Journal of Educational Psychology
2016, Vol. 108, No. 7, 1045–1059

0022-0663/16/$12.00 http://dx.doi.org/10.1037/edu0000104

Differential Prediction Generalization in College Admissions Testing

Herman Aguinis
Indiana University

Steven A. Culpepper
University of Illinois at Urbana–Champaign

Charles A. Pierce
University of Memphis

We introduce the concept of *differential prediction generalization* in the context of college admissions testing. Specifically, we assess the extent to which predicted first-year college grade point average (GPA) based on high-school grade point average (HSGPA) and SAT scores depends on a student's ethnicity and gender and whether this difference varies across samples. We compared 257,336 female and 220,433 male students across 339 samples, 29,734 Black and 304,372 White students across 247 samples, and 35,681 Hispanic and 308,818 White students across 264 samples collected from 176 colleges and universities between the years 2006 and 2008. Overall, results show a lack of differential prediction generalization because variability remains after accounting for methodological and statistical artifacts including sample size, range restriction, proportion of students across ethnicity- and gender-based subgroups, subgroup mean differences on the predictors (i.e., HSGPA, SAT-Critical Reading, SAT-Math, and SAT-Writing), and *SD*s for the predictors. We offer an agenda for future research aimed at understanding several contextual reasons for a lack of differential prediction generalization based on ethnicity and gender. Results from such research will likely lead to a better understanding of the reasons for differential prediction and interventions aimed at reducing or eliminating it when it exists.

Keywords: differential prediction, admissions testing, test fairness, test bias

Supplemental materials: http://dx.doi.org/10.1037/edu0000104.supp

As noted in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014), "the term predictive bias may be used when evidence is found that differences exist in the patterns of associations between test scores and other variables for different groups . . . one approach examines slope and intercept differences between two targeted groups . . . while another examines systematic deviations from a common regression line for any number of groups of interest" (pp. 51–52). Similarly, the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003) state that "slope and/or intercept differences between subgroups indicate predictive bias" (p. 32).

This article was published Online First January 21, 2016.

Herman Aguinis, Department of Management and Entrepreneurship, Kelley School of Business, Indiana University; Steven A. Culpepper, Department of Statistics, University of Illinois at Urbana-Champaign; Charles A. Pierce, Department of Management, Fogelman College of Business & Economics, University of Memphis.

The first and second authors contributed equally to this research. We thank Frank A. Bosco for assistance converting pdf files to raw data format.

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

The aforementioned widely adopted and standard definition of predictive bias, which is also labeled differential prediction, refers to a difference in the prediction of scores across subgroups and does not stipulate which group's scores are under- or overpredicted. In other words, differential prediction also exists when the prediction of criteria is different across groups such that the minority group "benefits" from overprediction. In fact, although not within the context of educational testing, lawsuits regarding reverse discrimination in preemployment testing such as the Ricci v. DeStefano et al. (2009) U.S. Supreme Court case are based on this logic because majority and minority applicants are protected under Title VII of the Civil Rights Act of 1964.

Aguinis, Culpepper, and Pierce (2010) revived the fairly dormant research domain of differential prediction and received substantial media attention, including coverage by USA *Today*, *The Economist*, *HR Magazine*, and many other outlets. Thus, research on this topic is important for educational psychology and other fields concerned with high-stakes testing, such as human resource management and industrial and organizational psychology, as well as society at large. Aguinis, Culpepper, et al. (2010) stated that there is an "important opportunity for . . . researchers to revive the topic of differential prediction and make contributions with measurable and important implications for organizations and society" (p. 675).

Following Aguinis, Culpepper, et al.'s (2010) call, several researchers have echoed the need for additional work regarding differential prediction in educational and preemployment contexts (Berry, Clark, & McClure, 2011; Berry, Sackett, & Sund, 2013;

Fischer, Schult, & Hell, 2013). Our study relies on a data-analytic approach similar to that used in investigations of validity generalization (i.e., the extent to which validity coefficients vary across studies) to introduce a new concept we label *differential prediction generalization*, which refers to the extent to which differential prediction varies across studies. Next, we offer a literature review and description of our study's rationale, goals, and contributions in relation to previous research.

Literature Review and Present Study

The potential existence of differential prediction by gender and ethnicity has been investigated for several decades. For example, Cleary (1966) investigated data from three colleges, Pfeifer and Sedlacek (1971) analyzed data from 13 institutions, and Temp (1971) investigated 13 institutions. More recently, Mattern and Patterson (2013) examined differential prediction of the SAT by relying on a larger database. In the majority of these studies, differential prediction has been found, on average, to be small such that tests overpredict grades for Black and Hispanic students (e.g., Mattern & Patterson, 2013) and underpredict grades for female students (e.g., Ancis & Sedlacek, 1997). The majority of this body of work has focused on understanding the degree of differential prediction in specific institutions or the average degree of differential prediction across institutions.

A related but different line of research has addressed the extent to which validity coefficients (e.g., correlation coefficient between test scores and a criterion such as college grades) generalize (i.e., are similar) across contexts. This line of inquiry was motivated by research conducted in the 1960s (e.g., Ghiselli, 1966; Guion, 1965) suggesting that validity coefficients change from context to context and, therefore, are situation-specific. In a seminal article challenging this situational specificity hypothesis, Schmidt and Hunter (1977) offered an analytic approach called validity generalization or psychometric meta-analysis, which involves first assessing the degree of variability of validity coefficients across studies and then calculating the extent to which such variability may be substantive (supporting situational specificity) or, instead, because of methodological and statistical artifacts (supporting validity generalization; Hunter & Schmidt, 2004). This two-step process is necessary because the observed variability of coefficients across contexts may be because of factors such as sampling error, measurement error, and range restriction (Aguinis & Pierce, 1998; Aguinis, Sturman, & Pierce, 2008).[1] In other words, these methodological and statistical artifacts can give the impression that there is a great deal of variability in correlation (i.e., validity) coefficients across studies, whereas in actuality this variability may be because of differences in sample size, measurement error, and range restriction.

Since the introduction of validity generalization procedures by Schmidt and Hunter (1977), several studies have been conducted examining correlations in the context of educational and preemployment testing. For example, Linn, Harnisch, and Dunbar (1981) conducted a validity generalization study of the LSAT and its relation with first-year grades and reported that the majority of the variance in observed validity coefficients was explained by methodological and statistical artifacts. Similarly, in two separate studies, Boldt (1986a, 1986b) conducted validity generalization analyses to understand whether the validity of the SAT and GRE generalizes across colleges and universities and the overall conclusion was that the correlation between these test scores and subsequent grades seems to generalize.

Considering our current knowledge about differential prediction and the separate but related body of work on validity generalization points to a knowledge gap regarding the extent of *differential prediction generalization*. This knowledge gap is important because, as noted by Linn (1978), "differences in prediction systems have a more direct bearing on issues of bias in selection than do differences in correlations" (p. 511). Specifically, validity generalization refers to whether the correlation between test scores and criteria is similar across contexts. In contrast, we conceptualize differential prediction generalization as the extent to which differential prediction (i.e., differences in regression coefficients across groups) is similar across contexts. Thus, differential prediction generalization is different from validity generalization and highly informative because, as noted by the *Standards for Educational and Psychological Testing*, "correlation coefficients provide inadequate evidence for or against a differential prediction hypothesis if groups or treatments are found to have unequal means and variances on the test and the criterion. It is particularly important in the context of testing for high-stakes purposes that test developers and/or users examine differential prediction and avoid the use of correlation coefficients in situations where groups or treatments result in unequal means or variances on the test and criterion" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, p. 66).

From a theoretical perspective, our interest in differential prediction generalization is motivated by several possible sociohistorical-cultural and social psychological explanations for why the use of test scores in educational and employment settings to predict performance can differ based on a test taker's ethnicity or gender and why differential prediction is unlikely to be similar (i.e., generalize) across contexts (Aguinis, Culpepper, et al., 2010; Berry et al., 2011; Culpepper & Davenport, 2009; Kobrin & Patterson, 2011; Pässler, Beinicke, & Hell, 2014). For example, these potential explanations include (a) stereotype threat (Brown & Day, 2006; Sackett, Hardison, & Cullen, 2004; Steele & Aronson, 1995; Walton, Murphy, & Ryan, 2015; Walton & Spencer, 2009); (b) lack of a common cultural frame of reference and identity across groups (Gould, 1999; Ogbu, 1993); (c) lack of a common framework for understanding and interpreting tests and the testing context (Grubb & Ollendick, 1986); (d) leniency effects favoring one group over another (Berry et al., 2013); (e) differential recruiting, mentoring, and retention interventions across groups (Berry et al., 2013); and (f) differential course difficulty across groups (Berry & Sackett, 2009). Given these factors, it seems unlikely that differential prediction would generalize across contexts and institutions. However, the possible presence of heterogeneity is an issue that has not been assessed systematically. For example,

[1] In addition to sampling error, measurement error, and range restriction, Hunter and Schmidt (2004) and others (Aguinis, Pierce, & Culpepper, 2009) have identified additional factors that increase the variance of validity coefficients across studies. These factors include scale coarseness, imperfect construct validity in the predictor and/or criterion variables, computational and other errors in data, and artificial dichotomization of continuous variables.

although Linn (1973) described differences in the extent of differential prediction across the 22 institutions included in his study, it is unclear the extent to which such variability was substantive in nature or because of methodological and statistical artifacts.

In sum, our study introduces the new concept of differential prediction generalization and investigates the potential presence of variability in ethnicity and gender-based differential prediction across contexts. We do so using data predicting first-year college grade point average (GPA) from SAT scores and high-school GPA.

Method

Data Collection Procedures and Participants

We obtained the raw data from Mattern and Patterson's (2013) Appendixes A-F, which include tables in a 384-page PDF document available at http://dx.doi.org/10.1037/a0030610.supp. We exported the data from these tables to Microsoft Excel using Able2Extract Pro 7.0 and SomePDF 1.0. Additional details regarding the data extraction algorithms and procedures are available from the authors upon request.

