AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Applied Psychology

ISSN: 0021-9010

2024, Vol. 109, No. 3, 402–414
https://doi.org/10.1037/apl0001152

Improving Our Understanding of Predictive Bias in Testing

Herman Aguinis[1] and Steven A. Culpepper[2]
[1] Department of Management, School of Business, The George Washington University
[2] Department of Statistics, University of Illinois Urbana-Champaign

Predictive bias (i.e., differential prediction) means that regression equations predicting performance differ across groups based on protected status (e.g., ethnicity, sexual orientation, sexual identity, pregnancy, disability, and religion). Thus, making prescreening, admissions, and selection decisions when predictive bias exists violates principles of fairness based on equal treatment and opportunity. First, we conducted a two-part study showing that different types of predictive bias exist. Specifically, we conducted a Monte Carlo simulation showing that out-of-sample predictions provide a more precise understanding of the nature of predictive bias—whether it is based on intercept and/or slope differences across groups. Then, we conducted a college admissions study based on 29,734 Black and 304,372 White students, and 35,681 Latinx and 308,818 White students and provided evidence about the existence of both intercept- and slope-based predictive bias. Third, we discuss the nature and different types of predictive bias and offer analytical work to explain why each type exists, thereby providing insights into the causes of different types of predictive bias. We also map the statistical causes of predictive bias onto the existing literature on likely underlying psychological and contextual mechanisms. Overall, we hope our article will help reorient future predictive bias research from *whether* it exists to the *why* of different types of predictive bias.

Keywords: fairness; diversity, equity, and inclusion; equal opportunity; test bias; affirmative action

As noted in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 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, 2018) define predictive bias as "systematic under- or overprediction of criterion performance for people belonging to groups differentiated by characteristics not relevant to criterion performance" (p. 49) and therefore "slope and/or intercept differences between subgroups indicate predictive bias" (p. 23). In other words, no predictive bias exists if the regression equations predicting criterion scores based on test scores are indistinguishable for the groups in question. For example, lack of predictive bias means that White and Black college applicants with the same SAT score are predicted to have the same college grade point average (GPA). On the other hand, predictive bias, also referred to as differential prediction, means that these two applicants are *not* predicted to have the same college GPA despite their identical SAT scores. Accordingly, making prescreening, admissions, and selection decisions in the presence of predictive bias violates fundamental principles of fairness based on equal treatment and equal opportunity (AERA et al., 2014; Camilli, 2013).

Organizational and societal interest in predictive bias and its implications for college admissions, as well as human resource selection and placement, is directly related to ongoing concerns about diversity, equity, and inclusion (DEI), which are considered part of the grand challenges of the 21st century (Eby, 2022; George et al., 2016). Moreover, "historically, testing and assessment has been linked to larger eugenics and white supremacy efforts that have tried to prove, through science, that African Americans and other non-whites are intellectually and culturally inferior" (Davis & Martin, 2018, p. 48). So, industrial–organizational (I–O) psychology research regarding predictive bias is far from a mere academic exercise. On the contrary, it has important and direct implications for college admissions policies and employment decisions that affect the lives of millions of people. Moreover, the June 2023 U.S. Supreme Court case on *Students for Fair Admissions v. Harvard* (2023) highlights the societal importance of these issues. Consistent with the call to arms by Rogelberg et al. (2022) that I–O psychology research should reach the public, Miles and Fassinger (2021) noted that "public psychology should … be conceptualized as outward-facing, socially engaged, discipline-wide, and focused on enhancing the public good via social justice aims such as equity, access, inclusion, justice, and safety for all" (pp. 1232–1233).

This article was published Online First October 12, 2023.
Herman Aguinis https://orcid.org/0000-0002-3485-9484
Steven A. Culpepper https://orcid.org/0000-0003-4226-6176
Herman Aguinis and Steven A. Culpepper contributed equally to this work. The authors have no known conflicts of interest to disclose.

The authors describe their analyses, all manipulations, and all measures in the study, and they adhered to the *Journal of Applied Psychology* methodological checklist. The R code they used is available upon request, and the data are available in the supplemental appendices of Mattern and Patterson (2013). Data were analyzed using R, Version 4.0.2. The study design, hypotheses, and analyses were not preregistered.

Correspondence concerning this article should be addressed to Herman Aguinis, Department of Management, School of Business, The George Washington University, Funger Hall, Suite 311, 2201 G Street Northwest, Washington, DC 20052, United States. Email: haguinis@gwu.edu

Predictive Bias Research to Date: Brief Review

Examining whether predictive bias exists has a long history dating back to Cleary (1968), who investigated data from three colleges, and Pfeifer and Sedlacek (1971), who analyzed data from 13 institutions. In most of these older studies, predictive bias has been found to be small and in the form of intercept but not slope differences such that tests overpredict performance for Black students.

Aguinis et al. (2010) challenged the conclusion that test bias in college admissions and preemployment testing is nonexistent and, if it exists, it only occurs regarding intercept-based differences that favor (i.e., overpredict the performance of) underrepresented group members. This conclusion was reached based on a simulation study of 15 billion 925 million individual samples of scores and more than 8 trillion 662 million individual scores. Specifically, Aguinis et al. (2010) provided evidence that the historically accepted moderated multiple regression procedure to assess test bias (i.e., Cleary, 1968) is itself biased: Slope-based bias is likely to go undetected, and intercept-based bias favoring underrepresented group members is likely to be found when it may not exist.

Aguinis et al. (2010) results prompted follow-up work by Mattern and Patterson (2013), who used data on over 475,000 students entering college between 2006 and 2008 to estimate slope and intercept differences in the college admissions context. Due to being employed by the College Board, Mattern and Patterson (2013) had access to College Board data (rather than simulations as done by Aguinis et al., 2010). Specifically, their study consisted of data for first-time, first-year undergraduates entering college in 2006, 2007, and 2008. They included 654,696 students and 177 unique institutions. Based on regression plots, they reported that college grades were consistently overpredicted for Black students, as had been reported in older studies.

Subsequently, Aguinis et al. (2016) introduced the concept of *differential prediction generalization*, which is the extent to which the type of predictive bias varies across samples and contexts. Using Mattern and Patterson's (2013) data, which are available in their Appendices A–F (i.e., a 384-page PDF document available at https://doi.org/10.1037/a0030610.supp), Aguinis et al. (2016) empirically examined whether predicted first-year college GPA based on high school GPA (HSGPA) and SAT scores depends on a student's ethnicity and whether this difference varied across contexts (i.e., different colleges and universities). Specifically, they 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. Results provided evidence for the lack of differential prediction generalization because variability in predictive bias across samples remained after accounting for methodological and statistical artifacts, including sample size, range restriction, proportion of students across ethnicity-based groups, subgroup mean differences on the predictors (i.e., HSGPA, SAT-Critical Reading [SAT-CR], SAT-Mathematics [SAT-M], and SAT-Writing [SAT-W]), and *SD*s for the predictors.

Providing additional supporting empirical evidence for Aguinis et al. (2016) findings of the presence of predictive bias and lack of differential predictive generalization, Berry et al. (2020) used meta-analysis and computational modeling and concluded that "cognitive ability tests can be expected to exhibit predictive bias against Hispanic applicants much of the time. However, some conditions did not exhibit underprediction" (p. 517). More specifically, their meta-analysis (i.e., Study 1) was based on 305 samples with Hispanic and White sample sizes of 9,917 and 74,428, respectively. Their computational modeling study included three steps: (a) correcting the Study 1 meta-analytic standardized mean difference (i.e., d value) between the subgroups on job performance (i.e., d_Y) for indirect range restriction, (b) using those corrected d_Y values to calculate intercept differences, and (c) testing Step 2 results for sensitivity to slope differences.

Subsequently, Sackett, Zhang, and Berry (2023) relied on 119 General Aptitude Test Battery validation studies for which they were able to obtain information not only on d_Y but also on d_X (i.e., standardized mean difference for the predictor) and the validity coefficient r_{XY}. They challenged Berry et al.'s (2020) underprediction results by concluding that "tests overpredict Hispanic performance … depending on assumptions made about artifact corrections" (p. 341). Specifically, Sackett, Zhang, and Berry (2023) did not apply any range restriction correction to their observed d_X value of .76 (which also differed from Berry et al.'s value of .83). They justified their choice by arguing that about 83% of the studies were concurrent, and "The lack of evidence of restriction in d_X fits this notion of little to no range restriction in this set of concurrent validity studies" (Sackett, Zhang, & Berry, 2023, p. 343). Subsequently, challenging Sackett et al.'s position that there is no need to implement a range restriction correction in concurrent validity studies, Oh et al. (2023) provided evidence that "unless all employees in all concurrent validation studies were hired randomly or the correlations among all selection procedures used are zero, it is hard to accept that [range restriction] has not occurred for any of the selection procedures" (p. 1308). Sackett, Berry, et al. (2023) responded that their

> Focus was on making range restriction corrections when conducting meta-analyses, where it is common for primary studies to be silent as to the prior basis for selection of the employees later participating in the concurrent validation study … As such, the applicant pool information needed for correction is typically not available. (p. 1311)

In another recent exchange, Dahlke and Sackett (2022) further challenged Aguinis et al. (2016) and Berry et al. (2020) conclusions about predictive bias and lack of differential prediction generalization. Instead, they supported the conclusion that predictive bias is small and consists of overprediction for underrepresented test takers. This conclusion was reached using a newly proposed predictive bias index called δ_{mod}. Dahlke and Sackett (2022) explicitly acknowledged that δ_{mod} is "ideally suited for use as an effect-size *complement* to the traditional significance tests performed in the Cleary framework" (p. 1997, emphasis added). However, despite this acknowledgment they explained that they applied "δ_{mod} to composites to revisit Aguinis et al. (2016) controversial finding that underprediction of racial/ethnic group performance is common" (p. 2009). They explained that "in all our analyses, we make use of recently developed standardized effect sizes for differential prediction" (Dahlke & Sackett, 2022, p. 1997). So, they did not use this index as just a "complement" but as a yardstick to assess the possible presence of predictive bias.

