On the Quantification and Generalizability of Differential Prediction in Selection Systems

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Jeffrey Alan Dahlke

IN PARTIAL FULTILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Paul R. Sackett, Advisor
Nathan R. Kuncel, Co-Advisor

May 2019

Acknowledgements

I have so many people to thank for helping me to get to this stage of my education and my career. I'm bound to leave people out, so I'd like to start by offering a blanket acknowledgement of thanks to all of my family, friends, mentors, teachers, and colleagues who have influenced me and helped me to develop into who I am today. That said, I have several specific acknowledgements to share regarding folks whose impacts were especially salient as I reflected on my graduate studies and dissertation research. These acknowledgements represent a journey, so I'll organize them chronologically.

First and foremost, I would like to thank my parents, Alan and Lori Dahlke, for their love and support throughout all of my personal, educational, and professional pursuits. I have always appreciated their trust in my judgment as I navigated through my undergraduate, master's, and doctoral studies. After all, it takes a special set of parents not to initiate a reality-check conversation with their son when he sets out to pursue a piano performance degree (still, I imagine they breathed a heavy sigh of relief when I found my way into a discipline with stronger career prospects!). They have been immensely supportive of me no matter where I roamed, making many trips to Mankato, Minneapolis, and even Germany to visit me. I'm also thankful to my sister, Kimberly Kroll, for putting up with me as her bratty little brother for all of these years.

Looking back on the path that led me to graduate studies in industrial and organizational (I-O) psychology, I owe a special thanks to Dr. Stuart Korshavn of St. Norbert College, as he was the one who helped me find this field in the first place. When I arranged a meeting with him to discuss the idea of pursuing graduate training in social

psychology (Stuart's own discipline) during my junior year, I certainly did not expect that he would end up directing me toward I-O psychology instead. I'm extremely grateful to Stuart for helping me to find I-O psychology and I've never looked back since he gave me the initial nudge that ultimately led me to where I am now.

My graduate studies in I-O psychology took place in two stages and, from the first stage, I owe much thanks to the faculty of the master's program at Minnesota State University, Mankato for setting me up for future success: Drs. Dan Sachau (my master's thesis advisor), Andi Lassiter, Kristie Campana, and Lisa Perez. Although the MNSU program awards terminal master's degrees, they raised the idea of pursuing a Ph.D. to me before I ever mentioned an interest in further education. I still remember sitting in the MSP airport in fall of 2013 waiting for a flight to visit alumni at DC-area consulting firms when Andi asked, of the blue, "so, have you thought about getting a Ph.D.?" I was only about two months into the program at that point, so this early vote of confidence in my potential meant a great deal and galvanized my commitment to earn a doctorate.

I owe a great deal of thanks to my doctoral advisor, Dr. Paul Sackett, who has played a pivotal role in my cultivation as a psychologist. Paul has been a constant mentor from my first project with him in the summer before my first semester at the University of Minnesota to the culmination of my dissertation work (and beyond). Paul's passion for research on issues of fairness and bias in selection has inspired me greatly and was a major influence in my dissertation work. I'm glad to have had Paul's unwavering support as I pursued my research ideas and as I searched for a career path that matched my values and goals. I hope our collaborations continue far into the future!

I also owe special thanks to my doctoral co-advisor, Dr. Nathan Kuncel. He was the one who called me to tell me I was accepted at the University of Minnesota and I'm so glad he has stuck with me as my co-advisor throughout my U of M journey and co-authored several papers with me. I am indebted to both Nathan and Paul for providing me with access to their College Board database for use in Study 3 of my dissertation (and in several other projects) – colleagues with data are good colleagues, indeed.

I am sincerely grateful to Drs. Michael Rodriguez and Richard Landers for serving on my doctoral examination committee. Each of them offered valuable insights and questions during our meetings, which made for very interesting discussions. I had the pleasure to take two courses with Michael in the Educational Psychology department and I was delighted to have him bring his expertise from outside of I-O psychology to my committee. I have enjoyed the opportunity to get to know Richard over the past year since he joined the University of Minnesota's I-O psychology faculty, and a part of me wishes my graduate studies weren't coming to an end so that I could have more time to learn from him about technology's implications for the world of work.

Last, but certainly not least, I would like to acknowledge my immense gratitude to the Human Resources Research Organization (HumRRO). I had the distinct privilege to receive HumRRO's 2018 Meredith P. Crawford Dissertation Fellowship, which helped to fund me during my dissertation year. I am deeply appreciative of the recognition and support I received from HumRRO as a Crawford fellow and I am thrilled to be joining the HumRRO staff as a Research Scientist in Alexandria, Virginia after graduation.

Abstract

Differential prediction analyses are important for personnel psychologists to determine whether the regression lines linking a predictor variable to a criterion/performance variable are comparable between a referent group and a legally protected focal group. Although many decades of research on cognitive tests has indicated that differential prediction does occur for racial/ethnic minority groups in the U.S. relative to Whites, the bulk of evidence has indicated that these differences result into the overprediction of Black and Hispanic individuals' performance from cognitive test scores, which does not indicate predictive bias against these groups. However, research published over the past decade by Aguinis Culpepper, and Pierce (2010; 2016) has questioned the accuracy and generalizability of past findings, arguing that the historic trends could have been caused by statistical artifacts. In a series of four studies, I present methodological advancements in the quantification of differential prediction and supply substantive analyses that refute the findings reported by Aguinis et al. (2010; 2016). Specifically, I (1) offer derivations of simplified effect-size estimation procedures for differential prediction analyses with accompanying standard-error estimators, (2) illustrate the effects of composite predictors on differential prediction effects, (3) demonstrate the generalizability of White-minority and male-female differential prediction in the post-secondary education admissions domain, and (4) present findings from a simulation study designed to identify which features of selection systems could cause statistical artifacts to bias the results of differential prediction analyses conducted on cognitive test scores.

Table of Contents

List of Tables

List of Figures

Introduction

People take high-stakes assessments at various points in their lives in the hopes that their scores will help them to access opportunities that will advance their education, careers, or both. For instance, high school students in the United States may take the SAT or ACT to compete for spots at colleges and universities; college students may take the GRE, LSAT, GMAT, or MCAT to pursue graduate or professional training; and job applicants may complete cognitive tests, personality inventories, simulations, interviews, or any number of other assessments as they vie for employment opportunities. In each of these example settings, an individual's score on an assessment plays some role in determining whether or not they will be selected for an opportunity and much potentially hangs in the balance. In high-stakes testing programs such as these, researchers and practitioners pay a great amount of attention to whether the assessments function similarly across subgroups of test takers. If the assessments function properly, opportunities will be offered to those who demonstrate the greatest potential to perform well, regardless of applicants' demographic backgrounds; however, if an assessment does *not* function properly, individuals could be unfairly denied opportunities through no fault of their own because the test does not relate to performance the same way across demographic groups. This is the problem that test developers and test users hope to avoid when they seek to answer the question "Does the same test score correspond to the same level of predicted performance across all relevant demographic groups?" If individuals from different groups who have the same test score have different levels of anticipated performance, it indicates that the test demonstrates "differential prediction;" if the

differences in prediction put historically underprivileged groups at a disadvantage, these differences indicate "predictive bias."

My present research examines several important issues related to the accuracy with which industrial and organizational (I-O) psychologists can detect differential prediction and draw interpretations about whether a predictor exhibits predictive bias. In a series of four studies, I develop ideas related to the accurate quantification and detection of differences in prediction. These studies include derivations of updated effect sizes for quantifying the magnitudes of differential prediction effects, demonstrations of how these effect sizes are affected by the formation of composite predictor variables, analyses to determine whether differences in prediction generalize across settings, and examinations of whether statistical artifacts such as range restriction and criterion measurement error can obscure subgroup differences in prediction. The remainder of the Introduction chapter presents foundational background information relevant to my series of investigations, including relevant statistical methods, historical findings regarding differential prediction, and summaries of prior research exploring the effects of statistical artifacts on predictive bias analyses. However, the most fundamental background information of all concerns what is meant by terms such as "fairness" and "bias."

Defining Fairness and Bias

"Fairness" is a desirable characteristic of any process that distributes resources, but what exactly it means for a process to be "fair" is not always so easily defined. Is it fair to allocate a particular resource based on merit, based on need, or equally, such that everyone gets the same amount? Clearly, there is no single answer that applies to all

scenarios: Fairness can vary in definition across cultures, situations, and individuals. The preferred definition of fairness in a given context can also depend on the importance of the resource in question. Just as the broad idea of fairness does not have a universal definition, psychologists have devised several criteria for what it means for a psychological test to be "fair." As stated in the Society for Industrial and Organizational Psychology's (SIOP) *Principles for the Validation and Use of Personnel Selection Procedures* (hereafter simply "the *Principles*"), "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" (2018, p. 38). The *Standards for Educational and Psychological Testing* (hereafter simply "the *Standards*"), a document compiled jointly by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), offers four possible meanings of fairness in the testing context: Equal group outcomes, equitable treatment during testing, comparable acccess to constructs, and lack of bias (AERA, APA, & NCME, 2014).

The first and simplest definition of fairness offered in the *Standards* (AERA, APA, & NCME, 2014) is that all groups receive equal outcomes, such as equal mean scores, equal pass rates, or equal rates of selection to receive opportunities. For example, an organization offering jobs to members of minority racial/ethnic groups at a lower rate than members of a majority group (a phenomenon known as "adverse impact") would be unfair according to this definition. This definition is rejected in the *Standards* because not

all instances of inequality are necessarily unfair; however, although not indicative of unfairness in-and-of themselves, unequal group outcomes should invite scrutiny of deeper issues associated with other conceptualizations of fairness. The second potential definition is equitable treatment in the testing process; for example, equitable treatment could include "testing conditions, access to practice materials, performance feedback, retest opportunities, and other features of test administration" (SIOP, 2018, p. 38). This definition of fairness is important for the test-administration process, but is not what is meant by "fairness" in the context of my present research. Third, fairness can be defined as "comparable access to the constructs measured by a selection procedure" (SIOP, 2018, p. 38); in other words, the measurement of an individual's standing on a construct should not be affected by their other attributes, such as demographic characteristics. The fourth definition of fairness offered in the *Standards* and echoed in the *Principles* is "lack of bias," which includes both measurement bias (also known as lack of measurement invariance) and predictive bias. My present research is focused on fairness operationalized as a lack of predictive bias, which means that a test is considered fair if "a common regression line can be used to describe the predictor-criterion relationship for all subgroups of interest" (SIOP, 2018, p. 39).

It is important to reiterate that, within the context of my present research, "bias" of predictor scores refers to predictive bias and should not be confused with any of the first three definitions of fairness outlined above, nor should it be confused with the concept of measurement bias. Measurement bias refers to "sources of irrelevant variance that result in systematically higher or lower scores for members of particular groups"

(SIOP, 2018, p. 42); in other words, measurement bias occurs when individuals from different groups who have the same latent or "true" score on a construct reliably receive different scores on an assessment built to assess that construct. My concentration on predictive bias means that my research focuses on differences between the subgroup regression lines that characterize the association between a predictor variable (e.g., test scores) and a criterion variable (e.g., college grades or job performance). Although measurement bias can be a precipitating factor that causes predictive bias to occur, it is a distinctly different issue; whereas measurement bias is concerned with the internal measurement quality of an assessment, predictive bias is concerned with how the scores on an assessment relate to external performance criteria.

Having established the operational definitions of fairness and bias that will be used in the present research, the next step is to clarify how researchers go about identifying evidence of predictive bias and how differential prediction is operationalized in statistical models. A number of competing definitions of bias were offered by psychologists during the 1960s and 1970s (e.g., Cleary, 1968; Cole, 1973; Darlington, 1971; Thorndike, 1971), but the modern definition that is endorsed in both the *Standards* (AERA, APA, & NCME, 2014) and the *Principles* (SIOP, 2018) was articulated by Cleary (1968).

