Differential Validity and Differential Prediction of Cognitive Ability Tests: Understanding Test Bias in the Employment Context

Christopher M. Berry

Department of Management and Entrepreneurship, Kelley School of Business, Indiana University, Bloomington, Indiana 47405; email: cmberry2@indiana.edu

Annu. Rev. Organ. Psychol. Organ. Behav. 2015. 2:435–63

The *Annual Review of Organizational Psychology and Organizational Behavior* is online at orgpsych.annualreviews.org

This article's doi:
10.1146/annurev-orgpsych-032414-111256

Keywords

test bias, differential validity, differential prediction

Abstract

Substantial mean score differences and significant adverse impact have long motivated the question of whether cognitive ability tests are biased against certain non-White subgroups. This article presents a framework for understanding the interrelated issues of adverse impact and test bias, with particular focus on two forms of test bias especially relevant for personnel selection: differential validity and differential prediction. Ethical and legal reasons that organizations should be concerned about differential validity/prediction are discussed. This article also serves as a critical review of the research literature on differential validity/prediction. The general conclusion is that available evidence supports the existence of differential validity/prediction in the form of correlation/slope and intercept differences between White and non-White subgroups. Implications for individuals and organizations are outlined, and a future research agenda is proposed highlighting the need for new, better data; new, better methods of testing for differential validity/prediction; and investigation of substantive factors causing differential validity/prediction.

INTRODUCTION

Cognitive ability tests are commonly used in organizations' personnel selection procedures because they strongly predict job performance and have considerable utility (Hunter et al. 2006, Schmidt & Hunter 1998, Schmitt 2014). At the same time, it is well known that African American and Hispanic American test takers score substantially lower on cognitive ability tests than White test takers (Roth et al. 2001). Thus, if cognitive ability tests play a significant role in hiring decisions, there will be an adverse impact against these non-White[1] subgroups. Adverse impact refers to differences between subgroups in hiring rates (see **Table 1** for this definition and an example of adverse impact, along with definitions and examples for other technical terms used throughout this article). When cognitive ability tests are used in personnel selection, their adverse impact usually takes the form of African American and Hispanic American applicants being hired at lower rates than White applicants. All of this naturally begs the question of why these non-White subgroups score lower. A particular test fairness concern is whether the tests are biased against non-White test takers. Of course, fairness is a social construct, and perceptions of fairness vary (SIOP 2003). However, most would probably deem it unfair if the reason non-White subgroup members score lower on an employment test is that the employment test is biased against them. Thus, there is great interest in whether cognitive ability tests are biased against non-White test takers, and there has been much disagreement about this issue over the years in organizational psychology and organizational behavior (e.g., Aguinis et al. 2010, Hunter & Schmidt 1978, Katzell & Dyer 1977, Mattern & Patterson 2013).

Adverse impact does not necessarily mean that the test itself is biased. Regardless, adverse impact is troubling in that it means that the use of cognitive ability tests in hiring systematically limits the employment opportunities of non-White subgroup members. Being denied employment for a desired job has profound personal, financial, and social impacts on persons. So, organizations interested in using cognitive ability tests for personnel selection are faced with an ethical dilemma: Are the strong predictive validity and utility of cognitive ability tests worth the negative personal, financial, and social impacts that these tests have on non-White subgroup members? Further, adverse impact is a significant legal liability for organizations in that it can be used as prima facie evidence of discrimination in equal employment opportunity lawsuits. The implication in such cases is that the selection procedure in question may be biased against non-White subgroup members, and it is up to the organization to then demonstrate the test is business related (i.e., related to job performance) and not biased against non-White test takers. So, besides having ethical implications, issues of adverse impact and test bias also play a major role in employment litigation.

This article presents a framework for understanding the interrelated issues of adverse impact and test bias, with a particular focus on two forms of test bias especially relevant to personnel selection: differential validity (subgroup differences in test validities) and differential prediction (subgroup differences in test–criterion regression equations). Additionally, the ethical and legal reasons why organizations should be concerned about differential validity and prediction are reviewed. Further, this article provides a synthesis of the current state of knowledge about differential validity and prediction. It concludes that the available evidence supports the existence of differential validity and prediction for cognitive ability tests but does not support the idea that statistical artifacts account for these differences. Thus, the case is made that future research in the

[1] The term non-White is used a number of times throughout this article to collectively refer to the African American and Hispanic American subgroups. This does not imply that these subgroups are always the same on all dimensions or that there are no other non-White subgroups. Rather, the term is used for the sake of brevity, in recognition of similar patterns of findings for the two subgroups for cognitive ability tests when compared with the White subgroup and in accordance with the American Psychological Association's preference for the term non-White rather than minority.

Table 1 Definitions and examples of technical terms used in this article

Term	Definition	Example[a]
Adverse impact	Differences in hiring rates between subgroups. The hiring rate is the percentage of applicants who were hired; this is also the selection ratio.	If 40 White and 10 non-White applicants were hired from a pool of 100 White and 50 non-White applicants, the White hiring rate would be 40% (40/100 = 0.40), and the non-White hiring rate would be 20% (10/50 = 0.20). The hiring rate for the non-White subgroup is only 50% that for the White subgroup (20%/40% = 0.50), signaling adverse impact against the non-White subgroup.
Differential item functioning	A form of measurement bias wherein individual items are biased such that, when ability is held constant, item difficulty or discrimination is different for subgroups.	When cognitive ability is held constant, 50% of non-Whites get an item correct, whereas 60% of Whites get the item correct. This item functions differently for the two subgroups in that it is more difficult for the non-White subgroup.
Differential prediction	Differences in regression equations between subgroups. This is often used as a synonym for predictive bias.	If the White regression equation resulting from regressing job performance on cognitive ability test scores is $Y' = 1.5 + 0.40X$ and the non-White regression equation is $Y' = 1.0 + 0.30X$, there is differential prediction in that the regression equations differ. In this example, the equations have both different intercepts (1.5 versus 1.0) and different slopes (0.40 versus 0.30), either of which would signal differential prediction.
Differential range restriction	Differences in the amount of range restriction between subgroups.	If the *u*-ratio (see definition below) on a cognitive ability test for the White subgroup were 0.70 and the *u*-ratio for the non-White subgroup were 0.60, the non-White subgroup would be more restricted in range on the cognitive ability test.
Differential validity	Differences in the correlation between the test and criterion (i.e., validity) between subgroups.	If the correlation between a cognitive ability test and job performance were 0.50 for the White subgroup but only 0.25 for the non-White subgroup, this would signal differential validity in that the test is less valid for the non-White subgroup.
Intercept differences	Differences in regression intercepts between subgroups; a form of differential prediction.	In the example for differential prediction above, the subgroups had different regression intercepts (1.5 versus 1.0), meaning that the subgroups' regression lines cross the *x*-axis (cognitive ability score of zero) at different points.

(*Continued*)

Table 1 (*Continued*)

Term	Definition	Example[a]
Measurement bias	When individuals who are identical on the construct measured by the test but who are from different subgroups have different probabilities of attaining the same observed score (Berry et al. 2011). Measurement invariance and differential item functioning (see definitions in this table) analyses are ways to test for two different forms of measurement bias.	If the factor structure of a cognitive ability test were different for White and non-White subgroups, or if members of the two subgroups had different probabilities of getting an item correct despite equal cognitive ability, this would signal measurement bias.
Measurement error	Random error affecting tests or criteria. Measurement error is one type of statistical artifact (see definition of observed validity below).	Differences between the observed cognitive ability test scores and true, latent cognitive ability that are a function of random influences such as idiosyncratic participant responses to particular test items or random changes in the testing environment.
Measurement invariance	An instance in which the factor structure of a test is the same (invariant) for two subgroups; lack of measurement invariance (i.e., differential functioning) is a form of measurement bias.	Imagine that a cognitive ability test measured two factors (verbal and quantitative), but the factor loadings for the verbal items were much higher for the non-White subgroup than for the White subgroup. This would be an instance of a lack of measurement invariance, as would any difference between the subgroups in test factor structure (e.g., different numbers of factors, different relationships between factors, different residuals).
Observed validity	The correlation between a test and a criterion, uncorrected for statistical artifacts.	The correlation between job incumbents' scores on a cognitive ability test and supervisor ratings of job performance would be an observed validity. The cognitive ability test and supervisor ratings are affected by measurement error. They will also commonly be affected by range restriction on the cognitive ability test. Both of these statistical artifacts systematically reduce the correlation between test scores and job performance relative to the correlation between the true constructs of cognitive ability and job performance. The observed validity is the correlation between these measures without attempting correction for these statistical artifacts.
Operational validity	The correlation between a test and a criterion, free from range restriction and measurement error in the criterion (but not the test).	If the observed validity of a cognitive ability test were corrected for range restriction and criterion measurement error (but not cognitive ability test measurement error), this would yield an estimate of the operational validity of the cognitive ability test.

(*Continued*)

Table 1 *(Continued)*

Term	Definition	Example[a]
Over/ underprediction	A form of differential prediction in which one subgroup's regression line relating the test and criterion lies above the other subgroup's regression line, resulting in overprediction of criterion performance for the subgroup with the lower regression line when the common regression line (regression line for both subgroups combined) is used to predict that subgroup's criterion performance.	See **Figure 1*b*** in which the White subgroup's regression line lies above the non-White subgroup's regression line. In this case, the common regression line lies between the two subgroups' regression lines and above the non-White subgroup's regression line. If the common regression line were used to predict criterion performance for the non-White subgroup, this would result in higher predicted criterion performance than if the non-White subgroup's regression line were used to predict criterion performance (overprediction). Similarly, in this example, the White subgroup's criterion performance would be underpredicted by the common regression line relative to the predicted criterion performance that would result from using the White subgroup's regression line to predict the White subgroup's performance.
Predictive bias	An instance in which, "for a given subgroup, consistent nonzero errors of prediction are made for members of the subgroup" (SIOP 2003, p. 32). This is often used as a synonym for differential prediction.	In the over/underprediction example above, using the common regression line to predict criterion performance would result in consistent nonzero errors of prediction (i.e., overprediction or underprediction) for both subgroups. Also, in the slope differences example below, use of the common regression line to predict performance would result in nonzero errors of prediction for each subgroup because each subgroup's regression line has a different slope (meaning the slope of each would also differ from the common pooled regression line). Similarly, in the differential prediction example above, the White and non-White subgroups have different regression equations; using the common pooled regression line equation (which would differ from each of the subgroup regression equations) would result in consistent nonzero errors of prediction.
Range restriction	An instance in which the range of scores on a test is restricted in a sample due to selection of the sample based on test scores or some variable related to test scores. Range restriction is a statistical artifact that systematically reduces the correlation between test and criterion scores.	Imagine a cognitive ability test has a standard deviation of 1.0 in an applicant pool. If only applicants scoring in the top 10% on cognitive ability were hired, then the standard deviation of cognitive ability test scores would be only about 0.39 in the hired incumbent sample. Because of this restricted range of cognitive ability test scores, correlations between the cognitive ability test and job performance would be smaller in the incumbent sample than in the entire applicant pool (if the entire applicant pool had hypothetically been hired).