The tables include variance-covariance matrices involving relations among SAT scores, first-year college GPA, high-school grade point average (HSGPA), and demographic variables (i.e., ethnicity and sex) for 176 colleges and universities (i.e., 348 unique cohorts). Specifically, these include participating colleges and universities that provided the College Board with GPA and these data were matched to College Board databases that include SAT scores and responses to the SAT questionnaire, which included self-reported HSGPA and demographic information. The data were collected by the College Board as part of a multiyear study between 2006 and 2008. Identical to Mattern and Patterson (2013), we treated each cohort (henceforth referred to as a "sample") as an individual data point. Sixty-one out of 339 (i.e., 17.99%), 48 out of 247 (i.e., 19.43%), and 50 out of 264 (i.e., 18.93%) institutions provided three samples for the female–male (FM), Black–White (BW), and Hispanic–White (HW) comparisons, respectively. Thus, the contribution of three samples by institutions is only a small portion of the total, which reduces the likelihood that dependency due to cohorts nested within institutions may have biased our results. To more formally assess the possibility of dependence in the data structure, we examined the variance attributed to cohorts nested within institutions and the result was only .4% of the total variability. In other words, this small amount of variance suggests that it is appropriate to treat each sample as an individual data point in our analyses because data dependence did not bias standard error estimates (Aguinis & Culpepper, 2015; Aguinis, Gottfredson, & Culpepper, 2013; Raudenbush & Bryk, 2002).

Mattern and Patterson (2013) reported that the institutions were diverse in terms of geographic region, public/private, size, and selectivity. In addition, Mattern and Patterson (2013) reported removing samples with fewer than 15 individuals in any of the ethnicity- or gender-based subgroups from their analyses. Accordingly, FM comparisons were made based on approximately 257,336 women and 220,433 men across 339 samples. BW comparisons were based on 29,734 Black and 304,372 White students across 247 samples. For the WH comparisons, analyses were based on 35,681 Hispanic and 308,818 White students across 264 samples.

Differential Prediction Analysis

Assessing the presence of differential prediction involves estimating the following three models (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014; Cleary, 1968; Society for Industrial and Organizational Psychology, 2003):

$$GPA = \beta_0 + \beta_1 HSGPA + \beta_2 SAT\text{-}CR + \beta_3 SAT\text{-}M + \beta_4 SAT\text{-}W + e \quad (1)$$

$$GPA = \beta_0 + \beta_1 HSGPA + \beta_2 SAT\text{-}CR + \beta_3 SAT\text{-}M + \beta_4 SAT\text{-}W + \beta_5 G + e \quad (2)$$

$$GPA = \beta_0 + \beta_1 HSGPA + \beta_2 SAT\text{-}CR + \beta_3 SAT\text{-}M + \beta_4 SAT\text{-}W + \beta_5 G + \beta_6 HSGPA \cdot G + \beta_7 SAT\text{-}CR \cdot G + \beta_8 SAT\text{-}M \cdot G + \beta_9 SAT\text{-}W \cdot G + e \quad (3)$$

Equation 1 includes the criterion GPA regressed on the predictors HSGPA, SAT-CR (SAT-Critical Reading), SAT-M (SAT-Math), and SAT-W (SAT-Writing). The model in Equation 2 differs from Equation 1 in that it includes a dummy variable G, which has two categories and is used to assess the FM, BW, or HW comparisons. The model in Equation 3 includes product terms that capture interaction effects on GPA (i.e., moderating effect of ethnicity and gender on the relation between the predictors and GPA) and can be written in matrix notation as follows:

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{e}_j \quad (4)$$

where, for sample j, $\mathbf{y}_j$ is a n_j dimensional vector of criterion scores (i.e., n_j is the size for sample j), $\mathbf{X}_j$ is a $n_j \times q$ matrix of predictor variables (i.e., $q = 9$ for Equation 3), $\boldsymbol{\beta}_j$ is a q dimensional vector of regression coefficients, and $\mathbf{e}_j$ is a n_j dimensional vector of errors. The goal of differential prediction analysis is to examine whether test scores differentially predict criteria for different groups by examining whether coefficients within $\boldsymbol{\beta}_j$ (i.e., β_5, β_6, β_7, β_8, and β_9 in Equation 3) are different from zero. Specifically, a nonzero regression coefficient associated with predictor G suggests the presence of intercept-based differential prediction and nonzero coefficients associated with the product terms suggests the presence of slope-based differential prediction.

Differential Prediction Generalization Analysis

We used multivariate meta-analytic regression modeling (MMA) to synthesize regression coefficients and assess the degree of variability in differential prediction across samples as described by Becker and Wu (2007) and Chen, Manning, and Dupuis (2012). The MMA procedure uses data from each sample (i.e., $\mathbf{b}_j$ and $Cov[\mathbf{b}_j|\mathbf{X}_j]$) to estimate a meta-analyzed mean, in addition to cross-sample variance components. Specifically, the random effects MMA model described by Chen et al. (2012) includes the following equation for $\mathbf{b}_j$:

$$\mathbf{b}_j = \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\delta}_j + \mathbf{e}_j \quad (5)$$

where $\mathbf{W} = [\mathbf{I}_q, \mathbf{W}_j]$ is a $q \times (q + p)$ block design matrix that includes a q dimensional identity matrix and a $q \times p$ matrix of sample-level variables to explain differences in $\mathbf{b}_j$. Furthermore, $\boldsymbol{\delta}_j$ is a vector of random effects for sample j defined as $\boldsymbol{\delta}_j \sim N_q$ ($\mathbf{0}_q$, $\mathbf{T}$) where $\mathbf{T}$ is a $q \times q$ between sample variance-covariance matrix that quantifies the amount of heterogeneity that exists across samples above and beyond sampling error (i.e., $\mathbf{e}_j$, which is an error with a multivariate normal distribution; Chen et al., 2012).

Methodological and statistical artifacts. The goal of differential prediction generalization analysis is to quantify the variability in differential prediction across samples. However, sampling error, range restriction, and measurement error are three factors that should be ruled out given that they usually account for the largest proportion of observed variance (Aguinis, 2001; Hunter & Schmidt, 2004). In fact, Schmidt and Hunter (1981) estimated that an average of 72% of the variance of validity coefficients observed across studies is the result of these artifacts and, moreover, sampling error alone accounts for 85% of the variance accounted for by artifacts. Accordingly, in our study, $\mathbf{W}_j$ includes sample size (i.e., to account for sampling error).

In addition to sampling error, range restriction can increase or decrease observed variability in relation to true variability (Murphy, 1993). Accordingly, as noted by Linn (1983), "it is essential that selection effects be considered if our correlational and regression analysis results are to be properly interpretable" (p. 13). Range restriction is pervasive in college admissions testing because the data examined include only those students who have been admitted and for whom GPA information is subsequently available. The standard corrections for range restriction require three assumptions: linearity between predictors and criterion, constant residual error variance, and criterion scores missing at random (MAR) (Mendoza, 1993; Mendoza, Bard, Mumford, & Ang, 2004). Under these assumptions, commonly employed corrections such as Lawley's multivariate correction (Birnbaum, Paulson, & Andrews, 1950; Lawley, 1944) yield unbiased estimates of population correlation coefficients.[2] Furthermore, simulation studies support the accuracy of the Lawley correction across different sample sizes, magnitude of predictor intercorrelations, and degree of selectivity (Muthén & Hsu, 1993; Sackett & Yang, 2000).

A relevant issue pertaining to our study is that if the MAR and linearity assumptions are satisfied, the restricted regression coefficients (i.e., estimates in the selected sample) equal the estimated unrestricted coefficients. Stated differently, if these assumptions are met, range restriction does not bias estimates of $\mathbf{B}_j$, and the least squares estimator for the restricted sample is identical to the estimator corrected for range restriction. For example, consider Lawley's procedure and let $\mathbf{S}_{xxj}$ denote a $q \times q$ variance-covariance matrix among the predictors (i.e., covariances among the predictors in Equation 3) and $\mathbf{S}_{xyj}$ be a q dimensional vector of covariances between the predictors in Equation 3 and GPA in the jth sample. If there is no range restriction, the q dimensional vector of coefficients for sample j in Equation 3 are estimated as $\mathbf{b}_j = \mathbf{S}_{xxj}^{-1}\mathbf{S}_{xyj}$. However, $\mathbf{S}_{xxj}$ and $\mathbf{S}_{xyj}$ differ from values in the unrestricted applicant pool and, similar to Mattern and Patterson (2013), researchers employ Lawley's correction, which uses sample j's $q \times q$ predictor variance-covariance matrix $\boldsymbol{\Sigma}_{xxj}$ from the applicant pool. This information is available because Mattern and Patterson reported $\mathbf{S}_{xxj}$ and also $\boldsymbol{\Sigma}_{xxj}$ for all students in the applicant pool. The q dimensional vector of range restriction corrected coefficients are defined as

$$\tilde{\mathbf{b}}_j = \boldsymbol{\Sigma}_{xxj}^{-1}\tilde{\boldsymbol{\Sigma}}_{xyj} \quad (6)$$

where the Lawley correction defines $\tilde{\boldsymbol{\Sigma}}_{xyj} = \boldsymbol{\Sigma}_{xxj}\mathbf{S}_{xxj}^{-1}\mathbf{S}_{xyj}$. As expected, the restricted coefficients equal the unrestricted coefficients. Specifically, $\tilde{\mathbf{b}}_j = \boldsymbol{\Sigma}_{xxj}^{-1}\tilde{\boldsymbol{\Sigma}}_{xyj} = \boldsymbol{\Sigma}_{xxj}^{-1}\boldsymbol{\Sigma}_{xxj}\mathbf{S}_{xxj}^{-1}\mathbf{S}_{xyj} = \mathbf{b}_j$, so that $\mathbf{b}_j = \tilde{\mathbf{b}}_j$ if the MAR and linearity assumptions are satisfied.