The Cleary Model of Predictive Bias

Psychologists have worked to develop a definition of predictive bias that relies on testable statistical relationships among subgroup membership, predictors, and criteria. The most widely accepted definition of predictive bias among I-O psychologists is based

on multiple moderated regression and was formulated by Cleary (1968); this definition was also contemporaneously recommended by Anastasi (1968). The Cleary model of bias specifies that a test is not biased against a focal group (i.e., a group that is purported to be historically underprivileged/disadvantaged; e.g., a racial minority group) if the unstandardized regression line characterizing the predictor-criterion relationship in the focal group is equal to regression line that describes the predictor-criterion association in a referent group (i.e., a group that is purported to be privileged/advantaged; e.g., a racial majority group). In other words, there is no evidence of predictive bias if the linear associations between predictor and criterion scores across groups share the same intercept and slope. Additionally, only differences in prediction between focal and referent subgroups' regression lines that lead to the underprediction (i.e., underestimation) of focal-group performance technically indicate bias *against* the focal group. Interpretations regarding predictive bias are based on empirical evidence regarding whether a given test score corresponds to the same level of expected performance, regardless of one's group membership. Given the Cleary model's basis in linear regression, a brief review of linear regression is in order to establish the fundamentals underlying the implementation of Cleary's method.

Background information on the usage of linear regression. In linear regression, observed criterion scores are decomposed into explained (i.e., "systematic") and unexplained (i.e., "error" or "residual") variance. When considering the criterion score of person i (denoted as Y_i), the explainable part of this score is expressed as $\hat{Y}_i$ and the unexplained part is expressed as e_i. In a simple linear regression model with a single

predictor, the "fitted" or "predicted" $\hat{Y}_i$ value is found using a prediction equation like the one shown in Equation 1,

$$\hat{Y}_i = b_0 + b_1 X_i \quad (1)$$

where b_0 is the intercept of the model and represents a constant value added to predicted criterion scores, X_i is the *i*th person's score on the predictor, and b_1 is the slope coefficient used to project predictor scores onto the criterion space to explain variance in observed criterion scores. The intercept and slope coefficients are computed in such a way that they explain as much of the variability in observed criterion scores as possible and represent the line of best fit (in the least-squares sense) through a bivariate distribution of predictor and criterion scores. Once the fitted $\hat{Y}_i$ values have been determined, the residual e_i scores are simply the part of Y_i that is left over after accounting for $\hat{Y}_i$, as shown in Equation 2.

$$e_i = Y_i - \hat{Y}_i \quad (2)$$

The correlation between the observed and fitted criterion values provides an indication of the regression model's fit to the data. By squaring this correlation, the resulting R^2 value (i.e., coefficient of determination) indicates what proportion of the observed variation in criterion scores is accounted for by the model, ranging from 0 (i.e., the predictor explains nothing about the criterion) to 1 (i.e., the predictor perfectly explains the criterion). R^2 can also be computed as a ratio of the variance of fitted criterion values to the variance of observed criterion values, or as one minus a ratio of the variance of residual scores to the variance of criterion scores (see Equation 3).

$$R^2 = \text{cor}\left(Y, \hat{Y}\right)^2 \qquad (3)$$

$$= \frac{\text{var}(\hat{Y})}{\text{var}(Y)}$$

$$= 1 - \frac{\text{var}(e)}{\text{var}(Y)}$$

When a regression equation is applied to the data set from which it was derived, the $\hat{Y}_i$ scores explain maximal variance in Y_i scores and the vector of residual scores will have a mean of zero. However, when a regression equation is applied to a new data set, the fit of the $\hat{Y}_i$ values will be suboptimal for that data set because the intercept and slope were influenced by idiosyncrasies of their original derivation data set; these idiosyncrasies of the derivation data do not translate to new data sets, which results in worse fit when the model is applied to new data (this is known as "shrinkage," because the R^2 value is only maximized for the derivation data and it is always smaller when the regression equation is applied to new data). Additionally, the mean of residuals can be non-zero when a regression equation is applied to a new data set. If the mean residual value observed when a regression equation is applied to different data deviates significantly from zero, it signals that the regression equation used is not sufficient to explain the association between *X* and *Y* in the new data set and that the new data set requires a different regression equation to adequately characterize its *X*-*Y* relationship. If two data sets can be described by the same slope but require significantly different intercepts, the differences between the data sets may be said to have a main effect on the criterion; however, if two data sets require significantly different slopes to describe the *X*-

Y association, the differences between the data sets may be said to moderate the effect of *X* on *Y*.

Regression models used in the Cleary framework. The logic of the Cleary model is based on the notion that, if a predictor yields unbiased predicted criterion scores, a single regression equation should be adequate to explain the *X*-*Y* relationships observed in data collected from different subgroups. According to Cleary (1968), "A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test is designed, consistent nonzero errors of prediction are made for members of the subgroup" (p. 115). Thus, if one computes a separate regression equation to explain the association between *X* and *Y* in each of two groups, the predictor is said to be unbiased if the subgroups have statistically indistinguishable slopes and intercepts. In order to test whether subgroups have unequal slopes or intercepts, the Cleary model relies on comparing the fit of three different regression models using multiple moderated regression. I describe each of these models below in order of increasing complexity and then describe the procedure for comparing and interpreting differences among the models in the following subsection.

The three regression models evaluated in the Cleary framework are built by progressively adding variables to a linear model. In Model 1 (depicted in Equation 4), one begins by simply regressing the criterion on the predictor of interest.

$$Y_i = b_0 + b_1 X_i + e_i \qquad 4$$

This model represents the Cleary model's null hypothesis that a single regression line describes both groups' data adequately; more complex models that include group membership information can be compared to this model to test the null hypothesis.

In Model 2 (depicted in Equation 5), one adds a dummy variable predictor indicating the group membership of each individual represented in the data set. This variable is commonly coded as "0" for members of the referent group (e.g., the majority group) and "1" for members of the focal group (i.e., a minority group). This dummy variable allows the analyst to determine whether group membership has a main effect on the criterion.

$$Y_i = b_0 + b_1 X_i + b_2 G_i + e_i \quad (5)$$

If there is a significant group main effect, it indicates that the intercept for the referent group (which, in this model is indicated by b_0) does not work for the focal group and the focal group requires its own intercept (which is computed as $b_0 + b_2$).

In Model 3 (depicted in Equation 6), one adds a third and final predictor variable representing the group-by-predictor interaction term. This variable is simply the product of the dummy variable *G* and the predictor *X*. The product variable's inclusion in the model allows the analyst to determine whether the referent and focal groups' data require different slopes to explain the *X*-*Y* relationship.

$$Y_i = b_0 + b_1 X_i + b_2 G_i + b_3 GX + e_i \quad (6)$$

If the interaction term explains significant variance, it indicates that the slope for the referent group (which is indicated by b_1) does not work for the focal group and the focal group requires its own slope (which is computed as $b_1 + b_3$).

Comparing regression models in the Cleary framework. After one has fit all three of the regression models described above, the models can be compared to each other to determine whether the models that include subgroup-specific information provide a better fit to the data than does the simple regression model from Model 1. These comparisons are made via hierarchical linear regression analyses (not to be confused with hierarchical linear modeling), which computes *F* tests for the differences in R^2 values between nested regression models. A significant *F* test indicates that the more complex model fits the data better than the simpler model.

Since the Cleary model was first introduced, various researchers have offered guidance regarding the order in which Models 1, 2, and 3 should be compared, with different sequencings of comparisons having important implications for statistical power. The earliest suggestions relied on what is known as "step-up" model comparisons in which models are compared in order of increasing complexity. For example, Bartlett, Bobko, Mosier, and Hannan (1978) recommended comparing Model 2 to Model 1 to determine whether there were intercept differences and then, if those models were significantly different, comparing Model 3 to Model 2 to determine whether there were slope differences. Lautenshlager and Mendoza (1986) noted that sequencing of tests in the step-up testing procedure has problematic implications for statistical power because "at each step all higher order effects not included in the model are pooled into the sum of squared error term (SSE), potentially decreasing the power of the sequential testing procedure" (p. 134). They argued that higher-order effects (i.e., the group-by-predictor interaction) should be included early in the model-comparison procedure to minimize the

error term and increase power. Lautenshlager and Mendoza also noted that the step-up procedure relies on null hypotheses that do not match the purpose of the Cleary model: Rather than testing the null hypothesis that subgroups have equal regression lines, the step-up procedure tests for differences in intercepts and slopes in a sequential fashion.

To overcome the limitations of step-up model testing, Lautenshlager and Mendoza (1986) formulated a step-down procedure that is the basis for modern applications of the Cleary model (cf. Aguinis, Culpepper, & Pierce, 2010; Rotundo & Sackett, 1999). First, to test whether there are any differences in prediction between the two groups, one must compare Models 1 and 3. If these models are significantly different, it means that the subgroups' data cannot be adequately described using a single regression line. At the very least, Model 3's superior fit means that the groups likely require separate intercepts, but they may also require different slopes. Next, to determine whether the group-membership dummy variable moderates the association between *X* and *Y* (i.e., the groups exhibit slope differences) or it simply has a main effect on *Y* (i.e., the groups exhibit intercept differences), one must compare Models 2 and 3. If Model 3 fits the data significantly better than Model 2, one can conclude that differential prediction has occurred in the form of slope differences. If Model 3 does not demonstrate superior fit, one must compare Model 2 to Model 1 to determine whether there are significant differences in intercepts. Note that, even if Model 3 differs from Model 1, it is possible that one would fail to detect differential prediction if Model 2 does not differ from Model 3 or Model 1.

If one cannot reject the null hypothesis that Models 1 and 3 are equal, the fact that the subgroups' data can be described by a single regression line means that the predictor yields unbiased predictions of performance for both groups. However, if one finds evidence of different intercepts or different slopes, all that can be concluded without further inquiry is that the groups do not have equal regression lines: The presence of differential prediction is not automatically indicative of predictive bias. To determine whether the differential prediction detected by the Cleary analyses represents predictive bias, one must examine the subgroup regression lines. This can be done by consulting the regression coefficients directly, but it is generally easier to interpret differential prediction by plotting the subgroup regression lines on a cartesian plane so as to visually inspect the patterns of differences. The following subsection describes forms of differential prediction and how one can determine which forms indicate predictive bias.

Forms of differential prediction. As described above, differential prediction occurs when the regression equations for two subgroups (e.g., the equations for majority and minority racial subgroups) are different, which means that a given score on a predictor variable is associated with different levels of predicted performance as a function of subgroup membership. However, not all differences between subgroup regression models indicate bias *against* the focal subgroup; some of these differences could be advantageous for the focal subgroup while others could be disadvantageous. Some researchers (e.g., Aguinis et al., 2010) view any differences in predicted performance between subgroups as problematic. According to this perspective, any differential prediction between subgroups indicates predictive bias against some group,

even if that group is the referent group. Other researchers (e.g., Cleary, 1968; Sackett, Schmitt, Kabin, & Ellingson, 2001) take the view that only disadvantageous differences wherein the focal subgroup's equation predicts better performance for focal-group members than does the referent group's equation (i.e., the referent model "underpredicts" focal-group performance) technically indicate bias against the focal group. I endorse the latter perspective and operationalize predictive bias as the underprediction of the focal subgroup's performance using the referent subgroup's regression line.

When differential prediction occurs in selection systems, it comes in one of two varieties: intercept differences or slope differences. Intercept differences occur when subgroup lines have substantially equal slopes but intersect with the *Y* axis at different points. Slope differences, on the other hand, occur when the subgroups' slopes are significantly different, regardless of differences in subgroup intercepts. Although some authors argue that the co-occurrence of slope and intercept differences constitutes a third form of differential prediction (cf. Aguinis et al., 2010), this is not a technically sound idea: It is poor statistical practice to interpret main effects in the presence of an interaction, and intercept differences are simply main effects of group membership. Therefore, if one detects slope differences, one need not be explicitly concerned with intercept differences, as the point at which the *Y* axis intersects with the *X* axis is arbitrary. When slopes differ, the significance of the intercept-difference coefficient in Model 3 may be dependent upon whether or not one has chosen to center the predictor scores and, if so, how one has chosen to perform the centering. Regardless of the form that differential prediction takes, the relative positions of the subgroup regression lines on

a coordinate plane are important because one must use these lines to determine which group is disadvantaged by the predictor's relationship with the criterion.

