Journal of Applied Psychology
2010, Vol. 95, No. 4, 648–680

0021-9010/10/$12.00 DOI: 10.1037/a0018714

Revival of Test Bias Research in Preemployment Testing

Herman Aguinis
Indiana University

Steven A. Culpepper
University of Colorado Denver

Charles A. Pierce
University of Memphis

We developed a new analytic proof and conducted Monte Carlo simulations to assess the effects of methodological and statistical artifacts on the relative accuracy of intercept- and slope-based test bias assessment. The main simulation design included 3,185,000 unique combinations of a wide range of values for true intercept- and slope-based test bias, total sample size, proportion of minority group sample size to total sample size, predictor (i.e., preemployment test scores) and criterion (i.e., job performance) reliability, predictor range restriction, correlation between predictor scores and the dummy-coded grouping variable (e.g., ethnicity), and mean difference between predictor scores across groups. Results based on 15 billion 925 million individual samples of scores and more than 8 trillion 662 million individual scores raise questions about the established conclusion that test bias in preemployment testing is nonexistent and, if it exists, it only occurs regarding intercept-based differences that favor minority group members. Because of the prominence of test fairness in the popular media, legislation, and litigation, our results point to the need to revive test bias research in preemployment testing.

Keywords: selection fairness, testing practices, employee selection, human resource management, staffing

Few topics in industrial and organizational (I/O) psychology and human resource management have generated more media attention than bias in preemployment testing (e.g., Abel, 2007; P. B. Brown, 2007; Marzulli, 2008). In addition, the topic of test bias receives immense public scrutiny in the legislation and litigation arenas (e.g., Berk, 1982; *Cormier v. P.P.G. Indus.,* 1981; *Hamer v. City of Atlanta*, 1989; Mehrens & Popham, 1992; Outtz, 2002; Reynolds & Brown, 1984; *United States v. City of Erie*, 2005). Test bias, also labeled *predictive bias* or *differential prediction*, occurs when the "slope or intercepts of the regression line relating the predictor [i.e., preemployment test] to the criterion [some measure of subsequent success, usually job performance] are different for one group than for another" (Society for Industrial and Organizational Psychology [SIOP], 2003, p. 32). In other words, as noted by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, and NCME], 1999), "no bias exists if the regression equations relating the test and the criterion are indistinguishable for the groups in question" (p. 79). The media attention is usually accompanied by emotionally laden polemics because the issue of test bias is entangled in broader societal issues such as individual liberties, civil rights, and social justice (Oswald, Saad, & Sackett, 2000; Reynolds, 1995). Although there is evidence that, as a whole, research in I/O psychology does not address many human-capital trends of interest to society at large (Cascio & Aguinis, 2008a), the topic of test bias is an exception given the continuous scholarly attention it has received over a period of 4 decades (e.g., Cleary, 1968; Culpepper & Davenport, 2009; Van Iddekinge & Ployhart, 2008).

Test bias is one of the issues in I/O psychology on which most researchers agree because findings seem consistent. The consensus in I/O psychology and related fields (e.g., education, human resource management) concerned with high-stakes testing is that, in the instances when it exists, test bias is found regarding intercept differences between groups in the form of overprediction of scores for minority group members (i.e., smaller intercept for the ethnic minority group compared to the majority group), but no differences are found regarding slopes across groups (e.g., Cole, 1981; Houston & Novick, 1987; Humphreys, 1986; Hunter, Schmidt, & Rauschenberger, 1984; Kuncel & Sackett, 2007; Linn, 1978; Rotundo & Sackett, 1999; Rushton & Jensen, 2005; Sackett, Schmitt, Ellingson, & Kablin, 2001; Sackett & Wilk, 1994; Schmidt & Hunter, 1981, 1998; we provide a detailed technical description of the issue of intercept versus slope differences in the next section). This conclusion has been reached regarding selection tools used in both work and other organizational settings to assess a heterogeneous set of constructs ranging from general mental abilities (GMA; e.g., Hartigan & Wigdor, 1989) to personality (e.g., Cortina, Doherty, Schmitt, Kaufman, & Smith, 1992; Saad & Sackett, 2002) and safety suitability (Te Nijenhuis & Van der Flier, 2004). Moreover, a similar conclusion has been reached

Herman Aguinis, Department of Management and Entrepreneurship, Kelley School of Business, Indiana University; Steven A. Culpepper, Department of Mathematical and Statistical Sciences, University of Colorado Denver; Charles A. Pierce, Department of Management, Fogelman College of Business and Economics, University of Memphis.

We thank Patrick F. McKay and Paul R. Sackett for comments on previous drafts.

Correspondence concerning this article should be addressed to Herman Aguinis, Department of Management and Entrepreneurship, Kelley School of Business, Indiana University, 1309 East 10th Street, Bloomington, IN 47405-1701. E-mail: haguinis@indiana.edu

regardless of which ethnic minority groups are compared to the majority group involved in the assessment of possible test bias. For example, results are based on studies conducted in the United States comparing the majority (i.e., Whites) to Latinos (e.g., Schmidt, Pearlman, & Hunter, 1980) and African Americans (e.g., Bartlett, Bobko, Mosier, & Hannan, 1978; Hunter & Schmidt, 2000). Also, the same conclusion has been reached regarding majority and minority groups classified according to socioeconomic status (e.g., Canivez & Konold, 2001) and entirely different ethnicity classifications outside of the United States. Examples include research conducted in the Netherlands comparing groups of Native Dutch (comparison group) with Turks, North Africans, Surinamese, Netherlands Antilleans, and former Yugoslavs (e.g., Te Nijenhuis & Van der Flier, 2000, 2004); research conducted in Israel including groups of individuals born in Israel (comparison group) with those born in Eastern countries (i.e., mainly Middle Eastern Arab countries and North Africa) and in Western countries (i.e., mainly Eastern and Central Europe; Reeb, 1976); and research conducted in South Africa including groups of non-Black Africans (comparison group) and Black Africans (Rushton, Skuy, & Bons, 2004).

The evidence about overprediction (i.e., favoring) of minority members' performance due to differences in intercepts and lack of differences regarding slopes seems so consistent, particularly for GMA testing, that a review of 85 years of human resource selection research concluded that "for predictive fairness, the usual finding has been a lack of predictive bias for minorities and women" (Schmidt & Hunter, 1998, p. 272). In fact, an official publication of SIOP (2003), the *Principles for the Validation and Use of Personnel Selection Procedures*, asserts that

> predictive bias has been examined extensively in the cognitive ability domain. For White–African American and White–Hispanic comparisons, slope differences are rarely found; while intercept differences are not uncommon, they typically take the form of overprediction of minority group performance. (p. 32)

Similarly, textbooks in I/O psychology (e.g., Cascio & Aguinis, 2005; Landy & Conte, 2007) also conclude that "when prediction systems are compared, differences most frequently occur (if at all) in intercepts" (Cascio & Aguinis, 2005, p. 192). It is thus no exaggeration to assert that the conclusion that test bias generally does not exist but, when it exists, it involves intercept differences favoring minority group members and not slope differences, is an established fact in I/O psychology and related fields concerned with high-stakes testing.

In the present study, we raise important questions and cast doubt about the established conclusions regarding test bias in preemployment testing and provide an alternative explanation for the consistent results reported over the past 40 years of research. This is certainly a tall order, but in our study we provide evidence to demonstrate that, although published studies on test bias date back to the late 1960s (e.g., Bartlett & O'Leary, 1969; Cleary, 1968), there is insufficient evidence to infer that there is no slope-based bias in preemployment tests. Moreover, we provide evidence indicating that the finding of intercept-based differences favoring members of the minority group could be the consequence of methodological and statistical artifacts. Next, we provide a detailed technical description of the concept of test bias and the procedure for assessing the possible presence of test bias.

Test Bias: Definition and Assessment

Test bias, also labeled differential prediction or predictive bias (Aguinis & Smith, 2007; Van Iddekinge & Ployhart, 2008), exists when regression lines linking test and criterion scores differ across relevant comparison groups (AERA, APA, and NCME, 1999; SIOP, 2003). If there is test bias, regression lines can differ based on intercepts, slopes, or both. As noted in the *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2003) and *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999), fairness is a social rather than a psychometric concept and there are at least four different ways to define the concept of fairness. Alternatively, test bias is a psychometric issue and has been defined so clearly that it has been adopted by the Uniform Guidelines on Employee Selection Procedures (1978) and also by the courts (e.g., *Cormier v. P.P.G. Indus.*, 1981; *Hamer v. City of Atlanta*, 1989; *United States v. City of Erie*, 2005). Although distinct, fairness and bias are closely related because "given the acceptance of the principle of individualized treatment based on individual merit, it appears unfair to overpredict or underpredict the performance of any individual or group of individuals" (Schmidt & Hunter, 1974, p. 1).

Formally assessing the presence of test bias is usually conducted using Cleary's (1968) regression model (e.g., Aguinis, 2004a; Bartlett & O'Leary, 1969; Campbell, 1996; Darlington, 1971; Grant & Bray, 1970; Hough, Oswald, & Ployhart, 2001; Maxwell & Arvey, 1993; Saad & Sackett, 2002). As noted in the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1999),

> when empirical studies of differential prediction of a criterion of members of different subgroups are conducted, they should include equations (or an appropriate equivalent) computed separately for each group or treatment under consideration or an analysis in which the group or treatment variables are entered as moderator variables. (Standard 7.6, p. 82)

Similarly, as noted in the *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2003), "testing for predictive bias involves using moderated multiple regression, where the criterion measure is regressed on the predictor score, subgroup membership, and an interaction term between the two. Slope and/or intercept differences between subgroups indicate predictive bias" (p. 32). Thus, the procedure for assessing possible test bias includes computing a series of regression equations regressing a criterion Y (i.e., typically a measure of job performance) on a predictor X (i.e., preemployment test scores); a second predictor, also labeled a moderator G (i.e., dummy coded variable when ethnicity includes two categories, but in general the number of dummy vectors is $k - 1$, where k is the number of groups; Aguinis, 2004a, Chapter 8); and a third predictor that is the product of X by G (i.e., product term carrying information about the interactive effect of X and G on Y). The following models are obtained and then compared per the procedure described next (Cascio & Aguinis, 2005; Lautenschlager & Mendoza, 1986; Rotundo & Sackett, 1999; Saad & Sackett, 2002):

$$Y = b_0 + b_1X + e \quad (1)$$

$$Y = b_0 + b_1X + b_2G + b_3XG + e \quad (2)$$

$$Y = b_0 + b_1X + b_2G + e, \quad (3)$$

where b_0 is the intercept; b_1, b_2, and b_3 are unstandardized regression coefficients; and e is a random error term. The first step involves comparing Equation 1 versus Equation 2 to test the null hypothesis that the increase in proportion of variance explained in Y by adding predictors G and XG is not different from zero. If this omnibus null hypothesis is rejected, there is evidence of test bias. At this point we know there is test bias but do not know whether test bias is due to differences between intercepts (G), slopes (XG), or both. To formally assess whether there are slope-based differences across the groups, we compare the R^2 (i.e., proportion of variance explained in Y) resulting from Equation 3 versus the R^2 resulting from Equation 2 (i.e., H_0: $\Delta\psi^2_{\text{slope}} = 0$). We use the symbol $\Delta R^2_{\text{slope}}$ to refer to the sample-based difference between R^2s. To formally assess whether there are intercept-based differences, we compare the R^2 resulting from Equation (3) versus the R^2 resulting from Equation 1 (i.e., H_0: $\Delta\psi^2_{\text{intercept}} = 0$). We use the symbol $\Delta R^2_{\text{intercept}}$ to refer to the sample-based difference between R^2s. In practice, if the null hypothesis $\Delta\psi^2_{\text{slope}} = 0$ is rejected, then one reaches the conclusion that test bias exists and there may not be a need to proceed with testing the null hypothesis $\Delta\psi^2_{\text{intercept}} = 0$ (unless one is interested in knowing whether overall bias is not only caused by slope differences but also by intercept differences). In our study, we are interested in the relative accuracy of the tests of each of these two null hypotheses, so we will assess the accuracy of conclusions regarding H_0: $\Delta\psi^2_{\text{intercept}} = 0$ regardless of conclusions regarding H_0: $\Delta\psi^2_{\text{slope}} = 0$.

Present Study

The established conclusion based on 40 years of research is that test bias is not found regarding slopes. When test bias is found, it is about differences in intercepts, but not in slopes, across groups. Moreover, in most cases, the direction of the intercept difference is such that minority group scores are overpredicted (i.e., larger intercept for the majority group). However, test bias also exists when the prediction of criteria is different across groups such that the minority group benefits from overprediction; lawsuits regarding reverse discrimination such as the *Ricci v. DeStefano* (2009) U.S. Supreme Court case are based precisely on this logic because both majority and minority applicants are protected under Title VII of the Civil Rights Act of 1964. In this study, we raise questions and cast doubt on these established conclusions about test bias in preemployment testing based on methodological and substantive reasons.

Reasons Why Slope-Based Differences Are Likely to Exist but Are Not Found

From a methodological perspective, research based on analytic developments, Monte Carlo simulations, and literature reviews has revealed that conclusions regarding the absence of slope differences across groups may not be warranted. That is, statistical power (i.e., the probability of detecting a slope-based difference across groups in the sample when it exists in the population) is typically insufficient. In practical terms, low power affects test bias assessment in that, if true differences exist, one may conclude incorrectly that a selection procedure predicts outcomes equally well for various groups—that is, that there are no slope differences. However, this sample-based conclusion may be incorrect. In fact, the selection procedure actually may predict outcomes differentially across subgroups. Such differential prediction may not be detected, however, because of the low statistical power inherent in test validation research (Schmidt & Hunter, 1981). Monte Carlo simulations (e.g., Aguinis & Stone-Romero, 1997) and comprehensive literature reviews (e.g., Aguinis, Beaty, Boik, & Pierce, 2005) conducted over the past decade lead to the conclusion that, unfortunately, much of the research accumulated over the past 40 years has attempted to test the null hypothesis of no differential prediction on the basis of studies too weak to detect possible differences (Katzell & Dyer, 1977). This body of research suggests that, even when slope-based test bias may be quite large in the population, the size of the sample-based observed effects is smaller due to the presence of statistical and methodological artifacts such as measurement error and range restriction (Aguinis et al., 2005). Thus, the concern is not only with statistical power and null hypothesis significance testing in general but also with the size of the observed slope-based test bias in relationship to its population counterpart.

An important culprit for low statistical power is that most validation research studies are conducted using small samples (Aguinis, 1995, 2004a). For example, Lent, Auerbach, and Levin (1971) reviewed 406 validity studies published in *Personnel Psychology* between 1954 and 1969 and found that the median sample size was 68. Monahan and Muchinsky (1983) reviewed articles in the human resource selection domain published in *Personnel Psychology* between 1950 and 1979 and found that the mean sample size for nine occupational groups ranged from 58 to 125, and the mean sample size for all occupational groups was 88. Salgado (1998) reported that the median sample size for 86 criterion-related validity studies published in the *Journal of Applied Psychology*, the *Journal of Occupational and Organizational Psychology*, and *Personnel Psychology* between 1983 and 1994 was 113. Russell et al. (1994) reported a median sample size of 103 for all validation studies of human resource selection systems published between 1964 and 1992 in the *Journal of Applied Psychology* and *Personnel Psychology*. Thus, sample sizes used in the vast majority of peer-reviewed human resource selection research are not sufficiently large to detect slope-based test bias (cf. Aguinis et al., 2005; Aguinis & Stone-Romero, 1997). Note that these are sample sizes in validation research in general, and not specifically in the subsumed area of test bias. Thus, our simulation study includes sample sizes that are as much as 10 times larger than these values.

Insufficient statistical power results from the use of small samples but is also due to the interactive effects of various statistical and methodological artifacts such as range restriction and unequal number of individuals across subgroups (Aguinis & Stone-Romero, 1997). Therefore, even differential prediction studies including very large samples may suffer from insufficient statistical power to detect slope-based differences (Aguinis et al., 2005). For example, unequal samples sizes across groups has an important detrimental effect on power because the effective total sample size for two independent-sample tests is the harmonic mean of the two subgroup sample sizes (Hsu, 1993). Also, measurement error (Busemeyer & Jones, 1983; Dunlap & Kemery, 1988; Evans, 1985) and range restriction in test scores (Aguinis & Stone-Romero, 1997) have important detrimental effects on statistical power. Most, if not all, differential prediction studies include more members of the majority than the minority group, restricted test

scores due to selection, and less than perfectly reliable measures for the predictor and criterion scores. Thus, although there may be an assumption that an unusually large sample size guarantees sufficient statistical power, this may not be the case.

As an example, consider a study by Rotundo and Sackett (1999). These authors conducted differential prediction analyses of data collected by the U.S. Employment Service during the years 1972–1987 with the goal of gathering concurrent validity evidence for the General Aptitude Test Battery as a predictor of supervisory ratings of performance. Rotundo and Sackett did not find evidence of slope-based test bias and stated that

> the sample size used in the present study was double the largest tabled value in the Stone-Romero and Anderson article, and the predictor reliabilities were in the .80 to .90 range. . . . We suspect that the power to detect a small effect size in the present study would be reasonably high. (p. 821)

Using information from their article, we were able to compute the statistical power of the Rotundo and Sackett analysis with the Aguinis, Boik, and Pierce (2001) analytically derived power approximation (equations used in calculating statistical power are included in Appendix A and the computer program is available at http://mypage.iu.edu/~haguinis/). In terms of input data for the power calculation, in the first analysis sample size was 17,020 for Whites and 1,212 for African Americans. We used the observed validity coefficients (.15 for Whites and .10 for African Americans), observed standard deviations for predictor scores (.94 for Whites and .83 for African Americans), and observed standard deviations for criterion scores (0.99 for Whites and 1.02 for African Americans). Test score reliability was not reported precisely, but we set it to .90 for both groups given that Rotundo and Sackett stated that "predictor reliabilities were in the .80 to .90 range" (p. 821). Reliability of criterion scores was not reported, but we set it to .52 for both groups given that this was a six-item measure of supervisory ratings (cf. Viswesvaran, Ones, & Schmidt, 1996). Range restriction was not reported, but we set it to .50 given that selection ratios of .50 or .60 seem to be normative for GMA tests (cf. Hunter & Hunter, 1984). The resulting statistical power was .101, which is much lower than the frequently used .80 benchmark (J. Cohen, 1988). In the second large sample size differential prediction analysis conducted by Rotundo and Sackett, input data included sample sizes of 17,020 for Whites and 6,296 for African Americans, observed validity coefficients (.15 for Whites and .14 for African Americans), observed standard deviations for predictor scores (.94 for Whites and .85 for African Americans), and observed standard deviations for criterion scores (.99 for Whites and .98 for African Americans). As in the first power analysis, we used .90 for predictor score reliabilities, .52 for criterion score reliabilities, and .50 for range restriction for both groups. The resulting statistical power was only .051, which means that the likelihood of detecting existing slope-based bias was only 5.1%.

Statistical power would increase to what is usually considered the acceptable level of .80 if some of the design and measurement characteristics are improved. For example, in the first analysis, increasing the sample size in the African American subgroup from 1,212 to 32,000, increasing the sample size in the Whites group from 17,020 to 90,000, and improving reliability in the criterion for both subgroups from .52 to .90 would yield a statistical power value of .80. In the second analysis, because the difference between subgroups is very small (i.e., difference in validity coefficients of only .01), even extreme improvements in terms of subgroup sample sizes and reliability would not lead to sufficient statistical power to detect slope-based test bias. For example, increasing the sample size of the African American subgroup from 6,296 to 50,000, increasing the sample size of the White subgroup from 17,020 to 120,000, and improving criterion score reliabilities to .90 in both groups would yield a power value that is still lower than .10.

We used the observed correlations and standard deviations to compute the targeted test bias size in each of the two aforementioned power analyses because these are meaningful targeted effect sizes. In the first analysis, the difference between validity coefficients is .05, and in the second analysis the difference is only .01. Because effect size is one of the major determinants of statistical power, this difference explains in part why power is larger for the first analysis compared to the second one and also why improving design and measurement characteristics would have a substantial impact on the power in the first analysis only. Nevertheless, our ability to detect slope-based differences that are seemingly small can be meaningful in the context of human resource selection, particularly in situations involving thousands of people (Cortina & Landis, 2009). Specifically, Aguinis and Smith (2007, 2009) demonstrated that the percentage of false negatives and false positives due to using a common regression line in the presence of test bias (i.e., "bias-based false positives and false negatives") is very large in many cases even if test bias is perceived to be small. As noted by Sackett, De Corte, and Lievens (2009),

> the Aguinis and Smith [2007] approach distinguishes between prediction errors due to imperfect validity and error made due to treating biased test as if it were unbiased (e.g., using a common regression line when, in fact, the regression lines for the groups under consideration differ). (p. 468)

In the specific case of the Rotundo and Sackett (1999; Tables 2–3) data, we used the Aguinis and Smith (2007) online calculator, which is available at http://mypage.iu.edu/~haguinis/, to understand whether or not the effect sizes we targeted in our power analysis were sufficiently meaningful to be detected. Again, for the first analysis, the difference in correlations between groups was only .05, which would in most contexts be considered small (Cortina & Landis, 2009). We used a desired selection cutoff score of 0 for the criterion given that the mean performance rating score (in standardized metric) is .03 for African Americans and .09 for Whites. Results indicate that if a common regression line was used instead of separate lines for each group, there would be 20.6% of false negatives in the African American subgroup and 1.42% of false positives in the White subgroup. In other words, given the size of the samples in the Rotundo and Sackett study, about 250 African Americans (out of a total of 1,212) would be denied employment incorrectly and about 242 (out of a total of 17,020) White applicants would be offered employment incorrectly. For the second analysis, for which there is a difference of only .01 in validity coefficients across groups, we also used a desired selection cutoff of 0 for the criterion as input in the Aguinis and Smith (2007) calculator. Errors due to using a common regression line instead of the subgroup-based regression lines would lead to 18.02% of false positives for the African American group and

14.4% of false negatives for the White group. Given the sample sizes, about 1,134 African Americans (out of a total of 6,296) would be incorrectly offered employment and about 2,451 Whites (out of a total of 17,020) would be rejected incorrectly. These are large numbers and these incorrect decisions, in spite of being based on effect sizes perceived to be small, are practically meaningful for the individuals and organizations involved. From an ethical standpoint, it may be argued that even if one individual is denied an opportunity unfairly or given an opportunity unfairly, this is one too many. The problem is obviously magnified when thousands of individuals are misclassified if a common regression line is used in selection decisions in the presence of test bias (cf. Aguinis & Smith, 2007, 2009).