The Present Study

Our article aims to provide new insights into the different types of predictive bias and explain the statistical causes and likely underlying psychological and contextual mechanisms that lead to each type. Our preceding brief literature review reveals a back-and-forth dialogue as researchers' focus and debates have been mostly centered on which statistic is better and whose conclusion about the presence or absence of predictive bias is more legitimate. There are good reasons for this dialogue, and good insights have emerged. However, much less attention has been devoted to the possible reasons for predictive bias. For example, in their study examining predictive bias against Latinx test takers, Berry et al. (2020) noted that "Given this evidence of predictive bias against Hispanic American job applicants, a natural question is what is causing this bias?" (p. 535). However, they did not provide an answer to their question. Instead, highlighting an obvious knowledge gap, they noted, "The main contribution of this study is not explaining what causes the predictive bias, but rather providing evidence that it exists" (Berry et al., 2020, p. 535). More recently, Landers et al. (2022) examined possible predictive bias when using game-based assessments (GBA). However, their goal was similarly to assess whether predictive bias exists, rather than why, by answering the following question: "Is differential prediction of GPA with a *g*-GBA similar to that of traditionally-measured *g* for race and gender within the academic sample?" (p. 1661).

The remainder of our article is structured as follows. First, we conducted a two-part study: a simulation and a college admissions study. In the simulation, we show the value of using out-of-sample predictions to understand not just whether predictive bias exists but, more importantly in advancing theory, the precise nature of predictive bias. In the college admissions study, we compared (a) 29,734 Black versus 34,372 White students and (b) 35,681 Latinx versus 308,818 White students and provide evidence about the existence of predictive bias based on intercepts and slopes. Second, we describe the different types of predictive bias by offering analytical work to explain why each type might exist. Specifically, we describe slope-based predictive bias and intercept-based predictive bias and the statistical causes for each type. Importantly regarding theory development, we also map the two types of predictive bias and their statistical causes onto existing research on likely underlying psychological and contextual reasons for the existence of each. Finally, we offer implications for theory, future research, and practice. Overall, we hope our article will help reorient future predictive bias research from *whether* it exists to the *why* of different types of predictive bias.

Two-Part Study: On the Nature of Predictive Bias

Simulation Study

The classic moderated multiple regression model for assessing the possible presence of predictive bias is as follows (Cleary, 1968):

$$Y = \beta_0 + \beta_1 X + \beta_2 M + \beta_3 XM + e, \tag{1}$$

where M represents a binary moderator such that $M = 1$ for members of the focal group and $M = 0$ for individuals in the reference group. We investigated the fit of different models suggesting different types of predictive bias by evaluating the accuracy of out-of-sample predictions (cf. Mendoza et al., 2004). Specifically, consider the following models:

$$\text{Model 1 (M1)}: Y = b_0 + b_1 X_i + e, \tag{2}$$

$$\text{Model 2 (M2)}: Y = b_0 + b_1 X_i + b_2 M_i + e, \tag{3}$$

$$\text{Model 3 (M3)}: Y = b_0 + b_1 X_i + b_2 M_i + b_3 X_i M_i + e. \tag{4}$$

The models differ in which predictor variables are included in Equation 1 (i.e., regression equation used to assess the possible presence of predictive bias). Each model refers to a different situation regarding the nature of predictive bias—different types. In M1, there is no predictive bias because the intercept and slope relating X to Y are the same for both groups. In M2, the groups differ in intercepts as indicated by the coefficient for the M dummy variable by the amount b_2. Finally, in M3, both M and XM are included in the model, and predictive bias is due to both group differences in intercepts and slopes. In this model, the coefficients for X and M in the presence of interactions are still meaningful and interpretable because they are defined as the average effect of one variable across the range of the other variables (Aguinis, 2004; Aiken & West, 1991; Cohen et al., 2003; Jaccard et al., 1990).[1]

Simulation Design

We generated testing data sets from the population and used the estimated coefficients from the training data sets for each of the three models to create predictions about the presence of different types of predictive bias. We then evaluated prediction accuracy (i.e., correct identification of a particular type of predictive bias) using cross-validation, given that it provides a procedure for identifying the best-fitting model without overfitting (Raju et al., 1999; Schmitt & Ployhart, 1999; Shao, 1993).

A standard approach for quantifying model fit is to compute the cross-validated −2 log-likelihood (−2*LL*). The −2*LL* for the normal error multiple regression model is $-2LL = n(\ln(2\pi) + \ln(RSS/n) + 1)$ where π is the well-known constant, $RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ is the residual sums of squares, and $\hat{Y}_i$ is the predicted value for observation i. The cross-validated $-2LL$ is computed by using parameter estimates from a training data set to calculate *RSS* with the test data set. We computed the cross-validated $-2LL$ for M1, M2, and M3.

In addition to computing the cross-validated $-2LL$, we also computed the Akaike information criterion (AIC; Akaike, 1974). We included AIC because there is theoretical evidence that it asymptotically equals the leave-one-out, cross-validated $-2LL$ for linear models (Stone, 1977). The AIC provides an estimate for the cross-validated $-2LL$, so it is a useful criterion for determining the best prediction model. Furthermore, besides the connection between

[1] Aiken and West (1991) wrote that "these effects should not be disregarded simply because they are not constant effects" (p. 39). Jaccard et al. (1990) noted that "some researchers find it meaningful on occasion to interpret main effects in the presence of statistical interaction if the main effect is viewed in terms of an average effect" (p. 34). Similarly, Cohen et al. (2003) wrote that "they represent the average effect of a predictor across the range of the [other] predictor" (p. 282). Also, Aguinis (2004), in a moderated multiple regression text devoted exclusively to categorical moderators (as is the case in predictive bias analysis), noted that "The presence of the interaction implies that this average [lower-order effect] was computed from heterogenous values" (pp. 35–36).

the AIC and cross-validation (cf. Hickman et al., 2022), the AIC is computationally easy to obtain with just the training data set. The AIC is defined as $-2LL + 2k$, where k denotes the number of model parameters. For instance, $k = 3$ for M1 (i.e., the model $\hat{Y}_i = b_0 + b_1X_i$ includes two regression coefficients and a residual variance) and $k = 4$ and 5 for M2 (i.e., $\hat{Y}_i = b_0 + b_1X_i + b_2M_i$) and M3 (i.e., $\hat{Y}_i = b_0 + b_1X_i + b_2M_i + b_3X_iM_i$), respectively.

Results and Discussion

Results in Table 1 show that assessing out-of-sample predictions using the cross-validated −2LL and AIC correctly captures the best-fitting model in providing accurate information about whether predictive bias exists and which specific type of predictive bias exists. Specifically, Table 1 reports the values of the cross-validated $-2LL$ and AIC divided by the sample size and shows that the model with the smallest AIC is generally the model with the smallest cross-validated $-2LL$. Consequently, results in Table 1 provide evidence that the AIC can be used to determine which model provides the best fit—meaning which specific type of predictive bias is present. For instance, Case 3 corresponds with the scenario where most group differences in predictions are due to intercept differences. For Case 3, both the cross-validated $-2LL$ and AIC correctly identified M2 (i.e., the model with X and M) as the best-fitting model for $n = 1{,}000$, which is interpreted as intercept differences being the type of predictive bias. In sum, results are informative not just regarding whether predictive bias exists but, importantly from a conceptual standpoint, regarding the precise type of predictive bias.

College Admissions Study

This study aimed to better understand the presence of different types of predictive bias in the illustrative case of racioethnic comparisons in college admissions decisions. Our assessment of the nature of predictive bias is based on an analysis of out-of-sample predictions described earlier.