When intercept differences occur, the subgroups' regression equations have equal slopes, but one group has consistently higher predicted performance than the other group (see Figure 1a). If subgroups have different intercepts and the focal group's intercept is higher, it indicates focal-group underprediction and constitutes evidence of predictive bias. When slope differences occur, the possible configurations of subgroup regression lines are innumerable. The most problematic slope differences occur when subgroup regression lines cross within the operational range of scores and the focal group's performance is overpredicted in part of the predictor-score range and underpredicted in the rest of the score range (see Figure 1b). If subgroup regression lines cross within the operational range of predictor scores, it constitutes evidence of predictive bias (Berry, 2015). If the subgroup lines do not cross, the focal group's performance may be entirely overpredicted (see Figure 1c) or entirely underpredicted (see Figure 1d), but the magnitude of the overprediction or underprediction will vary across the range of predictor scores. If the focal group's performance is consistently underpredicted within the operational range of predictor scores, this constitutes evidence of predictive bias.

Requirements for predictive bias analyses from the *Principles*. The fifth edition of the *Principles* (SIOP, 2018) specifies five requirements for predictive bias analyses; three of these recommendations contribute to the analysis practices recommended to this point. Unless these requirements are satisfied, it would be unwise for an analyst to rely on the Cleary model as defined above. The three requirements

related to already-reviewed concepts are described below; the remaining two requirements will be discussed momentarily in the context of their research origins.

First and most importantly, predictive bias analyses must use an unbiased criterion variable. If the criterion is biased in some way (e.g., members of focal groups are evaluated on different standards than members of the referent group, or criterion scores are based on qualitatively different types of performance for referent and focal groups), it is impossible to determine whether any observed differences in subgroup slopes or intercepts are due to problems with the predictor or simply problems with the criterion. Bias in the criterion can cause differential prediction to be detected when there is nothing wrong with the predictor; similarly, bias in the criterion can cause differential prediction to go undetected when a predictor is actually problematic. An example of this was reported by Saad and Sackett (2002), who used U.S. military data to examine differential prediction between males and females. They found that the "Effort and Leadership" performance dimension was overpredicted for females across 90% of the combinations of jobs and personality-based predictor variables they examined; this consistency of overprediction for a single criterion across different contexts and predictors signals that the criterion itself may be biased. Saad and Sackett noted that ratings on this dimension are based on performance in combat situations and that military policies restrict women's combat roles; as such, sex differences in performance opportunities are very likely to correspond to lower mean ratings for women than men on this dimension and contribute to the overprediction of female performance.

Second, the *Principles* (SIOP, 2018) specify that predictive bias analyses should be performed using samples that allow adequate statistical power to detect slope and intercept differences. Power in this context is not simply a matter of sample size, it is also a matter of the proportionality of the referent and focal groups in the analysis. Holding sample size constant, larger departures from a 50/50 split of cases between the referent and focal groups correspond to less powerful analyses. Settings in which the focal group in a predictive bias analysis is a clear minority group (relative to the number of referent-group cases) will require larger total sample sizes to achieve adequate power.

Third and finally, as predictive bias analyses are based on linear regression analyses, the assumption of homogeneity of error variances (i.e., normally distributed residuals with constant conditional variance across the predictor-score range) applies to all applications of the Cleary model. If this assumption is violated, significance tests associated with the regression models may not support valid inferences.

Definitions of Bias Rejected by Psychologists

As noted at the outset, there are many possible definitions of "fairness" in the testing context according to the *Standards* (AERA, APA, & NCME, 2014), only one of which deals with predictive bias. Based on the variety of ways in which psychologists have attempted to define fairness, it is perhaps no surprise that psychologists have not always agreed about how to define "test bias." Although Cleary's (1968) definition of predictive bias is preferred among personnel selection experts and has been recommended in both the *Standards* (AERA, APA, & NCME, 2014) and the *Principles* (SIOP, 2018), it is not the only definition of test bias that has been proposed. Several

other methods for operationalizing bias have been rejected because those methods do not align with the operational purpose of differential prediction analyses. The most prominent of these are described below to further clarify the Cleary-based definition of predictive bias used in the present research.

Quota definition. The simplest attempt at a definition of test bias is the "quota" definition, which states that a test is fair only if subgroups have equal mean scores (and, ideally, equal standard deviations), such that members of the subgroups would be selected at equal rates if selection were performed in a top-down fashion on the basis of test scores (Hunter, Schmidt, & Rauschenberger, 1977). This conceptualization of bias is identical to the idea of adverse impact in selection systems: A selection system exhibits adverse impact against a group (e.g., racial minority applicants or females) if members of that group are selected at a substantially lower rate than a referent group (e.g., White applicants or males). This definition fails to account for the fact that adverse impact can occur independently of any subgroup differences in predictor-criterion associations (e.g., subgroups can have equal regression lines but large predictor mean differences; it is also possible for there to be no mean differences on the predictor, yet wildly different subgroup regression lines exist). Although most selection professionals do aim to reduce or eliminate adverse impact in the interest of supporting workforce diversity, the idea that mean differences in predictor scores across subgroups indicate bias in the test is "unequivocally rejected within mainstream psychology" (Sackett, Borneman, & Connelly, 2008, p. 222).

Thorndike's constant ratio definition. According to Thorndike's (1971) definition of bias, "a test is fair if and only if the percentage of minorities selected with the test is equal to the percentage of minorities who would be successful if selection were conducted on a perfectly valid test or on the criterion measure itself" (Hunter et al., 1977, p. 245). In other words, Thorndike's definition stipulates that a test is unbiased only if the correlation between test scores and a dummy variable representing race is equal to the correlation between criterion scores and race; or, equivalently, a test is unbiased if there are equal magnitudes of standardized subgroup mean differences on the predictor and the criterion. Hunter et al. (1977) noted that this definition requires that tests overpredict minority performance unless there is perfect validity or there are no subgroup mean differences. This definition is inadequate because it is only concerned with patterns of mean differences without considering whether the predictor-criterion relationship is comparable across subgroups.

Equal validity definition. Similar to how the quota model is an inadequate definition of bias despite the fact that adverse impact is an important phenomenon in personnel selection, subgroup differences in predictor validity (i.e., "differential validity") are also not indicative of predictive bias despite being an important consideration in validation research. Differential validity occurs when predictor-criterion correlations computed within subgroups differ across subgroups, such that the predictor exhibits a different level of predictive strength as a function of group membership. The study of differential validity is important for predictor validation because it provides a coarse indication as to whether the construct assessed by the test is invariant across

subgroups in terms of its relationship with a criterion, under the assumption that the criterion variable is unbiased. However, the examination of differential validity requires that both the predictor and the criterion are converted to *Z* scores separately within each group, which means that subgroup variation in the means and variances of the predictor and the criterion are discarded in the process. Given that predictive bias is focused on the comparability of operational *unstandardized* subgroup predictor-criterion relationships, differential validity analyses cannot provide a useful indication of bias when considered in isolation.

Linn (1978) termed differential validity a "pseudoproblem" because differences in subgroup validity coefficients are primarily a distraction from the real issue of differential prediction addressed by the Cleary model. Subgroup validity coefficients can differ even when the slopes of unstandardized subgroup regression models are equal, or vice-versa. Consider that the slope in a simple single-predictor regression model can be computed using the formula shown in Equation 7,

$$b_1 = r_{XY}\frac{SD_Y}{SD_X} \tag{7}$$

where b_1 is a slope, r_{XY} is a validity coefficient, SD_Y is a criterion standard deviation, and SD_X is a predictor standard deviation. Validity differences only directly correspond to slope differences if subgroups have equal SD_Y/SD_X ratios. Subgroups could have equal validities yet different slopes if they have unequal SD_Y/SD_X ratios; similarly, subgroups could have different validity coefficients yet have equal slopes if their SD_Y/SD_X ratios offset the validity differences. Thus, although differential validity analyses can be important for understanding how the strength of the predictor-criterion relationship

differs across groups, focusing on the issue of differential validity at the expense of examining differential prediction is unwise because it does not convey whether differences in validity correspond to differences in predicted levels of performance across groups in operational selection systems.

Darlington's "Definition 3" and its special-case formulations. According to Darlington's (1971) definition of bias[1], "a test is fair only when the partial correlation between the cultural variable and a test with the criterion measure held constant is zero" (Hunter et al., 1977). In Darlington's model, a test is unbiased if a single regression line describes all subgroups' data when test scores are regressed on criterion scores. This is similar to the Clearly model, but with reverse regression: Instead of regressing the criterion on the predictor as in the Cleary model, the predictor is regressed on the criterion. This model is not used in modern differential prediction analyses because reverse regression makes it incompatible with the goals of operational selection systems: The primary goal of a selection system is to predict performance from test scores, not to predict test scores from performance, and a practical definition of predictive bias must be constructed accordingly. It is not useful for applied selection programs to quantify predictive bias using reverse regression models if, in actuality, operational predictions are made using conventional regressions of criteria on predictors. A proper test of predictive

---

[1] Technically, this is known as Darlington's "Definition 3," as he proposed four competing definitions. The number of this definition is omitted here because it is the only one of Darlington's definitions that I discuss.

bias must evaluate the equivalence of the subgroup regression models that would be used by selection professions to forecast performance from predictor scores, as is done in the Cleary model.

Jencks (1998) described a phenomenon he termed "selection system bias" that is quite similar to Darlington's definition of bias, but with a special emphasis on the substantive mechanisms that could give rise to bias. Jencks' selection system bias can occur when (1) performance is a function of both cognitive and non-cognitive attributes, (2) cognitive attributes are measured more easily than non-cognitive attributes (such that non-cognitive attributes are therefore omitted from the selection system, leaving the system to focus on cognitive attributes), and (3) the mean differences in cognitive attributes are larger than the mean differences in other attributes. Selection system bias is said to occur when the standardized mean difference in predictor scores is larger than the standardized mean difference in job performance, which can occur if a predictor fails to capture all of the determinants of performance.[2] Jencks also stated that "[s]election system bias… exists when blacks and whites who would perform equally well if they got a job have different chances of getting it. If a firm relies entirely on test scores to select workers, selection system bias will arise whenever the standardized racial gap in job performance is smaller than the standardized racial gap in test performance" (p. 77). In other words, selection system bias occurs when the probability of selection, conditional

[2] Note that this component of Jencks' (1998) definition is also similar to the Thorndike definition of bias.

on performance, is unequal between groups, which is effectively a restatement of the Darlington definition of bias.

A special case of Darlington's definition was proposed by Cole (1973), whose conditional probability model stated that a predictor is unbiased if subgroups have equal probability of selection given an equal probability of success. This is identical to the Darlington reverse-regression definition of bias when applied to dichotomous variables and it shares all of the problems of Darlington's definition. This dichotomous formulation of the Darlington definition was also articulated years earlier by Guion (1966).

Rationale for the modern preference of the Cleary model over competing definitions of bias. The models described above were all proposed as possible definitions of bias in the years following the passage of the Civil Rights Act of 1964 in an attempt to operationally define the characteristics of an unbiased selection system. As noted earlier, Cleary's (1968) model received the greatest degree of support and remains the basis for predictive bias analyses to this day. The Cleary model, and other analyses based upon its framework, tests whether those who have identical test scores have an equal anticipated likelihood of success, irrespective of group membership. Cleary's model is preferable to the rejective alternatives outlined above because (a) it accounts for the fact that mean differences on the predictor, considered in isolation, only indicate the adverse-impact potential of a predictor without providing any indication of how the predictor functions in a predictive model (cf. the quota definition), (b) equal mean differences on the predictor variable and the criterion variable do not indicate that there are comparable predictor-criterion relationships between groups (cf. Thorndike's model), (c) any differences that

may exist between subgroup-specific validity coefficients do not necessarily indicate that there are also differences between the slopes of subgroup's unstandardized linear regression formulas that would be used in operational prediction models (cf. the equal validity definition), and (d) the Cleary model applies regression analyses in a way that aligns with how linear models are applied in operational selection programs (i.e., the criterion variable is regressed on the predictor variable, rather than the predictor being regressed on the criterion as in the Darlington definition). In short, the Cleary model is favored because it offers the clearest way to test whether subgroup's unstandardized regression equations are equal when predicting a criterion from a predictor, which addresses the question of real interest when selection systems are examined for unbiased functioning across legally protected subgroups.