The prior discussion shows that the restricted regression coefficients equal the corrected coefficients when the MAR and linearity assumptions are satisfied. In contrast, the restricted standard errors are too small, which implies that inferences for regression coefficients $\boldsymbol{\beta}_j$ are incorrect (Aguinis & Stone-Romero, 1997; Culpepper, 2012b). Consequently, it is necessary to correct the sample standard deviation of GPA for range restriction to obtain a corrected covariance matrix of $\mathbf{b}_j$. Let s_j^2 be sample j's variance of college grades. Lawley's corrected variance $\tilde{\sigma}_j^2$ is estimated as

$$\tilde{\sigma}_j^2 = s_j^2 - \mathbf{S}_{xyj}^T\mathbf{S}_{xxj}^{-1}(\mathbf{I}_q - \boldsymbol{\Sigma}_{xxj}\mathbf{S}_{xxj}^{-1})\mathbf{S}_{xyj} \quad (7)$$

where T indicates a vector transpose and $\mathbf{I}_q$ is a q dimensional identity matrix. If college grades were collected for all applicants, $Cov(\mathbf{b}_j \mid \mathbf{X}_j) = \frac{\sigma^2}{N_j}\boldsymbol{\Sigma}_{xxj}^{-1}$ would be the variance-covariance matrix of $\mathbf{b}_j$ in the applicant pool conditioned upon the predictor matrix $\mathbf{X}_j$ with σ^2 as the criterion variance in the applicant pool and N_j as the number of applicants. However, college grades are collected for admitted and enrolled students only, so $\tilde{\sigma}_j^2$ must be used as an estimate of σ_j^2 and n_j is used rather than N_j, which implies that an estimate for the range restriction corrected variance-covariance matrix of the $\mathbf{b}_j$ for sample j is

$$Cov(\mathbf{b}_j \mid \mathbf{X}_j) = \frac{\tilde{\sigma}_j^2}{n_j}\boldsymbol{\Sigma}_{xxj}^{-1}. \quad (8)$$

In addition to sampling error and range restriction, measurement error in the criterion GPA also needs to be ruled out as a potential source of variability in differential prediction across samples. Criterion measurement error usually inflates observed variability of correlation coefficients across studies (Schmidt & Hunter, 1977). This effect has been documented regarding correlation coefficients but Cohen, Cohen, West, and Aiken (2003, pp. 56–57) showed that bivariate regression coefficients are unaffected by criterion measurement error. Extending the work by Cohen et al. (2003), Supplemental File A available online provides new deri-

[2] Although Mendoza (1993) argued that the MAR assumption is reasonable in the particular context of college admissions testing because decision-makers do not observe the missing criterion scores, the effects of violating the MAR, linearity, and homoscedasticity assumptions on differential prediction generalization analysis are unknown and would depend on the nature of the missing data pattern, the nonlinear relationship (i.e., concave or convex), and the nonconstant error pattern (Culpepper, 2015). Mattern and Patterson (2013) did not report results regarding compliance with these assumptions and, in addition, their dataset did not include sufficient information for us to conduct this assessment. Specifically, complete student records would be needed to test for compliance with the linearity and homoscedasticity assumptions and additional information from admissions offices would be needed to assess compliance with the MAR assumption. Thus, additional data and research are needed to address these issues.

vations and proof that correcting the criterion for measurement error using classical test theory does not affect the observed variance of differential prediction across samples in the multiple predictor case. Hence, correcting criterion measurement error in GPA would not change estimates of differential prediction variability.

Another methodological artifact that could affect the degree of observed differential prediction variability across samples is differential predictor measurement error. Mattern and Patterson (2013) reported reliability information for the predictors across all samples: .82, .91, .91, and .89 for HSGPA, SAT-Critical Reading (SAT-CR), SAT-Math (SAT-M), and SAT-Writing (SAT-W), respectively. Differential prediction variability may be due, at least in part, to differences in predictor reliability across institutions (i.e., the same population parameter may take on different sample-based values depending on the local degree of measurement error). However, it is not possible to correct for the potential effects of differential reliability on differential prediction variability without sample-level reliability information. Nevertheless, reliability estimates for all predictors are .80 or higher which, as noted by Lance, Butts, and Michels (2006), "appears to be Nunnally's (1978) recommended reliability standard for the majority of purposes cited in organizational research" (p. 206). Accordingly, it is unlikely that differential predictor reliability would be so large as to completely eliminate all differential prediction variability if it exists. Nevertheless, if the College Board makes these data available in the future, analyses considering sample-level measurement error will be possible.

Finally, there are additional factors that could account for observed variability in differential prediction across samples. Specifically, some of these factors include unequal number of test takers across groups (i.e., women vs. men, Blacks vs. Whites, Hispanics vs. Whites); subgroup mean differences regarding the predictors SAT-CR, SAT-M, SAT-W, and high-school GPA; and standard deviations (SDs) for the predictors (as suggested by Linn, 1983). Thus, we included each of these factors in our study.

Quantifying differential prediction variability. To quantify the degree of differential prediction variability across samples, we conducted a formal test using Cochran' Q statistic. Q is a statistic for evaluating the degree to which regression coefficients differ across samples and is computed by summing the squared deviations of each study's regression coefficient estimate from the overall meta-analytic estimate and weighting each study's contribution by its sample size. Hence, a statistically significant Q suggests the presence of heterogeneity beyond what is expected by chance (Aguinis & Pierce, 1998; Aguinis et al., 2008). In addition, we also conducted a variance decomposition analysis and report the percent of cross-sample variance that remains after sampling error; range restriction; proportion of test takers across ethnicity- and gender-based subgroups, subgroup mean differences on the predictors (i.e., SAT-CR, SAT-M, SAT-W, and HSGPA); and *SD*s for the predictors have been accounted for as possible sources of variance.

Implementing differential prediction generalization analysis. We conducted the following steps. First, we computed unstandardized regression coefficients, $\mathbf{b}_j$, from Equation 3 for each institution. Then, we corrected the variance-covariance matrix for $\mathbf{b}_j$ for range restriction using Equations 7 and 8. We implemented the MMA procedure as in Equation 5 for two models. Model 1 used $\mathbf{b}_j$ and $Cov(\mathbf{b}_j \mid \boldsymbol{X}_j)$ as discussed earlier as input for the MMA procedure. For Model 1 there were no sample-level variables included (i.e., $W = I_9$ and $W_j = 0$). Model 2 extended Model 1 by including the following sample-level predictors into W_j: inverse of sample size, proportion of test takers in reference group, subgroup mean differences regarding predictors (i.e., three SAT tests and HSGPA), and sample-level *SD*s for the four predictors. In the Results section, S_T refers to the standard deviation of unstandardized regression coefficients from the meta-analyzed mean coefficients. Furthermore, we also estimated S_b, which denotes the estimated *SD* of random effects (δ_j in Equation 5 for Model 2). We implemented the differential prediction generalization analysis with R (R Core Team, 2014) using the mvmeta (Gasparrini, Armstrong, & Kenward, 2012) and mvtmeta (Chen, 2012) packages.

Similarities and Differences in Data-Analytic Approach Between Mattern and Patterson (2013) and Present Study

We implemented the same range restriction correction as Mattern and Patterson that was described previously. However, there is an important difference between the data-analytic approach employed by Mattern and Patterson compared with our study. Specifically, our study implemented a novel differential prediction generalization analysis based on the multivariate meta-analytic regression modeling approach recommended by Becker and Wu (2007), who provided a detailed discussion concerning the merits of different approaches for meta-analyzing regression coefficients. We followed their recommendation because this approach considers the size of each sample explicitly and the effects of other factors (i.e., range restriction; proportion of students across ethnicity- and gender-based subgroups; subgroup mean differences for the predictors HSGPA, SAT-CR, SAT-M, and SAT-W; and *SD*s for the predictors) and, therefore, allows us to understand the extent to which observed variability in differential prediction is substantive or because of methodological and statistical artifacts.

Results

Corroboration of Mattern and Patterson (2013) Results

We first attempted to corroborate Mattern and Patterson's results based on multiple regression correlations (i.e., square root of R^2) for models with different subsets of the predictors and different types of corrections. This corroboration was necessary prior to our substantive analysis assessing differential prediction generalization to confirm the integrity of the database and that our differential prediction assessment procedure is identical to the one implemented by Mattern and Patterson.

Table 1 includes the multiple correlations reported by Mattern and Patterson (2013) based on observed (i.e., uncorrected) scores (R_{obs}), multiple correlation based on models using Lawley's correction for predictor and criterion range restriction (R_{RR}), multiple correlation based on models correcting for predictor and criterion range restriction and criterion measurement error (R_{RRME}), and multiple correlation based on models correcting for predictor and criterion range restriction and predictor and criterion measurement

Table 1
Comparison of Results in the Present Study With Results Reported in Tables 2, 5, and 6 of Mattern and Patterson (2013)

Predictor combinations	MP		Corroborated		MP		Corroborated		MP		Corroborated		MP		Corroborated	
	R_{obs}	SD_{obs}	R_{obs}	SD_{obs}	R_{RR}	SD_{RR}	R_{RR}	SD_{RR}	R_{RRME}	SD_{RRME}	R_{RRME}	SD_{RRME}	ρ	SD_{ρ}	ρ	SD_{ρ}
MP Table 2																
I																
A. SAT	.367	.070	.367	.070	.473	.075	.473	.075	.511	.083	.511	.083	.527	.084	.527	.083
B. SAT, female	.402	.068	.402	.063	.501	.072	.501	.067	.541	.079	.541	.074	.554	.080	.554	.075
C. SAT, female, ints.	.405	.067	.405	.062	.504	.070	.504	.066	.544	.077	.544	.073	.558	.077	.558	.074
II																
A. HSGPA	.370	.075	.370	.074	.473	.056	.473	.055	.511	.060	.511	.060	.566	.066	.566	.066
B. HSGPA, female	.382	.080	.382	.075	.481	.059	.481	.055	.519	.063	.519	.059	.571	.068	.571	.065
C. HSGPA, female, ints.	.383	.080	.383	.075	.482	.060	.482	.055	.521	.064	.521	.059	.573	.070	.574	.066
III																
A. HSGPA,	.468	.057	.468	.057	.565	.059	.565	.059	.610	.065	.610	.065	.643	.065	.643	.065
B. HSGPA, SAT, female	.482	.056	.482	.056	.575	.057	.575	.057	.621	.064	.621	.063	.650	.064	.650	.064
C. HSGPA, SAT, female, ints.	.486	.056	.486	.056	.579	.056	.579	.056	.625	.063	.625	.062	.654	.063	.654	.063
MP Table 5																
IV																
A. SAT	.366	.068	.366	.068	.467	.074	.467	.074	.504	.081	.504	.081	.521	.082	.521	.082
B. SAT, Black	.376	.068	.376	.068	.476	.074	.476	.074	.514	.082	.514	.082	.529	.082	.529	.082
C. SAT, Black, ints.	.379	.068	.379	.068	.480	.074	.480	.074	.518	.081	.518	.081	.536	.082	.536	.081
V																
A. HSGPA	.384	.069	.384	.069	.480	.052	.480	.052	.518	.056	.518	.056	.573	.062	.573	.062
B. HSGPA, Black	.408	.063	.408	.063	.503	.052	.503	.052	.543	.057	.543	.057	.591	.061	.591	.061
C. HSGPA, Black, ints.	.411	.063	.411	.063	.507	.051	.506	.051	.547	.056	.547	.056	.598	.062	.596	.064
VI																
A. HSGPA	.473	.054	.473	.054	.563	.056	.563	.056	.607	.062	.607	.062	.641	.062	.641	.062
B. HSGPA, SAT, Black	.479	.054	.479	.054	.568	.056	.568	.056	.613	.063	.613	.063	.645	.062	.645	.062
C. HSGPA, SAT, Black, ints.	.483	.054	.483	.054	.574	.056	.574	.056	.619	.062	.619	.062	.655	.063	.654	.063
MP Table 6																
VII																
A. SAT	.354	.067	.354	.070	.444	.072	.444	.074	.479	.078	.480	.081	.496	.079	.496	.081
B. SAT, Hispanic	.359	.068	.359	.070	.448	.073	.448	.074	.484	.080	.484	.081	.500	.080	.500	.081
C. HSGPA, Hispanic, ints.	.363	.067	.363	.069	.453	.072	.453	.073	.489	.079	.489	.080	.506	.079	.505	.083
VIII																
A. HSGPA	.373	.068	.373	.072	.462	.053	.462	.054	.498	.057	.498	.059	.552	.064	.552	.065
B. HSGPA, Hispanic	.392	.059	.392	.061	.478	.053	.478	.050	.516	.058	.516	.055	.566	.062	.566	.061
C. HSGPA, Hispanic, ints.	.394	.059	.394	.061	.480	.053	.480	.050	.518	.058	.518	.054	.570	.062	.570	.062
IX																
A. HSGPA	.464	.053	.464	.055	.546	.054	.546	.056	.589	.060	.589	.062	.623	.059	.623	.062
B. HSGPA, SAT, Hispanic	.468	.052	.468	.054	.549	.054	.549	.056	.593	.060	.593	.062	.627	.059	.627	.062
C. HSGPA, SAT, Hispanic, ints.	.471	.053	.471	.055	.554	.054	.554	.055	.599	.060	.599	.062	.634	.060	.633	.066