Data Set and Measures

We used the College Board data set Mattern and Patterson (2013) made available. The data set includes four predictors: HSGPA, SAT-CR, SAT-M, and SAT-W. In addition, the data set includes a categorical variable for Black versus White, and Latinx versus White comparisons, as well as interactions between the continuous predictors and categorical variable. Implementing the same procedure as Dahlke and Sackett (2022), we used an equal-weight composite including HSGPA, SAT-CR, SAT-M, and SAT-W.

Black–White (BW) comparisons are based on 29,734 Black and 304,372 White students across 247 samples, and Latinx–White (LW) comparisons are based on 35,681 Latinx and 308,818 White students across 264 samples. We use the term samples to mean institution cohorts given that the College Board data set includes information for 176 distinct institutions (i.e., colleges and universities) and includes between one to three cohorts per institution for 2006, 2007, and 2008.

Results and Discussion

The AIC model comparison results for BW and LW are plotted in Figure 1's top and bottom panels. This figure plots the difference in AIC between M2 and M1 (i.e., $AIC_{M2} - AIC_{M1}$) on the x-axis versus the difference in AIC for M3 and M1 on the y-axis (i.e., $AIC_{M3} - AIC_{M1}$) for all institution cohorts. Note that the differences $AIC_{M2} - AIC_{M1}$ and $AIC_{M3} - AIC_{M1}$ provide an indication of the relative fit of M2 (i.e., intercept-based predictive

Table 1

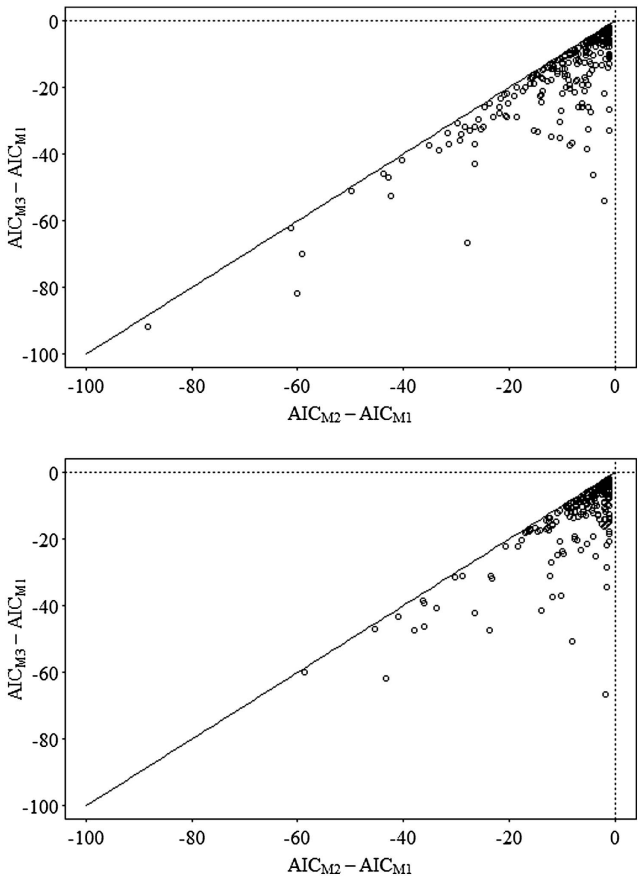
Results of Monte Carlo Simulation Assessing the Accuracy of Out-of-Sample Predictions With Cross-Validated −2 Log-Likelihood (−2LL) and Akaike Information Criterion (AIC) to Understand the Nature of Predictive Bias

				Cross-validated −2LL/n			AIC/n		
Case	n	β_2	β_3	M1	M2	M3	M1	M2	M3
1	1,000	−0.250	−0.069	2.845	**2.842**	2.843	2.841	**2.839**	**2.839**
	5,000	−0.250	−0.069	2.842	2.839	**2.838**	2.843	**2.839**	**2.839**
2	1,000	−0.222	−0.031	2.841	**2.839**	2.840	2.845	**2.843**	2.844
	5,000	−0.222	−0.031	2.841	**2.838**	**2.838**	2.842	**2.839**	**2.839**
3	1,000	−0.194	0.008	2.842	**2.840**	2.841	2.845	**2.843**	2.844
	5,000	−0.194	0.008	2.842	**2.839**	**2.839**	2.842	**2.838**	2.839
4	1,000	−0.167	0.046	2.841	**2.839**	2.840	2.840	**2.838**	2.839
	5,000	−0.167	0.046	2.842	**2.839**	**2.839**	2.841	**2.838**	**2.838**
5	1,000	−0.139	0.085	2.842	**2.840**	**2.840**	2.845	**2.843**	**2.843**
	5,000	−0.139	0.085	2.843	**2.840**	**2.840**	2.842	**2.839**	**2.839**
6	1,000	−0.111	0.123	2.841	2.839	**2.838**	2.846	**2.843**	**2.843**
	5,000	−0.111	0.123	2.843	2.840	**2.839**	2.843	2.840	**2.839**
7	1,000	−0.083	0.162	2.842	2.840	**2.839**	2.845	2.843	**2.842**
	5,000	−0.083	0.162	2.843	2.840	**2.838**	2.844	2.841	**2.839**
8	1,000	−0.056	0.201	2.846	2.844	**2.842**	2.846	2.844	**2.841**
	5,000	−0.056	0.201	2.844	2.842	**2.838**	2.845	2.842	**2.839**
9	1,000	−0.028	0.239	2.847	2.846	**2.842**	2.847	2.846	**2.842**
	5,000	−0.028	0.239	2.845	2.843	**2.838**	2.847	2.844	**2.840**
10	1,000	0.000	0.278	2.848	2.846	**2.840**	2.850	2.849	**2.843**
	5,000	0.000	0.278	2.848	2.845	**2.839**	2.848	2.845	**2.839**

Note. For all cases, $\beta_0 = 0$, $\beta_1 = 0.5$, $\sigma_0 = 1$, $\Delta\mu = -0.8$, and $p = 0.1$ for 1,000 replications. n = sample size. In M1, there is no predictive bias because the intercept and slope relating X to Y are the same for both groups. In M2, predictive bias is due to intercepts only. In M3, predictive bias is due to group differences in both intercepts and slopes. Bold values denote the smallest −2LL and AIC values (i.e., better fit). M = model.

Figure 1

College Admissions Study Results: Relative Frequency of Different Types of Predictive Bias for Black Versus White (Top Panel) and Latinx Versus White (Bottom Panel) Comparison



Note. M1 includes the composite C (i.e., no predictive bias per Equation 2), M2 includes C and the grouping variable M (i.e., intercept-based predictive bias per Equation 3), and M3 includes C, M, and the interaction term CM (i.e., intercept-based and slope-based predictive bias based on Equation 4). AIC = Akaike information criterion. The diagonal reference line indicates where $AIC_{M2} - AIC_{M1} = AIC_{M3} - AIC_{M1}$. Every point is below the reference line, suggesting that M3 provides the best fit. M = model.

bias) to M1 (i.e., no predictive bias) and M3 (i.e., intercept-based and slope-based predictive bias) to M1 and negative (positive) values provide evidence against (in favor of) M1. Consequently, the two panels in Figure 1 provide detailed visualizations regarding the nature of prediction bias across institution cohorts in the College Board data set based on inferential fit indexes (rather than merely descriptive statistics). For instance, Figure 1's top panel shows that both M2 and M3 have a better fit than M1 (i.e., $AIC_{M2} - AIC_{M1} < 0$ and $AIC_{M3} - AIC_{M1} < 0$) across all institution cohorts for the BW comparison. In other words, the model that includes both slope- and intercept-based predictive bias outperforms the model with no predictive bias.

The panels in Figure 1 also provide information about the relative fit of M2 versus M3. We included a diagonal, 45-degree line in Figure 1 to indicate where M2 and M3 showed similar improvement relative to M1 (i.e., the point where $AIC_{M2} - AIC_{M1} = AIC_{M3} - AIC_{M1}$). Accordingly, points plotted below the reference line correspond with institutions where M3 (i.e., predictive bias based on both slopes and intercepts) outperforms M2 (i.e., predictive bias based on intercepts only). Figure 1's top panel shows that M3 is the best-fitting model for the BW comparison in every institution cohort. Results in Figure 1's bottom panel show a similar pattern for the LW comparison. Specifically, M2 and M3 improve fit relative to M1 (i.e., no predictive bias), but M3 best fits the BW and LW comparisons.

Overall, results indicate that AIC for M3 was the smallest (i.e., best fit) for all 247 BW and 264 LW institution cohorts. This provides evidence that a model with slope- and intercept-based predictive bias resulted in the best fit. Furthermore, the AIC for M1 was the largest for all comparisons, which suggests limited evidence for the absence of predictive bias. Next, we describe the likely reasons for the two types of predictive bias.

Predictive Bias: Understanding Statistical Causes and Likely Underlying Psychological and Contextual Mechanisms

Statistical Causes

In this section, we unpack statistical causes for various types of predictive bias. This material is necessarily technical. The following section links the statistical causes with substantive psychological and contextual mechanisms.