Effect Sizes for Quantifying Magnitudes of Differential Prediction

The Cleary model, as described above, relies on a rather mechanical process of analyzing data to determine whether differential prediction is evident and, if so, whether it is indicative of predictive bias against the focal group. However, there is a problem with relying on such a mechanical process driven by statistical significance testing: It ignores the *magnitude* of differences between groups' regression models. Without considering the effect size associated with differences in prediction, it is possible that very small differences in prediction could trigger significant Cleary model results if an organization has a large enough employee sample to allow adequate statistical power, even if the magnitude of the differences in prediction would not considered "practically significant." It is helpful to have an effect-size metric to accompany any statistical

analysis so that the practical significance of the effect can be evaluated along with evidence of statistical significance; differential prediction analyses are no exception to this. In fact, as the final stage of their step-down implementation of the Cleary model, Lautenshlager and Mendoza (1986) recommended that researchers "Verify whether practically meaningful differences in prediction occur over [the] observed score range" (quote from their Figure 1, p. 136).

Nye and Sackett (2017) recently proposed a class of standardized effect sizes called "d_{Mod}" (which signifies the standardized mean difference, or *d*, between moderated regression models) to quantify categorically moderated regression effects. These effect sizes are based in the Cleary framework and can help researchers to understand the magnitude of differential prediction effects. The d_{Mod} class of effect sizes includes a signed effect-size measure, d_{Mod_Signed}, for which the sign on the result indicates whether a focal group's performance is overpredicted or underpredicted, on average. When d_{Mod_Signed} is used in settings in which subgroup regression lines cross within the operational range of predictor scores, instances of overprediction and underprediction can cancel out and d_{Mod_Signed} will reflect the net difference in prediction over the entire range of predictor scores. Expressed in slope-intercept form, Nye and Sackett's (2017) signed effect size is shown in Equation 8,

$$d_{Mod\_Signed} = \frac{1}{SD_{Y_1}} \int f_2(X)[X(b_{11} - b_{12}) + b_{01} - b_{02}]\mathrm{d}X \qquad 8$$

where SD_{Y_1} is the referent group's observed criterion standard deviation, f_2 is the normal-density function for the focal group's predictor distribution, b_{11} and b_{12} are the subgroup slopes for the referent and focal groups, respectively, and b_{01} and b_{02} are the subgroup

intercepts for the referent and focal groups, respectively. As d_{Mod_Signed} represents the standardized mean difference between predicted scores from two regression models, the interpretation of d_{Mod_Signed} is similar to the interpretation of Cohen's *d* except that d_{Mod_Signed} is framed in terms of predicted criterion scores rather than observed scores. A d_{Mod_Signed} value indicates the average difference in prediction between the focal regression line and the referent regression line, scaled in terms of the referent group's criterion standard deviation. An unsigned version of the effect size is also possible by integrating over the absolute-value differences between subgroup regression models (see Equation 9).

$$d_{Mod\_Unsigned} = \frac{1}{SD_{Y_1}} \int f_2(X) |X(b_{11} - b_{12}) + b_{01} - b_{02}| \mathrm{d}X \tag{9}$$

Although the signed effect size in Equation 8 indicates whether underprediction or overprediction is more common, it does not provide separate indices of underprediction and overprediction. The signed effect size accounts for the direction of the net differences in predictions, but it cannot convey the magnitude of underprediction if subgroup regression lines cross because at least some of the underprediction will be negated by overprediction; likewise, overprediction is offset by underprediction. To avoid the interpretive ambiguities associated with comparing d_{Mod_Signed} values, Dahlke and Sackett (2018) developed some special-case equations based on Nye and Sackett's (2017) d_{Mod_Signed} equation that quantify underprediction and overprediction separately.

As a pure index of underprediction, one can compute an effect size within the range of scores where underprediction occurs. Dahlke and Sackett (2018) denote this effect size as d_{Mod_Under} and it is computed using Equation 10,

$$d_{Mod\_Under} = \frac{1}{SD_{Y_1}} \int_{X:\, \hat{Y}_1 < \hat{Y}_2} f_2(X)[X(b_{11} - b_{12}) + b_{01} - b_{02}]\mathrm{d}X \quad (10)$$

where $X: \hat{Y}_1 < \hat{Y}_2$ indicates that the integral includes all *X* scores in the focal group's predictor distribution for which the predicted criterion score from the referent group's model (i.e., $\hat{Y}_1 = b_{01} + b_{11}X$) is lower than the predicted criterion score from the focal group's model (i.e., $\hat{Y}_2 = b_{02} + b_{12}X$). By the same logic that provides the basis for d_{Mod_Under}, one can obtain a pure index of overprediction by integrating differences in prediction over the range of scores where overprediction occurs. Dahlke and Sackett denote this effect size as d_{Mod_Over} and it is computed using Equation 11,

$$d_{Mod\_Over} = \frac{1}{SD_{Y_1}} \int_{X:\, \hat{Y}_1 > \hat{Y}_2} f_2(X)[X(b_{11} - b_{12}) + b_{01} - b_{02}]\mathrm{d}X \quad (11)$$

where $X: \hat{Y}_1 > \hat{Y}_2$ indicates that the integral includes all *X* scores for which the predicted criterion score from the referent group's model is higher than the predicted criterion score from the focal group's model. If there is no underprediction within the operational range of predictor scores, d_{Mod_Under} will be zero and d_{Mod_Over} will be equal to d_{Mod_Signed}. If there is *only* underprediction within the operational score range, d_{Mod_Over} will be zero and d_{Mod_Under} will be equal to d_{Mod_Signed}. The d_{Mod_Under} and d_{Mod_Over} directional effect sizes are useful for isolating the magnitudes of underprediction and overprediction. Due to the fact that Equations 10 and 11 represent non-overlapping segments of the complete integral from Equation 8, d_{Mod_Under} and d_{Mod_Over} can be added together to get d_{Mod_Signed} and their absolute values can be added together to get $d_{Mod\_Unsigned}$.

As an example of how d_{Mod_Signed}, d_{Mod_Under}, and d_{Mod_Over} are used, consider the hypothetical scenario depicted in Figure 2. The numerical values in this example are

arbitrary and were chosen only to illustrate the use of d_{Mod} effect sizes to quantify differential prediction in the presence of slope bias. The overall d_{Mod_Signed} for the example scenario is .200, which means that the focal group's performance is overpredicted by an average of .200 *SD*s when the referent line is used to make predictions. In this example, d_{Mod_Under} is -.045 and d_{Mod_Over} is .245, which means that the magnitude of overprediction is 5.44 times as large as the average magnitude of underprediction. Use of normal-density weights in the d_{Mod} equations means that the proportions of cases affected by underprediction (i.e., 25% in the example) and overprediction (i.e., 75% in the example) are automatically factored into the effect sizes, which ensures that the estimates of underprediction and overprediction reflect both the prevalence of differences in prediction as well as the magnitudes of the differences. This means that the d_{Mod} formulas give less weight to large differences in prediction that affect small proportions of focal-group members than to similarly large differences that affect large proportions of focal-group members. In this way, the practical impact of underprediction or overprediction is always apparent from a d_{Mod} effect size.

The d_{Mod} formulas offered by Nye and Sackett (2017) and Dahlke and Sackett (2018) require integrating over a distribution of differences between regression lines, which makes these effect sizes somewhat difficult to compute, as one must have access to software capable of integration to apply the formulas. In addition to being relatively inconvenient to compute, the use of integrals in the d_{Mod} formulas may deter those with limited calculus backgrounds from adopting these methods because of their uncertainty

about what the computations mean.[3] Most other effect sizes (e.g., Pearson correlations, Cohen's *d* values, odds ratios) can be computed using rather simple algebraic formulas; this not only facilitates computation of the effect sizes, but avoids reliance on more advanced math concepts that may dissuade potential users from applying the formulas. Given that the d_{Mod} effect sizes describe differences between two simple linear equations, it should be possible to reformulate the equations into simpler algebraic computational formulas that can be used more easily by a larger audience. I explore this possibility as I pursue my first research question:

> *Research Question 1*: How can d_{Mod} effect sizes for differential prediction be computed algebraically?

Of the d_{Mod} effect sizes, d_{Mod_Signed} is arguably the most important for quantifying differential prediction, as it quantifies overall magnitudes of differential prediction and it is therefore applicable to all differential prediction scenarios. Thus, it is important that researchers have access to methods for determining the standard error of this effect size. Nye and Sackett (2017) recommended using bootstrapping procedures to estimate the standard error of d_{Mod_Signed}, but bootstrapping is computationally demanding and is not a closed-form method to estimate a standard error. If d_{Mod_Signed} can be computed

---

[3] Two decades ago, Meehl (1998) lamented the "abysmally poor mathematical education that we require of our [psychology] students." Based on my personal observations, little seems to have changed in the intervening years; the quantitative training of psychology students remains focused on rather low-level mathematical skills. This state of affairs means that formulas involving math more complex than scalar algebra are likely to be unintelligible to a non-trivial portion of potential users.

algebraically, it should have a closed-form standard error estimator that can also be computed algebraically. The availability of a closed-form standard error formula for d_{Mod_Signed} would be an important advancement because it would facilitate the process of constructing confidence intervals around effect-size estimates and it would allow d_{Mod_Signed} estimates to be properly meta-analyzed. I derive procedures to estimate the standard error of d_{Mod_Signed} as I pursue my second research question:

Research Question 2: What is the standard error of the d_{Mod_Signed} effect size?

In deriving a standard error estimator for d_{Mod_Signed}, I chose not to pursue standard error estimators for d_{Mod_Under}, d_{Mod_Over}, and $d_{Mod\_Unsigned}$. I view the derivation of standard errors of these three effect sizes as secondary to the standard error of d_{Mod_Signed} because the inferential value of a standard error is clearest for d_{Mod_Signed}. Whereas d_{Mod_Signed} can take on both negative and positive values such that confidence intervals based on standard error estimates indicate whether d_{Mod_Signed} represents significant under- or over-prediction, the other d_{Mod} effect sizes are unidirectional and have fixed signs; these effect sizes therefore have skewed sampling distributions that will not be well-characterized by simple standard error estimates or conventional symmetric confidence intervals. Because of this, the uncertainty around estimates of d_{Mod_Under}, d_{Mod_Over}, and $d_{Mod\_Unsigned}$ is best determined via bootstrapping.

My research explores additional research questions related to d_{Mod} effect sizes, but these questions require additional context provided by prior research. Next, I summarize historic findings from the differential prediction literature to set the scene for the remainder of my research objectives.

Historical Evidence Regarding Differential Prediction

With the Cleary model serving as the preferred interpretive framework for predictive bias analyses, I-O psychologists in the United States have accumulated evidence regarding cognitive ability assessments' differences in prediction between Whites and racial/ethnic minority groups and between men and women. The fact that differential prediction information is primarily available for tests of cognitive ability in the United States for White-minority and male-female contrasts is due to the unique legal environment in the US, where race and sex are two of the most salient protected class attributes defined by federal law and where cognitive assessments are heavily scrutinized because of their roles as gate-keeping mechanisms to educational and employment opportunities (particularly in light of moderate-to-large mean differences in test scores among racial/ethnic groups). The evidence regarding both the directions of White-minority and male-female differences in prediction and the prevalence of intercept versus slope differences has been quite consistent across independent reviews of the literature, as well as in both education and employment contexts.