Note. MP = Mattern and Patterson (2013); HSGPA = high school grade point average; ints. = interactions; R_{obs} = multiple correlation based on observed (i.e., uncorrected) models; R_{RR} = multiple correlation based on models using Lawley's correction for predictor and criterion range restriction; R_{RRME} = multiple correlation based on models correcting for predictor range restriction and criterion measurement error; ρ = multiple correlation based on models correcting for range restriction and predictor and criterion measurement error (i.e., true validity model) using an errors-in-variables model. *SD* corresponds to the appropriate *SD* of correlations. Model A only includes first-order effects (i.e., Equation 1), Model B adds the dummy code G representing ethnicity- and gender-based comparisons (i.e., Equation 2), and Model C adds all interactions (i.e., "ints.") between continuous and categorical variables (i.e., Equation 3).

error using an errors-in-variables model (i.e., ρ; Culpepper, 2012a; Culpepper & Aguinis, 2011).[3] Table 1 includes several types of hierarchical regressions and all analyses are based on centered continuous predictors. For example, "I", "A" under "MP Table 2" corresponds to the FM comparison in Mattern and Patterson for a model that only includes SAT scores. In contrast, "III", "C" is a model that includes SAT and HSGPA variables, a gender reference variable, and all product terms between the categorical and continuous variables. Results shown in Table 1 indicate that the corroborated results are within minimal rounding error of Mattern and Patterson's results at each stage and after the implementation of each type of correction. Consequently, Table 1 provides evidence that the data, equations, and procedures we used to assess differential prediction are identical to those used by Mattern and Patterson.

Despite our ability to reproduce results, we found a few discrepancies that are likely typographical errors in Mattern and Patterson (2013) for the model including range restriction and criterion measurement error correction. In fact, we detected this same inconsistency in the Mattern and Patterson article for the FM, BW, and HW comparisons, which is highly improbable given that correcting for range restriction should lead to multiple correlation coefficients that are different from those based on observed data (e.g., Berry et al., 2013). In short, the only difference between our results and Mattern and Patterson's is that they may have mistakenly repeated the label "none" and copied the incorrect results in their Table 3 This discrepancy does not affect the differential prediction generalization results and conclusions reported herein because our analyses are based on their data and not results they reported in their Table 3

Differential Prediction Analysis

Table 2 reports range restriction corrected differential prediction results for the FM, BW, and HW comparisons (i.e., results from Model 1). Specifically, the EST column shows average (i.e., meta-analyzed) coefficients across samples. Results for the coefficients in Table 2 indicate small differences for the simple slope coefficients for the SAT subtests for the BW and HW comparisons. Also, coefficients reported in Table 2 provide evidence that the SAT-CR and SAT-M tests were more strongly related to college GPA for women in comparison to men. Table 2 also provides evidence of subgroup differences in intercepts across the three subgroup comparisons, as has been shown in the past. That is, women scored, on average, 0.15 grade points higher than men whereas Blacks and Hispanics earned GPAs that were, on average, 0.19 and 0.10 points lower than Whites, respectively. These results, which represent the average degree of differential prediction for slopes and intercepts across samples for the FM, BW, and HW comparisons are consistent with previous studies (e.g., Fischer et al., 2013; Mattern & Patterson, 2013).

Graphic Representation of Differential Prediction Across Samples

Prior to conducting differential prediction generalization analysis, we calculated differences in predicted GPA values, symbolized by $\Delta\hat{Y}$, for the FM, BW, and HW comparisons and present results in Figure 1. This figure offers a visual display of the variability of differential prediction across samples and plots the individual lines for each sample to provide a graphical representation of the regression coefficients that were modeled in the metaregression procedure (i.e., coefficients prior to corrections). In calculating values for $\Delta\hat{Y}$ for each predictor, the other predictor scores are assumed to be equal to their means and we plotted $\Delta\hat{Y}$ between −2 and 1.5 *SD*s around the predictor average. Thus, for example, for SAT-M, $\Delta\hat{Y} = \Delta\beta_0 + \Delta\beta_1 SAT\text{-}M$. The panels in Figure 1 include not only the aggregated degree of differential prediction across all samples (i.e., central tendency), but also the individual lines for each sample to provide an indication of dispersion across samples.

Figure 1 shows variability in subgroup prediction line differences prior to adjusting for statistical and methodological artifacts. Furthermore, Figure 1 shows that the direction of slope differences varies and that there are many samples for which GPA is either over- or underpredicted by as much as 0.25 on a 0 to 4.0 grade point scale and, in some cases, by 0.50 in the tails of predictor score distributions.

For pedagogical and illustrative purposes, Figure 2 plots the difference between predicted GPA values across subgroups, symbolized by $\Delta\hat{Y}$, for four prototypal scenarios to aid the interpretation of various types of differential prediction based on intercept and slope differences. Similar to Figure 1, Figure 2 plots $\Delta\hat{Y}$ prior to corrections for sample-level variables. Also similar to results plotted in Figure 1, for a given standardized predictor, z (i.e., HSPGA or SAT tests), $\Delta\hat{Y} = \Delta\beta_0 + \Delta\beta_1 z$ where $\Delta\beta_0$ and $\Delta\beta_1$ are intercept and slope differences, respectively, between the reference group coded as 0 (i.e., White, male) and the comparison group coded as 1 (i.e., ethnic minority, female). These illustrations are not average in terms of the amount and direction of differential prediction but, rather, exemplary for a considerable amount of samples. Also, to make comparisons easier, we used the same axis scales as in Figure 1.

First, consider Institution #61 in 2006 for the BW SAT-CR comparison, for which subgroup prediction equations are nearly equivalent (i.e., $\Delta\beta_0 = 0.006$ and $\Delta\beta_{SAT-CR} = 0.000$). The $\Delta\hat{Y}$ plot for Institution #61 is similar to a horizontal line with $\Delta\hat{Y} \cong 0$ for all values of z. Consequently, the plot for this institution is representative of those that include subgroups with similar intercepts and slopes. Next, consider Institution #136 in 2007 for the HW SAT-M comparison, for which the Hispanic intercept is approximately 0.20 units smaller than the White group (i.e., $\Delta\beta_0 = -0.198$ and $\Delta\beta_{SAT-M} = 0$). The $\Delta\hat{Y}$ plot for Institution #136 is horizontal, which indicates the absence of subgroup slope differences; however, $\Delta\hat{Y}$ is vertically shifted to the point where $\Delta\hat{Y} = -0.198$. In contrast, Institution #169 in 2008 for the HW SAT-W comparison includes subgroups that differ in slopes, but not intercepts where $\Delta\beta_0 = 0.014$ and $\Delta\beta_{SAT-W} = 0.004$. The extent to which institutions differ in slopes can be identified by the degree to which $\Delta\hat{Y}$ deviates from a horizontal line. For instance, Institution #169 has a $\Delta\hat{Y}$ plot with a positive slope that passes

[3] Corrections for range restriction and criterion measurement error affect R^2 values but, as noted above, they do not alter estimates of regression coefficients. The difference in R^2 values between uncorrected and corrected models is because of the fact that the Lawley procedure corrects the criterion variance and the correction for criterion measurement error divides the uncorrected R^2s by the root of the criterion reliability coefficient.

Table 2
Range Restriction Corrected Results of Differential Prediction Analysis for Female–Male, Black–White, and Hispanic–White Comparisons Using Meta-Analytic Regression Modeling

Variable	Female–Male				Black–White				Hispanic–White			
	EST	*SE*	Significance	S_b	EST	*SE*	Significance	S_b	EST	*SE*	Significance	S_b
HSGPA	.4394	.0066	***	.1073	.4635	.0079	***	.1153	.4548	.0076	***	.1130
SAT-CR	.0005	.0000	***	.0003	.0004	.0000	***	.0003	.0005	.0000	***	.0003
SAT-M	.0008	.0000	***	.0004	.0004	.0000	***	.0003	.0004	.0000	***	.0004
SAT-W	.0012	.0000	***	.0002	.0014	.0000	***	.0003	.0014	.0000	***	.0003
Reference	.1537	.0035	***	.0521	−.1883	.0078	***	.0919	−.1043	.0063	***	.0740
HSGPA ∗ Reference	−.0511	.0052	***	.0605	−.1388	.0114	***	.1400	−.0818	.0106	***	.1247
SAT-CR ∗ Reference	.0002	.0000	***	.0002	.0000	.0001		.0008	.0000	.0001		.0006
SAT-M ∗ Reference	.0003	.0000	***	.0003	.0001	.0001		.0007	.0001	.0001		.0007
SAT-W ∗ Reference	−.0001	.0000		.0002	−.0001	.0001		.0009	−.0001	.0001		.0008

Note. EST = fixed-effects coefficients; S_b = standard deviation of random effects ($\boldsymbol{\delta}_j$ in Equation 5). Criterion for all models: first-year college grade point average (GPA). Predictors: HSGPA: High school grade point average, SAT-CR: SAT Critical Reading, SAT-M: SAT Math, SAT-W: SAT Writing. Reference: Dummy variable representing subgroups and coded as 1 for women and 0 for men (female–male comparison), 1 for Black and 0 for White (Black–White comparison), and 1 for Hispanic and 0 for White (Hispanic–White comparison).
*** $p < .001$.

through the (0,0) point. Furthermore, we see that intercept differences are zero in Institution #169 by noting that the value of $\Delta\hat{Y}$ when $z = 0$ is zero. The fourth scenario, which refers to Institution #103 in 2007 for the BW SAT-M comparison, shows groups that differ in intercepts and slopes. For Institution #103, Blacks have a smaller intercept ($\Delta\beta_0 = -0.505$) and slope ($\Delta\beta_1 = -0.003$). Figure 2 shows that, for institution #103, $\Delta\hat{Y}$ is a downward sloping line indicating negative group differences in intercepts and slopes.