In addition to methodological reasons, we also expect to find slope-based test bias for several substantive reasons. The mechanisms leading to differences in mean test scores across groups are not necessarily the same as those leading to expected differences in slopes across groups. For example, Sackett, Kuncel, Arneson, Cooper, and Waters (2009) provided evidence that socioeconomic status is related to postsecondary admissions test scores but not to grades. So, socioeconomic status affects test scores but may not necessarily affect the relationship between test scores and criteria. In other words, although one may not be comfortable with the fact that higher socioeconomic status gives some applicants an advantage and one may not be comfortable with the associated societal consequences, "test scores contain meaningful information predictive of academic performance" (Sackett, Kuncel, et al., 2009, p. 2), but socioeconomic status does not necessarily moderate the relationship between test scores and outcomes. The expectation that there are slope-based differences across groups is not based on differences in socioeconomic status but, rather, on sociohistorical–cultural and social psychological explanations. Next, we provide examples of the types of mechanisms that may cause slope-based differences across ethnic-based groups.

The sociohistorical–cultural explanation relies on streams of literature outside of the field of psychology indicating that members of ethnic minority and ethnic majority groups do not share a similar cultural frame of reference and identity (Ogbu, 1993). Members of the minority group interpret discrimination against them as more or less permanent and institutionalized and develop "a folk theory of getting ahead which differs in some respects from that of Euro-Americans" (p. 495). These frames of reference have developed as a consequence of exclusion, segregation, and barriers in opportunity structure that lasted many generations. For example, several in-depth ethnographic studies reviewed by Ogbu (1993) suggest that African Americans do not believe they have the same chance of being successful compared to Euro-Americans with similar school credentials. In addition, minorities

> consciously and unconsciously perceive and interpret learning certain things and acting in certain ways they associate with their "oppressors," their "enemies," e.g., Euro-Americans, as threatening and therefore "resisted" . . . minorities perceive and interpret standard attitudes and behaviors in IQ and other test-taking situations as falling within the cultural frame of reference of Euro-Americans, not that of the minorities. (p. 501)

Moreover, there are family and community pressures to not "act White" (in the case of African American communities) or "act gringo" (in the case of Latino communities). This different cultural frame of reference leads some minority members to have lower expectancies regarding the probability that obtaining high test scores will lead to desirable rewards (Gould, 1999). In short, cultural frames of reference affect how tests and testing situations are interpreted. Thus, the meaning of test scores and the relationship between test scores and measures of performance are expected to differ across majority and minority ethnic groups (Grubb & Ollendick, 1986).

Social psychological explanations for why slope-based differences are expected across groups rely on the stereotype threat literature (Steele & Aronson, 1995; Walton & Spencer, 2009). Stereotype threat suggests that when the stereotype of a group to which a test taker belongs becomes salient, a test taker's concern about being evaluated negatively due to his or her placement in that group can lead to lowered levels of test performance. Although stereotype threat is only one of several contextual factors that affect test scores, it is "an important phenomenon with relevance to testing settings" (Sackett, Hardison, & Cullen, 2004, p. 11). In addition to affecting test scores, stereotype threat is likely to result in group-based slope differences because the effects of stereotype threat on minorities are unlikely to be identical on test (i.e., predictor) compared to performance (i.e., criterion) scores. As noted by R. P. Brown and Day (2006), "the extent to which stereotype threat influences predictive validity will depend on the degree to which stereotype threat differentially influences predictor and criterion scores" (p. 983). In short, differential effects of stereotype threat on test and criterion scores are expected to lead to slope-based differences for minority compared to majority group members.

Reasons Why Intercept-Based Differences Favoring Minority Group Members Are Found but Are Inflated or Not Likely to Exist

There is widespread consensus that "the consistent finding is overprediction [of performance], rather than underprediction, for Black and Hispanic students. . . . Findings for Blacks and Hispanics in the employment domain parallel those in educational admissions" (Sackett, Borneman, & Connelly, 2008, p. 223). In other words, if differences are found regarding intercepts, they are such that the intercept for the majority group is larger than the intercept for the minority group, favoring minority group applicants because, if a common regression line is used, minority group members' scores are overpredicted. In the present study, we cast doubt on this consistent finding regarding intercept differences favoring minority group applicants because of inflated Type I error rates. In other words, one may conclude that there are differences when they in fact they do not exist. Also, if a difference exists in the population, one may conclude that this difference is larger than it actually is (Terris, 1997). So, our analysis is based on null hypothesis statistical significance testing but also on the differences between population and observed intercept-based test bias.

Although they did not provide analytic or empirical proof, Linn and Werts (1971), and more recently Terris (1997), illustrated with examples and graphs that the test for intercept differences testing the null hypothesis H_0: $\Delta\psi^2_{\text{intercept}} = 0$ is likely to have inflated Type I error rates when test scores are measured with error (i.e., $\rho_{XX} < 1$) and when the test is correlated with the grouping variable (i.e., $\rho_{XG} > 0$). In addition, although not mentioned by Linn and

Werts, range restriction also interacts with slope-based differences across groups to affect Type I error rates. A correlation between X and G indicates a difference between groups regarding test scores (i.e., $\Delta\mu = \mu_{1x} - \mu_{0x} \neq 0$). Each of these conditions, test reliability less than 1.0, test score mean differences between minority and majority groups in favor of the majority group, and range restriction are the norm in the area of GMA testing (e.g., Roth, BeVier, Bobko, Switzer, & Tyler, 2001) and frequently observed when other types of preemployment testing are used (Hough & Oswald, 2000; Hough et al., 2001).

Appendix B includes a new analytic proof to explain why lack of perfect test score reliability and differences in test scores between groups in favor of the majority group can lead to the conclusion that a test is biased in favor of minority group members when in fact such a difference does not exist in the population of scores. This analytic proof has not been published elsewhere and explains the precise mechanism underlying the illustration provided by Linn and Werts (1971), particularly in the case when there is no slope-based bias across groups and there is range restriction. Note that Linn (1984) examined the impact of test score unreliability and group mean differences on observed over- or underprediction. In a similar analysis, Humphreys (1986) examined the impact of bias at the item level in terms of differences in item difficulty on group intercept differences. Millsap (1997, 1998, 2007) and Borsboom, Romeijn, and Wicherts (2008) extended earlier work by studying the relationship between factorial invariance, or measurement invariance, and intercept differences, or predictive invariance. Taken together, this body of research addressed the effect of measurement error on spurious intercept differences between groups that differ regarding test score means. Our new proof is unique and extends previous work in that it explains the effect of range restriction on the bias of intercept differences. Additionally, our proof is unique and extends previous research by estimating the degree to which Type I error rates are inflated for different combinations of test score measurement error, group mean differences, and range restriction. Consider the following equation showing the mechanism through which the size of the intercept-based difference is affected by unreliability, differences in test scores across groups, and range restriction (see Appendix B for the complete derivation of this approximation):

$$\Delta\psi^2_{\text{intercept}} = \frac{(r_{yG} - r_{xy}\sqrt{p_r(1-p_r)}\Delta\mu_r)^2}{1 - p_r(1-p_r)(\Delta\mu_r)^2}, \quad (4)$$

where $\Delta\psi^2_{\text{intercept}}$ is the unique contribution of G beyond X (cf. Equations 1–3), r_{yG} is the measurement-error attenuated and range restricted correlation coefficient between Y and G, r_{xy} is the measurement-error attenuated and range restricted correlation coefficient between test scores and the criterion, p_r is the proportion of the first group in the restricted samples, and $\Delta\mu_r$ is the mean difference in test scores between the groups in the restricted samples.

Equation 4 can be used to compute the degree of bias associated with G for a given level of measurement error and range restriction associated with a specific cut score. This estimation assumes there is no slope-based difference across the groups. As such, this estimation serves the purpose of showing the mechanism by which measurement error and differences in test scores across groups are likely to produce a bias in Type I error rates. Notwithstanding, we emphasize that this estimation does not include other factors, including slope-based bias and interaction effects that, as we show in our simulation, also have a substantial effect on the inflation of Type I error rates and bias in parameter estimation.

Equation 4 can be used to estimate the probability of committing a Type I error with the noncentral F-distribution (Mudholkar, Chaubey, & Lin, 1976; E. S. Pearson & Hartley, 1951). Specifically, let f denote the probability distribution function for the noncentral F-distribution and let the noncentral parameter be defined by $\lambda = \Delta\psi^2_{\text{intercept}}(n - 3)/(1 - \Delta\psi^2_{\text{intercept}} - r^2_{xy})$. The probability of identifying intercept differences is

$$P(F > F^*_{1-\alpha,\,1,\,n-3}) = \int_{F^*_{1-\alpha,\,1,\,n-3}}^{\infty} f(F,\,1,\,n-3,\,\lambda)dF, \quad (5)$$

where $F^*_{1-\alpha,\,1,\,n-3}$ is the critical value for one and $n - 3$ degrees of freedom at the $1 - \alpha$ confidence level.

Equations 4 and 5 enable an understanding of the effect of test score reliability, test score mean differences between groups, and range restriction on Type I error rates of tests of intercept differences in the absence of slope-based bias. In fact, one can input values into these equations to calculate how various situations involving the operation of these design artifacts increase the likelihood of the conclusion that there is intercept-based bias in favor of the minority group when this is actually not the case. Equation 4 also shows that if the mean test score is higher for the minority group as opposed to the majority group (i.e., the opposite of what is typically the case), then one would find that tests are biased in favor of the majority group (i.e., overprediction of majority group criterion scores).

Using Equations 4 and 5, consider the case of no intercept-based test bias (i.e., the null hypothesis H_0: $\Delta\psi^2_{\text{intercept}} = 0$ is true). A value of $\rho_{XY} = .50$ is justifiable given that this is a typical value for GMA tests (Schmidt & Hunter, 1998) and $p = .80$ given that it is typical to have about 20% of minority group members. In the unrealistic and almost impossible situation in which X is measured without error (i.e., $\rho_{XX} = 1$), if there is no difference in mean test scores across the groups (also an unrealistic assumption in the context of GMA testing), and the a priori Type I error rate is set at .05, Equation 4 yields $\Delta\psi^2_{\text{intercept}} = 0$ and the Type I error rate for a sample size of 250 is .05. This is exactly the value we should obtain given no true intercept-based differences. However, if instead of assuming that $\rho_{XX} = 1$ and $\mu_{1x} - \mu_{0x} = 0$, we use realistic values of $\rho_{XX} = .80$ and $\mu_{1x} - \mu_{0x} = 1.0$ (cf. Roth et al., 2001), solving Equation 5 results in $\Delta\psi^2_{\text{intercept}} = .002$. This effect may not be perceived as being very large. However, the associated Type I error rate for $N = 250$ is .117. In other words, a small degree of bias inflates the Type I error rate by more than twice the nominal level. Type I errors are inflated even more for larger samples. For example, considering $N = 1{,}500$ (cf. Sackett, Kuncel, Arneson, Cooper, & Waters, 2009, Table 1, p. 5), the associated Type I error for a bias of .002 is .458. Stated differently, given these parameter values and a very large sample size, we would expect researchers to identify intercept differences nearly 46% of the time due to chance alone. Finally, note that the degree of Type I error inflation estimated using our analytic solution assumes that there are no slope-based differences in the population. Results of our Monte Carlo simulation reported herein demonstrate that Type

I error rates are even higher and reach values in the .80s when slope-based differences exist.

In short, we provide a new analytic proof to support and provide a precise explanation for the illustrations discussed by Linn and Werts (1971) and Terris (1997). Specifically, we show that researchers are more likely to conclude incorrectly that performance is overpredicted for members of the minority group when the mean minority group test score is lower than the mean majority group test score and test scores are measured with less than perfect reliability, which are normative conditions in the context of GMA and other types of preemployment testing.

To summarize material we have discussed to this point, the established conclusion regarding slope-based bias in preemployment testing is that it is a rare occurrence. When bias exists, it is about intercept differences favoring minority group members, but not about slope differences. This finding, obtained by implementing the widely accepted Cleary (1968) test bias assessment procedure, has been replicated using various ethnic groups in the United States including African Americans and Latinos as well as different ethnic groups in other countries (e.g., Israel, South Africa, the Netherlands). However, these established conclusions are not consistent with recent methodological research showing that differential prediction assessment regarding slope-based differences is usually conducted with insufficient levels of statistical power, which can lead to the incorrect conclusion that bias does not exist (i.e., Type II errors). Also, these established conclusions are not consistent with expectations, mainly derived from theories outside of the field of I/O psychology, that sociohistorical–cultural and social psychological mechanisms are likely to lead to slope-based differences across groups. In addition, our new derivation provides analytic evidence that it is likely that the finding of intercept-based differences favoring minority group members is a result of a statistical artifact. In fact, given typical and frequently observed conditions in human resource selection research such as less-than-perfect test reliability and mean test score differences favoring the majority group, we would expect to find artifactual intercept-based differences favoring the minority group even if these differences do not exist in the population. Finally, it makes conceptual sense that intercept-based differences are smaller than they are believed to be or even nonexistent and that a slope-based difference exists. Specifically, if the expectation based on conceptual and statistical arguments is true and there is no intercept-based bias and, instead, there is slope-based bias, this would mean that observed performance differences between groups (i.e., lower average test and performance scores for members in the minority group) would be explained by slope-based differences. Thus, the presence of slope-based differences and the absence of intercept-based differences are consistent with meta-analytic findings regarding observed performance differences between groups (e.g., McKay & McDaniel, 2006; Walton & Spencer, 2009).

Figure 1 provides a graphic illustration of the previously established conclusion (Panel A) and what may actually be happening (Panels B and C) given the arguments and evidence we have discussed thus far. This figure shows situations including a slope-based difference and no intercept-based difference (Panel B) and a slope-based difference and a small intercept-based difference (Panel C). The situations illustrated in Panels B and C are consistent with the meta-analytic findings regarding differences in both test scores and performance in favor of the majority group (e.g., Walton & Spencer, 2009).

Given the arguments and evidence discussed thus far, we conducted a Monte Carlo simulation to examine established conclusions regarding bias in preemployment testing. We took advantage of advanced Monte Carlo methodology by using a research design with 3,185,000 unique cells and 15 billion 925 million individual samples. By implementing a powerful and sensitive methodology, we attempted to uncover patterns of relationships that may have remained hidden during the past 40 years while human resource selection researchers investigated test bias using samples that were subject to the detrimental effects of methodological and statistical artifacts.

Method

Overview

Our Monte Carlo simulation examined the effects of methodological and statistical artifacts on the accuracy of test bias assessment regarding intercept- and slope-based differences. Our simulation is the most exhaustive and comprehensive study to date to examine established conclusions regarding bias in preemployment testing. Factors manipulated included magnitude of intercept- and slope-based test bias, total sample size, proportion of minority group sample size to total sample size, predictor (i.e., preemployment test scores) and criterion (i.e., job performance) reliability, predictor range restriction, correlation between predictor scores and the dummy-coded grouping variable (e.g., ethnicity, gender), and group mean difference on predictor scores. We investigated a wide range of values for each of these factors and their effects on the observed magnitude of intercept- and slope-based bias as well as Type I error rates (when true bias was set to zero in the population) and statistical power (when true bias was set to larger than zero in the population) for detecting intercept- and slope-based bias. The parameter values included in our simulation were chosen to include the range of values that are typical and also have been reported in some of the most widely cited and influential sources for the established conclusions regarding test bias in preemployment testing (e.g., Dunbar & Novick, 1988; Hartigan & Wigdor, 1989; Houston & Novick, 1987; Hunter et al., 1984).

Manipulated Parameters

Total sample size. As is the case for any inferential test, sample size affects statistical power. In our study, total sample size was manipulated to vary from 100 to 1,000 in increments of 100. Several published reviews of the human resource selection literature ascertained that total sample size is usually around 100 (Lent, Auerbach, & Levin, 1971; Monahan & Muchinsky, 1983; Russell et al., 1994; Salgado, 1998). Accordingly, our simulation used 100 as the lower bound value and 10 times that number as the upper bound value to go substantially higher than the typical value.

Correlation between predictor and moderator: Proportion of individuals in the minority group and average group differences in test scores. Similar to Dunlap and Kemery (1988) and Evans (1985), we manipulated the correlation between the

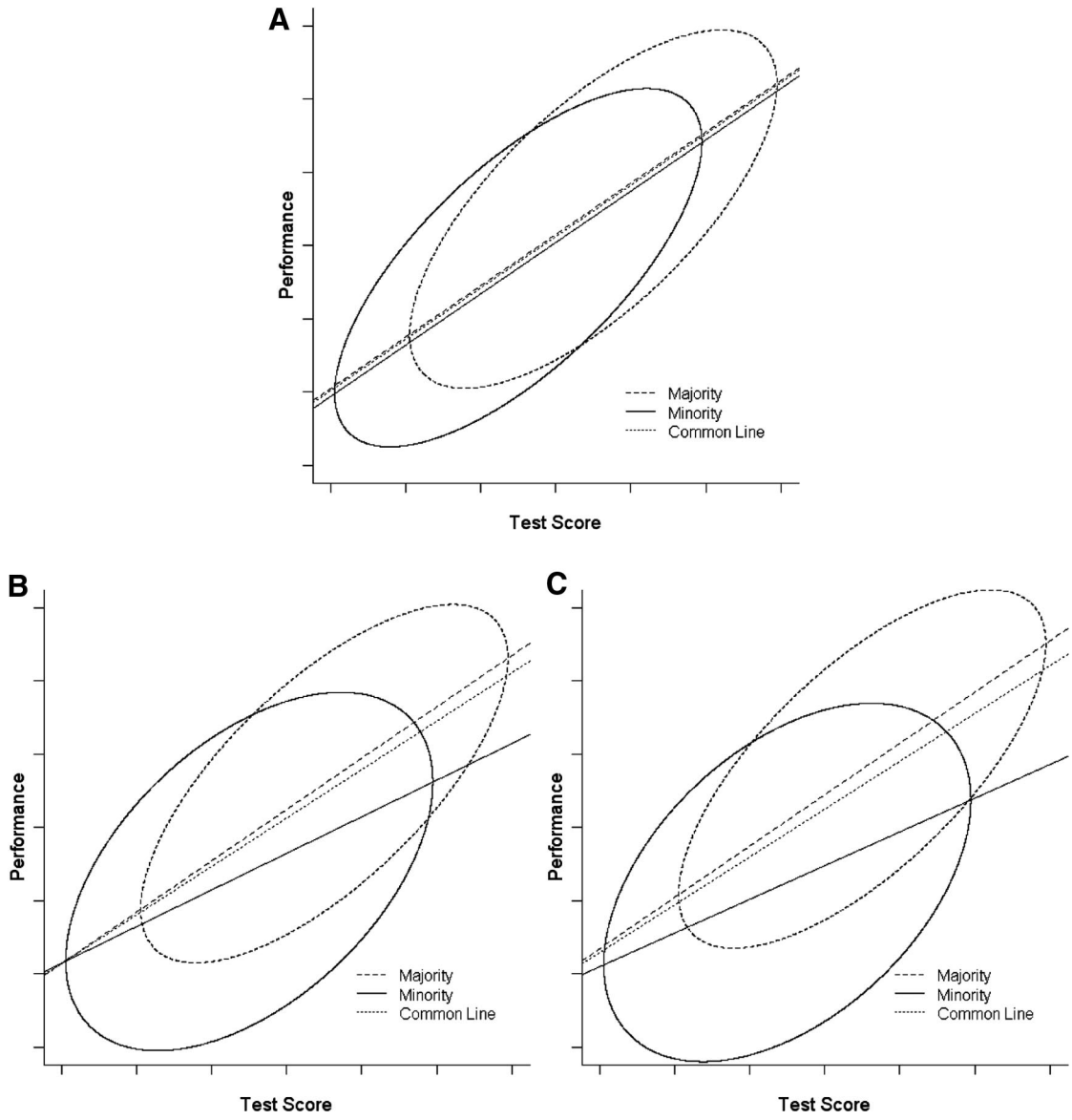

Figure 1. Illustration of the typical finding of no slope-based differences and intercept-based differences favoring the minority group (Panel A), and the possibility that there are slope-based differences (Panels B and C) together with no intercept-based differences (Panel B) and small intercept-based differences (Panel C).

predictor X and the categorical moderator G using the equation for the point-biserial correlation. Specifically, the correlation between X and G was modeled by generating X scores as follows:

$$X = \sqrt{1 - \rho_{XG}^2} Z_x + \rho_{XG} G, \tag{6}$$

where Z_x is derived from a standard normal distribution, G is a dummy variable that equaled zero for the focal group (e.g., ethnic minority group, women) and one for the reference group (e.g., the ethnic majority group, men), and ρ_{XG} is the correlation between X and G. Note that G, while a dichotomous variable, was also standardized (i.e., mean of zero and variance of one) to model the relationship between X and G and the relationship between Y and G. Also note that Dunlap and Kemery and Evans modeled a continuous moderator variable, but in our study the moderator variable G is a dichotomous variable. Thus, our procedure differs slightly from Dunlap and Kemery and Evans in that the point-biserial correlation between X and G is based upon mean differences in X between the two groups and the proportion of minority group sample size to total sample size. Specifically, the point-biserial correlation between X and G is a follows:

$$\rho_{XG} = \frac{(\mu_1 - \mu_0)\sqrt{p(1-p)}}{\sigma_X}, \tag{7}$$

where the correlation between X and G is determined by the product of the mean test score differences between the focal (μ_{1x})

and reference (μ_{0x}) groups and a function of the proportion of individuals in the focal group (p) divided by the standard deviation for X (i.e., σ_x).

In our study, $\sigma_x = 1$ for the focal and reference group for all conditions, the mean difference on X between the focal and reference groups ($\mu_{1x} - \mu_{0x}$) in standard deviation units varied from 0.0 to 1.0 in increments of .25 (Roth et al., 2001), and the proportion of individuals in the focal group varied from .1 to .5 in increments of .1. In total, we included 25 different combinations of p and $\mu_{1x} - \mu_{0x}$, and each combination yielded a unique value for ρ_{XG}. Table 1 includes the values for ρ_{XG} for each combination of p and $\mu_{1x} - \mu_{0x}$. In addition, results included in Table 1 confirm the effect of p and $\mu_{1x} - \mu_{0x}$ on ρ_{XG}; the population correlation is higher for larger values of $\mu_{1x} - \mu_{0x}$, and ρ_{XG} is nonlinearly related to p and reaches a maximum value when p equals .50.