We begin by revisiting the classical test theory model for the predictor variable (i.e., test scores). More specifically, we consider the case where the observed test score (or composite) is the sum of the true achievement score (i.e., true scores [T]) and measurement error (i.e., E),

$$X = T + E. \tag{5}$$

The typical assumption is that T and E are independent, indicating no systematic relationship between true scores and errors. Therefore, the observed variance for X is the sum of variances for the true scores and errors, $\sigma_X^2 = \sigma^2 + \sigma_E^2$. Furthermore, the observed mean is $\mu_X = \mu + \mu_E$, and it is generally assumed that the mean of the errors is zero (i.e., $\mu_E = 0$), so that over repeated measurements $\mu_X = \mu$.

The prediction model assumes that the criterion (i.e., Y) relates to true scores. We consider a general setting involving the following simple regression model,

$$Y = \beta_0 + \beta_1 T + e, \tag{6}$$

where β_0 and β_1 are regression coefficients and e is a prediction error with variance σ_e^2. A few additional details about the model in Equation 6 should be stated. First, we expect that true scores T positively relate to the criterion (i.e., $\beta_1 > 0$), although in some contexts a predictor (e.g., dark triad of personality) may be negatively related to the criterion without loss of generality. Second, T and e are not necessarily independent. True scores are more likely to relate to prediction errors because of the well-known omitted variable problem (Sackett et al., 2003). More specifically, e in Equation 6 includes all other variables that affect the criterion other than the true scores. For instance, in the context of racioethnic comparisons in college admissions testing, where T represents the true score for academic readiness and Y is college grades, e would include variables such as the opportunity to learn, first-generation

college status, levels of academic advising received, and college course rigor. We should therefore expect true scores to relate to other variables that also affect Y. In general, we let the covariance between T and e be denoted by σ_{Te}.

In high-stakes testing situations, T is unobserved, so criterion predictions are based upon the observed scores, X, using a model such as

$$Y = b_0 + b_1 X + \varepsilon. \tag{7}$$

The parameters in Equation 7 are estimated using least-squares regression, and results are often interpreted as proxies for the coefficients in Equation 6. Recall that the usual least-squares estimators for Equation 7 are:

$$\hat{b}_1 = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}, \tag{8}$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}, \tag{9}$$

where $\bar{X}$ and $\bar{Y}$ are the predictor and criterion means, $\hat{\sigma}_X^2$ is the sample estimate of σ_X^2, and $\hat{\sigma}_{XY}$ is the sample estimate of the covariance between X and Y.

We can define the least-squares estimators as functions of the true score parameters in Equation 6. For instance, the population covariance between the criterion and observed scores is

$$\sigma_{XY} = \beta_1 \sigma^2 + \sigma_{Te}, \tag{10}$$

because T and e possibly covary, but the measurement errors E and prediction errors e are assumed to be independent. Recall that $\sigma^2 = \sigma_X^2 r_{xx}$ where $r_{xx} = \sigma^2/\sigma_X^2$ is the reliability estimate (i.e., the share of true score variance in X), so $\sigma_{XY} = \beta_1 \sigma_X^2 r_{xx} + \sigma_{Te}$. The least-squares estimates are functions of sample means (i.e., $\bar{X}$, $\bar{Y}$, $\hat{\sigma}_X^2$, and $\hat{\sigma}_{XY}$). It is well known that sample means are likely to be close to the corresponding population means as the sample size increases, and it is customary to denote convergence in probability as:

$$\begin{aligned} \bar{X} &\xrightarrow{P} \mu_X \\ \bar{Y} &\xrightarrow{P} \mu_Y \\ \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2} &\xrightarrow{P} \frac{\sigma_{XY}}{\sigma_X^2}. \end{aligned} \tag{11}$$

Note that the model for X in Equation 5 implies that $\mu_X = \mu$. Therefore, the least-squares estimator for a given group converges in probability such that $\hat{b}_0$ and $\hat{b}_1$ are likely close to b_0 and b_1, which are the population parameters defined as:

$$\begin{aligned} b_1 &= \beta_1 r_{xx} + \frac{\sigma_{Te}}{\sigma_X^2} \\ b_0 &= \mu_Y - b_1 \mu. \end{aligned} \tag{12}$$

The relation between X and Y (i.e., b_1) is a function of the true score–criterion relationship β_1, the observed score reliability r_{xx}, the covariance ratio between T and e, and the observed score variance. Note that the ratio σ_{Te}/σ_X^2 corresponds with the share of the relationship between X and Y attributed to the association between true scores and omitted variables. We, therefore, refer to that share as $\nu = \sigma_{Te}/\sigma_X^2$. Equation 12 shows that larger values of β_1 produce larger b_1, reliability tends to attenuate the relationship, and associations between true scores and omitted variables can have a positive or negative impact on b_1 depending upon the sign of ν. Furthermore, the intercept in Equation 12 is a function of the criterion mean μ_Y, the population slope b_1, and the true score mean μ. In the case where X and Y are positively related, the intercept equals larger values with increases in μ_Y. Note that Equation 6 implies that $\mu_Y = \beta_0 + \beta_1 \mu$, so we can express b_0 as:

$$b_0 = \beta_0 + (\beta_1 - b_1)\mu. \tag{13}$$

Consequently, the role of μ on the value of b_0 depends upon the sign and magnitude of the differences between β_1 and b_1.

In predictive bias research, we are specifically interested in understanding the extent to which the prediction equations are invariant across groups. We, therefore, add group indices to the parameters above to distinguish between groups. The values for group $g = 0, 1$ are:

$$\begin{aligned} b_{1g} &= \beta_{1g} r_{xx,g} + \nu_g \\ b_{0g} &= \beta_{0g} + (\beta_{1g} - b_{1g})\mu_g. \end{aligned} \tag{14}$$

We let group differences in intercepts between the focal and reference group be $\Delta b_0 = b_{01} - b_{00}$ and group slope differences be $\Delta b_1 = b_{11} - b_{10}$. Consequently, the group differences in slopes and intercepts are:

$$\begin{aligned} \Delta b_1 &= (\beta_{11} r_{xx,1} - \beta_{10} r_{xx,0}) + \Delta\nu \\ \Delta b_0 &= \Delta\beta_0 + (\beta_{11} - b_{11})\mu_1 - (\beta_{10} - b_{10})\mu_0, \end{aligned} \tag{15}$$

where $\Delta\nu = \nu_1 - \nu_0$ and $\Delta\beta_0 = \beta_{01} - \beta_{00}$ are the differences in omitted variable effects and latent prediction intercepts. Note that Equation 14 implies that $\beta_{1g} - b_{1g} = \beta_{1g}(1 - r_{xx,g}) - \nu_g$, so we can update the expression for Δb_0 in Equation 15 to:

$$\begin{aligned} \Delta b_0 = \Delta\beta_0 &+ (\beta_{11}(1 - r_{xx,1}) \\ &- \nu_1)\mu_1 - (\beta_{10}(1 - r_{xx,0}) - \nu_0)\mu_0. \end{aligned} \tag{16}$$

We next discuss the values of the model parameters related to slope- and intercept-based predictive bias. As a preview, the left columns in Tables 2 and 3 include a summary of statistical causes for slope-based predictive bias and intercept-based predictive bias, respectively.

Statistical Factors Causing Slope-Based Predictive Bias

Equation 15 shows that slope-based predictive bias is due to group differences in latent true score–criterion slopes (i.e., $\Delta\beta_1 = \beta_{11} - \beta_{10}$), predictor reliability estimates (i.e., $\Delta r_{xx} = r_{xx,1} - r_{xx,0}$), and omitted variable effects (i.e., $\Delta\nu$). Note that the interpretation of $\Delta b_1 > 0$ (i.e., underprediction of the focal group) is symmetric with $\Delta b_1 < 0$ (i.e., overprediction of the focal group), and opposite interpretations are made by simply reversing the definition of focal and reference groups.

Consider the situation when the focal group has a smaller observed slope so that $\Delta b_1 < 0$, which, when holding intercept differences constant, has the effect of creating overprediction for focal group members for some range of values for X. Focal group slopes are smaller whenever $(\beta_{11} r_{xx,1} - \beta_{10} r_{xx,0}) + \Delta\nu < 0$.