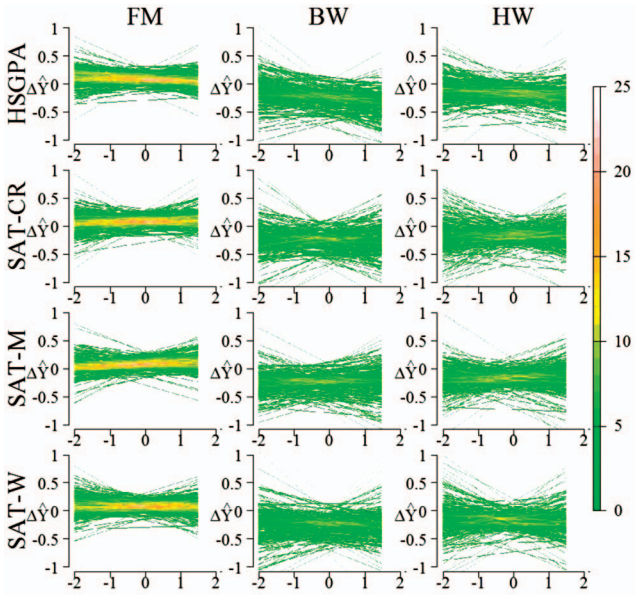

Figure 1. Variability in differential prediction across 348 samples of students in 176 colleges and universities. $\Delta\hat{Y}$ scores show differences between predicted first-year grade point average (GPA) scores across ethnicity- and gender-based subgroups based on models with scores corrected for range restriction. SAT-CR: SAT Critical Reading, SAT-M: SAT Math, SAT-W: SAT Writing, HSGPA: high school grade point average. The coloring indicates number of samples that overlap in subgroup prediction equation differences. FM: female versus male, BW: Black versus White, and HW: Hispanic versus White comparisons. The *x*-axes show predictor scores (i.e., HSGPA, SAT-CR, SAT-M, and SAT-W) and the *x*- and *y*-axes show scores in *SD* units.

Pervasiveness of Differential Prediction Across Samples

Figure 2 includes actual yet illustrative scenarios only. Accordingly, Table 3 includes more comprehensive information regarding the pervasiveness of differential prediction across samples. Specifically, Table 3 shows the percent of samples with intercept and slope differences different from zero for the three subgroup comparisons. We did not implement Bonferroni-type corrections to minimize a possible Type I error inflation because product terms capturing the interactions are correlated and such correction would result in overly conservative tests given the known insufficient statistical power in differential prediction analysis (Aguinis, Culpepper, et al., 2010; Bobko & Russell, 1994; Cronbach, 1987; McClelland & Judd, 1993). Moreover, as noted by Mattern and Patterson (2013), "Although the overall sample size was quite large, the average sample size per study was substantially smaller" (p. 142). Specifically, the average subgroup sample sizes were approximately 120 for African Americans and 135 for Hispanics, which are not uncommonly large (e.g., Aguinis & Stone-Romero, 1997). Much larger sample sizes are needed to achieve satisfactory statistical power (Aguinis, 2004a; Aguinis, Boik, & Pierce, 2001).

Table 3 shows that gender-based (i.e., FM) differential prediction occurred for slopes in 8.3%, 16.2%, and 4.1% of the samples for SAT-CR, SAT-M, and SAT-W, respectively. Considering results for the SAT-M, given that the FM comparison was based on a total of 477,769 students, approximately 77,399 (i.e., 16.2% of the total) attended an institution where SAT-M differentially predicted first-year college grades based upon gender. Black-White differences for slopes for HSGPA, SAT-CR, SAT-M, and SAT-W occurred in 39.7%, 19.4%, 13.4%, and 16.2% of the samples, which amounts to approximately 132,640, 64,817, 44,770, and 54,125 students out of a total of 334,106, respectively. In addition,

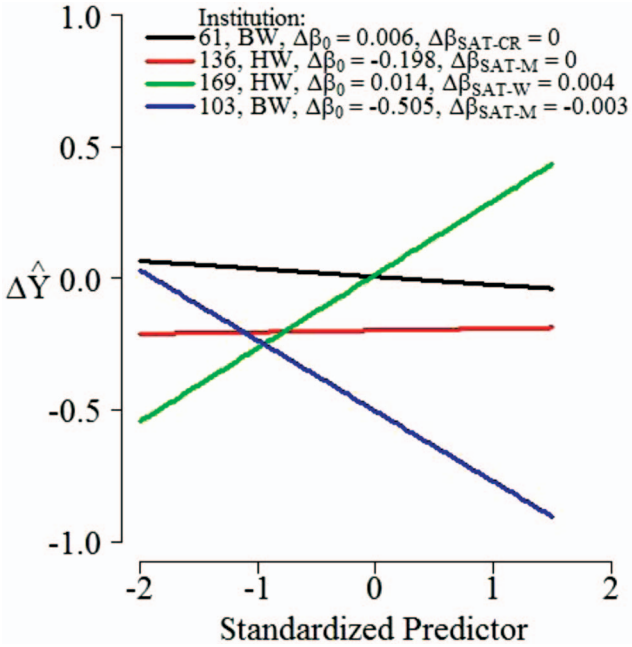

Figure 2. Prototypical scenarios based on actual samples showing no differential prediction and three forms of differential prediction. Institution #61: no differential prediction, Institution #136: differential prediction based on intercepts but not slopes, Institution #169: differential prediction based on intercepts and slopes, Institution #103: differential prediction based on slopes but not intercepts, $\Delta\hat{Y}$: subgroup-based differences in predicted criterion value (i.e., first-year college grade point average [GPA]), $\Delta\beta_0$ = subgroup-based differences in intercepts, and $\Delta\beta_1$ = subgroup-based differences in slopes. SAT-CR: SAT Critical Reading, SAT-M: SAT Math, SAT-W: SAT Writing. FM: female versus male, BW: Black versus White, and HW: Hispanic versus White comparisons. *x*- and *y*-axes show scores in *SD* units.

there were HW differences for HSGPA and the SAT subtests in 25.0%, 13.3%, 18.9%, and 15.5% of the samples, respectively, which suggests that approximately 86,125, 45,818, 65,110, and 53,397 students attended institutions where there is Hispanic–White differential prediction (out of a total of 344,499 students). Finally, Table 3 shows that differential prediction based on intercepts is even more pervasive: 80.8%, 61.9%, and 41.3% of samples for the FM, BW, and HW comparisons, respectively. In other words, there is differential prediction for the vast majority of samples for the FM comparison, for more than half for the BW comparison, and for just under half for the HW comparison.

Differential Prediction Generalization Analysis

Going beyond the reporting of the average degree of differential prediction across samples, Table 2 also includes the square root of the estimated *SD* of random effects for the nine regression coefficients for the three comparisons (i.e., the column labeled as "S_b"). S_b quantifies the extent of systematic differences in differential prediction across samples.

To assess the degree of differential prediction variability across samples, Table 4 includes results of a formal test pertaining to differential prediction generalization using Cochran's *Q* statistic. Recall that a statistically significant *Q* test suggests the presence of heterogeneity beyond what is expected by chance (Aguinis & Pierce, 1998; Aguinis, Sturman, & Pierce, 2008). Table 4 includes results for Model 1, which includes the nine predictor variables (i.e., five first-order effects and four product terms), and for Model 2, which includes Model 1 and the following additional sample-level predictors: inverse of sample size (to account for sampling error), proportion of test takers in reference group (i.e., to account for differences in the size of samples across ethnicity- and gender-based subgroups), subgroup mean differences regarding predictors (i.e., three SAT tests and HSGPA), and sample-level *SD*s for the four predictors. Results in Table 4 show that 13 out of the 15 *Q* tests are statistically significant. The only two statistically nonsignificant tests were the FM comparison for the SAT-W and SAT-CR tests. In other words, results in Table 4 indicate that (a) differential prediction based on HSGPA, SAT-CR, SAT-M, and SAT-W does not generalize for the BW and HW comparisons; (b) differential prediction based on HSGPA and SAT-M does not generalize for the FM comparison, and (c) there is differential prediction generalization based on the SAT-CR and SAT-W for the FM comparison.

In addition to *Q* statistics, Table 4's column labeled % shows the percent of variance in coefficients across samples that remains after accounting for methodological and statistical artifacts (i.e., variance decomposition based on S_b values from Model 2). More precisely, the rows for "Reference" show the percent of intercept-based differential prediction variance across samples remaining after accounting for methodological and statistical artifacts and the rows pertaining to two-way interactions show the percent of slope-based differential prediction variance across samples remaining after accounting for methodological and statistical artifacts. These results offer additional information about the extent of variability (i.e., degree of lack of generalization) for each test and subgroup comparison. Lack of differential prediction generalization was greatest for HSGPA for the BW comparison (about 34% of variance in coefficients across samples remains after methodological and statistical artifacts are taken into account), followed by the intercept for the BW and HW comparisons (about 29% of variance remaining for each), HSGPA for the HW comparison (about 28% of variance remaining), SAT-W for the BW comparison (about 20% of variance remaining), SAT-M for the HW comparison (about 19% of variance remaining), and the intercept for the FM comparison (also about 19% of variance remaining). Alternatively, for the SAT-W, only about 3% of variance in differential prediction across samples remains after artifacts are taken into account for the FM comparison.