(Continued)

Table 1 (*Continued*)

Term	Definition	Example[a]
Selection ratio	The percentage of applicants hired. It is also referred to as the hiring rate and is calculated as the number of hired applicants divided by total applicants.	If there were 100 applicants for a job and 40 of them were hired, the selection ratio would be 40%.
Slope differences	Differences in regression slopes between subgroups; a form of differential prediction.	In the differential prediction example above, the slopes of the regression lines relating cognitive ability tests and job performance differ for the White and non-White subgroups (0.40 versus 0.30). This means that a one-unit increase in test score is associated with a larger increase in job performance for the White subgroup than for the non-White subgroup. This also means the cognitive ability test is a better predictor of job performance for the White subgroup.
Test bias	"Any construct-irrelevant source of variance that results in systematically higher or lower scores for identifiable groups of examinees" (SIOP 2003, p. 32). Two particularly relevant forms of test bias are measurement bias and predictive bias.	Put simply, test bias refers to some aspect of the test causing it to work systematically differently across racial/ethnic subgroups. The measurement and predictive bias examples above are examples of instances of test bias.
Test fairness	"Fairness is a social rather than a psychometric concept. Its definition depends on what one considers to be fair. Fairness has no single meaning and, therefore, no single definition, whether statistical, psychometric, or social" (SIOP 2003, p. 31).	One person might deem it unfair that using cognitive ability tests for personnel selection results in lower hiring rates for non-Whites. Another person might deem this unfair only if the differences in hiring rates are a function of some form of bias against non-Whites in the test content.
u-Ratio	An index of range restriction wherein the restricted standard deviation of test scores is divided by the unrestricted standard deviation of test scores. *u*-Ratios less than 1.0 signal range restriction.	In the range restriction example above, the hired incumbent (restricted) standard deviation was 0.39, and the unrestricted applicant pool standard deviation was 1.0. In this case, the *u*-ratio is 0.39 (0.39/1.0 = 0.39).

[a] For all of the examples highlighting differences between subgroups (e.g., adverse impact, differential validity), the reader should assume that the differences are statistically significant.

cognitive ability test bias literature must draw upon newer and better data and look to different substantive explanations for differential validity and prediction.

BOUNDARY CONDITIONS

This review focuses primarily on cognitive ability testing in personnel selection, as the vast majority of test bias research has been related to cognitive ability tests. This is in great part due to cognitive ability tests exhibiting some of the largest and most consistent mean score differences between racial/ethnic subgroups (Roth et al. 2001). Other commonly used personnel selection procedures (e.g., employment interviews, personality tests, situational judgment tests) produce smaller, less consistent or even no mean differences between racial/ethnic subgroups (Bobko et al. 1999, Whetzel et al. 2008), so there has been little to no bias research for such procedures. This review also focuses mostly on African American–White and Hispanic American–White comparisons. This is not meant to imply that test bias is not a potential concern for other racial/ethnic subgroups (e.g., Asian Americans, Native Americans) or other non-race-based subgroups (e.g., men and women). Rather, it again reflects that the largest and most consistent mean score differences on cognitive ability tests have been found between the White subgroup and the African American and Hispanic American subgroups, and therefore the largest test bias research bases exist for these subgroups. Finally, this review focuses most directly on preemployment testing. Cognitive ability tests are also used for high-stakes selection purposes in other domains, such as college admissions and military selection and placement (Berry et al. 2011, Sackett et al. 2001). There are significant similarities between the cognitive ability tests, how they are used, and how they are validated across these three domains (Berry et al. 2011, 2013b). So, the present review draws upon the college admissions and military selection literatures, but only when they provide some insight that is particularly relevant to preemployment testing.

A FRAMEWORK FOR UNDERSTANDING THE INTERRELATED ISSUES OF ADVERSE IMPACT AND TEST BIAS

What follows is a framework for understanding the relationship between adverse impact and test bias. The interested reader may wish to also see Aguinis & Smith (2007), which touches on similar issues. Adverse impact and test bias (in its various forms) are interrelated concepts. Adverse impact is mostly a function of two factors: the mean test score difference between subgroups and the selection ratio (i.e., percentage of the applicant pool selected). All else equal, as the mean difference becomes larger, adverse impact will become greater because fewer applicants in the lower-scoring subgroup will meet the test score standard. Further, given a mean score difference between subgroups and all else equal, adverse impact will become greater as the selection ratio becomes smaller. So, adverse impact is typically more of an issue in highly competitive selection settings (e.g., a firefighter job opening for which there may be thousands of applicants) than in less competitive selection settings with large selection ratios (e.g., an unskilled labor opening). Therefore, adverse impact potential is greatest when the mean subgroup score difference is large (as with cognitive ability tests) and the selection ratio is small (as is typical in high-stakes personnel selection settings) (Sackett et al. 2008).

An important way in which these two factors differ is in the degree to which they are a function of the cognitive ability test. The selection ratio is not a direct function of the cognitive ability test; rather, it is a function of the number of job openings, the size of the applicant pool, and/or decisions about where to place the cut score. The mean difference between subgroups is more directly a function of the cognitive ability test (e.g., using a less cognitively loaded test would likely reduce the size of the mean difference). So, adverse impact inevitably leads to the question, why is there a mean score difference between racial/ethnic subgroups on the cognitive ability test? There are a number of possible answers that have little to do with test bias [e.g., true differences between the subgroups on the attribute(s) that the cognitive ability test measures or contextual/societal influences that contaminate test scores more for one subgroup].

Another possibility is test bias, such that the test itself is biased against the lower-scoring group and, therefore, is the cause of the score difference and adverse impact. A technical definition of test bias is "any construct-irrelevant source of variance that results in systematically higher or lower scores for identifiable groups of examinees" (SIOP 2003, p. 32). Put more simply, test bias refers to some aspect of the test causing it to work systematically differently across racial/ethnic subgroups. Test bias can manifest in two broad forms: measurement bias and predictive bias. Measurement bias refers to cases in which "individuals who are identical on the construct measured by the test. . .but who are from different subgroups have different probabilities of attaining the same observed score" (Berry et al. 2011, p. 883). This generally means that the factor structure of the test differs across subgroups (referred to as lack of measurement invariance in factor-analytic terms) and/or there are individual items for which subgroups have different probabilities of correct answers when latent cognitive ability is held constant (typically referred to as differential item functioning). Berry et al. (2011) reviewed the measurement invariance and differential item functioning literatures for cognitive ability tests and concluded (*a*) there was no consistent evidence that factor structures differed across racial/ethnic subgroups and (*b*) there were typically as many items biased in favor of non-White subgroups as biased against these subgroups, so differential item functioning tended to cancel out at the total test score level (see also Hough et al. 2001 and O'Neill & McPeek 1993 for similar reviews and conclusions about differential item functioning). So, the literature to date suggests that measurement bias is not a major factor contributing to adverse impact. Further, in a personnel selection context, perhaps the most crucial issue is whether test scores predict performance on the job. Measurement bias certainly can affect whether cognitive ability test scores predict job performance equally for each subgroup (e.g., if the items in a test were biased against one subgroup, the test would not measure ability as well, and test scores probably would not predict an unbiased job performance criterion as well, for that subgroup). However, measurement bias does not necessarily translate to subgroup differences in job performance prediction between subgroups (Berry et al. 2011).

The second major form of test bias is predictive bias, which refers to when "for a given subgroup, consistent nonzero errors of prediction are made for members of the subgroup" (SIOP 2003, p. 32). Put simply, predictive bias occurs when cognitive ability test scores do not relate to, or predict, job performance equally for White versus non-White subgroups. Predictive bias has been investigated using two related frameworks: differential validity and prediction. Differential validity refers to differences between subgroups in the validity coefficients (i.e., correlations between cognitive ability tests and job performance) for a cognitive ability test. The predictive bias concern with differential validity is whether the correlation between cognitive ability tests and job performance is lower for one subgroup, which would suggest the test is not as valid for that subgroup. Differential prediction refers to differences between subgroups in the regression equations predicting job performance from cognitive ability test scores; either regression slope or intercept differences signal predictive bias. **Figure 1** is a stylized depiction of differential prediction. The ovals represent the groups of scores for two subgroups: White and non-White (the White subgroup's scores tend to be higher on both the test and criterion). Each subgroup has its own regression line (the solid lines in the figure), and there is also a common regression line (dashed line) representing the regression line for both subgroups pooled together. US employment law requires that the common regression line, rather than the subgroup-specific regression lines, be used for

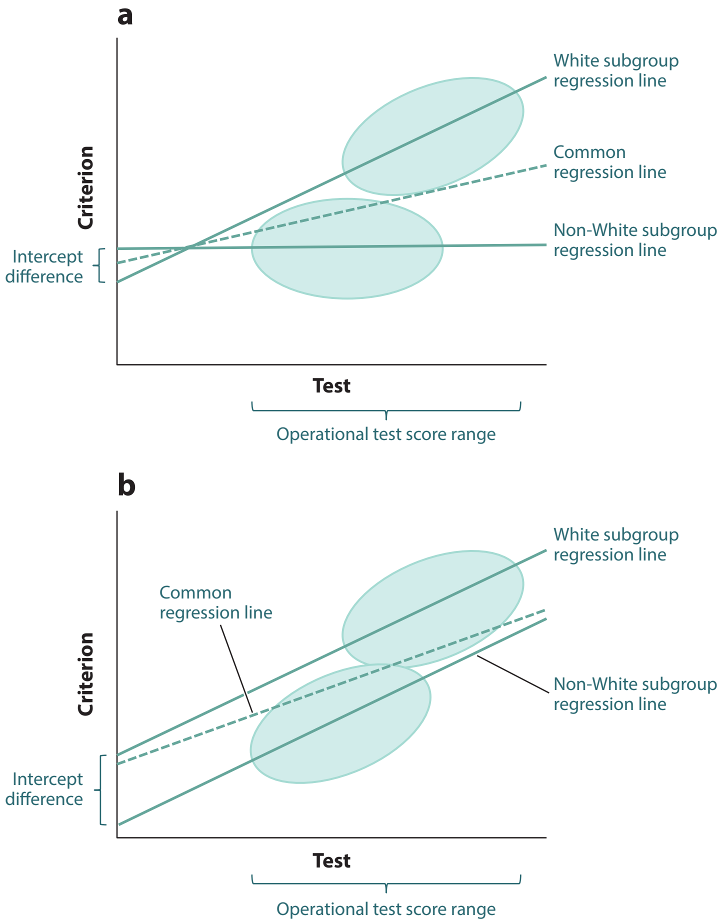
a
White subgroup regression line
Common regression line
Non-White subgroup regression line
Criterion
Intercept difference
Test
Operational test score range
b
White subgroup regression line
Common regression line
Non-White subgroup regression line
Criterion
Intercept difference
Test
Operational test score range