Bartlett, Bobko, Mosier, and Hannan (1978) summarized White-Black differential prediction findings from 1190 worker samples that had been analyzed in unpublished reports. Of these samples, 68 (5.71%) exhibited statistically significant White-Black slope differences and 214 (17.98%) exhibited statistically significant White-Black intercept differences. The rate of significant slope differences was trivially different from the 5% that would have been expected by chance, but the rate of intercept differences was much greater than chance. Of the 210 instances of interpretable intercept differences

(four instances were said to represent unclear patterns of differential prediction), 207 (98.57%) represented overprediction of Black individuals' performance and only three (1.43%) represented underprediction of Black individuals' performance. This finding that Black individuals' performance is overpredicted relative to White individuals' performance due to intercept differences is consistent with Cleary's (1968) original research on the topic.

Hartigan and Wigdor (1989) also studied White-Black differences in prediction and examined results from 72 samples of workers who took the General Aptitude Test Battery (GATB). They found that, of the 72 *t* statistics testing for White-Black slope differences, only two of these (2.78%) were statistically significant at the .05 level. Upon examining the distribution of slope-difference *t* statistics, Hartigan and Wigdor concluded, "there is a tendency for the slope to be greater for nonminorities than for blacks, but the differences are generally not large enough to be detected reliably in an individual study because of relatively small samples of people in each group" (pp. 180-181). Of the 70 samples without significant slope differences, 26 (37.14%) showed statistically significant intercept differences; 25 of these differences represented overprediction of Black individuals' performance and only one represented underprediction. In sum, the GABT did not exhibit reliable patterns of predictive bias against Black individuals. Hartigan and Wigdor's findings were very much in agreement with Bartlett et al.'s (1978) findings, such that slope differences were rare and occurred at roughly chance levels, whereas intercept differences were quite common and overwhelmingly indicated overprediction of Black individuals' performance.

Schmidt, Pearlman, and Hunter (1980) presented a systematic review of White-Hispanic differential prediction findings in the employment domain. Out of 220 samples included in their review, Schmidt et al. found that 2.27% showed significant White-Hispanic slope differences and 7.73% showed significant White-Hispanic intercept differences. From this, the authors concluded that slope differences are "certainly chance phenomena" (p .721) and, although the rate of significant intercept differences was only slightly higher than the 5% rate anticipated by chance alone, Schmidt et al. posited that more consistent intercept trends could be possible in future research.

Findings for White-Black and White-Hispanic differences in the educational admissions domain parallel those in the employment domain. Linn (1978) concluded that the academic performance of Black students tends to be overpredicted compared to the performance of White students. In his review of differential prediction in post-secondary education research, Linn (1973) found that although Black students' performance was generally overpredicted, overprediction tended to occur to a greater degree for those with high predictor score than for those with low predictor scores, despite non-significant slope differences. Thus, although White-Black differential prediction primarily takes the form of intercept differences, small non-significant slope differences can still result in slightly different magnitudes of overprediction across the predictor score range. Young (2001) replicated Linn's finding for overall White-Black differential prediction, reporting that the performance of Black students was reliably overpredicted. Young also found that the performance of Hispanic students tended to be overpredicted compared to the performance of White students.

The results for White-minority differential validity analyses in both employment and post-secondary education settings support the conclusion that, relative to White individuals, the performance of Black and Hispanic individuals is likely to be overpredicted and this overprediction is overwhelmingly due to intercept differences. As overprediction works to the advantage of individuals from minority backgrounds, the historic trends in differential prediction do not suggest bias against Black or Hispanic individuals. As such, the standing conclusion in the literature on high-stakes testing in the United States is that cognitive ability tests are not predictively biased against Black or Hispanic test takers.

Male-female differences in prediction, however, do tend to show that underprediction of females' academic performance is commonplace when standardized tests are examined for predictive bias separately from other predictors. Recent meta-analytic evidence from Fischer, Schult, and Hell's (2013) review of 130 studies indicates that standardized test scores underpredict women's college GPAs by an average of 0.24 points on a four-point GPA scale. Fischer et al.'s finding supports decades of prior research in which patterns of female underprediction were detected (e.g., Ramist, Lewis, & McCamley-Jenkins, 1994; Young, 2001). This male-female difference in prediction is primarily attributable to the fact that standardized tests only measure cognitive constructs whereas college academic performance is a function of both cognitive and non-cognitive factors; after one accounts for such non-cognitive factors, patterns of female underprediction on cognitive tests are reduced or eliminated, as described next.

Omitted Variables and the Importance of a Fully Specified Model

Although standardized tests do reliably underpredict women's performance when analyzed as stand-alone predictors, this does not necessarily mean that the tests are truly biased and should not be used; rather, it highlights the importance of accounting for variables that make unique contributions to the prediction of performance, above and beyond what is explained by test scores. Women's performance is only underpredicted if one fails to account for the non-cognitive determinants of performance on which men and women have mean differences. On average, women are slightly more conscientious than men, particularly in terms of orderliness (Feingold, 1994; Weisberg, DeYoung, & Hirsh, 2011). This means that the underprediction of women's academic performance could simply represent the residual male-female differences in GPAs that are attributable to differences in personality and other non-cognitive factors after accounting for cognitive ability. The failure to account for important determinants of performance on which the groups being compared exhibit mean differences is known as the "omitted variables problem" (Linn & Werts, 1971; Sackett, Laczo, & Lippe, 2003).

Keiser, Sackett, Kuncel, and Brothen (2016) studied male-female differences in prediction with respect to separate components of students' grades in an introductory psychology course. They examined students' overall course grades and their cumulative GPAs as criteria, but also broke down course grades into cognitive components (i.e., exam grades and quiz grades) and less-cognitive components that reflected a substantial degree of student discretion (i.e., discussion points and points earned from voluntary participation in research studies). Students' ACT scores allowed unbiased prediction of

males' and females' performance on cognitive course components, but ACT scores underpredicted females' GPAs and their performance on non-cognitive course-grade components. After accounting for students' scores on the Big Five personality traits, sex no longer had a significant relationship with any of the criteria and females' performance was no longer significantly underpredicted; of the Big Five, conscientiousness had the largest and most consistent relationships with criteria. Kling, Noftle, and Robins (2013) reported similar support for conscientiousness as a mediator of the effect of sex on academic performance. Results reported by Stricker, Rock, and Burton (1993) offer further support for the impact of non-cognitive individual differences on grades. Stricker et al. found that underprediction of women's grades was substantially reduced after the researchers accounted for students' academic preparation in high school, study habits, and attitudes about mathematics.

Cognitive test scores are seldom used as the sole predictor in a selection system, so the sex bias exhibited by cognitive test scores in isolation does not mean that the whole selection system of which they are a part will be biased. For example, admissions officers have access to information beyond students' test scores, such as personal statements, letters of recommendation, and high school transcripts. Not only are these sources taken into consideration when making holistic evaluations of college applicants, the information contained within these sources can offer cues about students' non-cognitive attributes (e.g., diligence and motivation). If standardized test scores were found to exhibit predictive bias, it would be ideal to quantify the non-cognitive attributes signaled in students' other application materials so that they could be included with test

scores in a regression model to determine whether a composite of cognitive and non-cognitive variables predicts without bias. A finding of significant differences in slopes or intercepts between groups can be indicative of a real limitation of a selection system, but one can only be sure that such a finding is not spurious if it results from a fully specified model (i.e., a model that includes all of the information used in the selection process).

If multiple predictors are used to make holistic evaluations about applicants, simply including all relevant predictors in the same regression model does not guarantee a clearly interpretable set of results. Sackett et al. (2003) noted that, "If selection is to be conducted on the basis of a composite of predictors, testing for differential prediction using the composite is the appropriate course of action" (p. 1053). Testing a single composite predictor for differential prediction ensures that the predictor subjected to the Cleary model is as similar to the operational usage of predictor data as possible. The recommendation to base predictive bias analyses on composite predictors, when appropriate to capture operational usage, is formally included in SIOP's fifth edition of the *Principles* (2018).

Research on the omitted variables problem has demonstrated that including all key determinants of performance in differential prediction analyses can help to avoid the spurious findings of focal-group underprediction that can occur when individual predictors that collectively comprise a multi-predictor selection system are examined separately for bias. However, neither the omitted variables problem nor the broader issue of combining predictors into composites has been described with respect to d_{Mod_Signed} effect sizes. When effect sizes are used in personnel selection research, it is important for

researchers to understand how combining multiple predictor variables into a composite predictor affects the magnitudes of effects, and d_{Mod_Signed} is no exception. I explore this issue as I pursue my third research question:

Research Question 3: How are d_{Mod_Signed} effect sizes affected by the formation of composite predictors?

There are many ways to form composite variables and a series of relatively recent articles demonstrated that Pareto-optimal weighting of predictor scores is an effective way to balance the competing objectives of maximizing validity and minimizing adverse impact (De Corte, Lievens, & Sackett, 2007; De Corte, Sackett, & Lievens, 2011; Druart & De Corte, 2012; Sackett, Corte, & Lievens, 2008; Song, Wee, & Newman, 2017). Pareto-optimal weighting is a method for finding optimal tradeoffs when combining information so as to make progress toward satisfying multiple competing goals. In personnel selection settings, Pareto analyses help researchers determine how much weight to give to each of several predictor variables in a selection system to achieve (a) the maximum validity possible for a given level of adverse-impact potential or (b) the minimum adverse-impact potential possible for a given level of validity. Pareto solutions represent compromises between the two extreme options of using multiple linear regression weights (i.e., the set of weights that maximizes the validity of the composite predictor) and using only the predictor with the smallest subgroup mean difference (i.e., assigning a weight of 1 to the predictor with the least adverse-impact potential and weights of 0 to all other predictors). With small compromises to validity, it is often possible to substantially reduce adverse-impact potential. However, although much is

known about the validity and adverse-impact implications of Pareto solutions, to date no one has examined how using these weights affects differential prediction. I explore this issue as I pursue my fourth research question:

> *Research Question 4*: How does the usage of Pareto-optimal weighting solutions affect d_{Mod_Signed} effect sizes?

Note that Research Question 4 is an extension of Research Question 3 that focuses on how one's choice to create a specific type of composite predictor variable affects d_{Mod_Signed} estimates.

Effects of Statistical Artifacts on Differential Prediction Analyses

The remainder of my research questions focus on the effects of statistical artifacts such as range restriction (i.e., selection effects) and measurement error on differential prediction analyses. Statistical artifacts have well-known and well-documented effects on the results of statistical tests; in nearly all cases, artifacts reduce statistical power by both attenuating effect sizes of observed effects and inflating standard error estimates. However, artifacts' effects on differential prediction analyses may be more complex than their effects on simpler analyses such as correlations and *d* values (cf. Aguinis et al., 2010). Given the ubiquity of measurement error and selection artifacts in organizational research, it is critical that psychologists understand whether and how these artifacts bias the test statistics and significance tests associated with differential prediction analyses. There are three primary ways in which we can reach a deeper understanding of artifacts' effects on differential prediction analyses: Simulations, analytic proofs, and large-scale

studies in which real-world data are properly corrected for artifacts. Each of these are considered below as I review past research and outline my present research objectives.

To date, the most comprehensive simulation study of the effects of statistical artifacts on the detection of predictive bias was conducted by Aguinis et al. (2010). Aguinis et al. manipulated measurement error and direct range restriction (DRR; i.e., top-down selection performed on the predictor of interest) to study Type I error rates and statistical power when the Cleary model of bias was applied to simulated samples with varying sample sizes and varying percentages of members from a minority group. Aguinis et al. found that statistical artifacts seriously decreased the statistical power of tests to detect slope differences. They also found that artifacts increased the Type I error rates of tests to detect intercept differences, such that intercept differences favoring the minority group via overprediction were likely to be detected erroneously in observed data sets. Taken together, the simulation results regarding both low rates of overlooking slope differences and high rates of mistakenly detecting overprediction from intercept differences matched trends reported in the historic differential prediction literature; this prompted Aguinis et al. to warn that the findings from 40 years of predictive bias research may be incorrect. They suggested that the long-observed White-minority differences in prediction could simply be due to influence from statistical artifacts and should therefore be re-evaluated using modern methods.