Predictor and criterion reliability. Our simulation manipulated test (ρ_{XX}) and criterion (ρ_{YY}) score reliability. Based on material included in the introduction, we expected to find a negative relationship between ρ_{XX} and Type I error rates for the intercept-based test. On the other hand, we expected to find a positive relationship between ρ_{XX} and statistical power for the slope-based test (Aguinis et al., 2001; Dunlap & Kemery, 1988; Stone-Romero & Anderson, 1994). We set reliabilities to values from .7 to 1.0 in increments of .05 with a total of seven levels. The equations for introducing measurement error are the following:

$$x = \sqrt{\rho_{xx}}X + \sqrt{1 - \rho_{xx}}e_x \quad (8)$$

$$y = \sqrt{\rho_{yy}}Y + \sqrt{1 - \rho_{yy}}e_y, \quad (9)$$

where x and y are test and criterion scores with measurement error, and X and Y are the true (i.e., measurement-error free) scores. For tests of GMA, reliabilities can be in the .90s (e.g., Wonderlic, 1999), but they can also be in the .80s for certain types of g-loaded tests. For example, Rotundo and Sackett (1999) used the General Aptitude Test Battery and reported that "predictor reliabilities were in the .80 to .90 range" (p. 821). Moreover, reliabilities for other types of preemployment tests such as situational judgment tests and personality tests can be in the .70s and .80s (e.g., Mumford, Van Iddekinge, Morgeson, & Campion, 2008). Thus, our simulation design included values to cover this range.

Table 1
Summary of Values for ρ_{XG} for Various Values of p and $\mu_{1x} - \mu_{0x}$

	p				
$\mu_{1x} - \mu_{0x}$	.10	.20	.30	.40	.50
.00	.00	.00	.00	.00	.00
.25	.08	.10	.11	.12	.13
.50	.15	.20	.23	.24	.25
.75	.23	.30	.34	.37	.38
1.00	.30	.40	.46	.49	.50

Note. The predictor X was standardized with a mean of zero and variance of 1. ρ_{XG} = correlation between test scores (X) and dummy-coded moderator (G) denoting the grouping variable (e.g., ethnicity: 1 = majority, 0 = minority; gender: 1 = men, 0 = women); p = proportion of minority group members relative to total sample size; μ_{1x} and μ_{0x} = average predictor scores for the majority and minority groups, respectively (in standardized scores).

Range restriction. We manipulated range restriction in the predictor X (i.e., truncation) given that previous simulation work has demonstrated its detrimental effect on statistical power of the slope-difference test (Aguinis & Stone-Romero, 1997). In addition, range restriction due to the use of a cut score has the effect of reducing the upward bias in Type I error rates for the intercept-difference test when there are no slope-based differences across groups, as is assumed in our new analytic solution explaining the phenomenon illustrated by Linn and Werts (1971). However, the effects of range restriction are quite the opposite in the presence of slope-based differences. Specifically, consider the case where group intercepts are equivalent but the slopes differ across groups. Results would indicate no intercept-based differences across groups in the unrestricted sample. However, consider that we only observe values above a given cut score. If slopes differ in the presence of range restriction, the groups in the restricted sample will appear to differ in intercepts. We manipulated range restriction using the procedure described by Aguinis and Stone-Romero (1997) where values on the predictor X were truncated. Moreover, range restriction was simulated to represent the proportion of X values that were used to estimate the regression model. The values of range restriction varied from .10 (severe range restriction—only the top 10% of scores in the distribution are retained as can be seen in highly selective contexts) to 1.00 (no range restriction—all scores in the distribution are retained) in increments of .10.

Magnitude of true intercept- and slope-based bias. The effect size for intercept-based and slope-based test bias was modeled to be equal in the population within each design cell. In this way, we are able to observe any differential effects of the methodological and statistical artifacts on the accuracy of the slope-based test in relationship to the intercept-based test by holding true test bias constant. If the true amount of test bias is not held constant, we are unable to examine the relative impact of the various statistical and methodological artifacts on sample-based conclusions about the presence and degree of bias. Specifically, the population change in Y variance explained ($\Delta\psi^2$) was simulated to be equal for both the first-order effect of the moderator G and the $X \times G$ interaction (see Equations 1–3). The criterion was generated in a manner similar to Evans's (1985) approach with the exception that the moderator G in this study is categorical rather than continuous. The criterion was generated as a normally distributed variable with a mean of zero and variance of one using the following equation:

$$Y = \sqrt{1 - 2\Delta\psi^2 - \rho_{xy}^2}Z_y + \rho_{xy}X + \sqrt{\Delta\psi^2}G + \sqrt{\Delta\psi^2}XG, \quad (10)$$

where the validity coefficient between X and Y was held constant at .50 (i.e., $\rho_{XY} = .50$ for all permutations; Schmidt & Hunter, 1998), XG is the product between X and G, $\Delta\psi^2$ is the same for G and XG in each design cell, Z_y is generated from the standard normal distribution, and Y is a normally distributed criterion with a mean zero and unit variance. The $\Delta\psi^2$ effect sizes were simulated to vary from values equal to 0 (i.e., no test bias in the population for either the intercept or slope) to .15 (i.e., 15% of variance in Y is explained by G above and beyond X, and 15% of variance in Y is explained by XG above and beyond X and G). Our simulation

increased $\Delta\psi^2$ in increments of .005 for values from 0 to .10 and in increments of .01 for values ranging from .10 to .15 for a total of 26 levels. The $\Delta\psi^2$ values were chosen to cover the range of values usually observed in applied psychology and management (Aguinis et al., 2005; McClelland & Judd, 1993). Moreover, as noted in the introduction, even small effects can be very meaningful in human resource selection contexts when incorrect decisions due to test bias affect thousands of individuals.

Dependent variables. Our simulation included two types of dependent variables. First, we computed statistical power and Type I error rates for the intercept-based and slope-based bias assessment test. Second, we computed differences between population and sample-based intercept-based and slope-based effects (i.e., $\Delta\psi^2_{\text{intercept}}$ vs. $\Delta R^2_{\text{intercept}}$ and $\Delta\psi^2_{\text{slope}}$ vs. $\Delta R^2_{\text{slope}}$). Thus, our study does not examine results regarding statistical significance exclusively but also examines bias in the estimation of effect sizes. Regarding the computation of statistical power and Type I error rates, power was the probability of rejecting the null hypothesis of no test bias when the null hypothesis was false, and Type I error rate was the probability of rejecting the null hypothesis of test bias when the null hypothesis was true. We set the nominal Type I error rate at .05.

Simulation Procedure

Our simulation included a full-factorial design crossing all values for the manipulated parameters. This resulted in 3,185,000 unique combinations of parameter values. Table 2 provides a summary of each manipulated parameter and its values. We drew 5,000 samples for each parameter value combination in the research design, resulting in a total of 15 billion 925 million individual samples including more than 8 trillion (i.e., 8,662,500,000,000) individual scores. Information regarding details of the computer program (e.g., language, time to execute) is available from the authors upon request.

Table 2
Summary of Parameter Values Included in the Simulation

Manipulated parameter	No. of levels	Low value	High value	Increment
$\Delta\psi^2$	26	.00	.15	[a]
RR	10	.10	1.00	.10
N	10	100	1,000	100
ρ_{XX}	7	.70	1.00	.05
ρ_{YY}	7	.70	1.00	.05
ρ_{XG}				
p	5	.10	.50	.10
$\mu_{1x} - \mu_{0x}$	5	.00	1.00	.25

Note. Combining all parameter values resulted in the generation of 15 billion 925 million individual samples in 3,185,000 unique design cells. $\Delta\psi^2$ = population-based test bias; *RR* = range restriction (i.e., truncation on *X*); ρ_{XX} = predictor scores reliability; ρ_{YY} = criterion scores reliability; ρ_{XG} = correlation between continuous predictor *X* and dichotomous moderator *G*; *p* = proportion of minority group sample size to total sample size; $\mu_{1x} - \mu_{0x}$ = average difference between majority and minority predictor scores.

[a] The test bias parameter $\Delta\psi^2$ increased in value by .005 from .00 to .10 and by .01 from .10 to .15.

Key Accuracy Checks

We checked the accuracy of the simulation procedure using the following two methods. First, we compared population $\Delta\psi^2_{\text{intercept}}$ and $\Delta\psi^2_{\text{slope}}$ values against their sample-based estimates $R^2_{\text{intercept}}$ and R^2_{slope} for *N* values of 1,000, based on 5,000 replications, and without introducing any methodological and statistical artifacts (i.e., $p = .5$, range restriction (*RR*) = 1.0, $\rho_{XG} = 0$, and $\rho_{XX} = \rho_{YY} = 1.0$). Results confirmed the accuracy of the simulation procedures: The (true) population and (estimated) sample-based test bias values were virtually identical across the entire range of effect sizes for the intercept-difference and slope-difference test. As a second set of analyses for checking the accuracy of the simulation, we investigated the extent to which the estimated Type I error rates for the intercept-difference and slope-difference tests were close to the nominal level of .05. We held reliabilities for *X* and *Y* constant at 1.0, *RR* = 1.0 (i.e., no range restriction), and $\mu_{1x} - \mu_{0x}$ (i.e., difference in mean test scores across groups) at zero. Results confirmed the high degree of accuracy of the simulation procedures: The average absolute value difference between the nominal Type I error rate of .05 and the empirically derived Type I error rates was .0020 for the intercept-difference test and .0025 for the slope-difference test. Detailed results and tables regarding these two types of key accuracy checks are available from the authors upon request.

Results

We report results in three sections. First, we present results on test bias assessment regarding slope-based differences. Recall that in the introduction we provided reasons why slope-based differences could exist but are not found. Second, we present results on test bias assessment regarding intercept-based differences. Recall that in the introduction we provided reasons why intercept-based differences favoring minority group members are found but may not necessarily exist and if they exist they could be smaller than they are believed to be. Finally, we describe results comparing the factors affecting the relative accuracy of slope-based bias assessment in relationship to intercept-based bias assessment.

Test Bias Assessment Based on Slope Differences Across Groups

First, it is informative to compute the mean statistical power across the 3,062,500 cells in our design for which there is true population slope-based test bias. Across these cells, the mean predictor (i.e., test scores) and criterion reliability is .85, test score range restriction is .55, total *N* is 550, proportion of minority group sample size to total sample size is .30, difference between majority and minority predictor scores is .50 *SD* units, correlation between test scores *X* and moderator variable *G* is .11, and population slope-based bias (i.e., $\Delta\psi^2_{\text{slope}}$) is .068 (i.e., 6.8% of variance in *Y* is explained by the $X \times G$ interaction above and beyond the effects of *X* and *G*). For this combination of parameter values, statistical power to detect slope-based bias is about as good as flipping a coin: only .56. Also, as is expected given the detrimental impact of methodological and statistical artifacts described in the introduction, the estimated sample-based $\Delta R^2_{\text{slope}}$ is substantially smaller than its population counterpart: It is only .0107, which is 15.7% the size of its population counterpart.

Tables 3–6 include results regarding statistical power and estimated sample-based effect sizes for a representative set of parameter value combinations (more comprehensive tables based on the 3,185,000 design cells are available from the authors). We created these tables to include values that are observed in various human resource selection contexts using different types of tests including GMA, personality, situational judgment, application blanks, and so forth. For example, although the reliability for GMA tests can be in the .90s, the (interrater) reliability for the selection interview is in the .70s (Conway, Jako, & Goodman, 1995) and the reliability of some *g*-loaded tests can be in the .80s (Rotundo & Sackett, 1999). Similarly, although the mean total sample size in human resource selection research is around 100, some human resource selection research studies include larger samples (Aguinis et al., 2005). Thus, taken together, results in Tables 3–6 provide good coverage of frequently observed situations.

Table 3 includes values for a total sample size of 300 or 400, reliabilities for tests scores and the criterion of .80 or .85, proportion of minority group members to total sample size of .20 or .30, average difference of .50 *SD* units between majority and minority test scores, range restriction of .30 or .40, and correlation between test scores and ethnicity of .20 or .23. Also, Table 3 shows the resulting power and sample-based effect sizes when the true population slope-based bias is $\Delta\psi^2_{slope} = .01$ and $\Delta\psi^2_{slope} = .02$. Slope-based test bias of this magnitude is sufficiently large to produce large and practically significant rates of prediction errors (i.e., false positives and false negatives; Aguinis & Smith, 2007, 2009). Overall, the average power for all combinations of parameter values shown in Table 3 and $\Delta\psi^2_{slope} = .01$ is only .089, and the observed sample-based effect size is .001, which is 10% the value of its population counterpart. The situation in terms of power and observed effect size does not improve much when $\Delta\psi^2_{slope}$ is increased to .02. Specifically, mean statistical power is .130 and

Table 3
Statistical Power and Observed Effect Size for Intercept- and Slope-Based Test Bias for Selected Set of Parameter Values (Part 1)

							Intercept differences				Slope differences			
							$\Delta\psi^2 = .01$		$\Delta\psi^2 = .02$		$\Delta\psi^2 = .01$		$\Delta\psi^2 = .02$	
N	ρ_{XX}	ρ_{YY}	p	$\mu_{1x} - \mu_{0x}$	ρ_{XG}	RR	Power	ΔR^2	Power	ΔR^2	Power	ΔR^2	Power	ΔR^2
300	.80	.80	.20	.50	.20	.30	.814	.024	.975	.044	.074	.000	.091	.001
400	.80	.80	.20	.50	.20	.30	.914	.024	.994	.043	.079	.000	.103	.001
300	.80	.80	.20	.50	.20	.40	.779	.022	.963	.040	.068	.001	.093	.001
400	.80	.80	.20	.50	.20	.40	.880	.022	.989	.039	.080	.001	.111	.001
300	.80	.85	.20	.50	.20	.30	.836	.025	.985	.047	.075	.001	.092	.001
400	.80	.85	.20	.50	.20	.30	.932	.026	.995	.046	.080	.000	.108	.001
300	.80	.85	.20	.50	.20	.40	.804	.023	.971	.042	.069	.001	.097	.001
400	.80	.85	.20	.50	.20	.40	.902	.023	.992	.042	.084	.001	.115	.001
300	.85	.80	.20	.50	.20	.30	.801	.023	.974	.043	.073	.001	.093	.001
400	.85	.80	.20	.50	.20	.30	.903	.023	.993	.042	.079	.000	.109	.001
300	.85	.80	.20	.50	.20	.40	.764	.021	.956	.038	.072	.001	.102	.001
400	.85	.80	.20	.50	.20	.40	.865	.021	.988	.038	.078	.001	.115	.001
300	.85	.85	.20	.50	.20	.30	.824	.024	.983	.045	.073	.001	.096	.001
400	.85	.85	.20	.50	.20	.30	.922	.025	.996	.045	.080	.001	.114	.001
300	.85	.85	.20	.50	.20	.40	.788	.022	.967	.041	.073	.001	.103	.001
400	.85	.85	.20	.50	.20	.40	.887	.022	.991	.041	.081	.001	.119	.001
300	.80	.80	.30	.50	.23	.30	.881	.029	.990	.053	.088	.001	.128	.002
400	.80	.80	.30	.50	.23	.30	.953	.029	.999	.053	.102	.001	.157	.002
300	.80	.80	.30	.50	.23	.40	.851	.027	.985	.047	.094	.001	.136	.002
400	.80	.80	.30	.50	.23	.40	.935	.026	.997	.047	.104	.001	.169	.002
300	.80	.85	.30	.50	.23	.30	.903	.031	.995	.056	.090	.001	.131	.002
400	.80	.85	.30	.50	.23	.30	.967	.031	1.000	.056	.106	.001	.165	.002
300	.80	.85	.30	.50	.23	.40	.874	.028	.991	.050	.099	.001	.145	.002
400	.80	.85	.30	.50	.23	.40	.950	.028	.998	.050	.112	.001	.179	.002
300	.85	.80	.30	.50	.23	.30	.866	.028	.989	.051	.089	.001	.134	.002
400	.85	.80	.30	.50	.23	.30	.950	.028	.998	.051	.104	.001	.165	.002
300	.85	.80	.30	.50	.23	.40	.841	.025	.980	.046	.099	.001	.150	.002
400	.85	.80	.30	.50	.23	.40	.926	.025	.996	.046	.112	.001	.182	.002
300	.85	.85	.30	.50	.23	.30	.889	.030	.992	.055	.093	.001	.141	.002
400	.85	.85	.30	.50	.23	.30	.961	.030	.999	.054	.109	.001	.175	.002
300	.85	.85	.30	.50	.23	.40	.866	.027	.988	.049	.103	.001	.158	.003
400	.85	.85	.30	.50	.23	.40	.943	.027	.997	.049	.117	.001	.190	.002
							M							
350	.83	.83	.25	.50	.22	.35	.880	.026	.988	.047	.089	.001	.130	.002

Note. $\Delta\psi^2$ = population-based effect size (i.e., test bias); N = total sample size; ρ_{XX} = test reliability; ρ_{YY} = criterion reliability; p = proportion of minority group sample size to total sample size; $\mu_{1x} - \mu_{0x}$ = average standard deviation-unit difference between majority and minority test scores; ρ_{XG} = correlation between test scores and moderator G; RR = range restriction; ΔR^2 = sample-based effect size (i.e., test bias) estimate.

the sample-based effect size is .002, which is also 10% the value of its population counterpart.

Results shown in Table 3 are consistent in terms of the low statistical power and the underestimation of population slope-based test bias. For example, consider the first line in Table 3, which refers to a total sample size of 300, *X* and *Y* reliabilities of .80, proportion of minority group members to total sample size of .20, average difference in mean test scores between the majority and minority group of .50 *SD*, a correlation between test scores *X* and the moderator *G* of .20, and range restriction of .30. For this situation, power is .074 when $\Delta\psi^2_{slope} = .01$ and .091 when $\Delta\psi^2_{slope} = .02$. Moreover, estimated effect sizes are only $\Delta R^2_{slope} = .000$ and .001, respectively.

Table 4 includes parameter values that, from a research design and measurement perspective, are more desirable than those in Table 3 because sample size is larger, reliabilities are higher, proportions of majority and minority group members are not as different from each other, and range restriction is not as severe. From a substantive perspective, the difference in mean test scores across the groups is not as large as in Table 3 and, similarly, the correlation between test scores and the moderator is low. Specifically, total sample size is 500 or 600, reliabilities for tests scores and the criterion are .90 or .95, proportion of minority group members to total sample size is .40 or .50, there is a difference between majority and minority mean test scores of 0 or .25 *SD*, range restriction is .50 or .60, and there is a correlation between test scores and the grouping moderator variable of 0, .12, or .13. Table 4 shows the resulting power and sample-based effect sizes when the population slope-based bias is $\Delta\psi^2_{slope} = .03$ and $\Delta\psi^2_{slope} = .04$. For $\Delta\psi^2_{slope} = .03$, across all combinations of parameter values, the power to detect slope-based test bias is .74 and the sample-based ΔR^2_{slope} is .01, which is about 30% the value of the population effect. For $\Delta\psi^2_{slope} = .04$, across all combinations of parameter values, the power to detect slope-based test bias is .86 and the sample-based ΔR^2_{slope} is .013. Selecting specific entries in Table 4 shows that, even for those conditions that are more conducive to detecting slope-based test bias compared to those in Table 3, power is still insufficient. For example, consider Line 33 in Table 4: Sample size is 500, reliabilities for test scores and the criterion are .90, 40% of the sample consists of minority group members, the difference between majority and minority mean test scores is .25 *SD* units, the correlation between test scores and the moderator is only .12, and range restriction is .50. The probability of detecting a population slope-based test bias effect of $\Delta\psi^2_{slope} = .03$ is only .61. Moreover, the estimate of the size of slope-based test bias is only .008, which is 26.7% the size of its population counterpart.

Table 5 includes a third set of results. Although less desirable from a research design and measurement standpoint, the parameter value combinations shown in Table 5 are not infrequent in the preemployment testing literature. Specifically, Table 5 includes a total sample size of 100 (which, as noted earlier, is approximately the average in human resource selection research), reliabilities for test scores and the criterion of .70 or .75 (which, as noted earlier, is not uncommon for certain types of tests such as situational judgment and interviews), proportion of minority group members to total sample size of .10, a mean difference between test scores for the majority and minority groups of .75 or 1.00, correlation between test scores and the moderator of .23 or .30, and range restriction of .10 or .20. For these parameter value combinations, Table 5 shows that slope-based test bias would virtually never be found when $\Delta\psi^2_{slope} = .005$. The probability of detecting slope-based test bias is never greater than .054. Moreover, sample-based effect sizes are estimated to be .000 or smaller in every case.

Finally, Table 6 includes combinations of parameter values that are most typical of GMA tests. Specifically, the average total sample size is 533 (ranging from 100 to 1,000), average test score reliability is .90 (ranging from .85 to .95), criterion reliability is .85, the proportion of members in the minority group to total sample size is .20 or .30, standardized mean test score differences favoring the majority group are .75 or 1.00, and the average range restriction value is .40 or .60. In Table 6, Part 1 includes true test bias of .01, Part 2 includes true test bias of .02, Part 3 includes true test bias of .03, and Part 4 includes true test bias of .04. Once again, results are consistent with those presented in the previous tables: Overall, statistical power is inadequate to detect slope-based bias and the degree of bias is underestimated. For example, the average values for statistical power are only .14, .23, .31, and .39 in Parts 1–4, respectively.

Relative impact of manipulated parameters on statistical power to detect slope-based test bias. We conducted a regression analysis in which the criterion was statistical power values (i.e., rejection rates of the null hypothesis $\Delta\psi^2_{slope} = 0$ when $\Delta\psi^2_{slope} > 0$) and the predictors were the eight parameters manipulated in the simulation and their two-way interactions. This analysis allows us to understand the relative impact of the eight manipulated parameters and their two-way interactions on the power to detect slope-based test bias. We standardized each predictor before computing two-way products, and we estimated standardized regression weights to facilitate the interpretation of the relative strength of the effects given the very different metrics used for the scales for each predictor (e.g., *N* ranging from 100 to 1,000 vs. reliability ranging from .70 to 1.00). Results shown in Table 7 (columns labeled *Power: Slope*) indicate that population effect size ($\beta = .51$), total sample size ($\beta = .43$), correlation between test scores and minority status ($\beta = .41$), proportion of minority group members relative to total sample size ($\beta = .27$), and range restriction ($\beta = .31$) are the largest first-order effects. For example, a 1 *SD* unit increase in the proportion of minority group members relative to total sample size leads to a .27 *SD* unit increase in statistical power, holding all other variables in the model constant. These effects corroborate and expand results reported by Aguinis and Stone-Romero (1997), albeit their simulation design was smaller than ours. In addition, although not investigated by Aguinis and Stone-Romero, results show that test score reliability ($\beta = .11$) and criterion score reliability ($\beta = .10$) also affect power. Results in Table 7 also show that, although smaller in magnitude, there are several significant two-way interactions. As noted by Aguinis and Stone-Romero, these results indicate that various methodological and statistical artifacts interact to decrease the power to detect slope-based bias.