Figure 1

Differential prediction in the form of (*a*) regression slope differences and (*b*) regression intercept differences (no slope differences).

making hiring decisions. Similar to differential validity, regression slope differences mean that the test does not predict job performance as strongly for one subgroup and, therefore, is not as valid a predictor for that subgroup. For example, **Figure 1*a*** displays an instance in which the slope of the regression line for the non-White subgroup is less than that for the White subgroup. If subgroups do not differ in regression slopes, a difference in regression intercepts suggests that one subgroup's regression line lies above the other subgroup's throughout the test score range and, therefore, that job performance is underpredicted for the group with the higher regression line. For example, **Figure 1*b*** displays an instance in which the White and non-White subgroups have equivalent slopes, but the White subgroup has the higher regression intercept. In this example, the common regression line lies between the two subgroups' regression lines but consistently above the non-White subgroup's regression line. Therefore, the common regression line overpredicts job performance for the non-White subgroup compared with the non-White subgroup's regression line (the White subgroup's job performance is underpredicted by the common regression line). Note that if the subgroups have regression lines with different slopes (as in **Figure 1*a***), the regression intercept becomes less meaningful as an indicator of differential prediction, as the slope difference means that the regression lines must intersect at some point. When slopes differ between subgroups, the predictive bias issue is where the regression lines intersect and which subgroup's job performance is underpredicted throughout the operational test score range. For example, the depicted operational test score range in **Figure 1*a*** is the range from the lowest-scoring test taker to the highest-scoring test taker. Although the regression lines cross, it is at a point outside the operational score range, meaning that regardless of the slope differences, the non-White subgroup's regression line lies below that of the White subgroup and, therefore, that the criterion performance of the non-White subgroup is still overpredicted throughout the operational score range.

Differential prediction analyses have typically been carried out using moderated multiple regression (MMR) wherein job performance is regressed on cognitive ability test scores, a race dummy variable, and a test–race interaction term (Aguinis et al. 2010, Lautenschlager & Mendoza 1986). However, a simpler and functionally equivalent method is to regress job performance on cognitive ability test scores separately for each subgroup, generating separate unstandardized regression equations for each subgroup. This separate regression equation method better highlights the similarities and differences between differential validity and prediction. For example, the regression equations for the White and African American subgroups, respectively, are as follows:

$$\hat{Y}_W = b_{0W} + b_{1W}X_W, \tag{1a}$$

$$\hat{Y}_A = b_{0A} + b_{1A}X_A, \tag{1b}$$

where Y is job performance, X is the cognitive ability test, b_0 is the regression intercept, and b_1 is the unstandardized regression slope. The comparable differential validity analysis would only generate separate correlation coefficients for the White and African American subgroups: r_W and r_A, respectively. Both the correlation coefficients (r) and the regression slopes (b) capture the direction and magnitude of the relationship between the cognitive ability test and job performance. The correlation and regression coefficients are related via the following:

$$b_1 = r\left(\frac{s_Y}{s_X}\right), \tag{2}$$

where s_Y and s_X are the standard deviations of job performance and the cognitive ability test, respectively. So, correlations and regression slopes differ as a function of s_Y/s_X. This means that differences between subgroups in correlations/validities will generalize to differences in regression slopes unless the standard deviation ratio differs by a compensatory amount. Equations 1*a* and 1*b* also demonstrate that differential prediction analyses allow for comparisons of subgroups' regression intercepts—information differential validity analyses do not provide.

Thus, differential prediction and differential validity differ (*a*) by the ratio of job performance and test standard deviations and (*b*) in whether they provide information about regression intercept differences. Clearly, differential prediction analyses provide a more complete picture of whether the relationship between cognitive ability tests and job performance differs by subgroup. One way to think about differential validity versus differential prediction is that differential validity will translate into differential prediction in the form of regression slope differences (signaling predictive bias) unless the ratio of job performance to cognitive ability standard deviations differs by a compensatory amount between subgroups. So, both differential validity and prediction provide important information about predictive bias.

To sum up and integrate these concepts, adverse impact and test bias (in its various forms of measurement and predictive bias) can be related. If adverse impact occurs when a cognitive ability test is used, a concern is whether the mean difference on the cognitive ability test is caused by test bias. This bias could be in the form of measurement bias, which could cause the mean score difference through the test either measuring a different construct for one subgroup (i.e., has a different factor structure, meaning a lack of measurement invariance) or containing a preponderance of items biased against one subgroup (i.e., differential item functioning). This test bias could also be in the form of predictive bias. If predictive bias is in the form of a weaker regression slope for the subgroup that scored lower on the test (the state of affairs depicted in **Figure 1*a***), then the test is not as valid a predictor for the lower-scoring subgroup, which is something that might be viewed as unfair. Differential prediction's comparison of subgroups' regression slopes is a direct test of this form of predictive bias. Differential validity in the form of a smaller correlation between cognitive ability tests and job performance for one subgroup signals this same form of predictive bias unless it can be accounted for by a compensatory difference between subgroups in job performance and test standard deviations. If subgroups do not differ in regression slopes (either because validities did not differ or because subgroup differences in standard deviations can account for validity differences), then predictive bias in the form of subgroup differences in regression intercepts remains a concern. In this case, if the lower-scoring subgroup has the higher regression intercept, that subgroup's regression line is higher; therefore, cognitive ability test scores underpredict job performance for that subgroup (this is opposite of the pattern in **Figure 1*b***, wherein the lower-scoring subgroup's job performance is overpredicted by the common regression line). If a subgroup both scores lower on the cognitive ability test and has its job performance underpredicted by that same test, this is obviously problematic from test bias and fairness standpoints. So, all of these forms of test bias are possible (but not the only possible) causes of adverse impact. Thus, it is of interest to know whether cognitive ability test bias exists, what causes it, and what implications it has for organizations. These issues, and particularly predictive bias, are the focus of the remainder of this review.

TEST BIAS AND THE ENSUING ETHICAL AND LEGAL ISSUES FACING ORGANIZATIONS

Organizations choosing to use cognitive ability tests for personnel selection have always faced an ethical dilemma. Often termed the diversity–validity dilemma (e.g., Ployhart & Holtz 2008, Schmitt 2014), the problem is that cognitive ability tests are one of the selection procedures with the strongest utility and validity for predicting job performance (Schmidt & Hunter 1998), but they are also one of the selection procedures with the largest mean score difference between White and non-White subgroups. So, use of cognitive ability tests for personnel selection is an efficient means for identifying applicants likely to perform well on the job, but at the same time this will hurt racial/ethnic diversity in the organization. Alternatively, using selection procedures with smaller White/non-White mean score differences can result in the hiring of more non-White applicants, but these selection procedures typically have weaker predictive validity. So, organizations are faced with the dilemma of choosing which they value more: racial/ethnic diversity or maximum prediction of job performance.

The diversity–validity dilemma has typically been framed with the implicit assumption that cognitive ability tests are unbiased. The existence of test bias would add another layer to this ethical dilemma. If cognitive ability tests' adverse impact against non-Whites was due completely to test bias (e.g., tests do not predict job performance, or substantially underpredict job performance, for these non-White subgroups), the dilemma would likely be resolved. It would be difficult for any ethical organization to justify the use of a test that limits employment opportunities for non-Whites simply due to bias in the test itself. However, even if test bias does exist, it is unlikely to be that simple; it is doubtful that test bias alone accounts for adverse impact. Rather, imagine scenarios in which there is small but consistent test bias for a cognitive ability test exhibiting adverse impact. For example, imagine there is adverse impact and the correlation between cognitive ability test scores and job performance is $r = 0.60$ for Whites and $r = 0.50$ for African Americans, and these validity differences generalize to regression slope differences (the empirical literature reviewed below suggests that this is a very plausible scenario). In this case, there is predictive bias in that the test does not predict job performance as strongly for the African American subgroup. However, test scores are strongly predictive of job performance for each subgroup, despite the validity difference, and it is unlikely that there is a more valid alternative, even for predicting African American job performance. One might term this the test bias–validity dilemma. If the organization wishes to substitute a different selection procedure exhibiting less/no predictive bias, a reduction in validity would likely result. How much test/predictive bias is enough to outweigh predictive validity? Organizations' answers to this ethical dilemma will likely differ as a function of many factors, including value for diversity versus validity, tolerance for risk, and even key stakeholders' individual differences (Hunter & Schmidt 1976, Kim & Berry 2015). To help organizations make informed decisions, it is important to know whether test bias exists and, if so, in what form and magnitude.

Besides the ethical issues that can arise from cognitive ability test bias, there are also significant legal risks. Race is a protected class under Title VII of the Civil Rights Act (42 U.S.C. §2000e et seq.), which means that racial discrimination in employment decisions is illegal in the United States. The form of racial discrimination most applicable to cognitive ability tests is disparate impact, which refers to selection practices that may not be intentionally discriminatory but that are discriminatory in operation. Thus, if the use of a cognitive ability test in personnel selection causes adverse impact against non-Whites, this would fall under the definition of disparate impact, regardless of whether there was intentional discrimination. Title VII of the Civil Rights Act describes the process for determining whether unlawful discrimination has occurred in a disparate impact case:

> An unlawful employment practice based on disparate impact is established under the title only if: (i) a complaining party demonstrates that a respondent uses a particular employment practice that causes a disparate impact on the basis of race... and the respondent fails to demonstrate that the challenged practice is job related for the position in question and consistent with business necessity. (42 U.S.C. §2000e-2, as amended by Pub. L. No. 102-166, §105)

So, once a plaintiff demonstrates that a cognitive ability test has created adverse impact, the defendant (i.e., the organization) must then demonstrate that the cognitive ability test is job related. The *Uniform Guidelines on Employee Selection Procedures* (43 F.R. 38290 et seq., August 25, 1978; "*Uniform Guidelines*") are US federal guidelines describing the type of evidence that must be presented in disparate impact lawsuits (McDaniel et al. 2011) (see Murphy & Jacobs 2012 for a review of issues surrounding adverse impact evidence). According to the *Uniform Guidelines*, defendants/organizations must provide criterion-related validity evidence for cognitive ability tests (content-related validity evidence can substitute for some selection procedures, but not cognitive ability tests; see *Uniform Guidelines*, §14.C.1), which means demonstrating that test scores correlate with or predict job performance. As part of this criterion-related validity evidence, defendants/organizations must provide "studies of unfairness where technically feasible" (§14.B.8), with unfairness defined as "when members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance" (§14.B.8.a). This definition essentially refers to differential prediction, so evidence regarding differential validity or differential prediction is required when there is adverse impact. An important point here is that unfairness studies must be carried out only if there is adverse impact, so differential validity and prediction are a legal concern for organizations only when a non-White subgroup is negatively affected. So, besides the ethical concerns over using a biased test, organizations should also be concerned about differential validity and prediction because if these forms of test bias are found along with adverse impact as a result of a cognitive ability test, the organization may be successfully sued for employment discrimination. This just further reinforces the importance of knowing whether differential validity and prediction are typical for cognitive ability tests.

