. (2013). Test of slope and intercept bias in college admissions: A response to Aguinis, Culpepper, and Pierce (2010). *Journal of Applied Psychology*, 98, 134–147. doi:10.1037/a0030610
- McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology*, 59, 927–953. doi:10.1111/j.1744-6570.2006.00059.x
- Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing race/ethnicity and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, 61, 153–172. doi:10.1111/j.1744-6570.2008.00109.x
- Potosky, D. P., Bobko, P., & Roth, P. L. (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. doi:10.1111/j.1468-2389.2005.00327.x
- Reilly, R. R., & Warech, M. A. (1993). The validity and fairness of alternatives to cognitive ability tests. In L. Wing & B. Gifford (Eds.), *Policy issues in employment testing* (pp. 131–224). Boston, MA: Kluwer.
- Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, 54, 297–330. doi:10.1111/j.1744-6570.2001.tb00094.x
- Roth, P. L., Switzer, F. S., Van Iddekinge, C. H., & Oh, I.-S. (2011). Toward better meta-analytic matrices: How input values can affect research conclusions in human resource management simulations. *Personnel Psychology*, 64, 899–935. doi:10.1111/j.1744-6570.2011.01231.x
- Sackett, P. R., De Corte, W., & Lievens, F. (2010). Decision aids for addressing the validity–adverse impact trade-off. In J. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 453–472). New York, NY: Routledge.
- Sackett, P. R., & Roth, L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology*, 49, 549–572. doi:10.1111/j.1744-6570.1996.tb01584.x
- Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. E. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, 56, 302–318. doi:10.1037/0003-066X.56.4.302
- Sackett, P. R., & Shen, W. (2010). Subgroup differences on cognitive tests in contexts other than personnel selection. In J. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 323–346). New York, NY: Routledge.
- Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, 85, 112–118. doi:10.1037/0021-9010.85.1.112
- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & de Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology*, 56, 573–605. doi:10.1111/j.1744-6570.2003.tb00751.x
- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology*, 88, 1068–1081. doi:10.1037/0021-9010.88.6.1068
- Schmidt, F. L., Berner, J. G., & Hunter, J. E. (1973). Racial differences in validity of employment tests: Reality or illusion? *Journal of Applied Psychology*, 58, 5–9. doi:10.1037/h0035408
- Schmidt, F. L., & Hunter, J. E. (1980). The future of criterion-related validity. *Personnel Psychology*, 33, 41–60. doi:10.1111/j.1744-6570.1980.tb02163x
- Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist*, 36, 1128–1137. doi:10.1037/0003-066X.36.10.1128
- Schmidt, F. L., Oh, I.-S., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology*, 59, 281–305.
- Schmidt, F. L., Shaffer, J. A., & Oh, I.-S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology*, 61, 827–868. doi:10.1111/j.1744-6570.2008.00132.x
- Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology*, 82, 719–730. doi:10.1037/0021-9010.82.5.719
- Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures*. Bowling Green, KY: Author.
- Synk, D. J., & Swarthout, D. (1987). *Comparison of Black and nonminority validities for the General Aptitude Test battery* (Research Report No. 51). Washington, DC: U.S. Department of Labor.
- Tonowski, R. F. (2011). The *Uniform Guidelines* and personnel selection: Identify and fix the right problem. *Industrial and Organizational Psychology*, 4, 521–525. doi:10.1111/j.1754-9434.2011.01384.x
- Whetzel, D., McDaniel, M., & Nguyen, N. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, 21, 291–308.
- Young, J. W. (1994). Differential prediction of college grades by gender and by ethnicity: A replication study. *Educational and Psychological Measurement*, 54, 1022–1029. doi:10.1177/0013164494054004019

Received January 12, 2012

Revision received May 20, 2013

Accepted June 14, 2013 ■