Table 7 (columns labeled *Type I error: Slope*) also includes regression results using Type I error rates as the criterion (i.e., rejection rates of the null hypothesis $\Delta\psi^2_{slope} = 0$ when $\Delta\psi^2_{slope} = 0$). As shown in Table 7, these regression weights are smaller than .00 for every first-order and two-way interaction effect, and four of the first-order effects are statistically nonsignificant at the .05 level. Thus, these results indicate that although the manipulated parameters

Table 4
Statistical Power and Observed Effect Size for Intercept- and Slope-Based Test Bias for Selected Set of Parameter Values (Part 2)

							Intercept differences				Slope differences			
							$\Delta\psi^2 = .03$		$\Delta\psi^2 = .04$		$\Delta\psi^2 = .03$		$\Delta\psi^2 = .04$	
N	ρ_{XX}	ρ_{YY}	p	$\mu_{1x} - \mu_{0x}$	ρ_{XG}	RR	Power	ΔR^2	Power	ΔR^2	Power	ΔR^2	Power	ΔR^2
500	.90	.90	.40	.00	.00	.50	1.0	.083	1.0	.110	.608	.008	.744	.011
600	.90	.90	.40	.00	.00	.50	1.0	.083	1.0	.110	.686	.008	.819	.011
500	.90	.90	.40	.00	.00	.60	1.0	.071	1.0	.093	.670	.009	.795	.012
600	.90	.90	.40	.00	.00	.60	1.0	.070	1.0	.092	.750	.009	.873	.012
500	.90	.95	.40	.00	.00	.50	1.0	.088	1.0	.116	.635	.009	.772	.011
600	.90	.95	.40	.00	.00	.50	1.0	.088	1.0	.116	.721	.009	.849	.011
500	.90	.95	.40	.00	.00	.60	1.0	.075	1.0	.098	.698	.010	.826	.013
600	.90	.95	.40	.00	.00	.60	1.0	.074	1.0	.098	.780	.010	.896	.013
500	.95	.90	.40	.00	.00	.50	1.0	.086	1.0	.113	.638	.009	.773	.011
600	.95	.90	.40	.00	.00	.50	1.0	.086	1.0	.113	.720	.009	.845	.011
500	.95	.90	.40	.00	.00	.60	1.0	.073	1.0	.095	.702	.010	.824	.013
600	.95	.90	.40	.00	.00	.60	1.0	.072	1.0	.095	.787	.010	.893	.013
500	.95	.95	.40	.00	.00	.50	1.0	.091	1.0	.120	.665	.009	.803	.012
600	.95	.95	.40	.00	.00	.50	1.0	.091	1.0	.120	.748	.009	.873	.012
500	.95	.95	.40	.00	.00	.60	1.0	.077	1.0	.101	.734	.011	.852	.014
600	.95	.95	.40	.00	.00	.60	1.0	.077	1.0	.101	.808	.011	.920	.014
500	.90	.90	.50	.00	.00	.50	1.0	.090	1.0	.120	.676	.010	.808	.013
600	.90	.90	.50	.00	.00	.50	1.0	.091	1.0	.120	.755	.010	.871	.013
500	.90	.90	.50	.00	.00	.60	1.0	.077	1.0	.101	.747	.011	.854	.014
600	.90	.90	.50	.00	.00	.60	1.0	.076	1.0	.101	.822	.011	.921	.015
500	.90	.95	.50	.00	.00	.50	1.0	.096	1.0	.127	.708	.010	.837	.014
600	.90	.95	.50	.00	.00	.50	1.0	.097	1.0	.127	.785	.010	.894	.014
500	.90	.95	.50	.00	.00	.60	1.0	.081	1.0	.107	.775	.012	.879	.015
600	.90	.95	.50	.00	.00	.60	1.0	.081	1.0	.107	.847	.012	.939	.016
500	.95	.90	.50	.00	.00	.50	1.0	.093	1.0	.123	.704	.010	.834	.014
600	.95	.90	.50	.00	.00	.50	1.0	.094	1.0	.123	.784	.010	.898	.014
500	.95	.90	.50	.00	.00	.60	1.0	.079	1.0	.104	.773	.012	.878	.015
600	.95	.90	.50	.00	.00	.60	1.0	.078	1.0	.104	.845	.012	.933	.016
500	.95	.95	.50	.00	.00	.50	1.0	.099	1.0	.131	.737	.011	.858	.014
600	.95	.95	.50	.00	.00	.50	1.0	.100	1.0	.131	.818	.011	.923	.014
500	.95	.95	.50	.00	.00	.60	1.0	.084	1.0	.110	.804	.013	.899	.016
600	.95	.95	.50	.00	.00	.60	1.0	.083	1.0	.110	.871	.013	.950	.017
500	.90	.90	.40	.25	.12	.50	1.0	.080	1.0	.105	.606	.008	.743	.010
600	.90	.90	.40	.25	.12	.50	1.0	.081	1.0	.105	.686	.008	.818	.010
500	.90	.90	.40	.25	.12	.60	1.0	.069	1.0	.090	.668	.009	.792	.011
600	.90	.90	.40	.25	.12	.60	1.0	.069	1.0	.090	.752	.009	.872	.012
500	.90	.95	.40	.25	.12	.50	1.0	.085	1.0	.111	.632	.008	.772	.011
600	.90	.95	.40	.25	.12	.50	1.0	.085	1.0	.111	.717	.008	.847	.011
500	.90	.95	.40	.25	.12	.60	1.0	.073	1.0	.095	.696	.010	.825	.012
600	.90	.95	.40	.25	.12	.60	1.0	.073	1.0	.095	.778	.010	.894	.012
500	.95	.90	.40	.25	.12	.50	1.0	.081	1.0	.106	.638	.008	.773	.011
600	.95	.90	.40	.25	.12	.50	1.0	.081	1.0	.106	.721	.008	.845	.011
500	.95	.90	.40	.25	.12	.60	1.0	.069	1.0	.090	.701	.010	.822	.012
600	.95	.90	.40	.25	.12	.60	1.0	.069	1.0	.090	.785	.010	.892	.013
500	.95	.95	.40	.25	.12	.50	1.0	.086	1.0	.112	.665	.009	.801	.011
600	.95	.95	.40	.25	.12	.50	1.0	.086	1.0	.112	.746	.009	.873	.011
500	.95	.95	.40	.25	.12	.60	1.0	.073	1.0	.095	.734	.010	.850	.013
600	.95	.95	.40	.25	.12	.60	1.0	.073	1.0	.095	.806	.010	.921	.013
500	.90	.90	.50	.25	.13	.50	1.0	.087	1.0	.114	.675	.009	.810	.012
600	.90	.90	.50	.25	.13	.50	1.0	.088	1.0	.114	.754	.009	.872	.012
500	.90	.90	.50	.25	.13	.60	1.0	.075	1.0	.098	.746	.011	.856	.014
600	.90	.90	.50	.25	.13	.60	1.0	.074	1.0	.097	.822	.011	.919	.014
500	.90	.95	.50	.25	.13	.50	1.0	.092	1.0	.121	.707	.010	.837	.013
600	.90	.95	.50	.25	.13	.50	1.0	.093	1.0	.121	.787	.010	.891	.013
500	.90	.95	.50	.25	.13	.60	1.0	.079	1.0	.103	.774	.011	.879	.015
600	.90	.95	.50	.25	.13	.60	1.0	.079	1.0	.103	.847	.012	.939	.015
500	.95	.90	.50	.25	.13	.50	1.0	.088	1.0	.115	.702	.010	.833	.013
600	.95	.90	.50	.25	.13	.50	1.0	.088	1.0	.115	.785	.010	.898	.013
500	.95	.90	.50	.25	.13	.60	1.0	.075	1.0	.098	.773	.011	.877	.015

Table 4 (*continued*)

							Intercept differences				Slope differences			
							$\Delta\psi^2 = .03$		$\Delta\psi^2 = .04$		$\Delta\psi^2 = .03$		$\Delta\psi^2 = .04$	
N	ρ_{XX}	ρ_{YY}	p	$\mu_{1x} - \mu_{0x}$	ρ_{XG}	RR	Power	ΔR^2	Power	ΔR^2	Power	ΔR^2	Power	ΔR^2
600	.95	.90	.50	.25	.13	.60	1.0	.075	1.0	.098	.846	.012	.932	.015
500	.95	.95	.50	.25	.13	.50	1.0	.093	1.0	.122	.735	.010	.858	.014
600	.95	.95	.50	.25	.13	.50	1.0	.094	1.0	.122	.817	.010	.922	.014
500	.95	.95	.50	.25	.13	.60	1.0	.080	1.0	.104	.805	.012	.897	.016
600	.95	.95	.50	.25	.13	.60	1.0	.079	1.0	.104	.870	.012	.950	.016
							M							
550	.925	.925	.45	.125	.0625	.55	1.0	.082	1.0	.108	.743	.010	.860	.013

Note. $\Delta\psi^2$ = population-based effect size (i.e., test bias); ρ_{XX} = test reliability; ρ_{YY} = criterion reliability; p = proportion of minority group sample size to total sample size; $\mu_{1x} - \mu_{0x}$ = average standard deviation-unit difference between majority and minority test scores; ρ_{XG} = correlation between test scores and moderator G; RR = range restriction; ΔR^2 = sample-based effect size (i.e., test bias) estimate.

have a large detrimental effect on statistical power, they do not have a noticeable detrimental effect on Type I error rates (i.e., incorrectly concluding that there is slope-based test bias).

In sum, results reported in Tables 3–6 demonstrate that slope-based test bias is unlikely to be detected given values of total sample size, range restriction, predictor and criterion scores reliability, proportion of minority group sample size to total sample size, and differences between test scores across groups that cover a range usually observed in the preemployment testing literature. In fact, these results indicate that slope-based test bias is likely to go undetected even when reliabilities are very high (i.e., .95), sample size is about 10 times the average reported in the preemployment test validation literature (i.e., 1,000), and differences in test scores between minority and majority group members is half the size usually reported for GMA testing (i.e., .50 *SD* units).

Test Bias Assessment Based on Intercept Differences Across Groups

It is informative to compute the sample-based effect size for intercept-based differences (i.e., $\Delta R^2_{\text{intercept}}$) across the entire simulation design given that the true average population test bias effect is $\Delta\psi^2_{\text{intercept}} = .065$. Across all 3,185,000 cells, the mean predictor and criterion scores reliability is .85, range restriction is .55, total N is 550, proportion of minority group sample size to

(*text continues on page 665*)

Table 5
Statistical Power and Observed Effect Size for Intercept- and Slope-Based Test Bias for Selected Set of Parameter Values (Part 3)

							Intercept differences ($\Delta\psi^2 = .005$)		Slope differences ($\Delta\psi^2 = .005$)	
N	ρ_{XX}	ρ_{yy}	p	$\mu_{1x} - \mu_{0x}$	ρ_{xg}	RR	Power	ΔR^2	Power	ΔR^2
100	.70	.70	.10	.75	.23	.10	.205	.012	.051	.000
100	.70	.70	.10	.75	.23	.20	.197	.012	.049	.000
100	.70	.75	.10	.75	.23	.10	.220	.013	.050	.000
100	.70	.75	.10	.75	.23	.20	.211	.013	.047	.000
100	.75	.70	.10	.75	.23	.10	.198	.011	.052	.000
100	.75	.70	.10	.75	.23	.20	.180	.011	.047	.000
100	.75	.75	.10	.75	.23	.10	.213	.012	.052	.000
100	.75	.75	.10	.75	.23	.20	.190	.012	.046	.000
100	.70	.70	.10	1.00	.30	.10	.213	.013	.054	.000
100	.70	.70	.10	1.00	.30	.20	.213	.013	.048	.000
100	.70	.75	.10	1.00	.30	.10	.228	.014	.052	.000
100	.70	.75	.10	1.00	.30	.20	.226	.013	.050	.000
100	.75	.70	.10	1.00	.30	.10	.201	.011	.053	.000
100	.75	.70	.10	1.00	.30	.20	.189	.011	.046	.000
100	.75	.75	.10	1.00	.30	.10	.214	.012	.053	.000
100	.75	.75	.10	1.00	.30	.20	.200	.012	.045	.000
							M			
100	.725	.725	.10	.875	.265	.15	.206	.012	.050	.000

Note. $\Delta\psi^2$ = population-based effect size (i.e., test bias); ρ_{XX} = test reliability; ρ_{YY} = criterion reliability; p = proportion of minority group sample size to total sample size; $\mu_{1x} - \mu_{0x}$ = average standard deviation-unit difference between majority and minority test scores; ρ_{XG} = correlation between test scores and moderator G; RR = range restriction; ΔR^2 = sample-based effect size (i.e., test bias) estimate.

Table 6
Statistical Power and Observed Effect Size for Intercept- and Slope-Based Test Bias for Selected Set of Prototypical Parameter Values in General Mental Abilities Testing

							Intercept differences		Slope differences	
N	ρ_{XX}	ρ_{yy}	p	$\mu_{1x} - \mu_{0x}$	ρ_{xg}	RR	Power	ΔR^2	Power	ΔR^2
					Part 1 ($\Delta\psi^2 = .01$)					
100	.85	.85	.20	.75	.30	.40	.351	.021	.050	.001
100	.85	.85	.20	.75	.30	.60	.319	.018	.064	.001
100	.85	.85	.20	1.00	.40	.40	.339	.020	.049	.000
100	.85	.85	.20	1.00	.40	.60	.320	.017	.062	.001
100	.85	.85	.30	.75	.34	.40	.401	.025	.058	.001
100	.85	.85	.30	.75	.34	.60	.356	.021	.071	.002
100	.85	.85	.30	1.00	.46	.40	.374	.022	.057	.001
100	.85	.85	.30	1.00	.46	.60	.349	.020	.069	.001
100	.95	.85	.20	.75	.30	.40	.302	.017	.056	.001
100	.95	.85	.20	.75	.30	.60	.267	.014	.066	.001
100	.95	.85	.20	1.00	.40	.40	.264	.014	.054	.001
100	.95	.85	.20	1.00	.40	.60	.242	.013	.065	.001
100	.95	.85	.30	.75	.34	.40	.339	.020	.066	.001
100	.95	.85	.30	.75	.34	.60	.303	.016	.078	.002
100	.95	.85	.30	1.00	.46	.40	.283	.016	.066	.001
100	.95	.85	.30	1.00	.46	.60	.269	.014	.074	.002
500	.85	.85	.20	.75	.30	.40	.946	.021	.094	.001
500	.85	.85	.20	.75	.30	.60	.915	.019	.117	.001
500	.85	.85	.20	1.00	.40	.40	.934	.020	.088	.001
500	.85	.85	.20	1.00	.40	.60	.913	.018	.114	.001
500	.85	.85	.30	.75	.34	.40	.971	.025	.132	.001
500	.85	.85	.30	.75	.34	.60	.941	.021	.169	.002
500	.85	.85	.30	1.00	.46	.40	.955	.022	.129	.001
500	.85	.85	.30	1.00	.46	.60	.936	.020	.157	.001
500	.95	.85	.20	.75	.30	.40	.900	.018	.104	.001
500	.95	.85	.20	.75	.30	.60	.850	.015	.125	.001
500	.95	.85	.20	1.00	.40	.40	.849	.014	.102	.001
500	.95	.85	.20	1.00	.40	.60	.811	.013	.121	.001
500	.95	.85	.30	.75	.34	.40	.935	.020	.149	.001
500	.95	.85	.30	.75	.34	.60	.884	.016	.187	.002
500	.95	.85	.30	1.00	.46	.40	.879	.016	.142	.001
500	.95	.85	.30	1.00	.46	.60	.833	.014	.177	.002
1,000	.85	.85	.20	.75	.30	.40	.999	.022	.140	.001
1,000	.85	.85	.20	.75	.30	.60	.997	.018	.171	.001
1,000	.85	.85	.20	1.00	.40	.40	.999	.020	.132	.001
1,000	.85	.85	.20	1.00	.40	.60	.996	.018	.164	.001
1,000	.85	.85	.30	.75	.34	.40	.999	.025	.228	.001
1,000	.85	.85	.30	.75	.34	.60	.999	.021	.284	.002
1,000	.85	.85	.30	1.00	.46	.40	.999	.022	.216	.001
1,000	.85	.85	.30	1.00	.46	.60	.999	.020	.268	.001
1,000	.95	.85	.20	.75	.30	.40	.996	.018	.167	.001
1,000	.95	.85	.20	.75	.30	.60	.988	.015	.204	.001
1,000	.95	.85	.20	1.00	.40	.40	.987	.015	.158	.001
1,000	.95	.85	.20	1.00	.40	.60	.979	.013	.193	.001
1,000	.95	.85	.30	.75	.34	.40	.998	.020	.271	.001
1,000	.95	.85	.30	.75	.34	.60	.993	.016	.331	.002
1,000	.95	.85	.30	1.00	.46	.40	.993	.016	.259	.001
1,000	.95	.85	.30	1.00	.46	.60	.986	.014	.315	.002
					M					
533	.90	.85	.25	.875	.375	.50	.738	.018	.138	.001
					Part 2 ($\Delta\psi^2 = .02$)					
100	.85	.85	.20	.75	.30	.40	.580	.038	.070	.001
100	.85	.85	.20	.75	.30	.60	.514	.032	.075	.002
100	.85	.85	.20	1.00	.40	.40	.546	.033	.070	.001
100	.85	.85	.20	1.00	.40	.60	.499	.030	.076	.001
100	.85	.85	.30	.75	.34	.40	.653	.044	.084	.002

Table 6 (*continued*)

							Intercept differences		Slope differences	
N	ρ_{XX}	ρ_{yy}	*p*	$\mu_{1x} - \mu_{0x}$	ρ_{xg}	*RR*	Power	ΔR^2	Power	ΔR^2
100	.85	.85	.30	.75	.34	.60	.571	.036	.096	.003
100	.85	.85	.30	1.00	.46	.40	.595	.037	.078	.002
100	.85	.85	.30	1.00	.46	.60	.547	.033	.094	.003
100	.95	.85	.20	.75	.30	.40	.531	.033	.068	.001
100	.95	.85	.20	.75	.30	.60	.457	.027	.078	.002
100	.95	.85	.20	1.00	.40	.40	.463	.027	.070	.001
100	.95	.85	.20	1.00	.40	.60	.418	.023	.076	.002
100	.95	.85	.30	.75	.34	.40	.594	.038	.086	.003
100	.95	.85	.30	.75	.34	.60	.510	.030	.109	.004
100	.95	.85	.30	1.00	.46	.40	.510	.029	.086	.002
100	.95	.85	.30	1.00	.46	.60	.448	.025	.106	.003
500	.85	.85	.20	.75	.30	.40	.998	.038	.142	.001
500	.85	.85	.20	.75	.30	.60	.993	.032	.188	.002
500	.85	.85	.20	1.00	.40	.40	.995	.033	.136	.001
500	.85	.85	.20	1.00	.40	.60	.992	.030	.177	.002
500	.85	.85	.30	.75	.34	.40	.999	.043	.227	.002
500	.85	.85	.30	.75	.34	.60	.998	.036	.288	.003
500	.85	.85	.30	1.00	.46	.40	.999	.036	.212	.002
500	.85	.85	.30	1.00	.46	.60	.995	.032	.273	.003
500	.95	.85	.20	.75	.30	.40	.996	.033	.169	.002
500	.95	.85	.20	.75	.30	.60	.987	.027	.213	.002
500	.95	.85	.20	1.00	.40	.40	.990	.027	.161	.001
500	.95	.85	.20	1.00	.40	.60	.977	.023	.202	.002
500	.95	.85	.30	.75	.34	.40	.998	.037	.256	.003
500	.95	.85	.30	.75	.34	.60	.994	.030	.326	.004
500	.95	.85	.30	1.00	.46	.40	.994	.029	.243	.002
500	.95	.85	.30	1.00	.46	.60	.986	.024	.310	.003
1,000	.85	.85	.20	.75	.30	.40	1.000	.038	.244	.001
1,000	.85	.85	.20	.75	.30	.60	1.000	.032	.311	.002
1,000	.85	.85	.20	1.00	.40	.40	1.000	.033	.232	.001
1,000	.85	.85	.20	1.00	.40	.60	1.000	.030	.291	.002
1,000	.85	.85	.30	.75	.34	.40	1.000	.043	.416	.002
1,000	.85	.85	.30	.75	.34	.60	1.000	.036	.510	.003
1,000	.85	.85	.30	1.00	.46	.40	1.000	.036	.388	.002
1,000	.85	.85	.30	1.00	.46	.60	1.000	.033	.485	.003
1,000	.95	.85	.20	.75	.30	.40	1.000	.033	.270	.001
1,000	.95	.85	.20	.75	.30	.60	1.000	.027	.355	.002
1,000	.95	.85	.20	1.00	.40	.40	1.000	.027	.262	.001
1,000	.95	.85	.20	1.00	.40	.60	1.000	.023	.338	.002
1,000	.95	.85	.30	.75	.34	.40	1.000	.037	.467	.003
1,000	.95	.85	.30	.75	.34	.60	1.000	.030	.576	.004
1,000	.95	.85	.30	1.00	.46	.40	1.000	.028	.447	.002
1,000	.95	.85	.30	1.00	.46	.60	1.000	.025	.549	.003
					M					
533	.90	.85	.25	.875	.375	.50	.840	.032	.229	.002
					Part 3 ($\Delta\psi^2 = .03$)					
100	.85	.85	.20	.75	.30	.40	.733	.053	.071	.002
100	.85	.85	.20	.75	.30	.60	.662	.044	.078	.002
100	.85	.85	.20	1.00	.40	.40	.689	.046	.070	.002
100	.85	.85	.20	1.00	.40	.60	.643	.041	.077	.002
100	.85	.85	.30	.75	.34	.40	.793	.060	.092	.003
100	.85	.85	.30	.75	.34	.60	.725	.050	.115	.004
100	.85	.85	.30	1.00	.46	.40	.738	.050	.087	.003
100	.85	.85	.30	1.00	.46	.60	.692	.044	.111	.004
100	.95	.85	.20	.75	.30	.40	.693	.047	.084	.002
100	.95	.85	.20	.75	.30	.60	.616	.038	.089	.003
100	.95	.85	.20	1.00	.40	.40	.620	.038	.082	.002
100	.95	.85	.20	1.00	.40	.60	.554	.032	.087	.002
100	.95	.85	.30	.75	.34	.40	.754	.053	.106	.004
100	.95	.85	.30	.75	.34	.60	.671	.043	.133	.005