IS DIFFERENTIAL VALIDITY TYPICAL FOR COGNITIVE ABILITY TESTS?

This section addresses whether differential validity is typical for cognitive ability tests. First, research focused on differences between subgroups in observed validities is reviewed. Next, research attending to the role that range restriction plays in causing observed validities to differ is reviewed.

Differences in Observed Validities

In their popular human resource selection textbook, Gatewood et al. (2008, p. 547) stated that "differential validity does not exist." Until recently, this was the strong conclusion of the employment testing literature (Schmitt 2014), with other influential articles making similar statements (e.g., Schmidt & Hunter 1981). This conclusion was based on a number of reviews of differential validity of cognitive ability tests carried out in the 1970s and early 1980s that did not find statistically significant differences between African American and White (Hunter & Schmidt 1978, Hunter et al. 1979, O'Connor et al. 1975) or between Hispanic American and White (Schmidt et al. 1980) observed validity coefficients any more often than would be expected by chance. Observed validity coefficients refer to the observed correlation between cognitive ability tests and job performance, uncorrected for any statistical artifacts (e.g., range restriction, criterion measurement error). The design in each of these reviews was to collect a number of different samples for which cognitive ability validity/correlation coefficients could be calculated separately for White and non-White subgroups, tally the percentage of the time across samples that the validity difference between subgroups reached statistical significance, and compare this percentage with what would be expected just due to chance (i.e., 5%, per $\alpha = 0.05$). This method of synthesizing results across studies has been termed vote counting and has been criticized by meta-analysts for often providing misleading results (Hunter & Schmidt 2004).

Berry et al. (2011) noted that the small non-White sample sizes in the studies included in the vote-counting reviews of differential validity, combined with the typically small-to-moderate size of White/non-White validity differences, made the results of those vote-counting reviews difficult to interpret (see also Berry 2007 for an in-depth critical review of the 1970s and 1980s vote-counting studies). Berry et al. (2011) suggested separately meta-analyzing White and non-White validities as a solution. Thus, Berry et al. (2011) meta-analyzed the observed correlations between cognitive ability tests and job performance separately for 143 samples of African American and White test takers in employment settings, finding that validity was slightly lower for African Americans ($r = 0.16$ versus $r = 0.19$) and that the difference in favor of Whites persisted across levels of job complexity. Berry et al. also meta-analyzed validities for African American–White comparisons and Hispanic American–White comparisons in college admissions and military settings (not enough data were available for Hispanic American–White comparisons in employment settings), and cognitive ability test observed validity was always higher for Whites, with differences larger than in employment settings. Berry et al. (2014b) were able to locate 35 Hispanic American–White differential validity studies carried out in employment settings and demonstrated that meta-analytic observed validity was slightly lower for Hispanic Americans ($r = 0.15$ versus $r = 0.18$). Although these differences are small in absolute magnitude, they represent sizable differences in percentage terms, with African American and Hispanic American validities approximately 16–17% lower than White validities. Further, these validities are observed validities, uncorrected for the attenuating artifacts of range restriction and measurement error. In all, in contrast to the vote-counting reviews, the meta-analytic reviews of differential validity suggest that observed cognitive ability test validity is somewhat lower for African Americans and Hispanic Americans than for Whites.

Does Range Restriction Explain Observed Differential Validity?

All of the differential validity results reviewed in the previous section compared subgroups' *observed* cognitive ability test validities. It is well established that when assessing the relationship between a test and a criterion, operational validity, and not observed validity, of the test is most relevant (Sackett et al. 2008). Operational validity refers to the relationship between the test and job performance absent the effects of range restriction and criterion measurement error. In the differential validity context, range restriction is an especially plausible explanation for differential validity, as the mean difference between White and non-White subgroups on cognitive ability tests (Roth et al. 2001) could cause only the highest-scoring non-Whites to be hired, resulting in more restriction of range for the non-White subgroups (i.e., differential range restriction).

Roth et al. (2014) concluded, based on four studies, that differential range restriction likely explains Berry et al.'s (2011) finding of differential validity. Study 1 included cognitive ability test scores from three entire applicant pools in employment settings and one entire applicant pool in a college admissions setting. Roth et al. applied various selection ratios to these applicant data sets and demonstrated that when the same selection ratio/cut score is applied to African American, Hispanic American, and White test takers, the cognitive ability test scores for the selected non-Whites were more restricted in range than those for the selected Whites. Studies 2, 3, and 4 used simulated data to demonstrate that (*a*) if there is no population-level differential validity, top-down selection with a common cut score for each subgroup causes observed validities to be lower for the non-White subgroups; (*b*) applying range restriction corrections separately for each subgroup using subgroup-specific range restriction estimates provides the most accurate results (this was a point also made by Berry et al. 2011 and then empirically demonstrated by Berry et al. 2013b); and (*c*) differential range restriction (i.e., greater range restriction for non-White subgroups) causes observed validity differences to be exaggerated if the non-White validity is truly lower at the unrestricted population level. Roth et al. (2014) concluded that the pattern of results in their range restriction conditions across the four studies was so similar to Berry et al.'s (2011) results (i.e., non-White observed validity slightly lower than White validity) that range restriction likely accounted for Berry et al.'s finding of differential validity. Roth et al. concluded that true, unrestricted population validities do not differ for racial/ethnic subgroups.

A more direct test of whether Berry et al.'s (2011) results can be explained by range restriction would include an examination of whether there is evidence of differential range restriction in the primary studies used in Berry et al.'s (2011, 2014b) differential validity meta-analyses. Doing exactly that, Berry et al. (2014b) obtained a primary data set from the United States Employment Service that contained sample-level information for 127 African American–White and 35 Hispanic American–White differential validity studies, representing 79.9% of the African American–White participants in Berry et al. (2011)'s employment differential validity meta-analysis and 100% of the Hispanic American–White participants in Berry et al.'s (2014b). The sample-level primary data allowed the estimation of *u*-ratios (restricted standard deviation divided by an estimate of the unrestricted standard deviation of cognitive ability tests; an index of range restriction) for all of these samples. The average *u*-ratios were 0.86, 0.85, and 0.89 for African Americans, Hispanic Americans, and Whites, respectively, indicating almost identical amounts of range restriction. Berry et al. (2014b) then calculated meta-analytic observed cognitive ability test validities for each subgroup, finding results similar to Berry et al.'s (2011) (i.e., small validity differences in favor of Whites). Finally, the observed validities were corrected for indirect range restriction (Hunter et al. 2006), and the validity differences in favor of Whites remained, meaning that range restriction could not account for differential validity. Berry et al. (2014b) also carried out similar analyses for Berry et al.'s (2011) college admissions and military samples and came to the same conclusion, that range restriction could not account for differential validity.

So, to summarize, Roth et al. (2014) used a series of simulations to suggest that Berry et al.'s (2011) differential validity findings were just a function of differential range restriction. However, it is not clear how applicable Roth et al.'s simulation assumptions are to real operational selection settings or how representative Roth et al.'s data were of Berry et al.'s (2011) data. An examination of actual differential validity data from more than one million subjects used by Berry et al. (2011, 2014b) did not support the idea that range restriction accounts for Berry et al.'s finding of differential validity. Of course, this does not mean that it is impossible for differential validity to be caused by differential range restriction in some settings. Roth et al. (2014) demonstrated convincingly that in settings in which the authors' assumptions are met (top-down selection using only cognitive ability test scores with no preferential selection), differential range restriction could create the illusion of differential validity even if there is none at the population level. However, Berry et al. (2014b) provide strong evidence that differential range restriction was not the sole cause of differential validity in the primary studies included in Berry et al.'s (2011) meta-analysis.

Conclusions About the Existence of Differential Validity

The available empirical evidence that has used modern meta-analytic methods suggests that the validity of cognitive ability tests is lower for African Americans and Hispanic Americans than for Whites. The most up-to-date validity estimates corrected for indirect range restriction suggest that, although cognitive ability tests have appreciable levels of validity for all racial/ethnic subgroups, Hispanic American validity is about 13% lower, and African American validity is about 18% lower, than White validity (Berry et al. 2014b). Unless job performance reliabilities are lower for non-White subgroups, something for which there is very little relevant evidence (Berry et al. 2011), the differences for operational validities would even be slightly larger. This general conclusion holds across employment, college admissions, and military settings.

However, there are a number of important issues and caveats about the available empirical differential validity evidence. For one, the empirical evidence is quite dated. Virtually all of the employment setting evidence from Berry and colleagues' meta-analytic reviews came from studies carried out before 1989, with the majority being carried out considerably earlier. Second, the majority of these studies came from the United States Employment Service's validity studies of the General Aptitude Test Battery (GATB). So, it remains unclear the degree to which these results generalize to the present day or to other cognitive ability tests. Third, there is a general lack of research on why validities might differ across races. There has been work investigating statistical explanations such as range restriction (Berry et al. 2014b, Roth et al. 2014), and Berry et al. (2011) did perform moderator analyses, which suggested that job complexity does not explain differential validity. However, there has been little to no research on possible substantive causes of differential validity, such as socioeconomic status, racial discrimination, or stereotype threat; this is discussed more below in the agenda for future research.

IS DIFFERENTIAL PREDICTION TYPICAL FOR COGNITIVE ABILITY TESTS?