Table 2

Summary of Statistical Causes and Likely Underlying Psychological and Contextual Mechanisms for Slope-Based Predictive Bias and Lack of Differential Prediction Generalization (i.e., Different Types of Predictive Bias Across Contexts)

Statistical causes	Likely psychological and contextual mechanisms
• Differences in latent true score–criterion slopes (i.e., $\Delta\beta_1 = \beta_{11} - \beta_{10}$), predictor reliability estimates (i.e., $\Delta r_{xx} = r_{xx,1} - r_{xx,0}$), and omitted variable effects (i.e., $\Delta\nu$; see Equation 15). • Focal group slopes are smaller whenever $(\beta_{11}r_{xx,1} - \beta_{10}r_{xx,0}) + \Delta\nu < 0$. Accordingly, the observed slopes for the focal group are smaller than the reference group whenever the values of $\Delta\beta_1$, Δr_{xx}, and $\Delta\nu$ combine so that $\Delta b_1 < 0$. • Holding Δr_{xx} and $\Delta\nu$ constant implies that $\Delta\beta_1 < 0$, which results in smaller observed slopes for the focal group. • Two variables affect $\Delta\nu$: (a) the group covariance between true scores and prediction errors (i.e., $\sigma_{Te,g}$) and (b) the group predictor variances (i.e., $\sigma^2_{X,g}$). It is also possible to find that $\Delta\nu < 0$ for more complicated patterns of values for one group's true score–prediction error covariances and predictor variances. For instance, perhaps the focal group is more heterogeneous so that $\sigma^2_{X,1} < \sigma^2_{X,0}$. In this case, it is possible for $\Delta\nu < 0$ if $\sigma_{Te,1}$ is even smaller than $\sigma_{Te,0}$ (see Equation 17).	• Lack of common cultural frame of reference and identity across groups introduces measurement error, and therefore $\Delta r_{xx} \neq 0$, leading to different degrees of slope-based predictive bias across contexts. • Differential recruiting, mentoring, and retention interventions across groups: DEI efforts result in different values for omitted variables across groups (i.e., $\Delta\nu \neq 0$) resulting in different levels of slope-based predictive bias across contexts. • Differential course difficulty across groups results in $\Delta\mu_Y \neq 0$, $\Delta\beta_1 \neq 0$, and $\Delta\nu \neq 0$, leading to different levels of slope-based predictive bias across contexts. • Additional mechanisms include differences across groups regarding course selection, dropout rates, perceived interest, admissions procedures and criteria, environmental opportunities, threats, stressors, and daily experiences of members of different racioethnic groups. These and other idiosyncratic processes and decisions affect $\Delta\beta_1$, Δr_{xx}, $\Delta\nu$, $\Delta\beta_0$, $\Delta\mu$, and $\Delta\mu_Y$, resulting in different levels of slope-based predictive bias across contexts.

Note. DEI = diversity, equity, and inclusion.

Accordingly, the observed slopes for the focal group are smaller than the reference group whenever the values of $\Delta\beta_1$, Δr_{xx}, and $\Delta\nu$ combine so that $\Delta b_1 < 0$. For instance, holding $\Delta\beta_1$ and $\Delta\nu$ constant, we see that $\Delta r_{xx} < 0$ causes smaller slopes for the focal group.

Dahlke and Sackett (2022) reported that $\Delta r_{xx} < 0$ when comparing the reliability coefficients of Black and White students. Furthermore, group differences in the relationship between the criterion and true scores impact the observed scores. Holding Δr_{xx} and $\Delta\nu$ constant implies that $\Delta\beta_1 < 0$ also translates into smaller observed slopes for the focal group.

Finally, the role of omitted variables also shapes the extent to which $\Delta b_1 < 0$. Specifically, recall that the differential omitted variable effects variable $\Delta\nu$ is

$$\Delta\nu = \frac{\sigma_{Te,1}}{\sigma^2_{X,1}} - \frac{\sigma_{Te,0}}{\sigma^2_{X,0}}. \quad (17)$$

Equation 17 shows that two variables affect $\Delta\nu$: (a) the group covariance between true scores and prediction errors (i.e., $\sigma_{Te,g}$) and (b) the group predictor variances (i.e., $\sigma^2_{X,g}$). If the group covariances between true scores and prediction errors are equal (i.e., $\sigma_{Te,1} = \sigma_{Te,0}$; i.e., the omitted variables play a similar role across groups), then $\Delta\nu < 0$ whenever the focal group has a larger observed predictor variance (i.e., $\sigma^2_{X,1} > \sigma^2_{X,0}$). If the predictor variances are equal, then $\Delta\nu < 0$ if $\sigma_{Te,1} < \sigma_{Te,0}$. It is also possible to find that $\Delta\nu < 0$ for more complicated patterns of values for one group's true score–prediction error covariances and predictor variances. For instance, perhaps the focal group is more heterogeneous so that $\sigma^2_{X,1} < \sigma^2_{X,0}$. In this case, it is possible for $\Delta\nu < 0$ if $\sigma_{Te,1}$ is smaller than $\sigma_{Te,0}$.

Statistical Factors Causing Intercept-Based Predictive Bias

Factors causing $\Delta b_0 > 0$ (i.e., underprediction of the focal group) have symmetrical effects compared to those resulting in $\Delta b_0 < 0$ (i.e., overprediction of the focal group) if we reverse the definition of focal and reference groups. Considering the case when the focal group has a smaller intercept with $\Delta b_0 < 0$, the focal group intercept is smaller than the reference group whenever

$$\Delta\beta_0 + (\beta_{11}(1 - r_{xx,1}) - \nu_1)\mu_1 - (\beta_{10}(1 - r_{xx,0}) - \nu_0)\mu_0 < 0. \quad (18)$$

After holding other factors constant, we can see that $\Delta\beta_0 < 0$ and $\Delta\mu < 0$ produce focal group overprediction when $\Delta b_0 < 0$. Equation 18 shows several ways for the underlying parameters to combine to produce smaller focal group intercepts because $\Delta\beta_0$, $\Delta\beta_1$, Δr_{xx}, and $\Delta\nu$ affect Δb_0. Moreover, the criterion mean for a given group g is $\mu_{Yg} = \beta_{0g} + \beta_{1g}\mu_g$. Accordingly, Equation 18 also shows that group differences in β_{0g}, β_{1g}, or μ_g also result in intercept-based predictive bias.

Likely Underlying Psychological and Contextual Mechanisms

In summary, statistical factors that cause slope-based predictive bias are $\Delta\beta_1 \neq 0$, $\Delta r_{xx} \neq 0$, and $\Delta\nu \neq 0$. Factors that cause intercept-based bias are $\Delta\beta_0 \neq 0$, $\Delta\mu \neq 0$, and $\Delta\mu_Y \neq 0$ (which is a direct function of $\Delta\mu_0$ and $\Delta\mu$), in addition to factors that result in slope-based predictive bias as well (i.e., $\Delta\beta_1 \neq 0$, $\Delta r_{xx} \neq 0$, and $\Delta\nu \neq 0$). In this section, we link statistical causes to likely underlying psychological and contextual mechanisms based on existing research, as summarized in Table 2 (slope-based predictive bias) and Table 3 (intercept-based predictive bias).

Lack of Common Cultural Frame of Reference and Identity Across Groups

Members of different groups based on racioethnic and other demographic classifications do not share a common cultural frame of reference and identity (Ogbu, 1993). Specifically, members of underrepresented groups often interpret discrimination against them as permanent and institutionalized, which drives their attitudes and behaviors. For example, cultural frames of reference affect how tests

Table 3

Summary of Statistical Causes and Likely Underlying Psychological and Contextual Mechanisms for Intercept-Based Predictive Bias and Lack of Differential Prediction Generalization (i.e., Different Types of Predictive Bias Across Contexts)

Statistical causes	Likely psychological and contextual mechanisms
• Holding other factors constant, $\Delta\beta_0 < 0$ and $\Delta\mu < 0$ will tend to produce focal group overprediction where $\Delta b_0 < 0$. • There are many possible ways for the underlying parameters to combine to produce smaller focal group intercepts because $\Delta\beta_0$, $\Delta\beta_1$, Δr_{xx}, and $\Delta\nu$ affect Δb_0 (see Equation 18). • The criterion mean for a given group g is $\mu_{Yg} = \beta_{0g} + \beta_{1g}\mu_g$. Accordingly, group differences in β_{0g}, β_{1g}, or μ_g will also result in intercept-based predictive bias (see Equation 18).	• Differential recruiting, mentoring, and retention interventions across groups: Different levels of DEI intensity and resource allocation are likely to result in different values for omitted variables across groups (i.e., $\Delta\nu \neq 0$) resulting in different levels of intercept-based predictive bias across contexts. • Differential course difficulty across groups results in $\Delta\mu_Y \neq 0$, $\Delta\beta_1 \neq 0$, and $\Delta\nu \neq 0$, leading to different levels of intercept-based predictive bias across contexts. • Differential grading leniency across racioethnic groups causes differences in criterion means (i.e., $\Delta\mu_Y \neq 0$), which results in intercept-based predictive bias. • Differential mean test scores across racioethnic groups are due to true differences in latent scores, construct-irrelevant variance due to language proficiency, and measurement bias, which contribute to $\Delta\mu \neq 0$ leading to different levels of intercept-based predictive bias across contexts. • Additional mechanisms include differences across groups regarding course selection, dropout rates, perceived interest, admissions procedures and criteria, environmental opportunities, threats, stressors, and daily experiences of members of different racioethnic groups. These and other idiosyncratic processes and decisions affect $\Delta\beta_1$, Δr_{xx}, $\Delta\nu$, $\Delta\beta_0$, $\Delta\mu$ and/or $\Delta\mu_Y$, resulting in different levels of intercept-based predictive bias across contexts.