(*table continues*)

Table 6 (*continued*)

							Intercept differences		Slope differences	
N	ρ_{XX}	ρ_{yy}	p	$\mu_{1x} - \mu_{0x}$	ρ_{xg}	RR	Power	ΔR^2	Power	ΔR^2
100	.95	.85	.30	1.00	.46	.40	.654	.040	.101	.003
100	.95	.85	.30	1.00	.46	.60	.600	.034	.128	.004
500	.85	.85	.20	.75	.30	.40	1.000	.053	.195	.002
500	.85	.85	.20	.75	.30	.60	1.000	.045	.239	.002
500	.85	.85	.20	1.00	.40	.40	1.000	.046	.181	.002
500	.85	.85	.20	1.00	.40	.60	.999	.041	.225	.002
500	.85	.85	.30	.75	.34	.40	1.000	.060	.324	.003
500	.85	.85	.30	.75	.34	.60	1.000	.050	.414	.004
500	.85	.85	.30	1.00	.46	.40	1.000	.049	.301	.003
500	.85	.85	.30	1.00	.46	.60	1.000	.044	.389	.004
500	.95	.85	.20	.75	.30	.40	.999	.047	.222	.002
500	.95	.85	.20	.75	.30	.60	.999	.039	.269	.003
500	.95	.85	.20	1.00	.40	.40	.999	.038	.209	.002
500	.95	.85	.20	1.00	.40	.60	.998	.033	.255	.002
500	.95	.85	.30	.75	.34	.40	1.000	.053	.372	.004
500	.95	.85	.30	.75	.34	.60	1.000	.043	.474	.005
500	.95	.85	.30	1.00	.46	.40	1.000	.040	.353	.003
500	.95	.85	.30	1.00	.46	.60	.999	.035	.458	.005
1,000	.85	.85	.20	.75	.30	.40	1.000	.052	.332	.002
1,000	.85	.85	.20	.75	.30	.60	1.000	.045	.432	.002
1,000	.85	.85	.20	1.00	.40	.40	1.000	.046	.312	.002
1,000	.85	.85	.20	1.00	.40	.60	1.000	.041	.406	.002
1,000	.85	.85	.30	.75	.34	.40	1.000	.060	.550	.003
1,000	.85	.85	.30	.75	.34	.60	1.000	.050	.693	.004
1,000	.85	.85	.30	1.00	.46	.40	1.000	.049	.519	.003
1,000	.85	.85	.30	1.00	.46	.60	1.000	.044	.659	.004
1,000	.95	.85	.20	.75	.30	.40	1.000	.046	.384	.002
1,000	.95	.85	.20	.75	.30	.60	1.000	.039	.497	.003
1,000	.95	.85	.20	1.00	.40	.40	1.000	.037	.362	.002
1,000	.95	.85	.20	1.00	.40	.60	1.000	.033	.472	.002
1,000	.95	.85	.30	.75	.34	.40	1.000	.053	.624	.004
1,000	.95	.85	.30	.75	.34	.60	1.000	.043	.762	.005
1,000	.95	.85	.30	1.00	.46	.40	1.000	.040	.599	.003
1,000	.95	.85	.30	1.00	.46	.60	1.000	.034	.738	.005
					M					
533	.90	.85	.25	.875	.375	.50	.892	.045	.307	.003
					Part 4 ($\Delta\psi^2 = .04$)					
100	.85	.85	.20	.75	.30	.40	.826	.065	.086	.002
100	.85	.85	.20	.75	.30	.60	.769	.055	.092	.003
100	.85	.85	.20	1.00	.40	.40	.787	.056	.083	.002
100	.85	.85	.20	1.00	.40	.60	.744	.050	.089	.003
100	.85	.85	.30	.75	.34	.40	.878	.074	.118	.004
100	.85	.85	.30	.75	.34	.60	.823	.062	.136	.005
100	.85	.85	.30	1.00	.46	.40	.828	.060	.114	.004
100	.85	.85	.30	1.00	.46	.60	.784	.054	.129	.005
100	.95	.85	.20	.75	.30	.40	.799	.059	.091	.003
100	.95	.85	.20	.75	.30	.60	.729	.048	.108	.003
100	.95	.85	.20	1.00	.40	.40	.732	.047	.090	.002
100	.95	.85	.20	1.00	.40	.60	.668	.040	.101	.003
100	.95	.85	.30	.75	.34	.40	.855	.066	.135	.005
100	.95	.85	.30	.75	.34	.60	.777	.054	.165	.006
100	.95	.85	.30	1.00	.46	.40	.768	.050	.126	.004
100	.95	.85	.30	1.00	.46	.60	.701	.043	.155	.006
500	.85	.85	.20	.75	.30	.40	1.000	.066	.249	.002
500	.85	.85	.20	.75	.30	.60	1.000	.057	.316	.003
500	.85	.85	.20	1.00	.40	.40	1.000	.057	.238	.002
500	.85	.85	.20	1.00	.40	.60	1.000	.052	.295	.003
500	.85	.85	.30	.75	.34	.40	1.000	.075	.421	.004
500	.85	.85	.30	.75	.34	.60	1.000	.063	.529	.006

Table 6 (*continued*)

							Intercept differences		Slope differences	
N	ρ_{XX}	ρ_{yy}	*p*	$\mu_{1x} - \mu_{0x}$	ρ_{xg}	*RR*	Power	ΔR^2	Power	ΔR^2
500	.85	.85	.30	1.00	.46	.40	1.000	.061	.395	.004
500	.85	.85	.30	1.00	.46	.60	1.000	.055	.501	.005
500	.95	.85	.20	.75	.30	.40	1.000	.060	.294	.003
500	.95	.85	.20	.75	.30	.60	1.000	.050	.368	.004
500	.95	.85	.20	1.00	.40	.40	1.000	.048	.275	.002
500	.95	.85	.20	1.00	.40	.60	1.000	.042	.351	.003
500	.95	.85	.30	.75	.34	.40	1.000	.067	.485	.005
500	.95	.85	.30	.75	.34	.60	1.000	.055	.585	.007
500	.95	.85	.30	1.00	.46	.40	1.000	.050	.465	.004
500	.95	.85	.30	1.00	.46	.60	1.000	.044	.560	.006
1,000	.85	.85	.20	.75	.30	.40	1.000	.067	.436	.002
1,000	.85	.85	.20	.75	.30	.60	1.000	.057	.549	.003
1,000	.85	.85	.20	1.00	.40	.40	1.000	.057	.411	.002
1,000	.85	.85	.20	1.00	.40	.60	1.000	.052	.522	.003
1,000	.85	.85	.30	.75	.34	.40	1.000	.075	.701	.004
1,000	.85	.85	.30	.75	.34	.60	1.000	.064	.827	.006
1,000	.85	.85	.30	1.00	.46	.40	1.000	.061	.665	.004
1,000	.85	.85	.30	1.00	.46	.60	1.000	.055	.796	.005
1,000	.95	.85	.20	.75	.30	.40	1.000	.060	.508	.003
1,000	.95	.85	.20	.75	.30	.60	1.000	.051	.625	.004
1,000	.95	.85	.20	1.00	.40	.40	1.000	.048	.484	.002
1,000	.95	.85	.20	1.00	.40	.60	1.000	.043	.597	.003
1,000	.95	.85	.30	.75	.34	.40	1.000	.067	.776	.005
1,000	.95	.85	.30	.75	.34	.60	1.000	.056	.874	.007
1,000	.95	.85	.30	1.00	.46	.40	1.000	.050	.747	.004
1,000	.95	.85	.30	1.00	.46	.60	1.000	.044	.860	.006
					M					
533	.90	.85	.25	.875	.375	.50	.926	.056	.386	.004

Note. $\Delta\psi^2$ = population-based effect size (i.e., test bias); ρ_{XX} = test reliability; ρ_{YY} = criterion reliability; *p* = proportion of minority group sample size to total sample size; $\mu_{1x} - \mu_{0x}$ = average standard deviation-unit difference between majority and minority test scores; ρ_{XG} = correlation between test scores and moderator *G*; *RR* = range restriction; ΔR^2 = sample-based effect size (i.e., test bias) estimate.

total sample size is .30, and the difference between majority and minority predictor scores is .50 *SD* units. Across all design cells in our simulation, mean $\Delta R^2_{\text{intercept}} = .11$, which is 69% larger than its population counterpart. As a related analysis, we selected the 122,500 cells in our design for which there is no true intercept-based bias in the population (i.e., $\Delta\psi^2_{\text{intercept}} = 0$). If intercept-based test bias assessment is accurate, the resulting Type I error rates should be close to the .05 nominal value. However, the average Type I error rate for these cells is .09. These results indicate that differences based on intercepts across groups are overestimated and, under conditions simulated in our study, Type I error rates are also inflated, suggesting that researchers are likely to reach the conclusion that differences exist favoring minority group members when these differences actually do not exist.

Tables 3–6 provide additional evidence in support of the conclusion that intercept-based differences favoring minority group members are overestimated in many situations. For example, consider results in Table 3. This table includes values for total sample size of 300 or 400, reliabilities for tests scores and the criterion of .80 or .85, proportion of minority group members to total sample size of .20 or .30, average difference of .50 *SD* units between majority and minority test scores, range restriction of .30 or .40, and correlation between test scores and ethnicity of .20 or .23. Across all conditions for which there is intercept-based test bias of $\Delta\psi^2_{\text{intercept}} = .01$, the estimated sample-based $\Delta R^2_{\text{intercept}}$ is .026. This means that, based on the sample results, one concludes that test bias is more than 2.5 times larger than it actually is in the population of scores. Also, across all values in Table 3, when $\Delta\psi^2_{\text{intercept}} = .02$, its sample-based counterpart is $\Delta R^2_{\text{intercept}} = .047$, which is also more than twice as large as its population counterpart. Not surprisingly, given this large degree of overestimation of the test bias effect, and the positive relationship between effect size and statistical power, Table 3 also shows that statistical power is .88 for $\Delta\psi^2_{\text{intercept}} = .01$ and .988 for $\Delta\psi^2_{\text{intercept}} = .02$.

Results shown in Tables 4–6 follow the same pattern. For all conditions, the sample-based estimates of test bias based on intercept differences across the groups consistently overestimate the true degree of test bias favoring minority group members in the population. For example, across all conditions in Table 4 for which $\Delta\psi^2_{\text{intercept}} = .03$, the average $\Delta R^2_{\text{intercept}} = .082$, and for all conditions for which $\Delta\psi^2_{\text{intercept}} = .04$, the average $\Delta R^2_{\text{intercept}} = .108$. Similarly, as shown in Table 5, for $\Delta\psi^2_{\text{intercept}} = .005$, the average $\Delta R^2_{\text{intercept}} = .012$. For Table 6, which includes values seen as prototypical in GMA testing, Type I error rates are inflated and intercept-based bias is overestimated (favoring the minority group) in a systematic fashion. In Table 6, Part 1 shows that the sample intercept-based bias is 80% larger than its population counterpart (i.e., .018 vs. .01), Part 2 shows that the sample intercept-based bias is 60% larger than its population

Table 7
Standardized Models Regressing Type I Error Rates and Statistical Power on Manipulated Parameters for Intercept-Based and Slope-Based Test Bias Assessment

	Type I error[a]		Power[b]	
Parameter	Intercept	Slope	Intercept	Slope
$\Delta\psi^2$			.242	.514
N	.097	−.001	.195	.427
ρ_{XX}	−.157	.000	−.025	.109
ρ_{YY}	.023	.000, *ns*	.039	.097
p	−.034	.000, *ns*	.024	.272
$\mu_{1x} - \mu_{0x}$	.084	.000, *ns*	−.018	−.275
RR	.022	.000	−.083	.314
ρ_{XG}	.117	.000, *ns*	.036	.409
$\Delta\psi^2 \times N$			−.214	.034
$\Delta\psi^2 \times \rho_{XX}$			.030	.009
$\Delta\psi^2 \times \rho_{YY}$			−.039	.009
$\Delta\psi^2 \times p$			−.051	.040
$\Delta\psi^2 \times (\mu_{1x} - \mu_{0x})$			−.023	−.012
$\Delta\psi^2 \times RR$			.086	.006
$\Delta\psi^2 \times \rho_{XG}$			.019	.013
$N \times \rho_{XX}$	−.087	.000	.014	−.003
$N \times \rho_{YY}$	.013	.000	−.032	−.002
$N \times p$	−.008	.000	−.037	.021
$N \times (\mu_{1x} - \mu_{0x})$	.037	.000, *ns*	−.006	−.009
$N \times RR$	.010	.000	.068	−.004
$N \times \rho_{XG}$	.068	.000, *ns*	.013	.011
$\rho_{XX} \times \rho_{YY}$	−.021	.000, *ns*	.004	.003
$\rho_{XX} \times p$	.014	.000, *ns*	−.002	−.011
$\rho_{XX} \times (\mu_{1x} - \mu_{0x})$	−.058	.000, *ns*	−.031	−.001, *ns*
$\rho_{XX} \times RR$	−.018	.000	−.013	−.006
$\rho_{XX} \times \rho_{XG}$	−.115	.000, *ns*	.002	.007
$\rho_{YY} \times p$	−.002	.000	−.007	−.001
$\rho_{YY} \times (\mu_{1x} - \mu_{0x})$	.010	.000, *ns*	−.001, *ns*	−.002
$\rho_{YY} \times RR$	.002	.000, *ns*	.014	−.005
$\rho_{YY} \times \rho_{XG}$	.016	.000, *ns*	.003	.003
$p \times RR$	−.005	.000	−.008	−.003
$p \times \rho_{XG}$	−.039	.000	−.015	−.180
$(\mu_{1x} - \mu_{0x}) \times RR$	.002, *ns*	.000, *ns*	.021	−.004
$(\mu_{1x} - \mu_{0x}) \times \rho_{XG}$	.066	.000, *ns*	−.014	−.009
$RR \times \rho_{XG}$	.028	.000	−.001, *ns*	.005

Note. Intercept = intercept-based test bias; slope = slope-based test bias; $\Delta\psi^2$ = population-based effect size (i.e., test bias); RR = range restriction; ρ_{XX} = test reliability; ρ_{YY} = criterion reliability; p = proportion of minority group sample size to total sample size; $\mu_{1x} - \mu_{0x}$ = average standard deviation-unit difference between majority and minority test scores; ρ_{XG} = correlation between test scores and demographic variable (G was coded using 1 for the majority group and 0 for the minority group).
[a] The regression models using Type I error rates as the criterion were computed using the 122,500 cells in the design for which test bias does not exist in the population (i.e., $\Delta\psi^2 = 0$), so there is no variance in this variable. [b] The regression models using statistical power as the criterion were computed using the 3,062,500 cells in the design for which test bias exists in the population. The coefficient for $p \times (\mu_{1x} - \mu_{0x})$ was not derived due to collinearity with $p \times \rho_{XG}$. For all regression coefficients, $p < .01$, except for those denoted as *ns* (i.e., statistically nonsignificant, $p > .05$).

counterpart (i.e., .032 vs. .02), Part 3 shows that the sample intercept-based bias is 50% larger than its population counterpart (i.e., .045 vs. .03), and Part 4 shows that the sample intercept-based bias is 40% larger than it is in the population (i.e., .056 vs. .04).

In sum, the degree of intercept-based test bias is consistently overestimated given a large range of conditions observed in the validation research literature. Overall, the smaller the population effect, the larger the degree of overestimation. For population effects of $\Delta\psi^2_{\text{intercept}} = .03$ or smaller, test bias is believed to be more than twice as large as it is in actuality. Moreover, when there is no intercept-based bias in the population, researchers are likely to conclude incorrectly that bias favoring minority group members actually exists.

Relative impact of manipulated parameters on Type I error rates to assess intercept-based test bias. Similar to the regression analyses described earlier regarding slope-based test bias, Table 7 shows regression results to clarify the relative impact of each of the manipulated parameters on the Type I error rates for assessing intercept-based test bias. In Table 7, the columns labeled *Type I error: Intercept* show that the factors that have the largest impact on Type I error rates are (a) reliability of test scores ($\beta = -.16$), (b) the correlation between test scores and minority status

($\beta = .12$), and (c) total sample size ($\beta = .10$). Each of these effects is in the predicted direction such that Type I error rates increase with lower test score reliability, a larger difference in mean test scores between groups favoring majority group members, and a larger total sample size.

An issue that has not yet been reported in the literature is whether the parameters we manipulated in the simulation have interactive effects on Type I error rates when assessing possible intercept-based test bias. Table 7 also shows that all but one two-way interaction was statistically significant. The model including first-order effects only resulted in $R^2 = .50$, and the addition of the two-way terms increased this value to $R^2 = .86$. For example, the coefficient for $\rho_{XX} \times RR$ is $\beta = -.02$, which means that for a 1 *SD* unit decrease in *RR*, the slope of Type I error rates on test score reliability increases by .02 *SD* units, holding all other variables in the model constant. Recall that a range restriction value of 1.00 means all scores are retained, whereas a range restriction value of .10 means only 10% of scores are retained. So, this two-way interaction effect indicates that the positive effect of test score unreliability on Type I error rates is amplified as selectivity increases.

Results in Table 7 also indicate that test score reliability and test score differences produce an upward bias in the Type I error rates for the intercept-differences test. Moreover, test score reliability and test score differences across groups interact with each other (i.e., $\rho_{XX} \times [\mu_{1x} - \mu_{0x}]$) as well as with other parameters in producing an upward bias in Type I error rates. For example, test score reliability interacts with proportion of minority members to total sample size, range restriction (as noted above), and criterion score reliability; and mean test score differences across groups interact with total sample size and reliability of criterion scores. All of these two-way interactions were in the expected direction such that the negative effect of test score reliability and the positive effect of test score differences across groups on Type I error rates are amplified as values for the other parameters increase.

In sum, results included in Tables 3–6 indicate that intercept-based test bias is likely to be overestimated under a large range of conditions of total sample size, range restriction, predictor and criterion scores reliability, proportion of minority group sample size to total sample size, and differences between test scores across groups that are frequently observed in the preemployment testing literature. In fact, these results indicate that intercept-based test bias favoring minority group members could be found even when bias does not exist in the population.

Comparison of Relative Accuracy of Intercept-Based and Slope-Based Test Bias Assessment

Results reported in Table 7 provide comparative information regarding the relative impact of methodological and statistical artifacts on the accuracy of slope-based and intercept-based test bias assessment. First, a perusal of Type I error rates suggests that when no bias exists in the population, the manipulated parameters have very little impact on the accuracy of the slope-based test. Of all of the regression coefficients (see columns labeled *Type I error: Slope*), the largest absolute value is .001 and four are statistically nonsignificant. On the other hand, however, Type I error rates for intercept-based bias assessment (see columns labeled *Type I error: Intercept*) are affected by the methodological and statistical artifacts we manipulated: Each of the first-order effects is statistically significant, and all but one of the two-way interaction effects is statistically significant and regression weights are as large as $|.16|$ (for test score reliability) and $|.12|$ (for correlation between test scores and minority status).

A perusal of results regarding statistical power suggests that methodological and statistical artifacts have a stronger effect on the power of the slope-based test compared to the intercept-based test. The average of the absolute value of the first-order regression coefficients for the slope-based test (column labeled *Power: Slope*) is .30, whereas the same average for the first-order effects predicting power for intercept-based test bias is .08. In other words, the impact of the parameters we manipulated on power is 3.75 times larger for the slope-based test compared to the intercept test.

Figure 2 includes a graphic representation of the effects of true population test bias, total sample size, range restriction, and minority group proportion on the statistical power to detect test bias when bias exists in the population. Each of the four panels in this figure illustrates the result that methodological and statistical artifacts have a more detrimental effect on the power of the slope test compared to the intercept test. Figure 2A shows that, across values of all other parameters, the intercept test reaches power of .90 when $\Delta\psi^2_{\text{intercept}} \cong .02$, whereas the slope test does not reach power of .80 when $\Delta\psi^2_{\text{slope}} = .15$. Similarly, Figure 2B shows that the intercept test reaches power of .90 when total sample size is $\cong 200$, but the slope test does not reach power of .80 when total sample size is 1,000. Figure 2A also illustrates that more stringent range restriction actually increases the power of the intercept test in the presence of slope-based differences. On the other hand, as has been shown in previous research, more severe range restriction decreases the power of the slope test (Aguinis & Stone-Romero, 1997). Lastly, Figure 2C illustrates that as the proportion of minority group sample size to total sample size approaches .5, power improves for both the intercept and the slope test. However, deviations from the .5 value have a more detrimental impact on the power of the slope test compared to the intercept test, as is the case with the other methodological and statistical artifacts.