Differential prediction of cognitive ability tests with respect to job performance has been examined extensively. The cognitive ability testing literature's predominant conclusions about differential prediction came from a number of MMR studies carried out in the 1970s and 1980s (e.g., Bartlett et al. 1978, Hartigan & Wigdor 1989, Schmidt et al. 1980). The general method in these MMR studies was similar to that of the vote-counting reviews of differential validity. That is, multiple data sets containing cognitive ability test and job performance scores were collected for samples consisting of African Americans and Whites or Hispanic Americans and Whites. Within each sample, job performance was regressed on cognitive ability test scores, a race dummy variable (capturing the subgroup intercept difference), and a test–race interaction term (capturing the subgroup slope difference). The percentage of the time that statistically significant differences in either regression slopes or regression intercepts occurred across those samples was tallied. The general conclusion of these MMR studies was that regression slopes do not differ significantly more often than would be expected by chance (i.e., $\alpha = 0.05$) for non-White and White subgroups, but that significant intercept differences are common, such that the White regression intercept often lies above the non-White regression lines (SIOP 2003). So, the typical finding is that White and non-White subgroups have different, but parallel, regression lines relating cognitive ability test scores to job performance (similar to the trend depicted in **Figure 1*b***). This means that there is differential prediction in that the use of a common regression line across subgroups will result in systematic nonzero errors of prediction for subgroups. However, the errors of prediction for the non-White subgroups will be in the form of overprediction of job performance and therefore will not disadvantage these non-White job applicants. This is the saving grace for cognitive ability tests from a predictive bias standpoint: The MMR studies suggest that they do not exhibit predictive bias *against* non-Whites by underpredicting their job performance.

Recent research has challenged the conclusion that Whites have the higher regression intercept. The most serious challenge to the intercept conclusion is Aguinis et al.'s (2010) demonstration that the test for subgroup differences in regression intercepts used in the MMR studies is itself biased in such a way as to exaggerate the size of intercept differences. Via mathematical proof, Aguinis et al. showed that the intercept test in MMR overestimates the intercept difference between subgroups when there is measurement error in the predictor, when there is range restriction, and when there is a mean difference between subgroups on the cognitive ability test. It is at least hypothetically possible to control for the first two factors via statistical corrections, but no such control for mean differences between White and non-White subgroups on cognitive ability tests is possible. Thus, the intercept differences test is confounded and biased. This calls into question the general conclusion of overprediction of non-White job performance, as that conclusion has been based on analyses using the biased intercept test.

In order to test whether the intercept conclusion from previous research holds versus is artifactual, new research must find a way to test this without using the biased MMR intercept test. The first attempt to address Aguinis et al.'s concern was carried out in a college admissions setting.

Mattern & Patterson (2013) drew upon a sample provided to them by the College Board of 348 SAT validity studies carried out at 177 universities and including more than 475,000 college students. Mattern & Patterson correlated college grade point averages, SAT scores, a race dummy variable, and an SAT–race interaction term within each validity study and then meta-analyzed these correlation matrices across validity studies. The College Board provided detailed population information, so Mattern & Patterson were able to correct each correlation in the meta-analytic matrices for statistical artifacts such as range restriction and criterion unreliability [although these corrections were not done within subgroups, as recommended by Berry et al. (2011, 2013b) and Roth et al. (2014)]. They then used the meta-analytic, artifact-corrected correlation matrices to regress college grades on SAT scores, the race dummy, and the SAT–race interaction term and used the resulting regression equations to plot the African American, Hispanic American, and White regression lines. In the regression plots, the White regression lines lay above the non-White regression lines, which means that non-White grades would be overpredicted throughout the SAT score range. Such observation of the regression plots does not rely on the intercept test Aguinis et al. demonstrated was itself biased. Thus, Mattern & Patterson's method represents one solution to the problem pointed out by Aguinis et al., with the conclusion being that non-White criterion (grade) performance is still overpredicted.

Berry & Zhao (2015) used a different approach to investigate differential prediction of cognitive ability tests in employment settings without using the biased intercept test identified by Aguinis et al. (2010). Drawing on the math of subgroup intercept differences, Berry & Zhao (2015) demonstrated that, when subgroups' regression slopes are equal, the difference in regression intercepts for two subgroups is a function of the following formula:

$$\Delta b_0 = d_Y - r d_X, \tag{3}$$

where Δb_0 is the regression intercept difference between the subgroups, r is the correlation coefficient between the cognitive ability test and job performance, and d_Y and d_X are the job performance and cognitive ability test standardized mean differences (i.e., d-values) between subgroups, respectively. Berry & Zhao then located meta-analytic estimates of r (Hunter et al. 2006), d_Y (McKay & McDaniel 2006), and d_X (Roth et al. 2001); corrected them for indirect range restriction and criterion measurement error; and inserted the corrected meta-analytic estimates into the above formula to determine whether cognitive ability test scores overpredicted or underpredicted job performance for African Americans. Across a wide range of conditions in computation models using the above formula and methods, Berry & Zhao (2015) demonstrated that the White subgroup generally has the higher intercept, meaning that African American job performance is typically overpredicted. Berry & Zhao also modeled a number of conditions wherein the African American regression slope was as much as 50% lower than the White slope, and the conclusion of overprediction of African American job performance always held throughout the operational score range. Thus, without using the biased intercept test, Berry & Zhao demonstrated that the meta-analytic literature to date suggests that cognitive ability tests should generally ovepredict the job performance of African Americans and, therefore, are not predictively biased against that subgroup. So, similar to Mattern & Patterson (2013), Berry & Zhao (2015) concluded that, even when avoiding the biased intercept test and making appropriate corrections for statistical artifacts, African American job performance is typically overpredicted by cognitive ability test scores. There is no such published research available for comparisons between Hispanic American and White job applicants.

Recent research has also challenged the conclusion from the MMR studies that White and non-White subgroups' regression slopes do not differ. For example, Aguinis and colleagues (e.g., Aguinis et al. 2005, 2010; Aguinis & Stone-Romero 1997) have repeatedly demonstrated that the significance test used in the MMR studies to test for slope differences has very low statistical power. Aguinis et al. (2010, p. 651) provided an enlightening example in which they demonstrated that an MMR differential prediction study by Rotundo & Sackett (1999) with what would be considered very large sample sizes for a differential prediction study (White and African American *N*'s of 17,020 and 1,212, respectively) only had power = 0.101 (i.e., a 10.1% chance) to detect the subgroup slope difference that existed in that study. This calls into question the results of the much smaller MMR studies upon which the conclusion of a lack of slope differences was based. However, it does not directly demonstrate that slope differences exist.

Berry et al. (2014a) leveraged meta-analysis as well as the resulting increase in statistical power to test whether regression slopes actually differed across African American, Hispanic American, and White subgroups. The authors noted that, when regressing job performance on cognitive ability test scores separately for each subgroup, the unstandardized regression slope differs from the correlation/validity coefficient only by the ratio of the job performance and cognitive ability test standard deviations (s_Y/s_X). So, differential validity would generalize to regression slope differences unless that standard deviation ratio differs between subgroups by a compensatory amount. Thus, Berry et al. (2014a) located hundreds of differential validity studies that also reported White and non-White subgroups' cognitive ability test and job performance standard deviations, including the vast majority of studies included in Berry et al.'s (2011, 2014b) differential validity meta-analyses. Berry et al. (2014a) first meta-analyzed observed cognitive ability test validities for each subgroup, demonstrating that differential validity existed, with non-White observed validities slightly lower than White observed validities.[2] Range restriction information was available for 127 African American–White and 35 Hispanic American–White samples (these are the same samples included in Berry et al. 2014b), and *u*-ratios for the three subgroups hardly differed (*u*'s between 0.85 and 0.89), which suggests that differential range restriction did not confound results. Berry et al. (2014a) next calculated the s_Y/s_X ratio separately for each subgroup and then divided the non-White ratio by the White ratio within each sample (the resulting ratio was greater than 1.0 if s_Y/s_X was greater for the non-White subgroup, meaning that at least part of the subgroup validity difference was due to subgroup differences in the s_Y/s_X ratio). The resulting ratios in employment settings were greater than 1.0 for both the African American–White (1.12) and Hispanic American–White (1.07) comparisons, but this accounted for only a small part of the subgroup validity differences. The same pattern of results was found in college admissions and military studies. In all, Berry et al. (2014a) concluded that the typical differential validity finding (i.e., greater observed validity for the White subgroup) generalizes to regression slopes, but the differences in slopes are likely slightly smaller than validity differences, as White and non-White subgroups differ slightly in their s_Y/s_X ratios.

There are at least two important critiques of the differential prediction research reviewed to this point. First, most of this research has dealt with *observed* relationships between cognitive ability tests and job performance (Berry & Zhao 2015 and Mattern & Patterson 2013 are the only exceptions). It is virtually always the case that cognitive ability test differential validity/prediction studies will be affected by indirect range restriction (Hunter et al. 2006). Indirect range restriction has systematic downward-biasing effects on the unstandardized regression coefficients used in

[2] This was actually true only for the studies that used the GATB as the cognitive ability test, which happened to be the majority of studies (80% of the total sample size). In the 29 studies carried out in the 1960s or 1970s that did not use the GATB, African American validity was higher than White validity.

differential prediction studies (Linn 1983), particularly the MMR studies reviewed above. Berry & Zhao (2015) demonstrated that indirect range restriction may lead to the conclusion that there is overprediction of job performance for non-Whites at the level of observed validities, when there is in fact underprediction of job performance at the level of operational validity (i.e., when indirect range restriction is accounted for). Thus, it is a crucial shortcoming that published differential prediction studies in the employment testing literature have generally completely ignored indirect range restriction. To put it bluntly, the field knows a great deal about whether cognitive ability tests exhibit differential prediction when using observed data from job incumbents biased by indirect range restriction, but it knows far less about what actually matters: whether cognitive ability tests exhibit differential prediction in applicant pools at the point when selection occurs (i.e., before range restriction has occurred). Berry et al. (2014a) at least ran analyses suggesting that subgroups did not differ in amounts of range restriction. Therefore, their finding that differential validity probably generalizes to differential prediction in the form of slope differences likely also extends to regression slopes corrected for indirect range restriction. Also, although Mattern & Patterson's (2013) SAT differential prediction study was carried out in a college admissions setting, they found smaller slopes and lower intercepts for non-White subgroups after accounting for indirect range restriction. Berry & Zhao's (2015) findings suggest that the White subgroup still generally has the higher intercept (meaning that non-White job performance is still overpredicted) when indirect range restriction is accounted for. So, the results of Berry et al. (2014a) and Mattern & Patterson (2013) suggest that there are slope differences once indirect range restriction is accounted for; the results of Berry & Zhao (2015) suggest that there are intercept differences in favor of the White subgroup once indirect range restriction is accounted for.

The second important critique is that, similar to the differential validity studies, all of the MMR differential prediction studies in employment settings reviewed to this point were carried out in the 1980s or earlier and mostly used the same cognitive ability test (the GATB). Even studies as recent as Berry et al. (2014a) and Berry & Zhao (2015) used GATB data from the 1980s. So, the data are old, and it is not clear whether results generalize to the present day or beyond the GATB. These data from the 1980s are available only because the United States Employment Service carried out and made publicly available a large number of studies of the validity of the GATB. I am not aware of companies in the private sector that have been or are currently willing to make similar data available. Regardless, new data using different cognitive ability tests are clearly needed for future research.