Although Aguinis et al.'s (2010) study included a large number of simulated samples representing many different conditions, their claims have been disputed and follow-up studies by other researchers have called Aguinis et al.'s findings into question.

For example, Mattern and Patterson (2013) assessed predictive bias using SAT test scores, high school GPAs (HSGPAs), and first-year college GPAs collected by the College Board from 477,679 SAT test takers (these individuals represented 177 post-secondary institutions and 339 unique college cohorts). They computed differential prediction regression models for White-Black, White-Hispanic, and male-female comparisons using four different methods of handling artifacts: They computed models based on (1) observed data, (2) data corrected for range restriction, (3) data corrected for range restriction and GPA measurement error, and (4) data corrected for range restriction, GPA measurement error, and predictor measurement error. Similar to prior research, Mattern and Patterson found that Black and Hispanic students' college GPAs were consistently overpredicted relative to White students' GPAs and that female students' GPAs were consistently underpredicted relative to male students' GPAs, even after artifacts were controlled.

In another response to Aguinis et al.'s (2010) call for a revival of predictive bias research, Berry and Zhao (2015) heeded Aguinis et al.'s warning that intercept tests conducted with the Clearly model may be biased and developed an unbiased method to test for intercept differences that did not rely on multiple moderated regression. Berry and Zhao used meta-analytic data that had been corrected for measurement error and range-restriction artifacts to quantify differential prediction and found that cognitive tests still overpredicted job performance for Black applicants in the vast majority of cases. These results agree with Mattern and Patterson's (2013) results and support the classic finding that minority performance tends to be consistently overpredicted. It is also important to

note that, whereas Aguinis et al. based their findings on a hypothetical simulation, the findings reported by Mattern and Patterson as well as Berry and Zhao were obtained using real data that were properly corrected for artifacts. Berry and Zhao's research is notable not only for their substantive verification of historic findings, but also the fact that the recent revision of SIOP's (2018) *Principles* now states that "the need to use an unbiased estimate of the intercept difference and operational validity parameters instead of observed parameters" (p. 41) is a requirement for predictive bias analyses, in reference to Berry and Zhao's new method.

Mattern and Patterson's (2013) article was accompanied by an online supplement that contained covariance matrices, mean vectors, and college-GPA reliability estimates from all of the samples and all of the subgroup contrasts used in their analyses; this included both samples of enrolled college students and colleges' SAT-taking applicant populations. Aguinis et al. (2016) harvested this information and set out to re-analyze the data to determine whether Mattern and Patterson's findings generalized across college cohorts. Aguinis et al. used Mattern and Patterson's data to correct enrolled-student data for range restriction and meta-analyzed the coefficients from artifact-corrected regression models. They found evidence for a lack of differential prediction generalization in terms of intercept and slope differences, as indicated by significant amounts of estimated parameter variance for regression coefficients after accounting for sampling error and statistical artifacts. However, three substantial issues with Aguinis et al.'s analyses of these data should be noted that limit the inferences one can draw from their findings.

First, Aguinis et al.'s (2016) regression models were computed in such a way that separate main-effect and group-by-predictor interaction terms were estimated for each of four predictors (i.e., SAT Mathematics, SAT Critical Reading, SAT Writing, and HSGPA) within a single moderated multiple regression model. By including multiple predictors and interactions in the same model, Aguinis et al.'s conclusions about intercept and slope differences are based not on operational predictor scores, but on the *residuals* of predictor scores after controlling for the effects of the other predictors and interactions. This means that the slope-difference effects observed in Aguinis et al.'s analyses do not actually represent the effects they are purported to characterize: One cannot conclude that a predictor exhibits differential prediction if what is really being analyzed is the leftover variance not shared with other predictors. The core problem here is not simply that multiple predictors were included in the same model (this would be an acceptable practice that avoids the omitted-variables problem; however it would be better yet to examine a single composite), but rather that multiple slope-difference tests were conducted simultaneously. Given that the slope-difference estimates from Aguinis et al.'s regression models are really analyses of whether predictors' residuals demonstrate differential prediction, these coefficients do not reflect how predictor data are used in real-world selection systems and therefore do not provide tests of operational differences in prediction. As noted earlier, the ideal way to analyze a multi-predictor selection system for predictive bias is to combine the predictors into a single composite variable rather than test the separate predictors (Sackett et al., 2003; SIOP, 2018).

Second, the choice to test several highly correlated predictors for differential prediction in the same model creates serious multicollinearity issues that undermine the stability of estimated regression coefficients. When highly correlated variables are included in the same regression model, the unique contribution of each variable is difficult to determine and small fluctuations in the predictors' variance-covariance matrix can correspond to large swings the magnitudes of the predictors' regression coefficients (this is sometimes referred to as the "bouncing betas" problem). Highly correlated predictors effectively compete to explain variance in the regression model and can end up with very different regression coefficients, despite being similar in predictive importance; furthermore, in order for the coefficient for one of the highly correlated predictors to be large, the coefficients for other predictors must be small. For example, in the White-Black SAT regression models analyzed by Aguinis et al. (2016), the SAT Critical Reading regression coefficients correlated -.53 with the SAT Writing coefficients and the slope-difference coefficients for SAT Reading correlated -.63 with the slope-difference coefficients for SAT Writing (see Table 1 for a full array of intercorrelations among regression coefficients). This high level of dependency among the distributions of regression coefficients indicates that multicollinearity created the false appearance of variability across samples. The inflated variability of coefficients across samples from multicollinearity is problematic in its own right because it makes it difficult to interpret the regression results; however, it is all the more problematic in Aguinis et al.'s (2016) study because they tallied the rates of significant results for the various regression

coefficients: This can yield highly misleading results, as some coefficients could be forced into statistical significance by multicollinearity.

The third problem with Aguinis et al.'s (2016) analyses is that their tests of generalizability were based on *Q* tests that indicate whether the residual variance of regression coefficients is significantly different from zero after accounting for sources of artifactual variance. This perspective on generalizability is different from what I-O psychologists often mean when they discuss generalizability through the lens of Hunter and Schmidt's (2004; Schmidt & Hunter, 2015) approach to meta-analysis. In the Hunter-Schmidt method of meta-analysis, one evaluates generalizability by constructing "credibility intervals" around estimated mean effects that show whether the upper or lower 10% of the random-effects parameter distribution includes zero; if zero lies below the 10^{th} percentile or above the 90^{th} percentile of the parameter distribution, an effect is said to generalize. Heterogeneity analyses such as the *Q* statistic that are based on significance testing do not support the same practical interpretations regarding generalizability as do credibility intervals; with enough samples in a meta-analysis, it is possible to obtain a significant *Q* statistic despite a rather small amount of residual variance and a credibility interval that does not include zero. To make proper inferences regarding the practical generalizability of an effect, it is necessary to operationalize generalizability in a way that supports practical interpretations.

I have access to a larger and more up-to-date database of college performance and predictor data from the College Board than was used by Mattern and Patterson (2013) and Aguinis et al. (2016) and I will use this database to examine the generalizability of

differential prediction, correcting for the issues in Aguinis et al.'s analyses described above. In addition to examining the generalizability of slope and intercept differences, I will use the d_{Mod_Signed} standard-error estimator that results from my exploration of Research Question 2 to meta-analyze d_{Mod_Signed} statistics and quantify the practical magnitude of differential prediction effects. My formal research question regarding generalizability of differential prediction effects is:

> *Research Question 5*: Do differences in prediction associated with predictors of college academic performance generalize when quantified as (a) d_{Mod_Signed} effect sizes, (b) intercept differences, or (c) slope differences?

Aguinis et al.'s (2010) simulation study has sparked renewed interest in predictive bias, in general, and the effects of statistical artifacts on bias analyses, specifically. My remaining research questions pertain to the effects of artifacts on differential prediction analyses that have previously gone unaddressed, both in Aguinis et al.'s study and in the published responses to that work. To set the scene for my final two research questions, I describe the effects of measurement error and range restriction on subgroup slope and intercept differences and offer a critique of Aguinis et al.'s handling of statistical artifacts in their simulation.

Measurement error. On the face of things, the findings reported by Aguinis et al. (2010) appear to have dire consequences for predictive bias analyses. However, the value of a selection-oriented simulation for informing practice is entirely dependent upon whether the simulation's parameters and procedures faithfully represent what is expected to happen in operational selection systems. The linkage between the simulation's

parameters and operational practices is where Aguinis et al.'s simulation falls short: Their analysis was focused on latent construct-level relationships between test scores and performance rather than on the operational relationships that are of real interest when studying selection issues. By operational relationships, I mean that predictor scores are analyzed as they exist within the applicant population (that is, without range restriction and with measurement error) because, "in actual test use we must use observed test scores to predict future job performance and cannot use applicants' (unknown) true scores" (Hunter & Schmidt, 2004, p. 126). Differential prediction analyses test whether applicants' performance can be forecasted in real-world selection systems without bias against legally protected groups; this means that differential prediction is fundamentally an issue for operational predictor-criterion relationships, not latent relationships. The matter of whether predictor scores free of measurement error correspond to the same levels of performance regardless of group membership is merely a hypothetical concern because operational predictor scores always contain some amount of error. The real issue is whether analyses based on observed predictor scores support the same conclusions regarding differential prediction as would the operational applicant data.

The importance of focusing on operational relationships in predictive bias analyses is stated in the fifth edition of SIOP's (2018) *Principles*, in which the requirements for predictive bias analyses now indicate that "analysis of predictive bias is appropriately conducted on predictors as operationally used" (p. 41). This has three critical implications, the significance of which cannot be understated: (1) predictive bias analyses performed on multi-predictor selection systems in which applicants are

holistically evaluated should be based on a composite predictor instead of the individual predictors (Sackett et al., 2003), (2) predictor variables should be analyzed in their observed-score metrics such that estimates of predictor-criterion relationships are not corrected for predictor measurement error, and (3) predictor-criterion relationships should be corrected for appropriate forms of range restriction to reflect the fact that operational usage of predictor scores entails consideration of the full range of applicants' predictor scores. Implication #2 regarding not correcting for predictor measurement error is the most critical for my present discussion of reliability-related artifacts; this implication is important because the effects of measurement error on regression slopes are well documented, but not always in such a way that the operational implications for differential prediction are clear. I describe these issues in greater detail after offering proofs regarding how a focus on operational bias analyses can dramatically change one's interpretation of recent articles criticizing predictive bias research. The proofs below that show how measurement error affects regression coefficients are my own derivations, but the core principles are not new; in fact, the underlying ideas have been known to researchers for several decades (cf. Linn and Werts, 1971).

In demonstrating the effects of measurement error on statistics involved in predictive bias analyses, I default to the notation used by Schmidt and Hunter (2015) in their work on psychometric meta-analysis. In this notation, the observed (i.e., "measured" or "manifest") variables corresponding to the predictor and criterion are represented by X and Y, respectively. Additionally, the latent (i.e., "true-score") constructs corresponding

to the predictor and criterion are represented by *T* and *P*, respectively, where *T* is the abbreviation of "test construct" and *P* is the abbreviation of "performance construct."