Note. DEI = diversity, equity, and inclusion.

and testing situations are interpreted. Stated differently, there are differences across racioethnic groups regarding how members interpret the meaning of test scores and the relation between test scores and performance measures (Grubb & Ollendick, 1986). For instance, some members of underrepresented groups likely have lower expectations about the likelihood that obtaining good test scores will lead to desirable outcomes such as admission to college (Gould, 1999). This mindset and frame of reference develop over long periods, and reasons for its formation include exclusion, segregation, and barriers to opportunities—actual and perceived. As summarized by Awad et al. (2016), "The skills and cognitive competencies measured in assessments most often reflect instruction received by individuals developing in the dominant culture, regardless of whether the test is norm or criterion referenced" (p. 289).

An additional illustration in a slightly different context is a study based on about 110,000 students from six large, public, research-intensive universities (Hatfield et al., 2022). Hatfield et al. reported that "White male students have the highest probability of graduating with a STEM degree when they start college with that *intention* at 48.4% [emphasis added]; however, underrepresented female students only have a probability of 35.3%" (p. 9). They stated,

> In an equitable education system, students with comparable high school preparation, intent to study STEM, and who get Cs or better in all their introductory STEM courses ought to have similar probabilities of attaining a STEM degree. This is not what we observe. (p. 8)

Taken together, the expectations above likely lower r_{xx} for the focal group (i.e., more measurement error). The result is that $\Delta r_{xx} \neq 0$, which leads to slope-based predictive bias.

Differential Recruiting, Mentoring, and Retention Interventions Across Groups

Many organizations actively engage in DEI initiatives. In many cases, these involve extra efforts and resources to recruit, mentor, and retain members of underrepresented racioethnic groups. For example, many colleges and universities offer precollege programs for high-school students (i.e., potential applicants) who are members of underrepresented groups. Subsequently, attendees at these programs are actively recruited as future students. Similarly, many consulting firms such as BCG and McKinsey offer summer programs for college students who are members of underrepresented groups as a conduit to offer them internships and full-time positions eventually. These and related recruiting interventions will likely become even more common since *Students for Fair Admissions v. Harvard* (2023) because universities can no longer make racioethnic-conscious admission decisions (Lu, 2023). Consequently, they are likely to implement other DEI processes involving outreach and recruiting to maintain diversity in the student body.

In some cases, DEI initiatives also involve extra tutoring and counseling opportunities before and after admitting students (Berry et al., 2013). According to Berry et al. (2013), these and other DEI initiatives mean that students' admission into and success in college can be a function of factors other than test scores. All of these unmeasured factors can be conceptualized as omitted variables. Based on our previous discussion, $\Delta\nu$ is one of the three major factors resulting in slope-based predictive bias. So, different DEI intensity levels and resource allocation likely result in $\Delta\nu \neq 0$ and slope- and intercept-based predictive bias across contexts.

Differential Course Difficulty Across Groups

Young (1990, 1991) provided evidence that when members of different groups choose to take courses that vary in difficulty (i.e., differential course difficulty), this choice may explain differences regarding subsequent GPA scores across groups (i.e., higher GPA resulting from taking easier courses). Accordingly, Young (1990) used item response theory to estimate students' latent abilities based on the level of course difficulty and doing so resulted in adjusted scores for the criterion (i.e., "item response theory-based" criterion scores) that were predicted with greater accuracy.

Thus, differential course difficulty is related to $\Delta\mu_Y \neq 0$ and a resulting intercept-based predictive bias. In addition, differential course difficulty across groups leads to $\Delta\beta_1 \neq 0$, one of the three major factors resulting in slope-based predictive bias because it affects the predictor–criterion relation. Moreover, differential course difficulty across racioethnic groups can also lead to $\Delta\nu \neq 0$, resulting in intercept-based predictive bias. Because the differential course difficulty phenomenon likely affects GPA across majors (e.g., 3.36 in education vs. 2.78 in chemistry; Lindsay, 2022) and colleges, this factor likely results in different types and levels of predictive bias across contexts.

Leniency Effects Favoring One Group Over Another

Another underlying mechanism for predictive bias is also related to factors affecting students' GPA, which is the criterion variable used in predictive bias analysis in college admission decisions. As discussed earlier, group differences in criterion means (i.e., $\Delta\mu_Y \neq 0$) result in intercept-based predictive bias.

Leniency effects occur when graders apply a "shifting standards" model across racioethnic groups and assign students in a specific group higher grades than they deserve based on their academic performance. As noted by Berry et al. (2013), "some graders may grade fairly, some graders may be biased against minority students, and other graders may apply a 'shifting standards' model and give minority students higher grades than deserved" (p. 357). The resulting consistent error variance in members of a specific group results in differences in criterion means. Moreover, the extent and direction of leniency likely vary from institution to institution, resulting in different degrees of predictive bias across contexts (i.e., lack of differential prediction generalization).

Factors Affecting Differential Mean Test Scores Across Groups

There is a documented difference in cognitive ability scores across racioethnic groups ($\Delta\mu \neq 0$), resulting in intercept-based predictive bias. The reasons include a combination of several psychological and contextual factors. For example, a true difference in latent scores, construct-irrelevant variance due to language proficiency (Shewach et al., 2017), and measurement bias (Culpepper et al., 2019). Therefore, each of these factors contributes to $\Delta\mu \neq 0$ and results in intercept-based predictive bias and different levels of bias across contexts.

Additional Mechanisms

Additional mechanisms are likely to result in predictive bias (Kruse, 2016). Specifically, these are related to underlying mechanisms leading to $\Delta\beta_1 \neq 0$, $\Delta r_{xx} \neq 0$, $\Delta\nu \neq 0$, $\Delta\beta_0 \neq 0$, $\Delta\mu \neq 0$, and $\Delta\mu_Y \neq 0$. For example, these include differences across racioethnic groups regarding course selection, dropout rates, perceived interest, perceptions of testing, and admissions procedures and criteria. As noted by Berry and Sackett (2009) regarding criterion scores, "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).

In addition, regarding perceived interest on the part of test takers, Pae (2012) examined gender-based differences, which may also apply to racioethnic differences. Specifically, differences in perceived interest in specific test items across groups explained a significant portion of variance in the magnitude of differential item functioning (DIF). The effect was quite substantial: Every one-unit increase in the gender difference in the examinees' perceived interest in the reading passage produced a 0.55-unit increase in the magnitude of gender DIF. In other words, the more interesting the item was to men compared to women, the greater the DIF effect favoring men, and vice versa (Pae, 2012).

As additional factors, as noted by Newman et al. (2022), differences across racioethnic groups "are the result of differences in the environmental opportunities, threats, stressors, and daily experiences of Black and White Americans (i.e., they are 'maltreatment effects'). Systemic racism has had real consequences in the academic domain" (p. 48). Also, Walpole et al. (2005) reported that, among their top concerns, Latinx and African American urban high school students mention cultural and racial test bias. Unsurprisingly, Fleming (2000) noted that students of color have negative perceptions of standardized tests. The extent to which these factors have a stronger or weaker effect across contexts also likely contributes to the differential impact on both slope- and intercept-based predictive bias.

Discussion

As noted in a recent *American Psychologist* article,

> Any discipline is embedded in its historical zeitgeist, and psychology in the first half of the 20th century reflected the prevailing social views of its most prominent founders, an account largely characterized by white men promulgating the racism and sexism of their times and using "science" to explain away injustice. (Miles & Fassinger, 2021, p. 1235)

This is a strong statement that may or may not apply to predictive bias research. However, we should be open-minded to challenges to "historical" findings. This is particularly relevant given that scientific findings about racioethnic differences in the meaning and functioning of standardized testing need to be contextualized within a historical background. For example, consider slavery (1619–1865); legal segregation in education, employment, and voting (1865–1960s); the Homestead Act distributing land primarily to White and European immigrant families and creating modern gaps in property ownership and access to education (1862–1930s), and post–World War II legislation that specifically excluded farm workers and maids from minimum wage protections, work hour regulations, and unions at a time when 60%–75% of the African American labor force were farm workers and maids (Newman et al., 2007). As Newman et al. (2007, p. 1082) noted, these events "greatly benefited Whites as a group, permitting the establishment of

a large White middle class capable of intergenerational transmission of wealth through access to capital and education."

Given undeniable racioethnic historical differences in privileges and rights, the absence of predictive bias seems counterintuitive. But, as Kehoe (2002, p. 104) noted, "a critical part of the dilemma is that general mental abilities-based tests are generally regarded as unbiased." A conclusion that predictive bias does not exist is comfortable because adverse impact against members of underrepresented groups is a more defensible position (Ployhart et al., 2017). Similarly, a conclusion that predictive bias exists, but it is small and favors underrepresented applicants through the overprediction of their future performance, is "generally not viewed as a problematic finding" (Dahlke & Sackett, 2022, p. 2007). So, it is important to understand what specific type of predictive bias may exist and when. This is a relevant issue not only for test users but also for test developers. If predictive bias exists, depending on its nature, it may be necessary to redesign those tests—or discontinue their use altogether, as about 1,900 colleges and universities have decided to do by embracing test-optional policies (FairTest, 2023).