Conclusions About the Existence of Differential Prediction

As noted above, a great deal of MMR differential prediction studies were carried out in the 1980s or earlier. These studies generally concluded that regression slopes do not differ between White and non-White subgroups but that the White subgroup often has the higher intercept. Because of low statistical power and a bias in the intercept difference test used in these MMR studies (Aguinis et al. 2010) and because these studies made no attempt to account for indirect range restriction, they are confounded and almost useless as evidence regarding the existence or form of differential prediction. Some recent differential prediction research has at least attempted to account for indirect range restriction, and the conclusion so far has been that regression slopes are at least slightly lower for non-White subgroups. This means that White and non-White regression lines will cross at some point, so it is important to determine where this point is and whether it is in the operational score range, as over/underprediction of job performance results will reverse above and below this intersection point. Research drawing on meta-analytic estimates (Berry & Zhao 2015) suggests that these intersection points would generally be very low in the score range and that African American job performance is likely generally overpredicted by cognitive ability test scores.

Also, differential prediction research using new data and cognitive ability tests is needed, as data have to this point mostly been available only from the US Employment Service's GATB validity studies carried out in the 1980s or earlier. Finally, similar to differential validity, most research has focused on how to test for differential prediction and whether it exists, with almost no research addressing why differential prediction might occur.

OVERALL CONCLUSIONS ABOUT DIFFERENTIAL VALIDITY AND PREDICTION

Table 2 summarizes conclusions about differential validity and prediction based on past research (what we knew), such conclusions based on more recent research (what we know now), and reasons for why our knowledge has been updated and improved. Important trends in recent research are the use of modern meta-analytic methods, corrections for statistical artifacts, and attention to statistical power. This recent research has challenged some conclusions based on past research (validity and slope differences) and has reinforced and refined others (intercept differences). The available evidence now suggests that cognitive ability tests do exhibit test bias in the form of predictive bias. Using a common regression line for White and non-White job applicants will typically result in nonzero errors of prediction, as White and non-White subgroups do not share a common regression line. The non-White regression line typically has a lower slope and intercept than the White regression line, meaning that cognitive ability tests do not predict job performance quite as strongly for the non-White subgroup members (although cognitive ability tests still predict job performance strongly for each subgroup). However, errors of prediction will tend to be such that non-White job performance is overpredicted, rather than underpredicted, by cognitive ability test scores. In fact, somewhat counterintuitively, the lower validity/slope actually will result in greater *overprediction* of non-White job performance in the high cognitive ability score range than if the non-White subgroup had a slope/validity equal to or greater than the White subgroup. This is because the lower slope and lower intercept mean that the non-White regression line will lie below the White regression line throughout the score range, but the lower slope means that the non-White line will be slightly flatter than the White line. For example, imagine a situation somewhere between panels *a* and *b* of **Figure 1** wherein the non-White subgroup's line has a flatter slope (like in **Figure 1*a***) but also lies below the White subgroup's line throughout the plot (like in **Figure 1*b***). Therefore, the gap between the lines will be larger at the high end of the cognitive ability score range (where high-stakes selection takes place) than at the low end, resulting in greater overprediction of job performance at the high end of the test score range.

IMPLICATIONS FOR INDIVIDUALS

Adverse impact has profound personal, financial, and social impacts on non-White persons because adverse impact limits employment opportunities for non-Whites. So, what effects do the validity/slope and intercept differences have on individuals? Aguinis & Smith (2007) demonstrated that using a biased selection test as if it is unbiased results in "bias-based selection errors," meaning false positives or false negatives that are a function of test bias. Even very small differences in validities or regression slopes can cause substantial differences between subgroups in rates of false-positive and false-negative hires (Aguinis & Smith 2007). Being denied a job simply because of bias in a cognitive ability test would obviously be a very unfortunate outcome for an individual. It would be particularly troubling if the subgroup scoring lower on the tests (i.e., non-Whites) had more false negatives due to test bias. When hiring is done at the high end of the cognitive ability score range, a higher rate of false negatives for non-White applicants is unlikely, as the lower slope

Table 2 Conclusions about differential validity and differential prediction based on past research (what we knew), such conclusions based on more recent research (what we know now), and reasons for why our knowledge has been updated and improved

Issue	Subissue	What we knew	What we know now	Why our knowledge has been updated and improved
Differential validity	NA	The validity of cognitive ability tests does not differ across African American, Hispanic American, and White subgroups. Any instances of differential validity occur at less than chance levels.	Observed and operational validities of cognitive ability tests are about 10–20% lower for African Americans and Hispanic Americans than for Whites.	Past research relied on statistical significance testing, vote-counting techniques, and comparison of observed validities. Recent research (e.g., Berry et al. 2011, 2014b) used meta-analytic techniques and corrections for statistical artifacts (particularly indirect range restriction) to compare observed and operational validities for African American, Hispanic American, and White subgroups.
Differential prediction	Slope differences	MMR analyses demonstrated that regression slope differences do not occur more often than chance for African American, Hispanic American, and White subgroups.	Regression slopes differ for African American, Hispanic American, and White subgroups. These slope differences are likely slightly smaller than validity differences.	Recent research (e.g., Aguinis et al. 2010) demonstrated that MMR analyses have very low power to detect slope differences and that the typical MMR studies have been underpowered. Large-sample studies (e.g., Mattern & Patterson 2013) and meta-analyses (e.g., Berry et al. 2014a) have suggested that small slope differences in favor of Whites are likely the norm.
	Intercept differences	MMR studies demonstrated that Whites typically have a slightly higher intercept than African Americans and Hispanic Americans, meaning that non-White job performance is typically overestimated.	The same conclusion generally holds, although the evidence for it is improved. Future intercept differences research should avoid the MMR intercept test and attend to the effects of indirect range restriction.	Aguinis et al. (2010) demonstrated that the intercept test used in the MMR studies is biased toward a conclusion of non-White overprediction. Recent large-scale studies (Berry & Zhao 2015, Mattern & Patterson 2013) have investigated intercept differences using methods other than the biased intercept test and have made corrections for indirect range restriction, and the conclusion of non-White overprediction still generally holds.

Abbreviations: MMR, moderated multiple regression; NA, not applicable.

and intercept for the non-White regression lines suggest cognitive ability tests result in the greatest *overprediction* of job performance in the high test score range (i.e., errors of prediction, and therefore selection, will tend to be in the form of false positives for the non-White subgroup members). Despite the lack of underprediction or greater bias-based false negative rates for non-Whites, some may still find it troubling that cognitive ability tests are a less valid predictor of job performance for the non-White subgroup members scoring lower on, and being adversely impacted by, the cognitive ability test. When coupled with the fact that White/non-White subgroup differences in job performance (McKay & McDaniel 2006) are substantially smaller than differences in cognitive ability test scores (Roth et al. 2001), this provides a basis for questioning the fairness of cognitive ability test use. Of course, fairness is a value judgment (e.g., Berry et al. 2014b, SIOP 2003), and others would have a basis for rebutting that cognitive ability test use is fair because test scores still predict job performance strongly for each subgroup and even overpredict, rather than underpredict, non-White subgroup members' job performance. The position one takes on this issue is a matter of opinion and likely has much to do with individual differences in deep-seated beliefs and values regarding issues such as hierarchy in society (Kim & Berry 2015).

IMPLICATIONS FOR EMPLOYMENT LITIGATION

When cognitive ability tests create adverse impact, organizations will need to be able to show that the test is job related through criterion-related validity evidence. The *Uniform Guidelines* already called for "fairness studies" but allowed organizations to avoid such studies if they were not technically feasible. Given recent research challenging the long-held belief that there is no predictive bias, it may become more difficult for organizations to continue using a challenged cognitive ability test just because a fairness study is not technically feasible. When fairness studies are carried out, the available research suggests it is likely that there will be differential validity, such that criterion-related validity will be lower for non-White subgroups. Lower validity coupled with lower scores for these historically disadvantaged subgroups could be problematic in the eyes of courts. However, it remains to be seen how big a validity difference will be required for it to become an issue in court. Is any validity difference an issue? Does the cognitive ability test just need to be adequately valid for each subgroup, even if there are validity differences? Does the cognitive ability test need only to outperform alternative tests for each subgroup? The standard the courts would use remains unclear.

Because the *Uniform Guidelines* and other authoritative documents on test bias [e.g., the Society for Industrial and Organizational Psychology's *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP 2003) and the American Psychological Association and others' *Standards for Educational and Psychological Testing* (AERA et al. 1999)] focus more on differential prediction than on differential validity, differential prediction is likely to be the greater legal concern in the face of adverse impact. I am aware of anecdotal evidence that Aguinis et al. (2010) is being used by plaintiffs' witnesses to suggest that defendants' typical MMR differential prediction analyses that use the biased intercept difference test are meaningless. I am not aware of this being on the record for affecting any court cases yet. However, this is likely to be a challenge for organizations attempting to carry out fairness studies moving forward.

ADDITIONAL ISSUES, RECOMMENDATIONS, AND DIRECTIONS FOR FUTURE RESEARCH

Table 3 provides a summary of the issues highlighted in this section.

Table 3 A summary of issues that still need to be addressed and specific research needs for each issue

Overall issue	Specific future research needs
Need for new data: Most differential validity and differential prediction data are old (1980s and earlier), have used the GATB, and do not provide adequate information about statistical artifacts.	Differential validity/prediction studies using new data. Differential validity/prediction studies using tests other than the GATB. Any studies providing information about statistical artifacts separately by subgroup. Even short research reports with little/no theory testing.
Publication bias: Test vendors have a vested interest in showing that their tests are not biased.	Methods for determining whether differential validity/prediction studies exhibit publication bias. Creative partnerships between researchers, practitioners, and organizations to help practitioners/organizations feel safe about honestly sharing data.
Development of unbiased differential prediction methods: The intercept test in MMR is biased and does not allow for range restriction corrections. Alternatives developed in recent research have their own shortcomings, particularly a lack of significance tests.	New methods of testing for slope and intercept differences that account for indirect range restriction and allow statistical significance testing.
How large must differences be to matter?: Recent research suggests that validity/slope differences exist, but there is no consensus on how large these differences must be to matter.	Research on why and when different stakeholders think a validity/slope difference is "big enough to matter," as this is likely to be a matter of opinion and values. Research on contextual influences and individual differences.
What causes predictive bias?: Much research has focused on methods for testing for predictive bias and whether it exists; very little research has focused on what causes predictive bias. Differences in test–criterion relationships between subgroups can be caused by any factor that differentially affects test or criterion scores of White versus non-White test takers.	Can range restriction account for predictive bias? A lot of research exists on this topic, but it is conflicted and uses less than ideal data. Can psychometric characteristics of the test or criterion (measurement error or measurement bias) account for predictive bias? Can contextual influences (e.g., stereotype threat, the complexity of the job, racism in the workplace) account for predictive bias? Do true differences between subgroups in the role cognitive ability plays in determining job performance account for predictive bias? There are doubtless many other potential causes that require investigation.
Test bias research for other subgroups and selection methods: Most test bias research has focused on cognitive ability tests and comparisons between African American and White applicants, with far fewer focused on comparisons between Hispanic Americans and Whites. Many other selection methods that exhibit subgroup differences, and many other subgroups, remain unaddressed.	More research comparing Hispanic American and White subgroups. Test bias research investigating comparisons for other historically disadvantaged subgroups, such as women, Native Americans, and immigrants. Test bias research investigating comparisons for other selection methods that exhibit subgroup differences, such as personality tests, employment interviews, and work samples.