Figure 3 includes graphic displays of the effects of test score reliability, minority group proportion, group mean differences regarding test scores, and total sample size on Type I error rates for the 122,500 cells in our design for which bias does not exist in the population (i.e., $\Delta\psi^2 = 0$ for both the intercept and the slope effect). Each of the four panels illustrates the result that methodological and statistical artifacts have an effect on the Type I error rates for the intercept test but not the slope test. For each of the four panels, Type I error rates remain close to the nominal value of .05 for the slope test regardless of the value of the manipulated parameters. In sharp contrast, Type I error rates for the intercept test deviate substantially from the nominal .05 value as parameters take on values frequently observed in human resource selection research and practice. For example, Figure 3A shows that when test score reliability is about .80, Type I error rate is about .10; Figure 3B shows that when minority group proportion is about .3, Type I error rate is about .10; and Figure 3C shows that when test score differences between groups is 1.0, Type I error rate is about .15. Figure 3A illustrates the phenomenon for which we provided a new analytic proof in the introduction. That is, Type I error rates for the intercept test remain at the nominal level when test score

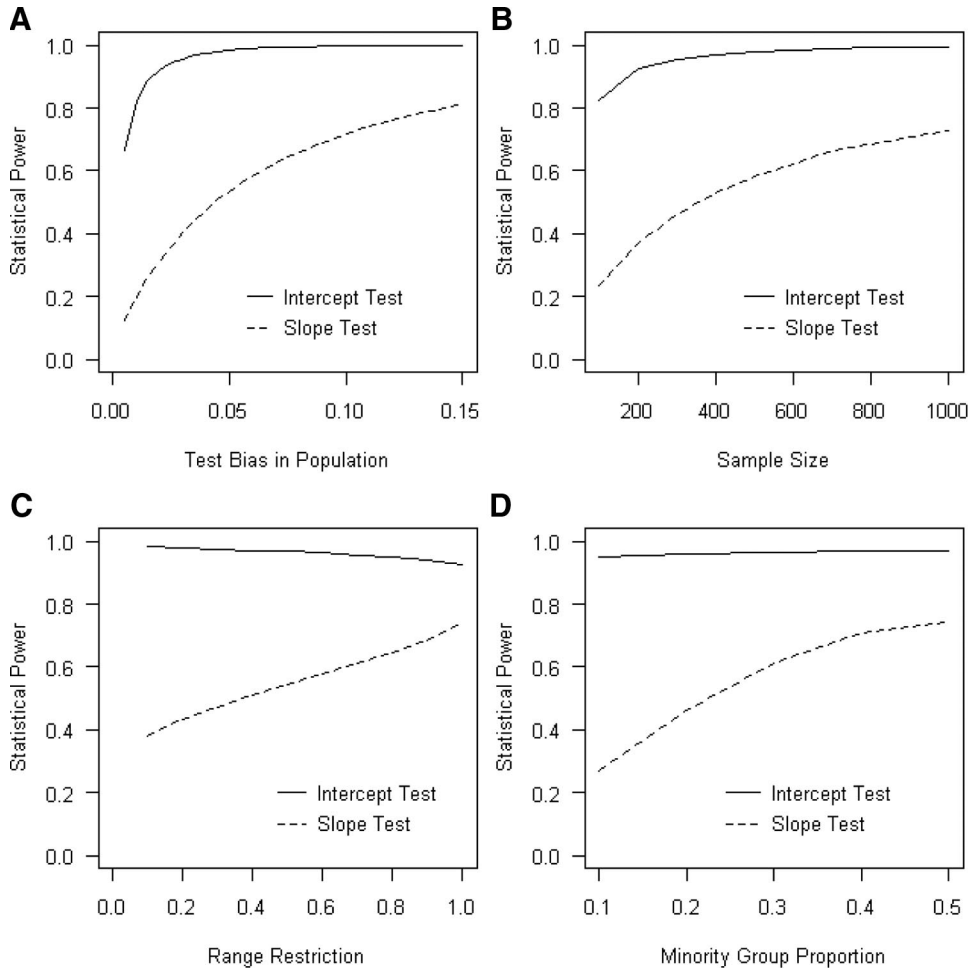

Figure 2. Relationships between statistical power to detect test bias and population test bias (Panel A), total sample size (Panel B), range restriction (Panel C), and minority group proportion (Panel D) based on the 3,062,500 design cells for which population test bias exists (i.e., $\Delta\psi^2 > 0$).

reliability is perfect. However, as measurement error is introduced in the test scores, Type I error rates increase leading to the incorrect conclusion that there is intercept-based bias favoring members of the minority group.

In sum, a comparison of the relative accuracy of intercept-based and slope-based test bias assessment indicates that the intercept-based test is more susceptible to inflation of Type I error rates (i.e., concluding there is bias when it does not exist or overestimating the presence of test bias) compared to the slope-based test. Alternatively, the slope-based test is more susceptible to the detrimental impact of methodological and statistical artifacts on statistical power (i.e., concluding there is no bias when it actually exists in the population) compared to the intercept-based test.

Additional Monte Carlo Simulation

As noted earlier, our analytic work in Appendix B explaining the mechanisms through which test score unreliability, differences in mean test scores between groups, and range restriction affect Type I error rates of intercept-based bias assessment assumes there is no slope-based bias. We conducted an additional simulation to show how range restriction and slope-based differences across groups affect results of the intercept-based test. For this additional Monte Carlo study, we generated data using the same approach described in the Method section for the main simulation with the exception that we set $\Delta\psi^2_{\text{intercept}} = 0$, and for all but 16 design cells we set $\Delta\psi^2_{\text{slope}} > 0$. Additionally, we included values of $\rho_{XX} = .90$ and .95, $\rho_{YY} = .90$ and .95, $p = .2$, and range restriction of .40 and .60 for *N*s of 250 and 1,000. Results are shown in Table 8.

In addition to conducting this second simulation, we used Equation 4 to estimate the degree of intercept-based bias analytically when $\Delta\psi^2_{\text{slope}} = 0$. Results are also included in Table 8. The columns labeled *Accuracy of analytic estimates* show the difference between the simulation-generated and analytically derived sample-based $\Delta R^2_{\text{intercept}}$ and the difference between the simulation-generated and analytically derived associated Type I error rate for the null hypothesis $\Delta\psi^2_{\text{intercept}} = 0$.

Results reported in Table 8 lead to two conclusions. First, Lines 1–8 show that, when true slope-based differences are zero, Type I error rates for a situation when intercept-based

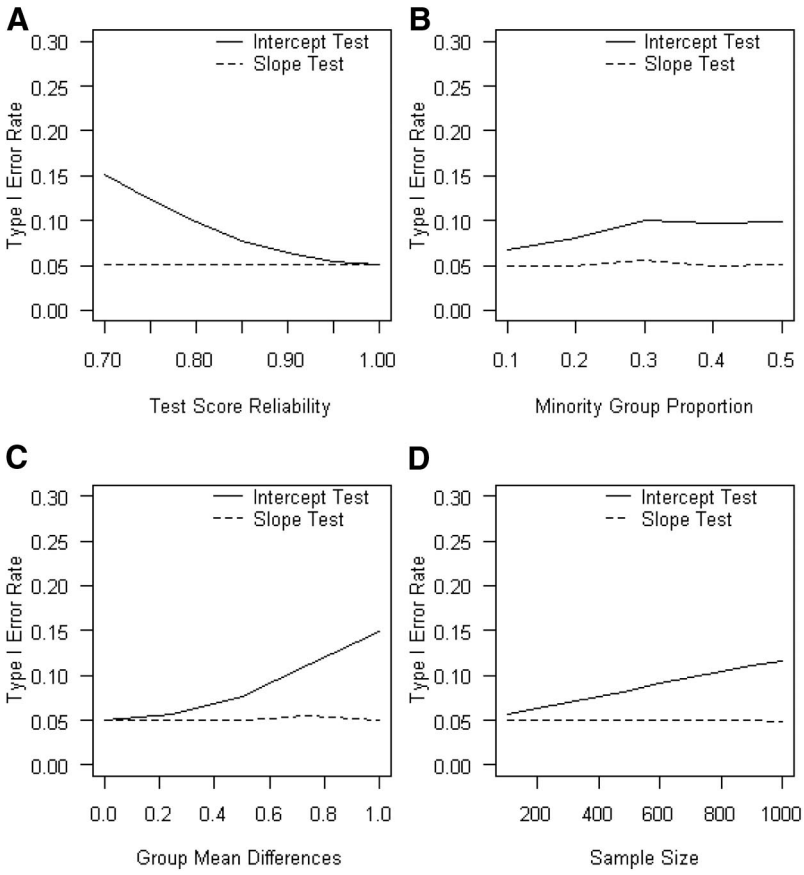

Figure 3. Relationships between Type I error rate and test score reliability (Panel A), minority group proportion (Panel B), group mean differences regarding test scores (Panel C), and total sample size (Panel D) based on the 122,500 design cells for which there is no test bias in the population (i.e., $\Delta\psi^2 = 0$).

differences are zero can be inflated but are not greater than .082. However, the remaining entries in this table show that as true slope-based differences range from $\Delta\psi^2 = .02$ to .04, which are still small values but fairly typical (Aguinis et al., 2005; McClelland & Judd, 1993), Type I error rates reach values as high as in the .80s, although there is a zero intercept-based difference across groups in the population. Consistent with this result, the sample-based test-bias effect is always positive for each condition, even though the true intercept-based test bias is zero in the population.

A second result worth noting from Table 8 is that Lines 1–8 show that the Type I error rates estimated using the analytic solution described in Appendix B are virtually identical to those derived from the simulation when slope-based bias does not exist in the population (which is an assumption of the analytic approximation). The columns labeled $\Delta(\Delta R^2)$ show the difference between simulation-generated and analytically derived intercept-based test bias. Across the 16 design cells included in Lines 1–8 in Table 8, the average difference between these values is only .000475. The columns labeled $\Delta(Prob.)$ show the difference between simulation-generated and analytically derived Type I error rates. Across the 16 design cells in Lines 1–8 in Table 8, the average of the absolute differences between these values is only .004.

Discussion

The goal of the present study was to revisit established conclusions regarding test bias in preemployment testing and provide an alternative explanation for the consistent results reported over the past 40 years of research. As noted in the *Principles for the Validation and Use of Personnel Selection Procedures*,

> predictive bias has been examined extensively in the cognitive ability domain. For White–African American and White–Hispanic comparisons, slope differences are rarely found; while intercept differences are not uncommon, they typically take the form of overprediction of minority group performance. (SIOP, 2003, p. 32)

In spite of these established conclusions, there are several reasons why slope-based bias may actually exist and why intercept-based bias favoring minority group members may be smaller than it is believed to be or not exist at all. Regarding the finding that no differences in slopes exist, Monte Carlo simulations and literature reviews have revealed that conclusions regarding the absence of slope differences across groups may not be warranted. More precisely, statistical power is typically inadequate. From a substantive standpoint, slope-based test bias is expected as a result of sociohistorical–cultural and social psychological explanations. Regarding intercept-based bias, we provide new analytic proof that

Table 8
Simulation Results for the Effects of Sample Size, Slope Differences, Criterion and Predictor Reliability, and Range Restriction on Type I Errors for Intercept-Based Test Bias Assessment

				$\rho_{XX} = .90$				$\rho_{XX} = .95$			
				Intercept-based differences		Accuracy of analytic estimates		Intercept-based differences		Accuracy of analytic estimates	
N	$\Delta\psi^2$	ρ_{YY}	*RR*	ΔR^2	*Prob.*	$\Delta(\Delta R^2)$	Δ(*Prob.*)	ΔR^2	*Prob.*	$\Delta(\Delta R^2)$	Δ(*Prob.*)
250	.00	.80	.40	.0040	.058	.0008	.0020	.0038	.052	.0008	.0010
250	.00	.80	.60	.0039	.056	.0008	−.0020	.0037	.056	.0008	.0040
250	.00	.90	.40	.0039	.056	.0007	−.0010	.0036	.047	.0006	−.0050
250	.00	.90	.60	.0038	.049	.0007	−.0100	.0038	.058	.0010	.0060
1,000	.00	.80	.40	.0012	.079	.0003	.0050	.0009	.049	.0001	−.0070
1,000	.00	.80	.60	.0012	.082	.0002	−.0010	.0010	.053	.0002	−.0050
1,000	.00	.90	.40	.0011	.072	.0001	−.0060	.0010	.057	.0002	.0000
1,000	.00	.90	.60	.0012	.081	.0002	−.0070	.0009	.057	.0001	−.0020
250	.01	.80	.40	.0058	.115			.0055	.107		
250	.01	.80	.60	.0046	.081			.0042	.067		
250	.01	.90	.40	.0059	.114			.0057	.118		
250	.01	.90	.60	.0047	.085			.0043	.074		
1,000	.01	.80	.40	.0032	.349			.0027	.283		
1,000	.01	.80	.60	.0019	.188			.0016	.138		
1,000	.01	.90	.40	.0035	.388			.0030	.335		
1,000	.01	.90	.60	.0021	.209			.0017	.161		
250	.02	.80	.40	.0073	.170			.0068	.156		
250	.02	.80	.60	.0051	.100			.0046	.084		
250	.02	.90	.40	.0077	.186			.0073	.169		
250	.02	.90	.60	.0051	.096			.0048	.091		
1,000	.02	.80	.40	.0045	.505			.0042	.477		
1,000	.02	.80	.60	.0023	.242			.0020	.209		
1,000	.02	.90	.40	.0051	.587			.0046	.527		
1,000	.02	.90	.60	.0026	.283			.0021	.216		
250	.03	.80	.40	.0088	.220			.0081	.193		
250	.03	.80	.60	.0056	.112			.0053	.109		
250	.03	.90	.40	.0094	.241			.0089	.227		
250	.03	.90	.60	.0058	.121			.0052	.110		
1,000	.03	.80	.40	.0060	.656			.0056	.616		
1,000	.03	.80	.60	.0028	.309			.0024	.252		
1,000	.03	.90	.40	.0067	.709			.0063	.679		
1,000	.03	.90	.60	.0030	.339			.0026	.286		
250	.04	.80	.40	.0103	.275			.0097	.247		
250	.04	.80	.60	.0059	.123			.0056	.118		
250	.04	.90	.40	.0111	.298			.0104	.279		
250	.04	.90	.60	.0063	.142			.0057	.129		
1,000	.04	.80	.40	.0076	.763			.0071	.736		
1,000	.04	.80	.60	.0032	.369			.0029	.324		
1,000	.04	.90	.40	.0084	.823			.0080	.801		
1,000	.04	.90	.60	.0035	.408			.0031	.359		

Note. The results are based upon 5,000 replications. The simulation was based upon $p = .2$, $\mu_{1x} - \mu_{0x} = 1.0$, and $\rho_{XY} = .5$. $\Delta\psi^2$ = population-based effect size for slope (i.e., test bias); ρ_{YY} = criterion reliability; *RR* = range restriction; ΔR^2 = sample-based effect size; *Prob.* = probability of statistically significant finding (i.e., Type I errors when $\Delta\psi^2 = 0$ and statistical power when $\Delta\psi^2 > 0$); $\Delta(\Delta R^2)$ = difference between effects from the Monte Carlo simulation and estimates derived using the analytic solution in Appendix B; Δ(*Prob.*) = difference between simulated Type I errors and analytically derived estimate.

researchers are more likely to find that performance is overpredicted for members of the minority group when the mean minority group test score is lower than the mean majority group test score and test scores are measured with less-than-perfect reliability. Both of these are conditions typically observed in human resource selection research and practice, particularly in the area of GMA testing.

We conducted a Monte Carlo simulation that generated 15 billion 925 million individual samples in 3,185,000 unique design cells. We chose to implement this type of comprehensive design to study test bias because, as noted by Linn (1978) more than 30 years ago, "the stakes are high, and the underlying issues are extremely emotional ones" (p. 507). Much like meta-analysis has changed our views on the validity of various selection tools (Schmidt & Hunter, 1998), Monte Carlo methodology is a tool that allows researchers to raise important questions about their current understanding of test bias in human resource selection.

If true slope-based differences exist and true intercept-based differences do not exist, our results suggest that researchers could make one or both of the following incorrect conclusions: (a) There is no slope-based test bias and (b) there is intercept-based test bias

favoring minority group members. Thus, our study provides an alternative explanation for the established conclusion in I/O psychology and human resource management that test bias is nonexistent and, if it exists, it occurs regarding intercept-based differences only. Specifically, our results suggest that the presence of methodological and statistical artifacts that are typically observed in human resource selection is a likely explanation for the consistent results obtained over the past 40 years of test-bias research.

Implications for Research and Practice

One important implication of our study is that, although ironic, it seems that for the past 40 years researchers have been trying to assess potential test bias with a biased procedure (Terris, 1997). Regarding the slope-based test, our results demonstrate that statistical power is consistently below the .80 standard (cf. J. Cohen, 1988). In other words, researchers are not likely to find slope-based test bias if it exists in the population. Although our introduction described several reasons why slope-based bias may exist, we do not know for certain that it does. However, what the present study demonstrates is that if slope-based bias does exist, it is very difficult if not impossible to detect in many human resource selection research contexts. Thus, it is likely that much past research has attempted to test the null hypothesis (i.e., no slope-based bias) using samples, measures, and research designs that had a very small chance of detecting bias if it existed. Our results are therefore not necessarily at odds with previous research because, for the most part, previous research has not adequately tested the slope-based bias hypothesis. Just like individual studies with flawed designs are not able to test the situational specificity hypothesis, individual studies with insufficient statistical power cannot test the differential prediction hypothesis either. This is because "studies into differential prediction that have been carried out up to now have shown some methodological weaknesses, such as small sample size, restriction of range, unreliability of assessment, and limited criteria" (Te Nijenhuis & Van der Flier, 1999, p. 172).

As early as the mid-1970s, Katzell and Dyer (1977) noted that it would be impossible to answer the question of differential validity and differential prediction unequivocally unless a rigorous test of the null hypothesis is conducted. At minimum, such a test would require the use of random samples of members from each of the groups under consideration and samples should be sufficiently large, measures sufficiently reliable, and so forth, so that statistical power to detect differential prediction is at least .80 (cf. J. Cohen, 1988). Our simulation, as well as the study by Aguinis and Stone-Romero (1997), indicates that these and other design and measurement issues have interactive effects on statistical power. Thus, one recommendation is to conduct a power analysis prior to data collection to make sure power will be sufficient to detect slope-based differences (if they exist). There are computer programs available online to conduct such a power analysis. These programs allow users to input information on several design and measurement issues such as anticipated total sample size, anticipated sample size in each subgroup, reliability for test and criterion scores, and anticipated range restriction (if any; see Aguinis, 2004a, for detailed descriptions and instructions on how to use each of these programs). In terms of selecting a targeted effect size for the power analysis, this choice can be facilitated using the Aguinis and Smith (2007) online calculator, which allows users to estimate false positives and false negatives associated with using a common regression line in the presence of slope-based test bias. Thus, one could use the Aguinis and Smith (2007) calculator to choose the effect size that leads to the largest, yet tolerable, amount of false positives and false negatives. Following these procedures, users would know they have statistical power to detect an effect size that is sufficiently large to be meaningful from a practical standpoint. From an ethical standpoint, some would argue that even if one person is misclassified due to test bias, then the test should not be used. So, this position suggests that an effect size leading to at least one individual being a false positive or false negative would be meaningful. Accordingly, the decision regarding what targeted effect size to use in the power analysis is subjective and should be made within each specific context (Aguinis et al., 2009). As noted by Sackett, De Corte, and Lievens (2009) in the context of the trade-off between validity and adverse impact,

> These decision aids do not tell the user what they should do as that is a matter of values . . . and not a technical issue. A trade-off that seems reasonable to some will be seen as inappropriate by others. (p. 469)

Regarding intercept-based bias, our analytic and empirical results demonstrate that differences favoring minority group members are likely to be found when they do not exist. Also, when they exist in the population, they are likely to be exaggerated in the samples. As expected based on the new analytic proof in Appendix B, as differences regarding mean test scores between the groups increase and test score reliability decreases, Type I error rates indicating that there are intercept-based differences favoring minority group members also increase. Thus, for many conditions in preemployment testing, one could conclude that there is intercept-based bias favoring minority group members when this is not true. Also, one could conclude that differences are larger than they are in actuality. We offer two suggestions to potentially remedy this situation: Improve test score reliability and decrease test score differences across groups. First, regarding the improvement of reliability, recent developments regarding researchers' understanding of various sources of measurement error (Schmidt, Le, & Ilies, 2003) have the potential to provide guidelines on how to improve the reliability of preemployment tests. Second, recent developments regarding how to decrease adverse impact by implementing interventions before, during, and after data collection (Outtz, 2009) can be fruitful in terms of decreasing mean test score differences across groups and, hence, improving the accuracy of intercept-based bias assessment.

As noted by Kehoe (2002), "a critical part of the dilemma is that GMA-based tests are generally regarded as unbiased" (p. 104). If test bias does not exist, then adverse impact against ethnic minorities is a defensible position that has formidable social consequences, and the field will continue to try to solve what seems to be an impossible dilemma between validity and adverse impact (Aguinis, 2004b; Ployhart & Holtz, 2008). A finding that tests are biased would create potential problems for all of the parties involved in preemployment testing research and practice because researchers would have to rethink how they design and implement preemployment tests. Given the science–practice disconnect in I/O psychology and human resource management (Cascio & Aguinis,

2008a), reviving the issue of test bias may provide a great opportunity for I/O psychology and management researchers to respond to society's need and interest in fairness in preemployment testing. It may provide an opportunity to change how I/O psychology and management research influences practice given that, to this point, "bias scholars have failed as influencers of social policy" (Cole, 1981, p. 1075). In other words, "considerably more thought and effort than have previously been the case should be devoted to the ingredients of test validation research; otherwise, the stream of excuses usually associated with unrewarding outcomes will continue" (Gael, Grant, & Ritchie, 1975, p. 411).

We have focused on test bias regarding groups based on ethnicity classifications primarily in the context of GMA testing because there is a voluminous literature related to these specific groups and type of preemployment testing. However, our results are also applicable when comparisons are based on gender instead of ethnicity and in the context of other types of preemployment tests (e.g., personality; Cortina et al., 1992). For example, Saad and Sackett (2002) examined possible gender-based differential prediction using personality tests and found that slope differences existed in only 5% of cases (i.e., a result that can be explained by chance alone), but some intercept-based differences were found favoring women. As a second illustration addressing a different type of preemployment test, Te Nijenhuis and Van der Flier (2004) explored possible differential prediction for majority and minority ethnic groups regarding safety suitability. Safety suitability is a multidimensional construct including a combination of cognitive functioning, attention functions, perceptual-motor abilities, and personality. Results suggested that there was no slope-based test bias.

Limitations and Suggestions for Future Research on Test-Bias Assessment

Our results are based on simulated data. Hence, although we made an effort to include a wide range of values for each of the parameters we included in the simulation, our results are bound by and hold only for the parameters and values we studied. Our results therefore point to what could be a plausible alternative explanation for the established conclusions regarding test bias, at least for the conclusions derived from studies with parameter values within the ranges we included in our simulation. Nevertheless, we think the evidence we provide is sufficiently compelling to lead to a revival of test bias research in preemployment testing.

To investigate what is actually happening, as opposed to what could be happening as we did in our study including simulated data, future researchers could investigate the question of possible differential prediction meta-analytically. This approach would be particularly fruitful regarding slope-based bias because meta-analysis, by virtue of accumulating data from several studies, would yield greater statistical power than any single study. As is described in detail in Appendix C, a meta-analytic investigation of test bias would be possible, but there are four challenges. First, regression coefficients from primary-level studies must be based on the same operationalizations of each of the predictors and criterion in the model. Second, regression coefficients in the primary-level studies must be computed using the exact same variables in the model because test bias estimates would change if, for example, a third predictor (i.e., another preemployment test) was added in the model (Sackett, Laczo, & Lippe, 2003). Third, to investigate the differential prediction question in a valid manner, researchers using meta-analysis would have to correct sample-based regression coefficients for methodological and statistical artifacts, but we are not aware of a procedure that would allow for such a correction in a meaningful manner. Raju, Fralicx, and Steinhaus (1986) proposed a meta-analytic approach to cumulate regression coefficients, but this approach is only useful for first-order effects and, thus, is not applicable for a meta-analysis of slope-based bias (i.e., meta-analyzing regression coefficients associated with the product term in our Equation 2). Finally, if instead of unstandardized regression coefficients, the meta-analysis cumulates standardized effects (e.g., f^2; Aguinis & Pierce, 2006), one would need information that is typically only reported in about 20% of published articles (Aguinis et al., 2005).