Discussion

Our results reveal that the conclusion that "findings indicated that the use of SAT and HSGPA results in minimal differential prediction" (Mattern & Patterson, 2013, p. 146) is only reached when we examine summary statistics collapsing across the 348 samples collected from the 176 colleges and universities. In contrast, differential prediction generalization analysis suggests that there is substantial variability in differential prediction across samples. In fact, subgroup differences in intercepts and slopes are quite large for many colleges and universities and sample-level variability remains after accounting for sampling error and other

Table 3
Pervasiveness of Range Restriction Corrected Differential Prediction Based on Intercepts and Slopes for Female–Male, Black–White, and Hispanic–White Comparisons

Variable	Female–Male (339 samples; 477,769 students)			Black–White (247 samples; 334,106 students)			Hispanic–White (264 samples; 344,499 students)		
	%	*N*	% ≥ TPA	%	*N*	% ≥ TPA	%	*N*	% ≥ TPA
HSGPA	.976	475,287		.992	332,443		.989	343,627	
SAT-CR	.295	220,203		.397	175,566		.394	183,023	
SAT-M	.490	344,011		.360	161,682		.356	170,312	
SAT-W	.605	399,730		.834	317,429		.837	328,262	
Reference	.808	450,604		.619	259,959		.413	194,088	
HSGPA ∗ Reference	.224	143,715	.024	.397	157,899	.093	.250	97,575	.080
SAT-CR ∗ Reference	.083	60,395	.425	.194	68,264	.721	.133	55,317	.667
SAT-M ∗ Reference	.162	91,421	.345	.134	52,285	.700	.189	60,086	.659
SAT-W ∗ Reference	.041	15,982	.192	.162	69,955	.356	.155	54,732	.352

Note. Criterion for all models: first-year college grade point average (GPA). Predictors: HSGPA = high school grade point average; SAT-CR = SAT Critical Reading; SAT-M = SAT Math; SAT-W = SAT Writing. Reference: Dummy variable representing subgroups and coded as 1 for women and 0 for men (Female–Male comparison), 1 for Black and 0 for White (Black–White comparison), and 1 for Hispanic and 0 for White (Hispanic–White comparison). % = percentage of samples showing individual regression coefficients different from zero ($p < .05$); *N* = number of students based on summing samples sizes of samples with coefficients different from zero; % ≥ TPA = percent of samples with a differential prediction effect as large as or larger than the test's predictive ability (i.e., reference group slope) regardless of statistical significance. All values are computed using the model in Equation 3.

methodological and statistical artifacts that could potentially inflate observed differential prediction variability (i.e., range restriction, proportion of test takers across ethnicity- and gender-based subgroups, subgroup mean differences on the predictors, and *SD*s for the predictors). The finding regarding overall lack of differential prediction generalization is new because past research has only provided evidence regarding validity generalization (i.e., Boldt, 1986a, 1986b; Linn et al., 1981), but not regarding differential prediction generalization (or lack thereof). The *Standards for Educational and Psychological Testing* note that "validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, p. 11). Accordingly, the result regarding overall lack of differential prediction generalization also has implications for validity because knowledge that differential prediction does not generalize requires interpretations of test scores within local contexts.

Implications for Theory and Future Research

Aggregating results based on samples for which there is over prediction for one subgroup and samples for which there is under prediction for the same subgroup leads to the conclusion that, across samples, differential prediction is virtually nonexistent. The British writer and politician Benjamin Disraeli (1804–1881) stated the following (Huff, 1954): "A man eats a loaf of bread, and another man eats nothing; statistics is the science that tells us that each of these men ate half a loaf of bread." The same issue of aggregation across heterogeneous units—samples of students from different colleges and universities in our particular case—explains why Mattern and Patterson's results suggest that differential prediction is "minimal."

The variability in observed differential prediction across samples is not explained fully by sampling error and other methodological and statistical artifacts that have accounted for the majority of variance in validity coefficients across studies in past research. Specifically, the lack of differential prediction generalization is not explained by criterion measurement error, range restriction, proportion of test takers in reference group, predictor *SD*s, and subgroup mean differences regarding predictors (i.e., SAT-CR, SAT-M, SAT-W, and high-school grade point average). For the FM comparison, HSGPA and SAT-M show the greatest lack of differential prediction generalization. For the BW comparison, HSGPA also shows the greatest lack of differential prediction generalization, followed by SAT-W, SAT-CR, and SAT-M. For the HW comparison, the greatest lack of differential prediction generalization was also observed for HSGPA, followed by SAT-M, SAT-W, and SAT-CR.

Taken together, results suggest that, as is the case in many areas in educational and organizational research (Rousseau, 1978), context should play an important role in future college admissions testing research. In particular, future research can investigate cross-level interaction effects (Aguinis et al., 2013; Mathieu, Aguinis, Culpepper, & Chen, 2012). Specifically, as mentioned in the Introduction, there are institution-level variables (i.e., Level 2 moderators) that likely affect the relationship between individual-level test scores and performance (i.e., a relationship between a level-one predictor and a level-one criterion). For example, why is it that for some contexts and tests there are prediction differences in favor of Black students whereas for others the opposite is true? Mattern and Patterson (2013) took the first and unprecedented step to make a substantial amount of data available, but their data did not include information on substantive institution-level factors. We hope the College Board and other test vendors, not only of college admissions tests but also employee selection tests, will make institution-level data available so that future research will be able to answer this and other related critical questions. In other words, we currently do not know which institution-level factors cause differential prediction, and which particular form of differential prediction, across contexts. Given our results, there is a need for future research to

Table 4
Range Restriction Corrected Results of Cochran's Q Test and Variance Decomposition for Differential Prediction Generalization Across Samples for Female–Male, Black–White, and Hispanic–White Comparisons

	Female–Male							Black–White							Hispanic–White						
	Model 1 (df = 338)		Model 2 (df = 328)		Variance decomposition			Model 1 (df = 246)		Model 2 (df = 236)		Variance decomposition			Model 1 (df = 263)		Model 2 (df = 253)		Variance decomposition		
Variable	Q	Significance	Q	Significance	S_T	S_b	%	Q	Significance	Q	Significance	S_T	S_b	%	Q	Significance	Q	Significance	S_T	S_b	%
Reference	1248	***	753	***	.08035	.03456	18.5	654	***	596	***	.15824	.08456	28.6	728	***	646	***	.13323	.07123	28.6
HSGPA * Reference	677	***	679	***	.14139	.05858	17.2	789	***	749	***	.22814	.13262	33.8	659	***	636	***	.23102	.12205	27.9
SAT-CR * Reference	384	*	361		.00098	.00022	5.2	510	***	483	***	.00201	.00083	17.2	456	***	433	***	.00192	.00058	9.0
SAT-M * Reference	498	***	418	***	.00088	.00021	5.6	448	***	404	***	.00150	.00058	15.0	529	***	525	***	.00174	.00076	19.3
SAT-W * Reference	341		327		.00096	.00017	3.1	518	***	483	***	.00201	.00090	20.2	481	***	462	***	.00196	.00079	16.2

Note. Q = Cochran's Q statistic assessing sample-level variability, df = degrees of freedom; S_T = SD of sample-based unstandardized regression coefficients around meta-analyzed mean coefficients from Model 1, and criterion for all models: first-year college grade point average. Predictors: HSGPA = high school grade point average; SAT-CR = SAT Critical Reading; SAT-M = SAT Math; SAT-W = SAT Writing. Reference: Dummy variable representing subgroups and coded as 1 for women and 0 for men (female–male comparison), 1 for Black and 0 for White (Black–White comparison), and 1 for Hispanic and 0 for White (Hispanic–White comparison). Model 1 includes the 9 predictor variables (i.e., 5 first-order effects and 4 product terms in Equation 3). Model 2 includes Model 1 and the following sample-level predictors: inverse of sample size, proportion of test takers in reference group, subgroup mean differences regarding predictors (i.e., three SAT tests and HSGPA), and sample-level standard deviations for the four predictors. Results for predictors HSGPA, SAT-CR, SAT-M, and SAT-W are not included in this table because they provide information on variability for the reference groups across samples and are not of substantive interest in terms of differential prediction generalization analysis. S_b = standard deviation of random effects ($\boldsymbol{\delta}_j$ in Equation 5 for Model 2), % = S_b^2/S_T^2 (i.e., percentage of variance in coefficients across samples remaining after accounting for range restriction), and tables including Model 2 coefficients and SEs are included in the supplementary online file (Tables S1–S3).
* $p < .05$. *** $p < .001$.

examine factors causing differential prediction to vary in magnitude and direction across contexts. Results of this research will likely lead to effective actions and interventions. To guide future research, we offer a more detailed description of how and why each of the mechanisms we listed in the Introduction may serve as possible explanations for the presence of differential prediction and differential prediction variability across institutions.

Stereotype threat. Stereotype threat is a situational phenomenon that occurs when individuals believe they face the prospect of being evaluated as a function of, and confirming, a negative stereotype about a group to which they belong (Steele & Aronson, 1995). According to Walton et al. (2015), standardized cognitive ability tests can induce stereotype threat among test takers who are members of underrepresented groups (e.g., women, members of ethnic minority groups). Referred to as the "latent-ability" hypothesis, stereotype threat can prevent such test takers from performing as well as they are capable; that is, some of their cognitive ability remains latent or hidden. Hence, test scores can show systematic differential prediction such that they underestimate the ability and potential performance of individuals from negatively stereotyped groups (Walton & Spencer, 2009). Walton et al. (2015) concluded that stereotype threat can affect ethnic minorities' scores on cognitive ability tests administered in evaluative settings (e.g., schools) and, thus, result in disproportionately negative effects on decisions regarding their selection. The magnitude of the effect of stereotype threat on differential prediction may, however, depend on the degree to which the threat affects predictor and criterion scores differentially across ethnicity-based subgroups (Brown & Day, 2006). In short, differential levels of stereotype threat are likely to lead to differential levels of differential prediction across institutions.

Lack of common cultural frame of reference and identity across groups. Members of different ethnicity-based subgroups do not share a common cultural frame of reference and identity (Ogbu, 1993). For example, ethnic minority group members may interpret discrimination against them as permanent and institutionalized. This frame of reference develops over long periods of time as the result of perceived or actual exclusion, segregation, and barriers to opportunities. It can make some ethnic minority group members have lower expectations about the likelihood that obtaining good test scores will lead to desirable outcomes such as admission to college (Gould, 1999). Stated differently, cultural frames of reference affect how tests and testing situations are interpreted. Hence, ethnicity-based subgroups differ in their interpretation of the meaning of test scores and the relation between test scores and performance measures (Grubb & Ollendick, 1986). Such ethnicity-based differences in cultural frames likely differ across contexts and institutions and, therefore, are another factor that likely leads to differential levels of differential prediction.