Abbreviations: GATB, General Aptitude Test Battery; MMR, moderated multiple regression.

Need for New Data

Most of the differential validity and prediction data are very old and do not provide adequate information about statistical artifacts such as range restriction and criterion reliability. Differential validity and prediction studies using new data and cognitive ability tests other than the GATB are sorely needed. A major obstacle will be getting these data from test vendors and organizations that use cognitive ability tests for personnel selection. Such companies have been reticent to provide this sort of data, presumably for fear of legal repercussions. However, I argue that, given the emerging change in the scientific literature's stance on differential validity/prediction (i.e., available evidence now suggests that they may exist) and the likely increased future requirement to produce evidence of test fairness when legally challenged (as discussed in the Implications for Employment Litigation section above), it is in companies' best interest to get ahead of this issue and take a lead in addressing it. If cognitive ability tests are not biased against racial/ethnic subgroups, the best course of action for these organizations is to publicly document this evidence. The College Board, the owner of the SAT, is an illustrative example. With so much public scrutiny of the SAT, the College Board's response over the years has generally been one of transparency. The College Board often releases reports on differential validity and prediction (e.g., Patterson & Mattern 2011) and even recently made publicly available large SAT validity study data sets (Mattern & Patterson 2013). Given the difficulties in getting organizations/vendors to provide relevant data, I encourage peer-reviewed journals to be open to publishing at least short reports based on any differential validity/prediction data researchers can obtain. At the least, these would be good fodder for later meta-analytic investigations.

Publication Bias

Of course, test vendors and organizations using cognitive ability tests have a vested interest in these tests being unbiased, so some amount of skepticism is warranted for any data these companies would release. Publication-bias analyses (e.g., Kepes et al. 2012) may be one way to mitigate some of these concerns. I am not aware of publication-bias analyses directly applicable to differential validity or differential prediction research. However, the logic for such analyses would be similar to that in more traditional publication-bias analyses. For example, a sign of publication bias (i.e., companies releasing only studies that show no differential validity/prediction and suppressing studies that do show it) would be small-sample studies consistently showing evidence of no differential validity/prediction, despite the expected effects of sampling error. Development of methods for detecting publication bias in differential validity/prediction contexts is a need for future research. Even better would be ways to keep such publication bias from happening in the first place. Perhaps creative researcher–practitioner partnerships could be part of the solution. Necessity is the mother of invention, and my hope is that as recognition continues to increase that differential validity/prediction is a potential problem, such creative partnerships will begin to emerge.

Development of Unbiased Differential Prediction Methods

Aguinis et al. (2010) demonstrated that the intercept difference test used in the MMR differential prediction analyses is biased. Further, the MMR analyses do not lend themselves well to corrections for statistical artifacts, most notably indirect range restriction. So, new methods of testing for differential prediction that do not rely on the biased intercept test and that account for indirect range restriction are needed. The study by Mattern & Patterson (2013) represents one attempt (see the discussion of Mattern & Patterson's work, above, for a description of their analytic method). Mattern & Patterson's method was an important step forward, but there are at least two issues with their analytic approach. First, because there are not separate correlation matrices for each subgroup (instead, there is a test–race interaction term), separate corrections for range restriction cannot be done for each subgroup. Instead, Mattern & Patterson corrected the correlation matrix for range restriction using an overall estimate of the unrestricted standard deviation of cognitive ability test scores that is the same for all subgroups. Berry et al. (2011, 2013b) and Roth et al. (2014) made the case that use of an overall range restriction correction, rather than subgroup-specific corrections, can lead to misestimation of unrestricted test–criterion relationships. Second, inspection of subgroups' regression lines in a plot does not allow for a test of statistical significance. So, this method is probably viable only for very large samples that minimize the effects of sampling error and not for the typical local validation study feasible for most organizations. Regardless of these issues, Mattern & Patterson's (2013) method is clearly superior to methods that simply use the biased intercept differences test and ignore the effects of indirect range restriction.

Berry & Zhao (2015) developed another method for testing for differential prediction wherein the test validity and the test and criterion *d*-values can be inserted into a formula to determine the subgroup intercept difference. This method avoids the biased intercept test identified by Aguinis et al. (2010), but has at least two important limitations. First, this formula assumes that subgroups' regression slopes do not differ, so it is applicable only when this is the case (Berry & Zhao had to use simulations and computation models to test whether their conclusions of overprediction would change if regression slopes differed). Second, like Mattern & Patterson's (2013) method, the Berry & Zhao (2015) method does not include a test for statistical significance. So, this method is also probably viable only for very large samples that minimize the effects of sampling error and not for the typical local validation study feasible for most organizations. Research developing new methods is needed.

How Large Must Differences Be to Matter?

Given available evidence, it is likely that cognitive ability tests exhibit predictive bias against non-Whites in that test validities and slopes are lower for these subgroups. However, how large must these differences in validities and slopes be to matter? For example, the observed validities in employment settings in Berry et al. (2011) were 0.16 and 0.19 for African Americans and Whites, respectively; they were 0.15 and 0.18 for Hispanic Americans and Whites, respectively, in Berry et al. (2014b). Using mostly the same data, Berry et al. (2014a) found similar validity differences and suggested that the slope differences would be slightly smaller. Although sizable in percentage terms, in absolute terms these are relatively small differences. Whether differences of this magnitude matter likely depends on one's perspective. For pure scientific inquiry, even small differences may be of interest to some who wish to know why those differences exist. For more applied purposes, such as using cognitive ability tests for high-stakes selection, the concern would be whether differences of this magnitude signal enough bias to declare cognitive ability test use unfair. Berry et al. (2014b, p. 30) made the point that

> [f]airness is, of course, a value judgment, and whether one judges that the slopes differ enough to make test use unfair would depend on one's perspective. For instance, some might believe that "differ enough" means the slope differences must be large enough to make the subgroup regression lines cross in the operational score range, thus resulting in underprediction of non-White performance in some part of the operational score range. Others might believe even the subgroup regression lines crossing is not a significant problem as long as the slopes are still steep for each group (i.e., test scores still predict performance for each group) and alternative predictors could not do any better. Others might believe any difference in slopes signal that the test is not working as well for its intended purpose for one subgroup and that this is a significant issue.

None of these positions is necessarily wrong, and trying to agree on a specific threshold for when a difference is big enough to matter is unlikely to be successful. Further, the position one takes is likely to be a function of various individual differences, including values, traits, and even deep-seated beliefs about inequality. For example, research by Kim & Berry (2015) suggests that social dominance orientation (Sidanius & Pratto 1999) influences whether one supports cognitive ability test use in college admissions and even influences the seemingly non-dominance-related reasons people endorse for their support, or the lack thereof. So, there may be at least as much scientific and practical value in understanding why and when different stakeholders think a validity/slope difference is "big enough to matter."

What Causes Predictive Bias?

Predictive bias research to date has focused mostly on (*a*) methods of testing for differential validity/prediction and (*b*) testing whether differential validity/prediction exists. Very little research has addressed what might cause differential validity/prediction. Thus, although I encourage continued research on (*a*) and (*b*), I also strongly encourage research on why differential validity/prediction exists. The factors that might cause differential validity and prediction are mostly the same, as both methods assess differences across subgroups in the relationship between cognitive ability tests and job performance. Berry et al. (2011; see the Possible Causes of Differential Validity section) provided an in-depth review of possible causes, so the interested reader is directed there; a brief summary is provided here. Differences between subgroups in the test–criterion relationship can be caused by any factor that differentially affects test or criterion scores of White versus non-White test takers. So, the existence of predictive bias does not de facto mean that the cognitive ability test is biased; it could just as well mean that the criterion is biased. Berry et al. (2011) listed four factors that could differentially affect test or criterion scores of Whites and non-Whites and thus cause differential validity/prediction: range restriction, psychometric characteristics of the test or criterion (measurement error or measurement bias), contextual influences [e.g., stereotype threat (Steele & Aronson 1995) or the complexity of the job, racism in the workplace], and true differences between subgroups in the role cognitive ability plays in determining job performance. For example, if measurement error affected the test scores of non-Whites more than those of Whites, the attenuation effect would be greater for non-Whites, causing differences between subgroups in the test–criterion relationship. There are doubtless factors other than those reviewed by Berry et al. (2011) that could cause differential validity/prediction. For example, self-efficacy plays a role in the relationship between ability and performance (e.g., Chen et al. 2001); if this role differs by subgroup, this could cause subgroup differences in test–criterion relationships. If racism affects job performance measurement more for non-Whites than for Whites, this could cause differences in test–criterion relationships. The important point is that any factor that affects test or criterion scores differently for two subgroups could cause the test–criterion relationship to differ for those subgroups. Given that the available empirical evidence supports the existence of differential validity/prediction, more research on causes of these phenomena is needed.

Test Bias Research for Other Subgroups and Selection Methods

The vast majority of test bias research in employment settings has focused on cognitive ability tests and comparisons between African American and White subgroups, with much less research focused on Hispanic Americans. However, test bias is a concern whenever there are mean score differences. Thus, investigations of test bias are relevant for any selection methods for which there are mean score differences across subgroups and for any subgroups that exhibit score differences.

There is some test bias research that has investigated sex-based differential validity/prediction of cognitive ability tests (e.g., Canivez & Konold 2001, Schult et al. 2013) and personality tests (Berry et al. 2013a, Saad & Sackett 2002). Given numerous other selection methods that exhibit subgroup differences, such as employment interviews (Huffcutt & Roth 1998) and work samples (Roth et al. 2003), and given that there are other historically disadvantaged subgroups (e.g., Native Americans, immigrants) that may score lower on certain selection methods, test bias research for these other methods and subgroups is needed.