Taken together, these four challenges may explain why, although researchers have conducted meta-analyses of the differential validity literature (e.g., Hunter, Schmidt, & Hunter, 1979), we are not aware of any meta-analysis of the differential prediction literature published in a peer-reviewed journal. Note that a finding of lack of differential validity (i.e., different validity coefficients across groups) does not provide evidence regarding lack of differential prediction (e.g., different slopes across groups). As highlighted by Hartigan and Widgor (1989), "the available reports comparing validities do not provide direct evidence regarding the possibility of differential prediction" (p. 178). This is the case because "equal correlations do not necessarily imply equal standard errors of estimate, nor do they necessarily imply equal slopes or intercepts" (Linn, 1978, p. 511). As noted by Bobko and Bartlett (1978), "a focus on subgroup validity differences distracts attention from the more global problems of test fairness and differential prediction" (p. 13). In short, "differences in prediction systems have a more direct bearing on issues of bias in selection than do differences in correlations" (Linn, 1978, p. 511).

We acknowledge the existence of large sample research addressing test bias within the context of college admission testing. These studies, which are mainly available in the form of technical reports published by the College Board, share two important characteristics. First, they address differential prediction in the context of college admission and not preemployment testing. So, in each of these studies, the criterion of interest is usually first-year grade point average (FYGPA), or some other type of grade point average (GPA; e.g., cumulative GPA). The meta-analytically derived observed correlation between GPA and job performance is only .16 and no greater than in the .30s when corrected for methodological and statistical artifacts (Roth, BeVier, Switzer, & Schippmann, 1996). A corrected correlation of .35 means that GPA and job performance share 12% of variance at the construct level and, hence, GPA and job performance do not seem to be interchangeable criteria (i.e., indicators of the same underlying latent construct; Binning & Barrett, 1989).

A second characteristic shared by the College Board technical reports is that they conclude that college admission tests tend to overpredict observed GPA for African American and Hispanic students. However, reports do not generally include estimates of intercept- or slope-based differences across subgroups or any type of test of statistical significance regarding hypotheses about intercept- or slope-based test bias. Instead, these reports focus on over- or underprediction of GPA without specifying whether this is

due to differences in intercepts, slopes, or both. For example, Bridgeman, McCamley-Jenkins, and Ervin (2000) conducted a study to investigate the impact of revisions in the content (and recentering) of the SAT. Their study included data for the old and revised versions of the SAT for the 1994 and 1995 entering classes of 23 colleges. Each of these colleges also provided FYGPA for all students in the 1994 and 1995 entering classes. Differential prediction was not assessed by examining intercept- or slope-based differences across subgroups. Instead,

> over/underprediction was analyzed by making predictions based on all students in a college and, then, . . . computing the difference between the predicted and actual FYGPA (predicted GPA minus actual GPA). The result is in grade-point units, with positive values indicating overprediction and negative values indicating underprediction. (Bridgeman et al., 2000, p. 3)

As a second illustration of the type of differential prediction analysis reported in College Board reports, Mattern, Patterson, Shaw, Kobrin, and Barbuti (2008) assessed the impact of changes made to the SAT (e.g., addition of a writing section). Mattern et al. analyzed SAT and FYGPA data from 151,316 students from 110 colleges and universities. Similar to Bridgeman et al., Mattern et al. did not report results regarding intercept- or slope-based test bias. Instead, they standardized FYPGAs within each institution and calculated regression equations within each college to assess the degree of over- or underprediction for each subgroup by using a common regression line. Next, identical to Bridgeman et al., Mattern et al. computed the average residual (i.e., predicted FYGPA minus observed FYGPA) for each subgroup within each college and then computed the average degree of over- or underprediction across all colleges. Also similar to Bridgeman et al., Mattern et al. concluded that African American and Hispanic students' FYGPA tends to be overpredicted. Examples of additional large sample size college admission testing reports following precisely the same procedure and reaching similar conclusions include Camara (2008) and Ramist, Lewis, and McCamley Jenkins (1994), among others. In short, these technical reports do not include information pertaining to slope-based differences across ethnicity-based subgroups, the size of such slope-based differences, or the statistical significance test for the null hypothesis of no slope-based test bias.

Although the College Board reports use GPA instead of job performance as the criterion and do not generally report information regarding intercept- or slope-based test bias, they do lead to the consistent conclusion that college admission tests overpredict grades for African American and Hispanic students. However, as noted by Young (2001), "it is accurate to say that the causes of this phenomenon are not yet completely known or understood" (p. 18). Thus, as implied by Young, there is a need for further test-bias research to understand whether, and when, test bias occurs, as well as the reasons why it occurs if it does.

Regarding additional specific suggestions for future research, we doubt that using alternative existing test-bias models will be fruitful. The Cleary (1968) model has prevailed in spite of the many competitors proposed (e.g., Schmidt & Hunter, 1974). The Cleary model has been accepted by researchers, professional organizations, and the legal system. Moreover,

> there seem to be no commonly accepted alternatives to the statistical approaches for instigating such [test bias] investigations. Yet what is even more distressing than the lack of explicit alternatives is an apparent inability of psychologists and psychometricians to articulate the relevant issues as they affect test takers from various racial and ethnic groups within the United States. (Helms, 1992, p. 1090)

Accordingly, we recommend a two-pronged approach. First, as suggested by Van Iddekinge and Ployhart (2008), future test bias assessments should "use power analysis to determine sample size required to draw valid inferences regarding differential prediction, and report the actual level of power for all relevant analyses" (p. 893). Van Iddekinge and Ployhart (Table 1, p. 914) summarized several suggestions and resources available to conduct such a power analysis. However, sample size is only one of several factors that have a detrimental effect on statistical power, so power calculations should include several additional factors (e.g., sample sizes across groups, measurement error, and so forth; Aguinis et al., 2001).

In addition, following Helms's (1992) recommendation, we suggest a new approach to human resource selection that examines how members of various groups are affected by testing and also approach testing from a different perspective. Specifically, Cascio and Aguinis (2008b), Ployhart (2006), and Ployhart, Schneider, and Schmitt (2006) have argued for a change in direction in human resource selection research including an expanded view of the staffing process that considers in situ performance and the role of time and context. In situ performance is the "specification of the broad range of effects—situational, contextual, strategic, and environmental—that may affect individual, team, or organizational performance" (Cascio & Aguinis, 2008b, p. 146). Combining the concept of in situ performance with a consideration of time and context and the recommendation offered by Helms is likely to lead to a better understanding of why and conditions under which "tests often function differently in one ethnic group population than the other" (Katzell & Dyer, 1977, p. 143). In other words, combining these areas of research may provide interesting insights regarding conditions under which individuals may perceive testing differently, how these differences may be related to their identity and cultural backgrounds, and how these perceptions may have an impact on test scores (cf. Walton & Spencer, 2009) as well as the relationship between test scores and performance.

Closing Comments

We are aware that we have set a tall-order goal of reviving research on test bias in preemployment testing in the face of established conclusions in the fields of I/O psychology, management, and others concerned with high-stakes testing. Our results indicate that the accepted procedure to assess test bias is itself biased: Slope-based bias is likely to go undetected and intercept-based bias favoring minority group members is likely to be found when in fact it does not exist (if slope-based bias exists and intercept-based bias does not exist in the population). Preemployment testing is often described as the cradle of the I/O psychology field (e.g., Landy & Conte, 2007). Accordingly, our study has important implications for the field as well as ethical, legal, and organizational implications (Oswald et al., 2000). The present study creates an important opportunity for I/O psychology and management researchers to revive the topic of test bias and make contributions with measurable and important implications for organizations and society.

References

Abel, D. (2007, September 12). Lawsuit challenges fairness of police test. *The Boston Globe.* Retrieved from http://www.boston.com/news/local/articles/2007/09/12/lawsuit_challenges_fairness_of_police_test/

Aguinis, H. (1995). Statistical power problems with moderated multiple regression in management research. *Journal of Management, 21,* 1141–1158.

Aguinis, H. (2004a). *Regression analysis for categorical moderators.* New York, NY: Guilford Press.

Aguinis, H. (Ed.). (2004b). *Test score banding in human resource selection: Legal, technical, and societal issues.* Westport, CT: Praeger.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90,* 94–107.

Aguinis, H., Boik, R. J., & Pierce, C. A. (2001). A generalized solution for approximating the power to detect effects of categorical moderator variables using multiple regression. *Organizational Research Methods, 4,* 291–323.

Aguinis, H., & Pierce, C. A. (2006). Computation of effect size for moderating effects of categorical variables in multiple regression. *Applied Psychological Measurement, 30,* 440–442.

Aguinis, H., & Smith, M. A. (2007). Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. *Personnel Psychology, 60,* 165–199.

Aguinis, H., & Smith, M. A. (2009). Balancing adverse impact, selection errors, and employee performance in the presence of test bias. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 403–423). New York, NY: Routledge.

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.

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. (2009). Customer-centric science: Reporting research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods.* Advance online publication. doi:10.1177/1094428109333339

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Barr, D. R., & Sherrill, E. T. (1999). Mean and variance of truncated normal distributions. *The American Statistician, 53,* 357–361.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31,* 233–241.

Bartlett, C. J., & O'Leary, B. S. (1969). A differential prediction model to moderate the effect of heterogeneous groups in personnel selection and classification. *Personnel Psychology, 22,* 1–18.

Berk, R. A. (Ed.). (1982). Introduction. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 1–8). Baltimore, MD: Johns Hopkins University Press.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478–494.

Birnbaum, Z. W., Paulson, E., & Andrews, F. C. (1950). On the effect of selection performed on some coordinates of a multi-dimensional population. *Psychometrika, 15,* 191–204.

Bobko, P., & Bartlett, C. J. (1978). Subgroup validities: Differential definitions and differential prediction. *Journal of Applied Psychology, 63,* 12–14.

Borsboom, D., Romeijn, J., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods, 13,* 75–98.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning test* (College Board Research Report No. 2000-1, ETS RR No. 00-1). New York, NY: College Board.

Brown, P. B. (2007, December 29). How smart is your manager? *The New York Times,* p. C5.

Brown, R. P., & Day, E. A. (2006). The difference isn't Black and White: Stereotype threat and the race gap on Raven's Advanced Progressive Matrices. *Journal of Applied Psychology, 91,* 979–985.

Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin, 93,* 549–562.

Camara, W. (2008). *Score trends, SAT validity and subgroup differences.* Retrieved from http://professionals.collegeboard.com/profdownload/pdf/sat_validity_and_Harvard_update_6-08.pdf

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

Canivez, G. L., & Konold, T. R. (2001). Assessing differential prediction bias in the Developing Cognitive Abilities Test across gender, race/ethnicity, and socioeconomic groups. *Educational and Psychological Measurement, 61,* 159–171.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle Rock, NJ: Pearson Prentice Hall.

Cascio, W. F., & Aguinis, H. (2008a). Research in industrial and organizational psychology from 1963 to 2007: Changes, choices, and trends. *Journal of Applied Psychology, 93,* 1062–1081.

Cascio, W. F., & Aguinis, H. (2008b). Staffing twenty-first-century organizations. *Academy of Management Annals, 2,* 133–165.

Civil Rights Act of 1964, Pub. L. No. 88-352, 78 Stat. 241.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cohen, A. C. (1950). Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. *The Annals of Mathematical Statistics, 21,* 557–569.

Cohen, A. C. (1951). Estimation of parameters in truncated Pearson frequency distributions. *The Annals of Mathematical Statistics, 22,* 256–265.

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

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36,* 1067–1077.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80,* 565–579.

Cormier v. P. P. G. Indus., Inc., 519 F. Supp. 211 (W.D. La. 1981).

Cortina, J. M. (2003). Apples and oranges (and pears, oh my!): The search for moderators in meta-analysis. *Organizational Research Methods, 6,* 415–439.

Cortina, J. M., Doherty, M. L., Schmitt, N., Kaufman, G., & Smith, R. G. (1992). The "Big Five" personality factors in the IPI and MMPI: Predictors of police performance. *Personnel Psychology, 45,* 119–140.

Cortina, J. M., & Landis, R. S. (2009). When small effect sizes tell a big story, and when large effect sizes don't. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in the organizational and social sciences* (pp. 287–308). New York, NY: Routledge.

Culpepper, S. A., & Davenport, E. C. (2009). Assessing differential prediction of college grades by race/ethnicity with a multilevel model. *Journal of Educational Measurement, 46,* 220–242.

Darlington, R. B. (1971). Another look at "cultural fairness." *Journal of Educational Measurement, 8,* 71–82.

Dunbar, S. B., & Novick, M. R. (1988). On predicting success in training

for men and women: Examples from Marine Corps clerical specialties. *Journal of Applied Psychology, 75,* 545–550.

Dunlap, W. P., & Kemery, E. R. (1988). Effects of predictor intercorrelations and reliabilities on moderated multiple regression. *Organizational Behavior and Human Decision Processes, 41,* 248–258.

Evans, M. G. (1985). A Monte Carlo study of the effects of correlated method variance in moderated multiple regression. *Organizational Behavior and Human Decision Processes, 36,* 305–323.

Gael, S., Grant, D. L., & Ritchie, R. J. (1975). Employment test validation for minority and nonminority telephone operators. *Journal of Applied Psychology, 60,* 411–419.

Gould, M. (1999). Race and theory: Culture, poverty, and adaptation to discrimination in Wilson and Ogbu. *Sociological Theory, 17,* 171–200.

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

Grubb, H. J., & Ollendick, T. H. (1986). Cultural-distance perspective: An exploratory analysis of its effect on learning and intelligence. *International Journal of Intercultural Relations, 10,* 399–414.

Hamer v. City of Atlanta, 872 F.2d 1521 (11th Cir. 1989).

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47,* 1083–1101.

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

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

Houston, W. M., & Novick, M. R. (1987). Race-based differential prediction in Air Force technical training programs. *Journal of Educational Measurement, 24,* 309–320.

Hsu, L. M. (1993). Using Cohen's tables to determine the maximum power attainable in two-sample tests when one sample is limited in size. *Journal of Applied Psychology, 78,* 303–305.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71,* 327–333.

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

Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law, 6,* 151–158.

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.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications for direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91,* 594–612.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 41–99). New York, NY: Plenum Press.

Katzell, R. A., & Dyer, F. J. (1977). Differential validity revived. *Journal of Applied Psychology, 62,* 137–145.

Kehoe, J. F. (2002). General mental ability and selection in private sector organizations: A commentary. *Human Performance, 15,* 97–106.

Kuncel, N. R., & Sackett, P. R. (2007). Selection citation mars conclusions about test validity and predictive bias. *American Psychologist, 62,* 145–146.

Landy, F. J., & Conte, J. M. (2007). *Work in the 21st century: An introduction to industrial and organizational psychology* (2nd ed.). Malden, MA: Blackwell.

Lautenschlager, G. J., & Mendoza, J. L. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement, 10,* 133–139.

Lent, R. H., Auerbach, H. A., & Levin, L. S. (1971). Research design and validity assessment. *Personnel Psychology, 24,* 247–274.

Linn, R. L. (1978). Single group validity, differential validity and differential prediction. *Journal of Applied Psychology, 63,* 507–512.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21,* 33–47.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8,* 1–4.

Marzulli, J. (2008, November 12). City, Feds, trying to settle firefighter-test bias suit. *New York Daily News,* p. 4.

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential validity and prediction of the SAT* (College Board Research Report No. 2008-4). New York, NY: College Board.

Maxwell, S. E., & Arvey, R. D. (1993). The search for predictors with high validity and low adverse impact: Compatible or incompatible goals? *Journal of Applied Psychology, 78,* 433–437.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114,* 376–390.

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

Mehrens, W. A., & Popham, W. J. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education, 5,* 265–283.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2,* 248–260.

Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33,* 403–424.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72,* 461–473.

Monahan, C. I., & Muchinsky, P. M. (1983). Three decades of personnel selection research: A state-of-the-art analysis and evaluation. *Journal of Occupational Psychology, 56,* 215–225.

Mudholkar, G. S., Chaubey, Y. P., & Lin, C. (1976). Some approximations for the noncentral-F distribution. *Technometrics, 18,* 351–358.

Mumford, T. V., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2008). The team role test: Development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology, 93,* 250–267.

Ogbu, J. U. (1993). Differences in cultural frame of reference. *International Journal of Behavioral Development, 16,* 483–506.

Olsen, R. J. (1980). Approximating a truncated normal regression with the method of moments. *Econometrica, 48,* 1099–1105.

Oswald, F. L., Saad, S., & Sackett, P. R. (2000). The homogeneity assumption in differential prediction analysis: Does it really matter? *Journal of Applied Psychology, 85,* 536–541.

Outtz, J. L. (2002). The role of cognitive ability tests in employment selection. *Human Performance, 15,* 161, 172.

Outtz, J. L. (Ed.). (2009). *Adverse impact: Implications for organizational staffing and high stakes selection.* New York, NY: Routledge.

Pearson, E. S., & Hartley, H. O. (1951). Charts of the power function for analysis of variance tests, derived from the noncentral F-distribution. *Biometrika, 38,* 112–130.

Pearson, K. (1903). Mathematical contributions to the theory of evolution—XI. On the influence of natural selection on the variability and

correlation of organs. *Philosophical Transactions of the Royal Society of London, Series A, 200,* 1–66.

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

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61,* 153–172.

Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: Contemporary practice and theory* (3rd ed.). Mahwah, NJ: Erlbaum.

Raju, N. S., Fralicx, R., & Steinhaus, S. (1986). Covariance and regression slope models for studying validity generalization. *Applied Psychological Measurement, 10,* 195–211.

Raju, N. S., Pappas, S., & Williams, C. P. (1989). An empirical Monte Carlo test of the accuracy of the correlation, covariance, and regression slope models for assessing validity generalization. *Journal of Applied Psychology, 74,* 901–911.

Ramist, L., Lewis, C., & McCamley Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (College Board Report No. 93-1, ETS RR No. 94-27). New York, NY: College Board.

Reeb, M. (1976). Differential test validity for ethnic groups in the Israel Army and the effects of educational level. *Journal of Applied Psychology, 61,* 253–261.

Reynolds, C. R. (1995). Test bias in the assessment of intelligence and personality. In D. H. Saklofske & M. Zeidner (Eds.), *International handbook of personality and intelligence* (pp. 545–573). New York, NY: Plenum Press.

Reynolds, C. R., & Brown, R. T. (Eds.). (1984). Bias in testing: Introduction to the issues. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 1–39). New York, NY: Plenum Press.

Ricci v. DeStefano, 129 S. Ct. 2658 (2009).

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

Roth, P. L., BeVier, C. A., Switzer, F. S., III, & Schippmann, J. S. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology, 81,* 548–556.

Rotundo, M., & Sackett, P. R. (1999). Effect of rater race on conclusions regarding differential prediction in cognitive ability tests. *Journal of Applied Psychology, 84,* 815–822.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law, 11,* 235–294.

Rushton, J. P., Skuy, M., & Bons, T. A. (2004). Construct validity of Raven's Advanced Progressive Matrices for African and non-African engineering students in South Africa. *International Journal of Selection and Assessment, 12,* 220–229.

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

Saad, S., & Sackett, P. R. (2002). Investigating differential prediction by gender in employment-oriented personality measures. *Journal of Applied Psychology, 87,* 667–674.

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

Sackett, P. R., De Corte, W., & Lievens, F. (2009). Decision aids for addressing the validity–adverse impact trade-off. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 453–472). New York, NY: Routledge.

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American–White differences on cognitive tests. *American Psychologist, 59,* 7–13.

Sackett, P. R., Kuncel, N. R., Arneson, J. J., Cooper, S. R., & Waters, S. D. (2009). Does socioeconomic status explain the relationship between admissions tests and post-secondary academic performance? *Psychological Bulletin, 135,* 1–22.

Sackett, P. R., Laczo, R. M., & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: The omitted variables problem. *Journal of Applied Psychology, 88,* 1046–1056.

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

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

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

Salgado, J. F. (1998). Sample size in validity studies of personnel selection. *Journal of Occupational and Organizational Psychology, 71,* 161–164.

Schmidt, F. L., & Hunter, J. E. (1974). Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. *American Psychologist, 29,* 1–8.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36,* 1128–1137.

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

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods, 8,* 206–224.

Schmidt, F. L., Oh, I., & 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., Pearlman, K., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Journal of Applied Psychology, 33,* 705–724.

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

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 35,* 4–28.

Stone-Romero, E. F., & Anderson, L. E. (1994). Relative power of moderated multiple regression and the comparison of subgroup correlation coefficients for detecting moderating effects. *Journal of Applied Psychology, 79,* 354–359.

Te Nijenhuis, J., & Van der Flier, H. (1999). Bias research in the Netherlands: Review and implications. *European Journal of Psychological Assessment, 15,* 165–175.

Te Nijenhuis, J., & Van der Flier, H. (2000). Differential prediction of immigrant versus majority group training performance using cognitive ability and personality measures. *International Journal of Selection and Assessment, 8,* 54–60.

Te Nijenhuis, J., & Van der Flier, H. (2004). The use of safety suitability tests for the assessment of immigrant and majority group job applicants. *International Journal of Selection and Assessment, 12,* 230–242.

Terris, W. (1997). The traditional regression model for measuring test bias is incorrect and biased against minorities. *Journal of Business and Psychology, 12,* 25–37.

Uniform Guidelines on Employee Selection Procedures, 43 Fed. Reg. 38290 (Aug. 25, 1978).

United States v. City of Erie, 411 F. Supp. 2d 524 (W.D. Pa. 2005).

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the

criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology, 61,* 871–925.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557–574.

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science, 20,* 1132–1139.

Wonderlic. (1999). *Wonderlic Personnel Test and Scholastic Level Exam: User's manual.* Libertyville, IL: Author.

Yang, H., Sackett, P. R., & Nho, Y. (2004). Developing a procedure to correct for range restriction that involves both institutional selection and applicants' rejection of job offers. *Organizational Research Methods, 7,* 442–455.

Young, J. W. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis* (College Board Research Report No. 2011-6). New York, NY: College Board.