Zwick (2019) wrote that "Our role as measurement specialists should not be to defend tests at all costs. Instead, we should be the judicious evaluators of tests and their applications" (p. 39). In other words,

> The question asked should be about the moral and professional responsibility for fairness of those involved in all stages of the assessment process. The public has little patience when one state organization passes the blame to another when something goes wrong. As Cronbach observed back in the 1980s, those who validate tests either before or after the tests are taken have a responsibility to "review whether a practice has appropriate consequences for individuals and institutions, and particularly to argue against adverse consequences" (Cronbach, 1988, p. 3). (Nisbet & Shaw, 2019, p. 623)

I–O psychology has much to contribute to research-based evidence to inform practices and policies that play an important societal role. As a conduit to hopefully help achieve this lofty goal, we discuss the implications of our results for theory and practice, limitations, and future research directions.

Implications for Theory and Practice

An important implication of our results is that predictive bias can be based on differences regarding slopes, intercepts, or both. We showed that factors that cause slope-based predictive bias are $\Delta\beta_1 \neq 0$, $\Delta r_{xx} \neq 0$, and $\Delta\nu \neq 0$. We also showed that factors that cause intercept-based bias are $\Delta\beta_0 \neq 0$, $\Delta\mu \neq 0$, and $\Delta\mu_Y \neq 0$ (which is a direct function of $\Delta\beta_0$, $\Delta\beta_1$, and $\Delta\mu$), in addition to factors that result in slope-based predictive bias as well (i.e., $\Delta\beta_1 \neq 0$, $\Delta r_{xx} \neq 0$, and $\Delta\nu \neq 0$). In terms of likely underlying psychological and contextual factors, we offered specific explanations mapped onto the statistical causes based on the existing literature: (a) lack of common cultural frame of reference and identity across groups; (b) differential recruiting, mentoring, and retention interventions across groups; (c) differential course difficulty across groups; (d) leniency effects favoring one group over another; (e) factors affecting differential mean test scores across groups; and (f) additional mechanisms (e.g., differences across groups regarding course selection, dropout rates, perceived interest, perceptions of testing, admissions procedures, and criteria; differences in the environmental opportunities, threats, stressors, and daily experiences of members of different racioethnic groups).

Regarding implications for practice, for a college or job applicant whose performance has been underpredicted due to their racioethnic status and, consequently, their application has been rejected, it is no consolation that the overall size of the difference between regression lines may not be perceived as "very large." As an illustration of such consequences, in a recently published introspective article in which scholars of color explored bias in academe, the lead author described experiences with standardized testing and revealed that "As a high school valedictorian and honors college graduate, I had never faced an educational setback like this before and ostensibly the biggest barrier was my quantitative GRE test score" (Holmes et al., 2022, p. 2). Predictive bias is of scientific and practical importance, and very personal as well.

Also, regarding practical implications, 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" (AERA et al., 2014, p. 49). Accordingly, predictive bias should be of concern regardless of which racioethnic group (e.g., Black, White, Latinx) is under or overpredicted. The reason is that

> 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. (AERA et al., 2014, p. 50)

Limitations and Future Research Directions

As mentioned earlier, some of the underlying mechanisms may not be due to testing but to historical and societal issues. As such, testing may reflect, rather than distort, a societal reality. Nevertheless, addressing these issues is particularly pressing, given that Jonson et al. (2019) documented a troubling science–practice misalignment. They reviewed 18 intelligence and achievement tests and found that nine of the 17 fairness *Standards* were practiced only rarely, four were practiced occasionally, and only four were practiced frequently. So, very few of the 17 fairness *Standards* are practiced frequently, and just under half occurred frequently or occasionally. So, there is a persistent science–practice gap that must be addressed (Cascio & Aguinis, 2008).

Our article is just a beginning of a journey. Much work remains to be done to advance our knowledge on why predictive bias exists and varies across contexts—instead of just focusing on whether predictive bias is present or absent. The starting point may be examining the relative importance of the underlying psychological and contextual process we described earlier. However, other factors not yet examined are likely to emerge. Overall, such future research would benefit from adopting a methodological framework that simultaneously considers measurement and prediction invariance.

Specifically, the established Cleary approach for assessing predictive bias based on multiple regression requires additional investigation, given that it assumes the absence of measurement error, which is virtually always violated. In predictive bias assessment, measurement error exists and differs across groups in

many situations. There is preliminary evidence that this dual measurement–prediction invariance methodological framework can lead to novel insights that inform theory and practice. For example, Culpepper et al. (2019) found that nearly a quarter of the statistically significant observed intercept differences were not statistically significant at the .05 level once predictor measurement error was accounted for. Moreover, measurement invariance was rejected for the SAT-M subtest at the .01 level for 74.5% and 29.9% of cohorts for Black versus White and Hispanic versus White comparisons, respectively. In addition, Black students with the same standing on a common factor had observed SAT-M scores that were nearly a third of a standard deviation lower than for comparable Whites. In sum, we suggest that future research on predictive bias considers measurement and predictive bias simultaneously to assess the causal mechanisms including (a) psychological and contextual issues leading to (b) differences in latent predictor scores that in turn affect the (c) nature and size of observed predictive bias.

Conclusions

In 1989, the U.S. Congress held a 1-day hearing to assess the level and effects of bias based on race and gender differences affecting standardized tests (Sex and Race Differences on Standardized Tests, 1989). In this hearing, the focus was on examining the role of standardized tests concerning educational and employment opportunities for women and members of underrepresented racioethnic groups. The hearing included testimony and statements from 14 witnesses. Sadly, none had an I–O psychology background. However, we hope the situation will change in the future. Our goal was to go beyond the existing predictive bias literature mainly focused on a technical debate of which index to use to learn whether predictive bias exists. We offered analytical and empirical evidence about the existence of different types of predictive bias and described statistical causes for the presence of each. In addition, we mapped likely underlying psychological and contextual mechanisms onto the statistical causes based on the existing literature hoping that our article will help advance this research stream from *whether* it exists to the *why* of predictive bias. Overall, we hope our article will stimulate future research and interventions regarding predictive bias, given that this topic offers a unique and important opportunity for the field of I–O psychology to play a leadership role and make important contributions to society. As Hall (1917, p. 11) noted in the first issue of the *Journal of Applied Psychology*, following this path would "show that applied psychology can render the most valuable service."

References

Aguinis, H. (2004). *Regression analysis for categorical moderators*. Guilford Press.

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology*, *95*(4), 648–680. https://doi.org/10.1037/a0018714

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2016). Differential prediction generalization in college admissions testing. *Journal of Educational Psychology*, *108*(7), 1045–1059. https://doi.org/10.1037/edu0000104

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage Publications.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

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

Awad, G. H., Patall, E. A., Rackley, K. R., & Reilly, E. D. (2016). Recommendations for culturally sensitive research methods. *Journal of Educational & Psychological Consultation*, *26*(3), 283–303. https://doi.org/10.1080/10474412.2015.1046600

Berry, C. M., & Sackett, P. R. (2009). Individual differences in course choice result in underestimation of the validity of college admissions systems. *Psychological Science*, *20*(7), 822–830. https://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*(3), 345–359. https://doi.org/10.1007/s10869-012-9284-3

Berry, C. M., Zhao, P., Batarse, J. C., & Reddock, C. (2020). Revisiting predictive bias of cognitive ability tests against Hispanic American job applicants. *Personnel Psychology*, *73*(3), 517–542. https://doi.org/10.1111/peps.12378

Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research and Evaluation*, *19*(2–3), 104–120. https://doi.org/10.1080/13803611.2013.767602

Cascio, W. F., & Aguinis, H. (2008). Research in industrial and organizational psychology from 1963 to 2007: Changes, choices, and trends. *Journal of Applied Psychology*, *93*(5), 1062–1081. https://doi.org/10.1037/0021-9010.93.5.1062

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*(2), 115–124. https://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.). Lawrence Erlbaum.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum.