Recall from Equation 1 that the definition of a simple linear regression slope can be expressed a product of the correlation between the criterion and the predictor and the ratio of the criterion *SD* to the predictor *SD*. The formula for the regression slope for true-score variables (i.e., $b_{1[T]}$) is shown in Equation 12.

$$b_{1[T]} = r_{TP}\frac{SD_P}{SD_T} \tag{12}$$

The corresponding formula for the regression slope for observed variables (i.e., $b_{1[X]}$) is shown in Equation 13.

$$b_{1[X]} = r_{XY}\frac{SD_Y}{SD_X} \tag{13}$$

The observed-score correlation and standard deviations from Equation 13 are functions of true-score correlations, true-score standard deviations, and reliability coefficients. Measurement error inflates the variance of observed-score variables relative to the variance of true-score variables; given that the reliability of *X* can be expressed as $r_{XX'} = SD_T^2/SD_X^2$, the expected value of the observed standard deviation of *X* is equal to the standard deviation of *T* divided by $\sqrt{r_{XX'}}$, as shown in Equation 14.

$$SD_X = \frac{SD_T}{\sqrt{r_{XX'}}} \tag{14}$$

The same type of measurement-error process that causes SD_X to represent an inflated version of SD_T also causes SD_Y to represent an inflated version of SD_P, as shown in Equation 15.

$$SD_Y = \frac{SD_P}{\sqrt{r_{YY'}}} \quad (15)$$

The observed-score correlation is an error-attenuated version of its true-score counterpart, such that the expected value of r_{XY} is the product of r_{TP} and the square roots of the reliability estimates of X ($r_{XX'}$) and Y ($r_{YY'}$), as the square root of a reliability coefficient provides an index of measurement quality indicating the correlation between true scores and observed scores; the magnitude of a correlation is deflated by measurement error in proportion to the magnitude by which the standard deviations of the covariates are inflated by measurement error. This definition of r_{XY} is shown in Equation 16.

$$r_{XY} = r_{TP}\sqrt{r_{XX'}}\sqrt{r_{YY'}} \quad (16)$$

With the correspondence between true-score and observed-score correlations and *SD*s established, Equation 17 shows that the expected value of $b_{1[X]}$ is simply equal to $b_{1[T]}$ times r_{XX} and that r_{YY} does not factor into the relationship between $b_{1[X]}$ and $b_{1[T]}$.

$$\begin{aligned} b_{1[X]} &= r_{XY}\frac{SD_Y}{SD_X} \quad (17)\\ &= r_{TP}\sqrt{r_{XX'}}\sqrt{r_{YY'}}\frac{\left(\frac{SD_P}{\sqrt{r_{YY'}}}\right)}{\left(\frac{SD_T}{\sqrt{r_{XX'}}}\right)}\\ &= r_{TP}\sqrt{r_{XX'}}\sqrt{r_{YY'}}\frac{SD_P\sqrt{r_{XX'}}}{SD_T\sqrt{r_{YY'}}}\\ &= r_{TP}\frac{SD_P r_{XX'}}{SD_T}\\ &= b_{1[T]}r_{XX'} \end{aligned}$$

Therefore, criterion measurement error does not bias the expected value of slope estimates in regression models, but predictor measurement error does. This is notable

because predictor measurement error is not corrected in operational estimates of statistics, but criterion measurement is corrected. Thus, the commonly applied corrections to estimate operational statistics from observed statistics have no influence on regression slopes (these corrections do, however, affect standard errors).

Due to the relationship shown in Equation 17, when the predictor is measured with error it must be the case that the observed slopes of the referent group ($b_{11[X]}$) and the focal group ($b_{12[X]}$) will always be smaller than the true-score slopes of the referent group ($b_{11[T]}$) and the focal group ($b_{12[T]}$), respectively. These inequalities are shown in Equation 18.

$$b_{11[X]} \leq b_{11[T]} \quad 18a$$

$$b_{12[X]} \leq b_{12[T]} \quad 18b$$

Based on Equation 18, it is possible to show that, if *X* is measured with error and is equally reliable in both the referent and focal groups, the difference between the subgroups' observed-score slopes will always be smaller in absolute value than the difference between their true-score slopes. If the referent group's true-score slope is larger than the focal group's true-score slope (i.e., $b_{11[T]} > b_{12[T]}$), then the positive $b_{11[X]} - b_{12[X]}$ difference will be smaller than or equal to $b_{11[T]} - b_{12[T]}$, as shown in Equation 19.

$$\left(b_{11[X]} - b_{12[X]}\right) = r_{XX'}\left(b_{11[T]} - b_{12[T]}\right) \leq \left(b_{11[T]} - b_{12[T]}\right) \quad 19$$

Similarly, if the referent group's true-score slope is smaller than the focal group's true-score slope (i.e., $b_{11[T]} < b_{12[T]}$), then the negative $b_{11[X]} - b_{12[X]}$ difference will be greater than or equal to $b_{11[T]} - b_{12[T]}$ in raw value, as shown in Equation 20.

$$\left(b_{11[X]} - b_{12[X]}\right) = r_{XX'}\left(b_{11[T]} - b_{12[T]}\right) \geq \left(b_{11[T]} - b_{12[T]}\right) \quad 20$$

Given that measurement error of the criterion variable cannot bias slope estimates, it also cannot bias subgroup differences in slopes (i.e., interaction coefficients); in the context of measurement error, only unreliability of the predictor variable can bias regression coefficients.

The fact that criterion measurement error cannot bias slopes also means that criterion measurement error cannot bias intercepts, as intercepts in single-predictor regression models are defined as a function of the slope, criterion mean, and predictor mean, and the expected values of variables' means are unaffected by measurement error. The definitions of the true-score intercepts for the referent group ($b_{01[T]}$) and the focal group ($b_{02[T]}$) are shown in Equation 21.

$$b_{01[T]} = \bar{Y}_1 - b_{11[T]}\bar{X}_1 \quad 21a$$

$$b_{02[T]} = \bar{Y}_2 - b_{12[T]}\bar{X}_2 \quad 21b$$

The corresponding definitions of the observed-score intercepts for the referent group ($b_{01[X]}$) and the focal group ($b_{02[X]}$) are shown in Equation 22.

$$b_{01[X]} = \bar{Y}_1 - b_{11[X]}\bar{X}_1 \quad 22a$$

$$b_{02[X]} = \bar{Y}_2 - b_{12[X]}\bar{X}_2 \quad 22b$$

Due to the fact that predictor measurement error causes observed-score slopes to be flatter than true-score slopes, predictor measurement error causes observed-score intercepts to be higher than true-score intercepts; this inequality is shown in Equation 23.

$$b_{01[X]} \geq b_{01[T]} \quad 23a$$

$$b_{02[X]} \geq b_{02[T]} \quad 23b$$

If there are positive referent-focal mean differences on both the predictor and criterion, as is typically the case, and the subgroups have equal slopes and predictor reliabilities, then Equation 24 shows that the difference between subgroups' observed-score intercepts will be larger than the difference between their true-score intercepts.

$$\left(b_{01[X]} - b_{02[X]}\right) \geq \left(b_{01[T]} - b_{02[T]}\right) \quad 24$$

This is easily proven by substituting the intercept symbols for their definitions. Note that $b_{01[T]} - b_{02[T]}$ can be expanded to the definition shown in Equation 25,

$$\begin{aligned} b_{01[T]} - b_{02T} &= (\bar{Y}_1 - \bar{Y}_2) - \left(b_{11[T]}\bar{X}_1 - b_{12[T]}\bar{X}_2\right) \\ &= (\bar{Y}_1 - \bar{Y}_2) - b_{1[T]}(\bar{X}_1 - \bar{X}_2) \end{aligned} \quad 25$$

where $b_{1[T]} = b_{11[T]} = b_{12[T]}$. Correspondingly, $b_{01[X_a]} - b_{02[X_a]}$ can be expressed as shown in Equation 26.

$$\begin{aligned} b_{01[X]} - b_{02[X]} &= (\bar{Y}_1 - \bar{Y}_2) - r_{XX'}\left(b_{11[T]}\bar{X}_1 - b_{12[T]}\bar{X}_2\right) \\ &= (\bar{Y}_1 - \bar{Y}_2) - r_{XX'}b_{1[T]}(\bar{X}_1 - \bar{X}_2) \end{aligned} \quad 26$$

The computational expressions of intercept differences in Equations 25 and 26 make it clear that, under the conditions outlined above, the expected observed-score intercept differences must always be larger than the expected true-score intercept differences. This is because the $b_{1[T]}(\bar{X}_1 - \bar{X}_2)$ term is regressed closer to zero by the $r_{XX'}$ coefficient in the definition of observed intercept differences, as shown in Equation 27.

$$(\bar{Y}_1 - \bar{Y}_2) - r_{XX'}b_{1[T]}(\bar{X}_1 - \bar{X}_2) \geq (\bar{Y}_1 - \bar{Y}_2) - b_{1[T]}(\bar{X}_1 - \bar{X}_2) \quad 27$$

In addition to their explication of the omitted variables problem, Linn and Werts (1971) described the same effect of predictor measurement error on slopes and intercepts as was demonstrated above. They noted that it would be problematic if two measures of

the same construct produced conflicting differential prediction results due to differences in reliability, particularly as reductions in reliability correspond to increases in magnitudes of expected observed intercept differences. Linn and Werts stated,

> Although the effect of unreliability is in the direction of making the test "look good" in the sense that it decreases the likelihood of observing an underprediction for the low scoring group, the magnitude of the effect is relatively small for tests with reliabilities in the range typically observed for standardized aptitude tests (p. p. 3).

The fact that measurement error can help make tests "look good" by increasing the likelihood of observing overprediction is certainly true. However, given that there is no way to remove the measurement error from operational test scores in the decision-making stage of a selection process, the practical implications of this phenomenon for selection practitioners are limited to the test-development stage. Selection experts already prize high reliability coefficients for the sake of maximizing predictive validity of test scores and supporting high-quality selection decisions, so it is highly unlikely that any professionally developed test intentionally games the system of predictive bias analyses by somehow engineering reliability levels that are just low enough to produce overprediction while still achieving useful levels of validity. As Linn and Werts noted in the quote above, the reliability levels of cognitive tests tend to be quite high, which leaves little room for measurement error to artificially create overprediction.

Quite apart from the magnitude of impact that predictor reliability has on observed intercept differences, the fact remains that the goal of predictive bias analyses is

to determine whether the predictor data available during the operational selection decision-making process relate to levels of actual post-hire performance. There is no reason that applicants' hypothetical "true scores" should factor into predictive bias analyses. The proper parameters against which the accuracy of predictive bias analyses should be determined are not the true-score parameters, but rather the unrestricted (that is, not range-restricted) observed-score predictor parameters with a perfectly reliable criterion.

Given that the fifth edition of the *Principles* (SIOP, 2018) specifically advocates analyzing predictor scores as operationally used, Aguinis et al.'s (2010) choice to compare their simulation results to true-score parameters makes the practical implications of their findings quite misleading.[4] In predictive bias analyses, true-score parameters are abstractions that represent a purely hypothetical scenario: They indicate what differences in prediction would be like if the predictor were measured perfectly, which never happens. As real-world tests of predictive bias are based on operational data, it is inappropriate to use true-score parameters as the benchmark in predictive bias simulations. To support inferences regarding operational usage of the Cleary model, it is necessary to compare simulation results to their corresponding operational parameters, not their true-score parameters. Beyond the problems with Aguinis et al. (2010) using

---

[4] It is important to note that Aguinis et al.'s (2010) research was conducted before SIOP undertook its recent revision of the *Principles*, so they should not necessarily be held accountable to the new recommendations retroactively. However, the logic of focusing on operational parameters rather than true-score parameters should still have been apparent.

inappropriate parameters to define baseline differential prediction values, Culpepper, Aguinis, Kern, and Millsap (2019) have recently presented a method for assessing predictor measurement invariance and differential prediction within a single analysis. As Culpepper et al.'s method involves computing differential prediction analyses based on individuals' estimated true scores rather than their operational predictor scores, this method is likely to provide researchers with inaccurate insights in the differential prediction occurring within operational selection systems. In light of the problematic published works by Aguinis et al. and Culpepper et al., the issue of how to properly evaluate differential prediction to support valid operational inferences is both timely and important.