Appendix A

Theorem From Aguinis et al. (2001) Used to Compute Statistical Power

Statistical power for the *F* test for assessing slope-based test bias is:

$$\text{Power} \approx \Pr\left[\left(\frac{k-1}{N-2k}\right)F^{1-\alpha}_{k-1,N-2k}\sum_{j=1}^{k}\frac{\sigma^2_{y,j}(1-\rho^2_j\alpha_{x,j}\alpha_{y,j})}{\alpha_{y,j}}H_j - \sum_{j=1}^{k-1}\omega_j G_j \leq 0\right],$$

where ω_j is the *j*th eigenvalue of $(\mathbf{C'DC})^{-1}\,\mathbf{C'VC}$;

$$\mathbf{D} = \text{Diag}\left[\frac{\alpha_{x,j}(n_j+1)}{(n_j-1)^2\delta_j\sigma^2_{x,j}};\ j=1,\ldots,k\right];$$

$$\mathbf{V} = \text{Diag}\left[\frac{\sigma^2_{y,j}\alpha_{x,j}(1-\rho^2_j\alpha_{x,j}\alpha_{y,j})(n_j+1)}{\alpha_{y,j}(n_j-1)^2\delta_j\sigma^2_{x,j}};\ j=1,\ldots,k\right];$$

and G_j for $j = 1, \ldots, k-1$ and H_j for $j = 1, \ldots, k$ are independently distributed chi-squared random variables. Specifically, $H_j \sim \chi^2\,(n_j - 2)$ for $j = 1, \ldots, k$ and $G_j \sim \chi^2(1, \lambda_j)$ for $j = 1, \ldots, k-1$, where λ_j is a noncentrality parameter;

$$\lambda_j = \frac{(\mathbf{u}_j'\mathbf{C}'\boldsymbol{\beta}_1)^2}{2\mathbf{u}_j'\mathbf{C}'\mathbf{VC}\mathbf{u}_j}$$

and $\mathbf{u}_j$ is the *j*th eigenvector of $(\mathbf{C'DC})^{-1}\,\mathbf{C'VC}$.

Appendix B

Analytic Proof Describing Why Measurement Error and Differences in Test Scores Across Groups Produce Bias in the Intercept-Based Difference Test

We begin by analytically examining the effect of *G* on *Y* after controlling for *X* (cf. Equations 1–3). Assuming no group intercept differences, we can express the unique contribution of *G* beyond *X*, $\Delta\psi^2_{\text{intercept}}$, using the formula for part correlations. Specifically, the effect size for intercept tests is shown below by squaring the part correlation between *Y* and *G*:

$$\Delta\psi^2_{\text{intercept}} = \frac{(\rho_{YG} - \rho_{XY}\rho_{XG})^2}{(1-\rho^2_{XG})} \tag{B1}$$

where ρ_{XY} is the validity coefficient (i.e., correlation between test scores and the criterion). If *X* and *Y* are measured with error, ρ_{YG}, ρ_{XY}, and ρ_{XG} will be attenuated, which can be accounted for in Equation B1 by substituting $\rho_{yG} = \rho_{YG}\sqrt{\rho_{YY}}$, $\rho_{xy} = \rho_{XY}\sqrt{\rho_{XX}\rho_{YY}}$, and $\rho_{xG} = \rho_{XG}\sqrt{\rho_{XX}}$ for ρ_{XG}. Furthermore, if the null hypothesis is true (i.e., $\Delta\psi^2_{\text{intercept}} = 0$) and *X* is measured without error (i.e., $\rho_{XX} = 1$), $\rho_{yG} = \rho_{YG}\sqrt{\rho_{YY}} = \rho_{XY}\rho_{XG}\sqrt{\rho_{YY}}$.

Range restriction is another important artifact that has been shown to reduce the power and effect sizes of tests of slope differences (Aguinis & Stone-Romero, 1997). We can incorporate the effect of range restriction on tests of intercept differences by finding values for the range restricted correlations (i.e., the ranged restricted values of ρ_{yG}, ρ_{xy}, and ρ_{xG}) in Equation B1. First, note that if *x* is standardized, $\rho_{xG} = \sqrt{p(1-p)}\Delta\mu$ (which is the definition of the point biserial correlation), where *p* is the proportion of the sample containing the first group and $\Delta\mu$ is the true difference in group means. In order to find the range restricted correlation between *x* and *G*, ρ_{xGr}, we must find p_r and $\Delta\mu_r$. In the unrestricted case where *X* is error free, $\mu = p\mu_1 + (1-p)\mu_0 = 0$ (where μ_1 and μ_0 are the group means) and, by definition, $\Delta\mu = \mu_1 - \mu_0$, so that the values of μ_0 and μ_1 that satisfy these two equations are $\mu_0 = -p\Delta\mu$ and $\mu_1 = (1-p)\Delta\mu$. If *X* is measured with error, the observed group means, μ_{0x} and μ_{1x}, are defined by $\mu_{x0} = -p\Delta\mu\sqrt{\rho_{XX}}$ and $\mu_{1x} = (1-p)\Delta\mu\sqrt{\rho_{XX}}$. Similarly,

(*Appendices continue*)

we can identify an expression for group variances on X_i using $\Delta\sigma^2 = \sigma^2_{X1} - \sigma^2_{X0}$ (σ^2_{X1} is the variance for the majority group and σ^2_{X0} the minority group) and

$$\sigma^2_X = n^{-1}\sum_{i=1}^{n} X_i^2 = n^{-1}\left(\sum_{i=1}^{n_1} X_i^2 + \sum_{i=1}^{n_0} X_i^2\right) = p[\sigma^2_{X1} + (\mu_1)^2] + (1 - p)[\sigma^2_{X0} + (\mu_0)^2] = 1.$$

Solving values for σ^2_{X1} and σ^2_{X0} using the two equations yields $\sigma^2_{X0} = 1 - (1 - p)(\mu_0)^2 - p[(\mu_1)^2 + \Delta\sigma^2]$ and $\sigma^2_{X1} = (1 - p)\Delta\sigma^2 + 1 - [p(\mu_1)^2 + (1 - p)(\mu_0)^2]$. Substituting the values for the squared means and setting $\Delta\sigma^2 = 0$ produces the following expressions for the true group variances: $\sigma^2_{X0} = \sigma^2_{X1} = 1 - p(1 - p)(\Delta\mu)^2$. Additionally, if X is measured with error, $\rho^2_{XG} = p(1 - p)(\Delta\mu)^2$ will be attenuated by ρ_{XX}, which suggests that $\sigma^2_{x0} = \sigma^2_{x1} = 1 - \rho_{XX}p(1 - p)(\Delta\mu)^2$.

Now consider a typical selection situation where a sample only includes subjects with observed scores greater than a given cut score, x^*. This type of range restriction, or truncation, in x alters the overall distribution of x, as well as each groups' distribution. Specifically, the first two moments for x restricted on values greater than x^* are defined by the truncated normal distribution (Barr & Sherrill, 1999; A. C. Cohen, 1950, 1951; Olsen, 1980) and are represented as $\mu_{xr} = E\{x|x > x^*\} = \mu + \sigma\lambda(\alpha) = \sigma\lambda(\alpha)$ and $\sigma^2_{xr} = \sigma^2\{x_i|x_i > x^*\} = \sigma^2[1 - \delta(\alpha)]$, where the r subscript denotes range restricted values and α, $\lambda(\alpha)$, and $\delta(\alpha)$ are defined as,

$$\alpha = \frac{x^* - \mu}{\sigma}$$

$$\lambda(\alpha) = \frac{\phi(\alpha)}{1 - \Phi(\alpha)}$$

$$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha],$$

where $\phi(\alpha)$ is the standard normal density defined at the cut score, x^*, and $\Phi(\alpha)$ is the cumulative distribution function. Similarly, we can identify the first two moments for each group. For the first group, $\mu_{x1r} = E\{x_1|x_1 > x^*\} = \rho_{XX}(1 - p)\Delta\mu + \sigma_{x1}\lambda(\alpha_1)$ and $\sigma^2_{x1r} = \sigma^2\{x_1|x_1 > x^*\} = \sigma^2_{x1}[1 - \delta(\alpha_1)]$ where $\alpha_1 = (x^* - \mu_{x1})/\sigma_{x1}$. Likewise, the moments for the other group are $\mu_{x0r} = E\{x_0|x_0 > x^*\} = \rho_{XX}p\Delta\mu + \sigma_{x0}\lambda(\alpha_0)$ and $\sigma^2_{x0r} = \sigma^2\{x_{i0}|x_{i0} > x^*\} = \sigma^2_{x0}[1 - \delta(\alpha_0)]$ where $\alpha_0 = (x^* - \mu_{x0})/\sigma_{x0}$. Let $\Delta\mu_r = \mu_{x1r} - \mu_{x0r}$ and $\Delta\sigma^2_r = \sigma^2_{x1r} - \sigma^2_{x0r}$ denote the observed differences between the groups' truncated means and variances, respectively.

Recall that $p = E\{G_i\}$ denotes the proportion in the first group in the nonrestricted sample. We must also derive an expression for the sample proportion in the range restricted sample. Let n refer to the size of the total sample, n_r refer to the size for the restricted sample, and p_r refer to the restricted proportion. Next, n_r is simply the sum of the number of individuals with $x > x^*$. That is, $n_r = pn(1 - \Phi(\alpha_1)) + (1 - p)n(1 - \Phi(\alpha_0))$ and $p_r = E\{G|x > x^*\} = pn(1 - \Phi(\alpha_1))/n_r$. Additionally, the variance for the range restricted G is $\sigma^2_{Gr} = \sigma^2\{G_i|x_i > x^*\} = p_r(1 - p_r)$. We can use the results above to compute the correlation between x and G in the restricted sample as

$$\rho_{xGr} = \sqrt{p_r(1 - p_r)}\Delta\mu_r.$$

Previous research documents formulas for correcting correlations in the presence of range restriction (Birnbaum, Paulson, & Andrews, 1950; Hunter, Schmidt, & Le, 2006; K. Pearson, 1903; Sackett & Yang, 2000; Schmidt, Oh, & Le, 2006; Yang, Sackett, & Nho, 2004). We use Birnbaum et al.'s (1950) results to compute the range restricted correlations between the predictors and the observed Y, y. Let $\mathbf{\Sigma_{xx}}$ and $\mathbf{\Sigma_{xxr}}$ denote the unrestricted and restricted variance–covariance matrices among x and G, respectively, and $\mathbf{\Sigma_{xy}}$ and $\mathbf{\Sigma_{xyr}}$ represent vectors with the unrestricted and restricted covariances between x and G and y_i. Additionally, σ^2_{yr} represents the restricted variance for the observed y. We can use $\mathbf{\Sigma_{xx}}$, $\mathbf{\Sigma_{xxr}}$, and $\mathbf{\Sigma_{xy}}$ to find $\mathbf{\Sigma_{xyr}}$ and σ^2_{yr}. Specifically, Birnbaum et al. noted the following expressions hold if linearity is met and the conditional variances and covariances are independent of predictor values:

$$\mathbf{\Sigma_{xyr}} = \mathbf{\Sigma_{xxr}}\mathbf{\Sigma_{xx}^{-1}}\mathbf{\Sigma_{xy}}$$

$$\sigma^2_{yr} = \sigma^2_y + \mathbf{\Sigma'_{xyr}}\left(\mathbf{\Sigma_{xxr}^{-1}} + \mathbf{\Sigma_{xxr}^{-1}}\mathbf{\Sigma_{xx}}\mathbf{\Sigma_{xxr}^{-1}}\right)\mathbf{\Sigma_{xyr}}.$$

The vector of measurement-error attenuated, range restricted correlations between the predictors and y_i, r_{yr}, is computed as

$$r_{yr} = [r_{yG}\ r_{xy}] = \sigma_{yr}^{-1}\mathbf{\Sigma_{xyr}}\left[\text{diag}\left(\mathbf{\Sigma_{xxr}}\right)\right]^{-1},$$

where r_{yG} and r_{xy} are the error-attenuated, range restricted correlations between y and G and y and x. Consequently, we can estimate the degree of bias in G by updating Equation B1 as follows:

$$\Delta\psi^2_{\text{intercept}} = \frac{(r_{yG} - r_{xy}\sqrt{p_r(1 - p_r)}\Delta\mu_r)^2}{1 - p_r(1 - p_r)(\Delta\mu_r)^2}$$

Appendix C

Analytic Description of Challenges in Investigating Differential Prediction Meta-Analytically

There are several challenges in conducting a meta-analysis of the differential prediction literature. These challenges may explain why, although researchers have conducted meta-analyses of the differential validity literature (e.g., Hunter et al., 1979), we are not aware of any meta-analysis of the differential prediction literature published in a peer-reviewed journal.

Before discussing these challenges, consider that a meta-analysis of the differential prediction literature would rely on

primary-level studies (denoted by k where $k = 1$ to K) reporting the following regression equation:

$$y_{ik} = b_{0k} + b_{1k}x_{ik} + b_{2k}G_{ik} + b_{3k}x_{ik}G_{ik} + e_{ik}, \quad \text{(C1)}$$

where i ($i = 1$ to N) indices individuals. For the current argument, assume that every primary-level differential prediction study provides estimates of intercept (b_{2k}) and slope (b_{3k}) differences and that the studies estimate the same functional relationship between the outcome and test scores and group membership as specified in Equation C1.

If estimates of b_{2k} and b_{3k} are available for each of the K studies, then regression equations can be developed to model primary-level study differences in b_{2k} and b_{3k}. Moreover, it would be necessary to weight the K b_{2k} and b_{3k} coefficients by the study sample size using weighted least squares. Raju et al. (1986) and Raju, Pappas, and Williams (1989) described a procedure for meta-analyzing first-order regression coefficients, but their procedure does not include a consideration of meta-analyzing regression coefficients for the product term b_{3k} in Equation C1.

Equations C2 and C3 below show the calculations that would be used for aggregating both intercept- and slope-based differences across studies:

$$b_{2k} = \gamma_{02} + \gamma_{12}b_{3k} + \sum_{v=2}^{V}\gamma_{v2}(d_{kv} - c_v) + u_{k2} \quad \text{(C2)}$$

$$b_{3k} = \gamma_{03} + \sum_{v=2}^{V}\gamma_{v3}(d_{kv} - c_v) + u_{k3}, \quad \text{(C3)}$$

where the V d_{kv} represent methodological and statistical artifacts that affect observed differences in b_{2k} and b_{3k} across the K studies. For example, d_{kv} should include primary-level study estimates of range restriction, reliability of x_i and y_i, sample standard deviations of x_i and y_i, proportion of minorities in the sample (or p), and so forth. Additionally, in Equation C3 γ_{03} is an intercept, the V γ_{v3} are coefficients that measure the effect of the V design characteristics on the observed slope-based effect (i.e., b_{3k}), and u_{k3} is a random error term. Similarly, γ_{02} and V γ_{v2} are the regression coefficients for the model for assessing intercept-based bias (i.e., b_{2k}). Lastly, as shown in our simulation results, it is important to examine the effect of slope differences on intercept differences, so γ_{12} represents the impact of b_{3k} on b_{2k}. Also, both in Equations C2 and C3, the V c_v are constants used to center d_{kv}.

A critical issue to consider is that a meta-analysis investigating test bias would attempt to estimate the average slope (b_{3k}) and intercept (b_{2k}) differences across the K studies after controlling for methodological and statistical artifacts such as measurement error and range restriction, which are known to bias intercept- and slope-based differences. If the meta-analytically derived slope-based bias is not corrected for the detrimental effects of methodological and statistical artifacts, the meta-analysis would not be informative regarding the size of the true population effect. The issue of incorporating methodological and statistical artifact corrections in the meta-analysis is particularly important in the case of the estimation of slope-based bias because the presence of artifacts decreases sample-based effects in relationship to their population counterparts (Aguinis et al., 2005). Additionally, in contrast to meta-analytic research using correlations, there are no current approaches for disattenuating product-term slope coefficients (i.e., b_{3k}) for measurement error.

The parameters of interest in Equations C2 and C3 are γ_{02} and γ_{03}, which are the average intercept-based and slope-based estimates of test bias across the K studies. The benefit of using regression is the ability to control for the V study design characteristics to adjust γ_{02} and γ_{03}. That is, γ_{02} and γ_{03} are the intercepts that represent the average b_{2k} and b_{3k} after accounting for differences in design characteristics. Consequently, hypothesis tests can be conducted using γ_{02} and γ_{03} (namely, H_0: $\gamma_{02} = 0$ and H_0: $\gamma_{03} = 0$) to test whether groups differ in intercepts and slopes across the K studies. However, the estimated values of γ_{02} and γ_{03} are dependent upon values of V c_v. Stated differently, if V $c_v = 0$ the intercepts represent the average b_{2k} and b_{3k} when all of the predictors (i.e., the design characteristics) are zero. However, it is not meaningful to conduct hypothesis testing for γ_{02} and γ_{03} for when the V design artifacts are zero because, for example, no study will use x_i with reliabilities equal to zero or range restriction values equal to zero.

Recall that b_{2k} and b_{3k} are biased in all K primary-level studies due to the presence of measurement error, selection effects, and other statistical and methodological artifacts. Researchers are more interested in the average b_{2k} and b_{3k} in the absence of measurement error and selection effects. Consequently, the V predictor must be centered by theoretically important values. For example, let ρ_{xxk} and RR_k represent the reliability of x_i and degree of range restriction in study k. In order to adjust for the presence of measurement error and selection, ρ_{xxk} and RR_k should be centered by 1, which represents a situation where there are no measurement error or selection effects. Similarly, every important study design characteristic could be centered by carefully chosen values that represent the absence of an artifact. In this case, it would be possible to estimate the average intercept and slope differences when design artifacts are not attenuating or inflating the estimates, and it would be straight forward to test whether γ_{02} and γ_{03} differ from zero. Moreover, average values for γ_{02} and γ_{03} are the meta-analytically derived population estimates for the degree of intercept-based and slope-based test bias. However, note that the estimated γ_{02} and γ_{03} values would be based upon values of the V predictors that do not exist in the sample (e.g., ρ_{xxk} and $RR_k = 1$), so estimates would be produced based on predictor values that do not exist in the samples.

(*Appendices continue*)

Given the description above, we now discuss four major challenges that researchers face in attempting to review the differential prediction literature meta-analytically. First, regression coefficients are referenced to the specific metrics of the scales used in each study. For example, assume that a researcher is interested in conducting a meta-analysis of the gender-based differential prediction literature for the relationship between cognitive abilities tests and job performance. If Study 1 used a test with a possible range of 1–100 and a supervisory rating of performance as a criterion with a range of 1–7, $b_{31} = 10$ means that a 1-point increase in test scores is associated with a 10-point difference in the slope of Y on X across groups. If Study 2 used a test with a possible range of 1–10 and monthly sales as a criterion with a range of \$0–\$300,000, $b_{32} = 10$ also means that a 1-point increase in test scores is associated with a 10-point differences in the slope of Y on X across groups. However, the meaning of b_{31} is different from the meaning of b_{32}. For Study 1, a 10-point difference may be practically meaningful given a maximum of 100 point on the Y scale. On the other hand, for Study 2, a 10-point difference may not be practically meaningful given a maximum of \$300,000 on the Y scale. In short, the first challenge is that cumulating regression coefficients would result in the well-known problem of "mixing apples and oranges" in meta-analysis (Cortina, 2003). As noted by Raju et al. (1989),

> without the common metrics for the criterion and predictor variables, it is almost impossible to interpret credibility intervals of the type used with the correlation model. The use of the new models for studying VG, therefore, requires that scales for the predictor and criterion variables be comparable across studies. Such a requirement will naturally limit the applicability of the covariance and regression slope models. (p. 903)

Thus, a meta-analysis of the differential prediction literature would be possible if regression coefficients from primary-level studies were based on the exact same operationalizations of each of the predictors, and criterion, in the model.

A second challenge is that even if primary-level studies use the same measures and all available subgroup-based statistics are available for a meta-analysis, regression coefficients change across studies when the regression models do not include exactly the same predictors due to the "omitted variables" phenomenon (Sackett et al., 2003). That is, each primary-level study could include additional and also different predictors in the model, so the estimated b_{2k} and b_{3k} represent different population estimates across studies. For example, Sackett et al. (2003) used Army Project A data to examine possible race-based differential prediction of personality measures. When the regression model included three terms only (i.e., first-order effect of personality, first-order effect of race, and the personality test by race product term), intercept-based bias was found for 45 analyses and slope-based bias was found for seven analyses. In contrast, when the regression equation included the Armed Services Vocational Aptitude Battery general factor as an additional predictor, conclusions regarding personality-based test biased changed for 32 of intercept-based cases and for three slope-based cases. In short, the second challenge is that the size and statistical significance of the regression coefficients in Equation 2 in our article change based on the inclusion of additional predictors in the model. Thus, a meta-analysis of the differential prediction literature would be possible if it cumulates regression coefficients from equations including exactly the same predictors in each primary-level study.

The third challenge is that, as noted above, the estimated values of γ_{02} and γ_{03} are dependent upon values of $V\,c_v$. In other words, if $V\,c_v = 0$ the intercepts represent the average b_{2k} and b_{3k} when all of the predictors are zero. Clearly, it is not meaningful to conduct hypothesis testing for γ_{02} and γ_{03} for situations when V is zero. For example, no study will use x_i with reliabilities equal to zero or range restriction values equal to zero (which would mean that all applicants are selected). Because b_{2k} and b_{3k} are biased in all primary-level studies due to the presence of measurement error and selection effects, V must be centered by theoretically important values. Then, it would be possible to estimate the average intercept- and slope-based bias when methodological and statistical artifacts are not attenuating or inflating the estimates and it would be straight forward to test whether γ_{02} and γ_{03} differ from zero. Thus, a meta-analysis of the differential prediction literature would be possible, but results would suggest the degree of intercept-based and slope-based bias for values for design artifacts (e.g., measurement error, range restriction) that are not realistic (e.g., every single applicant is selected).

Finally, a fourth challenge is that a possible way to overcome the first challenge is to compute subgroup-based regression coefficients using subgroup-based correlation coefficients and standard deviations or to conduct the meta-analysis using standardized over- or underprediction of various criteria (e.g., f^2; Aguinis & Pierce, 2006). However, standard deviations for subgroup-based predictor and criterion scores are typically reported in only about 20% of published articles (Aguinis et al., 2005, p. 96). Thus, a meta-analysis of the differential prediction literature would only be possible if primary-level studies report the necessary subgroup-based statistics (e.g., means, standard deviations, and sample size).

Received January 11, 2009
Revision received November 20, 2009
Accepted December 14, 2009 ■