Culpepper, S. A., Aguinis, H., Kern, J. L., & Millsap, R. (2019). High-stakes testing case study: A latent variable approach for assessing measurement and prediction invariance. *Psychometrika*, *84*(1), 285–309. https://doi.org/10.1007/s11336-018-9649-2

Dahlke, J. A., & Sackett, P. R. (2022). On the assessment of predictive bias in selection systems with multiple predictors. *Journal of Applied Psychology*, *107*(11), 1995–2012. https://doi.org/10.1037/apl0000996

Davis, J., & Martin, D. B. (2018). Racism, assessment, and instructional practices: Implications for mathematics teachers of African American students. *Journal of Urban Mathematics Education*, *11*(1–2), 45–68. https://doi.org/10.21423/jume-v11i1-2a358

Eby, L. T. (2022). Reflections on the Journal of Applied Psychology in times of change. *Journal of Applied Psychology*, *107*(1), 1–8. https://doi.org/10.1037/apl0001000

FairTest. (2023, July 7). *1,900 accredited, 4-year colleges & universities with ACT/SAT-optional or test-free testing policies for Fall 2023*. https://fairtest.org/test-optional-list/

Fleming, J. (2000). Affirmative action and standardized test scores. *The Journal of Negro Education*, *69*(1–2), 27–37. https://www.jstor.org/stable/2696262

George, G., Howard-Grenville, J., Joshi, A., & Tihanyi, L. (2016). Understanding and tackling societal grand challenges through management research. *Academy of Management Journal*, *59*(6), 1880–1895. https://doi.org/10.5465/amj.2016.4007

Gould, M. (1999). Race and theory: Culture, poverty, and adaptation to discrimination in Wilson and Ogbu. *Sociological Theory*, *17*(2), 171–200. https://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*(4), 399–414. https://doi.org/10.1016/0147-1767(86)90042-8

Hall, G. S. (1917). Practical relations between psychology and the war. *Journal of Applied Psychology*, *1*(1), 9–16. https://doi.org/10.1037/h0070238

Hatfield, N., Brown, N., & Topaz, C. M. (2022). Do introductory courses disproportionately drive minoritized students out of STEM pathways? *PNAS Nexus*, *1*(4), Article pgac167. https://doi.org/10.1093/pnasnexus/pgac167

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, *107*(8), 1323–1351. https://doi.org/10.1037/apl0000695

Holmes, O., IV, Smith, A. N., Loyd, D. L., & Gutiérrez, A. S. (2022). Scholars of color explore bias in academe: Calling in allies and sharing affirmations for us by us. *Organizational Behavior and Human Decision Processes*, *173*, Article 104204. https://doi.org/10.1016/j.obhdp.2022.104204

Jaccard, J. J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression*. Sage Publications.

Jonson, J. L., Trantham, P., & Usher-Tate, B. J. (2019). An evaluative framework for reviewing fairness standards and practices in educational tests. *Educational Measurement: Issues and Practice*, *38*(3), 6–19. https://doi.org/10.1111/emip.12259

Kehoe, J. F. (2002). General mental ability and selection in private sector organizations: A commentary. *Human Performance*, *15*(1–2), 97–106. https://doi.org/10.1080/08959285.2002.9668085

Kruse, A. J. (2016). Cultural bias in testing: A review of literature and implications for music education. *Update: Applications of Research in Music Education*, *35*(1), 23–31. https://doi.org/10.1177/8755123315576212

Landers, R. N., Armstrong, M. B., Collmus, A. B., Mujcic, S., & Blaik, J. (2022). Theory-driven game-based assessment of general cognitive ability: Design theory, measurement, prediction of performance, and test fairness. *Journal of Applied Psychology*, *107*(10), 1655–1677. https://doi.org/10.1037/apl0000954

Lindsay, S. (2022, October 15). *What's the average college GPA, by major?* https://blog.prepscholar.com/average-college-gpa-by-major

Lu, A. (2023, July 6). After Supreme Court ruling, DEI work gets more challenging and crucial, experts say. *The Chronicle of Higher Education*. https://www.chronicle.com/article/after-supreme-court-ruling-dei-work-gets-more-challenging-and-crucial-experts-say

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*(1), 134–147. https://doi.org/10.1037/a0030610

Mendoza, J. L., Bard, D. E., Mumford, M. D., & Ang, S. C. (2004). Criterion-related validity in multiple-hurdle designs: Estimation and bias. *Organizational Research Methods*, *7*(4), 418–441. https://doi.org/10.1177/1094428104268752

Miles, J. R., & Fassinger, R. E. (2021). Creating a public psychology through a scientist–practitioner–advocate training model. *American Psychologist*, *76*(8), 1232–1247. https://doi.org/10.1037/amp0000855

Newman, D. A., Hanges, P. J., & Outtz, J. L. (2007). Racial groups and test fairness, considering history and construct validity. *American Psychologist*, *62*(9), 1082–1083. https://doi.org/10.1037/0003-066X.62.9.1082

Newman, D. A., Tang, C., Song, Q. C., & Wee, S. (2022). Dropping the GRE, keeping the GRE, or GRE-optional admissions? Considering tradeoffs and fairness. *International Journal of Testing*, *22*(1), 43–71. https://doi.org/10.1080/15305058.2021.2019750

Nisbet, I., & Shaw, S. D. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, *26*(5), 612–629. https://doi.org/10.1080/0969594X.2019.1586643

Ogbu, J. U. (1993). Differences in cultural frame of reference. *International Journal of Behavioral Development*, *16*(3), 483–506. https://doi.org/10.1177/016502549301600307

Oh, I.-S., Le, H., & Roth, P. L. (2023). Revisiting Sackett et al.'s (2022) rationale behind their recommendation against correcting for range restriction in concurrent validation studies. *Journal of Applied Psychology*, *108*(8), 1300–1310. https://doi.org/10.1037/apl0001078

Pae, T.-I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing*, *29*(4), 533–554. https://doi.org/10.1177/0265532211434027

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*(4), 253–261. https://doi.org/10.1111/j.1745-3984.1971.tb00934.x

Ployhart, R. E., Schmitt, N., & Tippins, N. T. (2017). Solving the supreme problem: 100 years of selection and recruitment at the Journal of Applied Psychology. *Journal of Applied Psychology*, *102*(3), 291–304. https://doi.org/10.1037/apl0000081

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement*, *23*(2), 99–115. https://doi.org/10.1177/01466219922031220

Rogelberg, S. G., King, E., & Alonso, A. (2022). How we can bring I-O psychology science and evidence-based practices to the public. *Industrial and Organizational Psychology*, *15*(2), 259–272. https://doi.org/10.1017/iop.2021.142

Sackett, P. R., Berry, C. M., Lievens, F., & Zhang, C. (2023). Correcting for range restriction in meta-analysis: A reply to Oh et al. (2023). *Journal of Applied Psychology*, *108*(8), 1311–1315. https://doi.org/10.1037/apl0001116

Sackett, P. R., Laczo, R. M., & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: The omitted variables problem. *Journal of Applied Psychology*, *88*(6), 1046–1056. https://doi.org/10.1037/0021-9010.88.6.1046

Sackett, P. R., Zhang, C., & Berry, C. M. (2023). Challenging conclusions about predictive bias against Hispanic test takers in personnel selection. *Journal of Applied Psychology*, *108*(2), 341–349. https://doi.org/10.1037/apl0000978

Schmitt, N., & Ployhart, R. E. (1999). Estimates of cross-validity for stepwise regression and with predictor selection. *Journal of Applied Psychology*, *84*(1), 50–57. https://doi.org/10.1037/0021-9010.84.1.50

Sex and Race Differences on Standardized Tests. (1989). *Oversight hearings before the subcommittee on civil and constitutional rights of the committee on the judiciary. House of representatives, one hundredth congress, first session*. https://eric.ed.gov/?id=ED312276

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494. https://doi.org/10.1080/01621459.1993.10476299

Shewach, O. R., Shen, W., Sackett, P. R., & Kuncel, N. R. (2017). Differential prediction in the use of the SAT and high school grades in predicting college performance: Joint effects of race and language. *Educational Measurement: Issues and Practice*, *36*(3), 46–57. https://doi.org/10.1111/emip.12150

Society for Industrial and Organizational Psychology. (2018). *Principles for the validation and use of personnel selection procedures*.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 44–47. https://doi.org/10.1111/j.2517-6161.1977.tb01603.x

Students for Fair Admissions v. Harvard (2023, US Supreme Court No. 20–1199). https://www.supremecourt.gov/opinions/22pdf/20-1199_hgdj.pdf

Walpole, M., McDonough, P. M., Bauer, C. J., Gibson, C., Kanyi, K., & Toliver, R. (2005). This test is unfair: Urban African American

and Latino high school students' perceptions of standardized college admission tests. *Urban Education*, *40*(3), 321–349. https://doi.org/10.1177/0042085905274536

Young, J. W. (1990). Adjusting the cumulative GPA using item response theory. *Journal of Educational Measurement*, *27*(2), 175–186. https://doi.org/10.1111/j.1745-3984.1990.tb00741.x

Young, J. W. (1991). Gender bias in predicting college academic performance: A new approach using item response theory. *Journal of Educational Measurement*, *28*(1), 37–47. https://doi.org/10.1111/j.1745-3984.1991.tb00342.x

Zwick, R. (2019). Fairness in measurement and selection: Statistical, philosophical, and public perspectives. *Educational Measurement, Issues and Practice*, *38*(4), 34–41. https://doi.org/10.1111/emip.12299

Received February 12, 2022
Revision received August 14, 2023
Accepted August 16, 2023 ■