Leniency effects favoring one group over another. With respect to college students' grades and their GPA, leniency effects can occur when graders apply a "shifting standards" model and assign some minority students higher grades than they deserve (Berry et al., 2013). The resulting error variance in some minority students' GPA can affect the relation between cognitive ability test scores and GPA. Because this shifting of standards is unlikely to be homogenous across institutions, it is another contextual factor

likely to create variability in the degree of differential prediction across institutions.

Differential recruiting, mentoring, and retention interventions across groups. To meet affirmative action goals, many academic institutions make extra efforts to recruit, mentor, and retain ethnic minority students—this is also the case regarding women in fields in which they are underrepresented (e.g., STEM: science, technology, engineering, and math). These extra efforts could include using different admissions standards, offering extra tutoring, and providing counseling opportunities while in college (Berry et al., 2013). According to Berry et al. (2013), if institutions implement these efforts, then students' admission into and success in college can be a function of factors other than their cognitive ability, which could reduce the relation between their cognitive ability test scores and GPA. Because such efforts clearly differ across institutions, it could also be a factor leading to different degrees of differential prediction.

Differential course difficulty across groups. Finally, differential prediction may also be explained, at least in part, by differential course difficulty across gender- or ethnicity-based subgroups. For example, Berry and Sackett (2009) determined that differential course difficulty may explain differences regarding GPA scores and, moreover, this phenomenon may lead to a decrease in the resulting validity coefficient. Because differences in course difficulty are unlikely to be homogenous across institutions, it is also unlikely that the degree of differential prediction is homogeneous across institutions.

More broadly, there are additional issues regarding the use of GPA as the criterion that may lead to differential prediction variability across institutions. For example, these include differential course selection, drop-out rates, and institutional selection criteria at the local level, among others. As summarized by Berry and Sackett (2009), "College GPA certainly reflects academic performance to some degree, but there are also well-known sources of construct-irrelevant variance in GPA—particularly instructors' grading idiosyncrasies . . ." (p. 822). Hence, these and other idiosyncrasies associated with a student's GPA, which are likely to vary across institutions, may account, at least in part, for the lack of differential prediction generalization found in our study.

Implications for Practice

Results regarding overall lack of differential prediction generalization imply that SAT scores and HSGPA seem to function differently across some subgroups and institutions in predicting first-year college GPA. These results have important implications for practice given that, since 2005, between 1.4 million and 1.6 million students have taken the SAT annually, more than 1.66 million students have done so in the class of 2012 (College Board, 2013), and about 1.7 million students have taken it during the year 2013 (Lewin, 2014).

Results included in Table 3, and our earlier discussion, provide evidence regarding the pervasiveness of differential prediction. However, to gain a fuller understanding of practical significance, it is also important to consider the magnitude of the effect (Aguinis, Werner, et al., 2010). Table 3 includes the percentage of samples with subgroup slope differences that exceed the magnitude of the test's predictive ability for the reference group (i.e., slopes between predictors and criterion). For instance, the reference group slope for SAT-CR was smaller than the subgroup differences in 42.5% of FM, 72.1% of BW, and 66.7% of HW comparisons. In contrast, Table 3 shows that fewer than 10% of samples had reference group slopes for HSGPA that were less in magnitude than the subgroup difference.

Although the aforementioned results regarding the prevalence and magnitude of differential prediction provide evidence regarding practical significance, results have important implications even if differential prediction were smaller and existed in only a handful of samples. The reason is that more than 1.5 million students and their families are affected annually by decisions based on students' scores. Moreover, for a test taker whose GPA has been underpredicted for a desired college because of her ethnicity or his gender, it is no consolation that on average, and across institutions, differential prediction is minimal. In short, our results regarding practical significance show that differential prediction should be taken seriously and this is the reason why the *Standards for Educational and Psychological Testing* "emphasize that fairness to all individuals in the intended population of test takers is an overriding, foundational concern, and that common principles apply in responding to test-taker characteristics that could interfere with the validity of test score interpretation" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, p. 49). Moreover, "a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct . . . characteristics of all individuals in the intended population, including those associated with race, ethnicity, gender . . . must be considered throughout all stages of development, administration, scoring, interpretation, and use so that barriers to fair assessment can be reduced" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, p. 50).

Our results suggest that lack of differential prediction when using HSGPA and SAT tests cannot be assumed in making college admissions decisions. Depending on the institution and its local practices (e.g., admissions, grading, affirmative action policies), and various contextual and societal factors, it is possible that there may be differential prediction—and the form of such differential prediction is unlikely to be the same across samples. In terms of practice, institutions that rely on SAT and HSGPA for admissions and other types of decisions (e.g., scholarship allocations) would be well served by conducting a local differential prediction study to understand whether it exists and its nature. Only through an assessment of the presence of differential prediction together with future research aimed at understanding the reasons for various types of differential prediction will we be able to minimize it and, hopefully, eliminate it. Moreover, the finding that there is differential prediction may call into question the use of a particular test in a particular institution. In short, sample-level variability is too substantial to rely on results that are aggregated across institutions for determining whether differential prediction exists at any one institution.

One possibility in terms of practice would be to use a specific institution-based regression equation in making GPA predictions, but there are three important caveats. First, a local differential prediction study relies on data from one institution only and, consequently, sample size may be small. Accordingly, because of

a small sample and accompanying insufficient statistical power, such institution-based differential prediction analysis is likely to conclude that there is no differential prediction even if such differential prediction exists (Aguinis, Culpepper, et al., 2010). Thus, a power analysis is necessary before one can reach a conclusion of no differential prediction with confidence (Aguinis, 2004a). An additional recommendation is to use data from more than one cohort of students—particularly for the case of smaller institutions. But, such aggregation requires homogeneity of cohorts and contextual process that may account for differential prediction in a particular institution. Second, even if a local differential prediction study involves adequate statistical power, the resulting coefficients are influenced by statistical and methodological artifacts (e.g., sampling error, range restriction). Hence, they should be corrected so that the best estimates of population coefficients are used (Hunter & Schmidt, 2004). Third, one of the five anonymous reviewers included in the *Journal of Educational Psychology* review team that evaluated the original and subsequent nine revisions of our manuscript commented that the substantive factors we described as possible sources of differential prediction could be described as "institutional biases." Hence, this reviewer noted that the recommendation about conducting a local institutional-level differential prediction analysis might legitimize these institutional biases.

Regardless of whether an institution-level or other regression equation is used, a possible solution to address the existence of differential prediction would be to not use a common line and, instead, use different regression lines across subgroups. This practice used to be fairly typical (Schmidt & Hunter, 2004), possibly reflecting practitioners' belief regarding the existence of differential prediction. However, with the passage of the Civil Rights Act of 1991, the legal defensibility of this within-group norming has come into question and, in fact, it is generally illegal without a consent decree (Aguinis, 2004b, Cascio & Aguinis, 2011). Thus, the current legal context in the United States highlights the urgency to conduct additional research involving academic-practitioner collaborations that will hopefully result in a greater understanding of why and how differential prediction occurs.

Finally, our analyses involved an examination of differential prediction by assessing each individual predictor. We followed this approach because the goal of differential prediction analysis is to understand whether test score-performance relations vary across groups—for each test used in the decision making process (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014C). However, as noted by an anonymous reviewer, "what if, for a single school, predictive bias is found for 1 or 2 predictors (e.g., SAT-CR and SAT-M), but not the other predictors such that when the total application score is computed, the bias from SAT-CR and SAT-M is virtually cancelled out?" Although this is a possibility in some cases, our position is that, based on professional standards (i.e., American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014; Society for Industrial and Organizational Psychology, 2003), the goal of differential prediction analysis is to understand the role of each test and, therefore, differential prediction analysis should be interpreted at the test level of analysis.

Limitations

Our results and conclusions should be interpreted within the context of several limitations because of data unavailability that we mentioned earlier. Specifically, we were unable to assess the potential impact of violating the linearity, MAR, and constant variance assumptions. In addition, we were unable to assess the potential impact of bias in the criterion scores (i.e., GPA). Finally, we were unable to correct for the potential effects of differential reliability of predictors across samples.

Conclusion

Our introduction of the new concept called *differential prediction generalization*, which combines previous work on differential prediction and validity generalization, leads to the conclusion that the degree and nature of differential prediction vary across samples. Such differences remain after some methodological and statistical artifacts that affect the observed variance of differential prediction across institutions are taken into account. Thus, the lack of differential prediction generalization is not because of artifacts such as sampling error, criterion measurement error, and range restriction. Moreover, our results suggest that hundreds of thousands of individuals attend institutions for which there is differential prediction of first-year GPA and, consequently, scores are under or over predicted based on a student's ethnicity and gender when a common regression line is used to make admissions and other decisions. Because predictions of GPA are used by many institutions to make admissions, scholarship, and other important decisions that affect the lives of students and their families, there is an important need for future research aimed at understanding the reasons for differential prediction and differential prediction variability across institutions.

References

Aguinis, H. (2001). Estimation of sampling variance of correlations in meta-analysis. *Personnel Psychology, 54,* 569–590. http://dx.doi.org/10.1111/j.1744-6570.2001.tb00223.x

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

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

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. http://dx.doi.org/10.1177/109442810144001

Aguinis, H., & Culpepper, S. A. (2015). An expanded decision making procedure for examining cross-level interaction effects with multilevel modeling. *Organizational Research Methods, 18,* 155–176. http://dx.doi.org/10.1177/10944281145636

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95,* 648–680. http://dx.doi.org/10.1037/a0018714

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management, 39,* 1490–1528. http://dx.doi.org/10.1177/0149206313478188

Aguinis, H., & Pierce, C. A. (1998). Testing moderator variable hypotheses meta-analytically. *Journal of Management, 24,* 577–592. http://dx.doi.org/10.1177/014920639802400501

Aguinis, H., Pierce, C. A., & Culpepper, S. A. (2009). Scale coarseness as a methodological artifact: Correcting correlation coefficients attenuated

from using coarse scales. *Organizational Research Methods, 12,* 623–652. http://dx.doi.org/10.1177/1094428108318065

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. http://dx.doi.org/10.1037/0021-9010.82.1.192

Aguinis, H., Sturman, M. C., & Pierce, C. A. (2008). Comparison of three meta-analytic procedures for estimating moderating effects of categorical variables. *Organizational Research Methods, 11,* 9–34. http://dx.doi.org/10.1177/1094428106292896

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods, 13,* 515–539. http://dx.doi.org/10.1177/1094428109333339

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ancis, J. R., & Sedlacek, W. E. (1997). Predicting the academic achievement of female students using the SAT and noncognitive variables. *College and University, 72,* 1–8.