CONCLUSIONS

When mean score differences on a selection test between subgroups cause adverse impact, a natural concern is that the test is biased. In personnel selection settings, predictive bias in the form of differential validity or differential prediction is of particular concern. Although the long-standing conclusions in organizational psychology and organizational behavior had been that, for cognitive ability tests, (*a*) differential validity and differential regression slopes do not exist, and (*b*) differential regression intercepts favor non-Whites by overpredicting job performance, these conclusions have been challenged in recent research that has used more modern analytic approaches. At this point, the available empirical evidence suggests that there are at least small validity and regression slope differences between White and non-White subgroups, but the general conclusion regarding intercept differences probably still holds. Thus, predictive bias likely affects cognitive ability test scores to some degree, creating possible legal concerns for organizations using cognitive ability tests for personnel selection. Availability of quality, current data is a major issue facing this research literature, so I hope that this review, and the likelihood of predictive bias that it highlights, stimulates future research and perhaps even the release of current data by organizations and test vendors.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

AERA (Am. Educ. Res. Assoc.), APA (Am. Psychol. Assoc.), NCME (Natl. Counc. Meas. Educ.). 1999. *Standards for Educational and Psychological Testing*. Washington, DC: AERA

Aguinis H, Beaty JC, Boik RJ, Pierce CA. 2005. Effect size and power in assessing moderating effects of categorical variables using multiple regression: a 30-year review. *J. Appl. Psychol.* 90:94–107

Aguinis H, Culpepper SA, Pierce CA. 2010. Revival of test bias research in preemployment testing. *J. Appl. Psychol.* 95:648–80

Aguinis H, Smith MA. 2007. Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. *Pers. Psychol.* 60:165–99

Aguinis H, Stone-Romero EF. 1997. Methodological artifacts in moderated multiple regression and their effects on statistical power. *J. Appl. Psychol.* 82:192–206

Bartlett CJ, Bobko P, Mosier SB, Hannan R. 1978. Testing for fairness with a moderated multiple regression strategy: an alternative to differential analysis. *Pers. Psychol.* 31:233–41

Berry CM. 2007. *Toward an understanding of evidence of differential validity of cognitive ability tests for racial/ethnic subgroups*. PhD Thesis, Univ. Minn., Minneap.

Berry CM, Barratt CL, Dovalina CL, Zhao P. 2014a. Can racial/ethnic subgroup criterion-to-test standard deviation ratios account for conflicting differential validity and differential prediction evidence for cognitive ability tests? *J. Occup. Organ. Psychol.* 87:208–20

Berry CM, Clark MA, McClure TK. 2011. Racial/ethnic differences in the criterion-related validity of cognitive ability tests: a qualitative and quantitative review. *J. Appl. Psychol.* 96:881–906

Berry CM, Cullen MJ, Meyer JM. 2014b. Racial/ethnic subgroup differences in cognitive ability test range restriction: implications for differential validity. *J. Appl. Psychol.* 99:21–37

Berry CM, Kim A, Wang Y, Thompson R, Mobley WH. 2013a. Five-factor model personality measures and sex-based differential prediction of performance. *Appl. Psychol.* 62:13–43

Berry CM, Sackett PR, Sund A. 2013b. The role of range restriction and criterion contamination in assessing differential validity by race/ethnicity. *J. Bus. Psychol.* 28:345–59

Berry CM, Zhao P. 2015. Addressing criticisms of existing predictive bias research: Cognitive ability test scores still overpredict African Americans' job performance. *J. Appl. Psychol.* 100:162–79

Bobko P, Roth PL, Potosky D. 1999. Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Pers. Psychol.* 52:561–89

Canivez GL, Konold TR. 2001. Assessing differential prediction bias in the Developing Cognitive Abilities Test across gender, race/ethnicity, and socioeconomic groups. *Educ. Psychol. Meas.* 61:159–71

Chen GW, Casper WJ, Cortina JM. 2001. The roles of self-efficacy and task complexity in relationships among cognitive ability, conscientiousness, and work-related performance: a meta-analytic examination. *Hum. Perform.* 14(3):209–30

Gatewood RD, Field HS, Barrick MR. 2008. *Human Resource Selection*. Mason, OH: South-West. 6th ed.

Hartigan JA, Wigdor AK. 1989. Differential validity and differential prediction. In *Fairness in Employment Testing: Validity Generalization, Non-White Issues, and the General Aptitude Test Battery*, ed. JA Hartigan, AK Wigdor, pp. 172–88. Washington, DC: Natl. Acad. Press

Hough LM, Oswald FL, Ployhart RE. 2001. Determinants, detection and amelioration of adverse impact in personnel selection procedures: issues, evidence and lessons learned. *Int. J. Sel. Assess.* 9:152–94

Huffcutt AI, Roth PL. 1998. Racial group differences in employment interview evaluations. *J. Appl. Psychol.* 83:179–89

Hunter JE, Schmidt FL. 1976. Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychol. Bull.* 83:1053–71

Hunter JE, Schmidt FL. 1978. Differential and single-group validity of employment tests by race: a critical analysis of three recent studies. *J. Appl. Psychol.* 63:1–11

Hunter JE, Schmidt FL. 2004. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. New York: Sage. 2nd ed.

Hunter JE, Schmidt FL, Hunter R. 1979. Differential validity of employment tests by race: a comprehensive review and analysis. *Psychol. Bull.* 86:721–35

Hunter JE, Schmidt FL, Le H. 2006. Implications of direct and indirect range restriction for meta-analysis methods and findings. *J. Appl. Psychol.* 91:594–612

Katzell RA, Dyer FJ. 1977. Differential validity revived. *J. Appl. Psychol.* 62:137–45

Kepes S, Banks GC, McDaniel M, Whetzel DL. 2012. Publication bias in the organizational sciences. *Organ. Res. Methods* 15:624–62

Kim A, Berry CM. 2015. Individual differences in social dominance orientation predict support for the use of cognitive ability tests. *J. Personal.* 83:14–25

Lautenschlager GJ, Mendoza JL. 1986. A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Appl. Psychol. Meas.* 10:133–39

Linn RL. 1983. Pearson selection formulas: implications for studies of predictive bias and estimates of educational effects in selected samples. *J. Educ. Meas.* 20:1–15

Mattern KD, Patterson BF. 2013. Test of slope and intercept bias in college admissions: a response to Aguinis, Culpepper, and Pierce (2010). *J. Appl. Psychol.* 98:134–47

McDaniel MA, Kepes S, Banks GC. 2011. The *Uniform Guidelines* are a detriment to the field of personnel selection. *Ind. Organ. Psychol.* 4:494–514

McKay PF, McDaniel MA. 2006. A reexamination of Black-White mean differences in work performance: more data, more moderators. *J. Appl. Psychol.* 91:538–54

Murphy KR, Jacobs RR. 2012. Using effect size measures to reform the determination of adverse impact in equal employment litigation. *Psychol. Public Policy Law* 18:477–99

O'Connor EJ, Wexley KN, Alexander RA. 1975. Single-group validity: fact or fallacy? *J. Appl. Psychol.* 60:352–55

O'Neill KA, McPeek WM. 1993. Item and test characteristics that are associated with differential item functioning. In *Differential Item Functioning*, ed. PW Holland, H Wainer, pp. 255–76. Hillsdale, NJ: Erlbaum

Patterson BF, Mattern KD. 2011. *Validity of the SAT for predicting first-year grades: 2008 SAT validity sample*. Stat. Rep. No. 2011–5, Coll. Board, New York

Ployhart RE, Holtz BC. 2008. The diversity-validity dilemma: strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Pers. Psychol.* 61:153–71

Roth PL, Bevier CA, Bobko P, Switzer FS, Tyler P. 2001. Ethnic group differences in cognitive ability in employment and educational settings: a meta-analysis. *Pers. Psychol.* 54:297–330

Roth PL, Huffcutt AI, Bobko P. 2003. Ethnic group differences in measures of job performance: a new meta-analysis. *J. Appl. Psychol.* 88:694–706

Roth PL, Le H, Oh I, Van Iddekinge CH, Buster MA, et al. 2014. Differential validity for cognitive ability tests in employment and educational settings: not much more than range restriction? *J. Appl. Psychol.* 99:1–20

Rotundo M, Sackett PR. 1999. Effect of rater race on conclusions regarding differential prediction in cognitive ability tests. *J. Appl. Psychol.* 84:815–22

Saad S, Sackett PR. 2002. Investigating differential prediction by gender in employment-oriented personality measures. *J. Appl. Psychol.* 87:667–74

Sackett PR, Borneman MJ, Connelly BS. 2008. High stakes testing in higher education and employment: appraising the evidence for validity and fairness. *Am. Psychol.* 63:215–27

Sackett PR, Schmitt N, Ellingson JE, Kabin MB. 2001. High-stakes testing in employment, credentialing, and higher education: prospects in a post-affirmative-action world. *Am. Psychol.* 56:302–18

Schmidt FL, Hunter JE. 1981. Employment testing: old theories and new research findings. *Am. Psychol.* 36:1128–37

Schmidt FL, Hunter JE. 1998. The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. *Psychol. Bull.* 124:262–74

Schmidt FL, Pearlman K, Hunter JE. 1980. The validity and fairness of employment and educational tests for Hispanic Americans: a review and analysis. *Pers. Psychol.* 33:705–24

Schmitt N. 2014. Personality and cognitive ability as predictors of effective performance at work. *Annu. Rev. Organ. Psychol. Organ. Behav.* 1:45–65

Schult J, Hell B, Passler K, Schuler H. 2013. Sex-specific differential prediction of academic achievement by German ability tests. *Int. J. Sel. Assess.* 21:130–34

Sidanius J, Pratto F. 1999. *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*. New York: Cambridge Univ. Press

SIOP (Soc. Ind. Organ. Psychol.). 2003. *Principles for the Validation and Use of Personnel Selection Procedures*. Bowling Green, OH: SIOP. 4th ed.

Steele CM, Aronson J. 1995. Stereotype threat and the intellectual test performance of African Americans. *J. Personal. Soc. Psychol.* 69:797–811

Whetzel DL, McDaniel MA, Nguyen NT. 2008. Subgroup differences in situational judgment test performance: a meta-analysis. *Hum. Perform.* 21:291–309

Annual Review of
Organizational
Psychology and
Organizational Behavior

Volume 2, 2015

Contents

Errata

An online log of corrections to *Annual Review of Organizational Psychology and Organizational Behavior* articles may be found at http://www.annualreviews.org/errata/orgpsych.