Becker, B., & Wu, M. (2007). The synthesis of regression slopes in meta-analysis. *Statistical Science, 22,* 414–429. http://dx.doi.org/10.1214/07-STS243

Berry, C. M., Clark, M. A., & McClure, T. K. (2011). Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology, 96,* 881–906. http://dx.doi.org/10.1037/a0023222

Berry, C. M., & Sackett, P. R. (2009). Individual differences in course choice result in underestimation of college admissions system validity. *Psychological Science, 20,* 822–830. http://dx.doi.org/10.1111/j.1467-9280.2009.02368.x

Berry, C. M., Sackett, P. R., & Sund, A. (2013). The role of range restriction and criterion contamination in assessing differential validity by race/ethnicity. *Journal of Business and Psychology, 28,* 345–359. http://dx.doi.org/10.1007/s10869-012-9284-3

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. http://dx.doi.org/10.1007/BF02289200

Bobko, P., & Russell, C. J. (1994). On theory, statistics, and the search for interactions in the organizational sciences. *Journal of Management, 20,* 193–200. http://dx.doi.org/10.1177/014920639402000111

Boldt, R. F. (1986a). *Generalization of GRE General Test validity across departments* (GRE Board Professional Report No. 82-13P). Princeton, NJ: Educational Testing Service.

Boldt, R. F. (1986b). *Generalization of SAT validity across colleges* (College Board Report No. 86–3, ETS RR No. 86–24). New York, NY: College Entrance Examination Board. http://dx.doi.org/10.1002/j.2330-8516.1986.tb00179.x

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. http://dx.doi.org/10.1037/0021-9010.91.4.979

Cascio, W. F., & Aguinis, H. (2011). *Applied psychology in human resource management* (7th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Chen, H. (2012). mvtmeta: Multivariate meta-analysis (R package version 1.0). Retrieved from http://CRAN.R-project.org/package=mvtmeta

Chen, H., Manning, A. K., & Dupuis, J. (2012). A method of moments estimator for random effect multivariate meta-analysis. *Biometrics, 68,* 1278–1284. http://dx.doi.org/10.1111/j.1541-0420.2012.01761.x

Civil Rights Act of 1964, P. L. 88–352, 78 Stat. 241 (1964).

Civil Rights Act of 1991, P. L. 102–166, 105 Stat. 1071 (Nov. 21, 1991).

Cleary, T. A. (1966, June). *Test bias: Validity of the Scholastic Aptitude Test for Negro and White students in integrated colleges* (Educational Testing Service Research Bulletin, RB-66–31). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1966.tb00529.x

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124. http://dx.doi.org/10.1111/j.1745-3984.1968.tb00613.x

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

College Board. (2013). *SAT pressroom*. Retrieved from http://press.collegeboard.org/sat

Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analyses recently proposed. *Psychological Bulletin, 102,* 414–417. http://dx.doi.org/10.1037/0033-2909.102.3.414

Culpepper, S. A. (2012a). Evaluating EIV, OLS, and SEM estimators of group slope differences in the presence of measurement error: The single indicator case. *Applied Psychological Measurement, 36,* 349–374. http://dx.doi.org/10.1177/0146621612446806

Culpepper, S. A. (2012b). Using the criterion-predictor factor model to compute the probability of detecting prediction bias with ordinary least squares regression. *Psychometrika, 77,* 561–580. http://dx.doi.org/10.1007/s11336-012-9270-8

Culpepper, S. A. (2015). An improved correction for range restricted correlations under extreme, monotonic quadratic nonlinearity and heteroscedasticity. *Psychometrika*. Advance online publication. http://dx.doi.org/10.1007/s11336-015-9466-9

Culpepper, S. A., & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods, 16,* 166–178. http://dx.doi.org/10.1037/a0023355

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. http://dx.doi.org/10.1111/j.1745-3984.2009.00079.x

Fischer, F. T., Schult, J., & Hell, B. (2013). Sex-specific differential prediction of college admission tests: A meta-analysis. *Journal of Educational Psychology, 105,* 478–488. http://dx.doi.org/10.1037/a0031956

Gasparrini, A., Armstrong, B., & Kenward, M. G. (2012). Multivariate meta-analysis for non-linear and other multi-parameter associations. *Statistics in Medicine, 31,* 3821–3839. http://dx.doi.org/10.1002/sim.5471

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

Gould, M. (1999). Race and theory: Culture, poverty, and adaptation to discrimination in Wilson and Ogbu. *Sociological Theory, 17,* 171–200. http://dx.doi.org/10.1111/0735-2751.00074

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. http://dx.doi.org/10.1016/0147-1767(86)90042-8

Guion, R. M. (1965). *Personnel testing*. New York, NY: McGraw-Hill.

Huff, D. (1954). *How to lie with statistics*. New York, NY: Norton.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Kobrin, J. L., & Patterson, B. F. (2011). Contextual factors associated with the validity of SAT scores and high school GPA for predicting first-year college grades. *Educational Assessment, 16,* 207–226. http://dx.doi.org/10.1080/10627197.2011.635956

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organiza-*

tional Research Methods, 9, 202–220. http://dx.doi.org/10.1177/1094428105284919

Lawley, D. N. (1944). IV: A note on Karl Pearson's selection formula. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences, 62,* 28–30.

Lewin, T. (2014, March 5). A new SAT aims to realign with schoolwork. *The New York Times*. Retrieved from http://www.nytimes.com/2014/03/06/education/major-changes-in-sat-announced-by-college-board.html

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43,* 139–161. http://dx.doi.org/10.3102/00346543043002139

Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology, 63,* 507–512. http://dx.doi.org/10.1037/0021-9010.63.4.507

Linn, R. L. (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement, 20,* 1–15. http://dx.doi.org/10.1111/j.1745-3984.1983.tb00185.x

Linn, R. L., Harnisch, D. L., & Dunbar, S. B. (1981). Validity generalization and situational specificity: An analysis of the prediction of first-year grades in law school. *Applied Psychological Measurement, 5,* 281–289. http://dx.doi.org/10.1177/014662168100500301

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology, 97,* 951–966. http://dx.doi.org/10.1037/a0028380

Mattern, K. D., & Patterson, B. F. (2013). Test of slope and intercept bias in college admissions: A response to Aguinis, Culpepper, and Pierce (2010). *Journal of Applied Psychology, 98,* 134–147. http://dx.doi.org/10.1037/a0030610

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114,* 376–390. http://dx.doi.org/10.1037/0033-2909.114.2.376

Mendoza, J. (1993). Fisher transformations for correlations corrected for selection and missing data. *Psychometrika, 58,* 601–615. http://dx.doi.org/10.1007/BF02294830

Mendoza, J., Bard, D. E., Mumford, M., & Ang, S. C. (2004). Criterion-related validity in multiple-hurdle designs: Estimation and bias. *Organizational Research Methods, 7,* 418–441. http://dx.doi.org/10.1177/1094428104268752

Murphy, K. R. (1993). The situational specificity of validities: Correcting for statistical artifacts does not always reduce the trans-situational variability of correlation coefficients. *International Journal of Selection and Assessment, 1,* 158–162. http://dx.doi.org/10.1111/j.1468-2389.1993.tb00104.x

Muthén, B. O., & Hsu, J. W. Y. (1993). Selection and predictive validity with latent variable structures. *British Journal of Mathematical and Statistical Psychology, 46,* 255–271. http://dx.doi.org/10.1111/j.2044-8317.1993.tb01015.x

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Ogbu, J. U. (1993). Differences in cultural frame of reference. *International Journal of Behavioral Development, 16,* 483–506. http://dx.doi.org/10.1177/016502549301600307

Pässler, K., Beinicke, A., & Hell, B. (2014). Gender-related differential validity and differential prediction in interest inventories. *Journal of Career Assessment, 22,* 138–152. http://dx.doi.org/10.1177/1069072713492934

Pfeifer, C. M., Jr., & Sedlacek, W. E. (1971). The validity of academic predictors for black and white students at a predominantly white university. *Journal of Educational Measurement, 8,* 253–261. http://dx.doi.org/10.1111/j.1745-3984.1971.tb00934.x

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ricci v. DeStefano et al. (2009). U.S. 07–1428 (U.S. Supreme Court, 2009).

Rousseau, D. M. (1978). Characteristics of departments, positions, and individuals: Contexts for attitudes and behavior. *Administrative Science Quarterly, 23,* 521–540. http://dx.doi.org/10.2307/2392578

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. http://dx.doi.org/10.1037/0003-066X.59.1.7

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85,* 112–118. http://dx.doi.org/10.1037/0021-9010.85.1.112

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540. http://dx.doi.org/10.1037/0021-9010.62.5.529

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research. *American Psychologist, 36,* 1128–1137. http://dx.doi.org/10.1037/0003-066X.36.10.1128

Schmidt, F. L., & Hunter, J. E. (2004). SED banding as a test of scientific values in I/O psychology. In H. Aguinis (Ed.), *Test-score banding in human resource selection: Legal, technical, and societal issues* (pp. 151–173). Santa Barbara, CA: Praeger.

Society for Industrial and Organizational Psychology, Inc. (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, 69,* 797–811. http://dx.doi.org/10.1037/0022-3514.69.5.797

Temp, G. (1971). Validity of the SAT for blacks and whites in thirteen integrated institutions. *Journal of Educational Measurement, 8,* 245–251. http://dx.doi.org/10.1111/j.1745-3984.1971.tb00933.x

Walton, G. W., Murphy, M. C., & Ryan, A. M. (2015). Stereotype threat in organizations: Implications for equity and performance. *Annual Review of Organizational Psychology and Organizational Behavior, 2,* 523–550. http://dx.doi.org/10.1146/annurev-orgpsych-032414-111322

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. http://dx.doi.org/10.1111/j.1467-9280.2009.02417.x

Received April 18, 2014
Revision received October 21, 2015
Accepted November 15, 2015 ■