It is in part because of the inequivalences illustrated above that Aguinis et al. (2010) reported low power for slope-difference tests and inflated Type I error rates for intercept-difference tests. Aguinis et al.'s findings regarding reduced power for slope-difference tests emerged in part because the reliability of the predictor variable was manipulated and varied, which resulted in observed slope-difference parameters that were smaller than the unrestricted true-score slope-difference parameters. Similarly, manipulating the reliability of the predictor variable resulted in observed intercept-difference parameters that were larger than the unrestricted true-score intercept-difference parameters. These reliability effects resulted in inflated magnitudes of intercept differences and deflated magnitudes of slope differences relative to the true-score parameters against which Aguinis et al. evaluated their simulation results. The overpowered tests of intercept differences and underpowered tests of slope differences

observed in Aguinis et al.'s simulation were due in large part to the biasing effects of predictor reliability on the differences between subgroup regression parameters.

The proof outlined above has critical implications for understanding prior work on the relationship between measurement error and differential prediction. Selection researchers are generally most interested in the relationship between observed predictors (X) and latent performance (P) because operational predictor scores contain measurement errors and nothing can be done to remove the influence of those errors from the operational usage of test scores. Beyond their attempts to design and use more reliable predictors, selection professionals accept the measurement error in operational test scores as an unavoidable nuisance and do not attempt to correct for it when computing validity coefficients in the interest of estimating the validity of operational predictor scores. Measurement error in criterion variables, however, is an artifact that selection professionals do correct for because they are interested in answering the question, "how well would we predict performance if performance were measured perfectly?" It is no fault of the predictor variable that the criterion variable is measured imperfectly, so we must correct for criterion measurement error whenever doing so will help us to understand how a predictor functions in a selection system. In the case of regression analyses, it is fortunate that criterion measurement error does not bias estimates of slopes or intercepts, only the standard errors associated with them.

In their simulation study, Aguinis et al. (2010) manipulated the reliability of both predictors and criteria and this choice can help to explain why they reached the conclusions they did. Ideally, Aguinis et al. would have held predictor reliability constant

in their simulation to represent the fact that operational predictor variables contain error and the quality of predictor measurement is already determined by the time predictor scores are used in decision making. The proper question to ask in a simulation of predictive bias is not "how well do observed relationships estimated from sample data approximate the unrestricted true-score relationships?" but rather "how well do observed relationships estimated from sample data approximate the unrestricted operational relationships?" The difference between these two questions is of paramount importance, as a failure to carefully and thoughtfully consider which statistical artifacts impact operational data and which do not can result in a misleading indication of the state of affairs impacting real-world selection programs. Although predictor measurement error does impact regression slopes, the fact that predictor measurement error should go uncorrected in operational analyses makes predictor reliability irrelevant in applied considerations of predictive bias.

Range restriction. Range restriction is a phenomenon that commonly occurs in selection programs and is present when the variance of a variable is smaller in a subgroup selected from an applicant pool than in the complete applicant pool. As variance has a critical role in all statistical analyses, the restriction of variability in the selected group has an attenuating effect on many commonly computed effect sizes (e.g., correlations, Cohen's *d* values) and also inflates the standard errors of estimated statistics. Range restriction comes in two general forms: Direct range restriction (DRR) and indirect range restriction (IRR). The distinction between these two types of range restriction is a matter of which variable was involved in the selection process that created the range restriction:

DRR occurs when the predictor being analyzed was explicitly used in selection, whereas IRR occurs when some other predictor was used to make selection decisions. Although DRR and IRR both have the effect of reducing the variance of a variable, there are key differences between these range-restriction mechanisms that have important implications for predictive bias analyses.

DRR occurs when individuals are selected into (or screened out of) a sample on the basis of their scores on a variable of interest. For example, if one were validating a cognitive ability test for the prediction of job performance and only applicants with scores in the top 50% of the test score distribution were selected into the organization, DRR would have occurred because the predictor used in the predictive validity analysis was the sole basis for top-down selection decisions. Compared to IRR, DRR has a stronger biasing effect on effect sizes because no other type of selection procedure can have as big of an impact on the variance of the predictor of interest as does DRR. In other words, DRR is the most efficient way to reduce/truncate a predictor variable's variance.

In contrast to DRR, which occurs via explicit selection on the predictor of interest, IRR occurs when individuals are selected into (or screened out of) a sample on the basis of their scores on a third variable that is correlated with the predictor and/or criterion of interest. In this case, DRR occurs to a variable that is not included in one's analysis, which indirectly restricts the variance of one's focal variables by virtue of their association with the directly range restricted variable. For example, if one were validating a cognitive ability test for the prediction of job performance and selection decisions were based on interview scores that were positively correlated with cognitive ability test

scores, the cognitive ability test scores would be affected by IRR. Due to the positive cognitive ability-interview correlation, those with higher interview scores also tend to have higher cognitive ability scores. This means that selecting applicants with high interview scores results in reduced variance in cognitive ability scores among the selected group because applicants with lower cognitive ability scores are screened out of the sample at a higher rate than individuals with higher ability scores.

The most important difference between DRR and IRR to consider in predictive bias analyses, beyond their structural differences, is the fact that DRR does not bias the parameters associated with regression lines describing a predictor's relationship with a criterion in a bivariate-normal population (Sackett & Yang, 2000), but IRR can and typically does alter the parameters of a regression line. As accurate estimate of regression effects is of central importance in differential prediction analyses, it will be helpful to consider in greater detail how different types of selection procedures affect regression estimates. There are only two selection scenarios that can take place without biasing regression slopes: (1) DRR of the predictor of interest and (2) random selection (i.e., IRR in a scenario where the correlations between the selection variable and both the predictor and criterion of interest are zero). If IRR influences a differential prediction analysis such that the selection variable is correlated with the predictor and/or criterion of interest, IRR will certainly have a biasing effect on estimates of subgroup differences in intercepts and/or slopes. This biasing effect will be more pronounced when there are subgroup mean differences on *Z*, which gives rise to differential range restriction between groups

and correspondingly different magnitudes of IRR effects on the subgroup regression parameters.

Illustrations of the different effects of DRR and IRR on predictor-criterion relationships are shown in Figure 3. The simulated data depicted in Figure 3 represent an unrestricted applicant pool of 10,000 cases (panel A), a directly range-restricted subset of the applicant pool consisting of 5,000 cases (panel B), and an indirectly range-restricted subset of the applicant pool consisting of 5,000 cases (panel C). The sample size used in this illustration was chosen to be so large that the effects of range restriction on the regression coefficients could not be attributed to sampling error. All unrestricted correlations among the predictor, criterion, and third variable used to induce IRR were set to .50. Panel A of Figure 3 shows the unrestricted data for which the regression line has an intercept of .00 and a slope of .50. Panel B shows directly range-restricted data in which only those cases with scores in the top 50% of the predictor distribution were selected; the regression coefficients were not biased by selection, but the validity estimate was attenuated. Panel C shows indirectly range-restricted data in which only those cases with scores in the top 50% of the distribution of the third variable were selected. IRR attenuated the predictor's validity estimate and also biased the regression coefficients; the indirectly range-restricted intercept was .21 and the slope was .42. In a more complex scenario featuring multiple subgroups, IRR could easily bias the subgroups' regression coefficients by different amounts such that the unrestricted pattern of differences in intercepts and slopes would be muddied, obscured completely, or even reversed in direction.

Aguinis et al.'s (2010) use of DRR as the mechanism by which range restriction was induced in their simulated samples severely limits the applied value of their findings. I do not disagree that the introduction of statistical artifacts poses problems for the statistical power and error rates of moderated regression procedures, but I do argue that use of DRR as a range-restriction mechanism is an oversimplification of the issue at hand. It is common for organizations to use multiple pieces of information to make selection decisions, which necessarily results in IRR rather than DRR. As noted earlier, DRR has no biasing effect on population regression lines, but IRR does. Therefore, in order to illuminate the effects of realistic range restriction artifacts on predictive bias results, a simulation must model range restriction as IRR rather than as DRR. The fact that IRR can bias regression parameters whereas DRR cannot makes IRR a much bigger threat to the accuracy of differential prediction analyses than is DRR. I am unaware of any systematic study examining the effects of IRR on the conclusions drawn from differential prediction analyses.

Joint effects of measurement error and range restriction. Had Aguinis et al.'s (2010) simulation focused on the effects that statistical artifacts have on researchers' ability to detect operational differential prediction trends, they likely would have reached different conclusions regarding the effects of artifacts on the power and error rates of statistical tests used within the Cleary framework. This is because criterion measurement error and DRR both increase the standard errors of regression estimates, but neither artifact biases the regression parameters. Aguinis et al.'s manipulation of predictor measurement error in their simulation *did* bias regression parameters, however, which is

an important contributing factor to their pattern of findings. It is important that future simulations of artifacts' impacts on differential prediction analyses induce only those artifacts that are relevant for characterizing the differences between operational and observed data, rather than those that characterize the differences between true-score and observed data.

In order to address key problems in Aguinis et al.'s (2010) simulation and provide evidence regarding how range restriction and criterion measurement error impact operational differential prediction analyses, I conducted a new simulation in which the results of differential prediction analyses performed on observed data are compared to operational parameters. The goals of this simulation were twofold: To describe how statistical artifacts (particularly indirect range restriction) affect regression parameters and to describe how these artifacts affect statistical power and Type I error rates in differential prediction analyses. These objectives are described in my final two research questions:

> *Research Question 6*: Which parameters of applicant populations and selection systems are most predictive of biased estimation of (a) d_{Mod_Signed} effect sizes, (b) intercept differences, and (c) slope differences?
>
> *Research Question 7*: Which parameters of applicant populations and selection systems have the biggest impact on the ability of researchers to detect (a) intercept differences and (b) slope differences?

Overview of Studies

The remainder of this manuscript is structured around four studies designed to address the research questions introduced above. Study 1 is focused on the derivation of simplified d_{Mod} formulas and a closed-form standard-error estimator for d_{Mod_Signed}. This study addresses Research Question 1 and Research Question 2.

Study 2 uses meta-analytic mean correlations and White-Black *d* values to demonstrate the effects of forming predictor composites on magnitudes of d_{Mod_Signed} estimates. This study examines three different methods for forming composites: Unit weighting (i.e., giving equal weight to all predictors), regression weighting, and Pareto-optimal weighting. This study addresses Research Question 3 and Research Question 4.

Study 3 uses a large database from the College Board to examine the generalizability of differential prediction effects in the post-secondary education context using first-year GPA as the criterion variable and HSGPA and SAT subtests as predictors. This study addresses Research Question 5, but also re-addresses Research Question 3, as Study 3 combines HSGPA and the three SAT subtests into various composite predictors to conceptually replicate the trends demonstrated in Study 2.

Study 4 is a simulation study that examines the effects of direct range restriction (DRR), indirect range restriction (IRR), and criterion measurement error on the accuracy with which differential prediction effects can be estimated and detected. This study addresses Research Question 6 and Research Question 7.

Study 1: Algebraic Standardized Effect Sizes for Differential Prediction with Standard Error Estimates

In this study, I expand upon the formulation of standardized effect sizes for quantifying the effects of binary categorical moderators reported by Nye and Sackett (2017) and Dahlke and Sackett (2018). These effect sizes are known as d_{Mod} because they are interpreted like Cohen's *d* effect sizes, but computed using predicted dependent-variable scores based on moderated regression equations rather than observed dependent-variable scores. Although d_{Mod} effect sizes can be computed in any setting where the regression of a criterion on a predictor is moderated by a dichotomous variable, these d_{Mod} effect sizes mark an important development in predictive bias research, specifically, because they can summarize differential prediction in standard deviation units (rather than only interpreting differential prediction using significance testing). Thus, these effect sizes can quantify differences in prediction in a consistent metric across subgroup comparisons by identifying a single referent group to use in all analyses.

Integration-Based Formulas for d_{Mod} Effect Sizes Presented in Prior Studies

As indicated in the Introduction chapter, parametric formulas for computing d_{Mod} effect sizes have been presented previously (see Dahlke & Sackett, 2018; Nye & Sackett, 2017) and those formulas relied on integrating a function over the distribution of predictor scores to arrive at an estimate of the average difference in prediction between two groups' regression lines. In this integration-based approach, each score in the focal group's predictor distribution is used as an input to the simple linear regression formulas that describe the focal and referent groups' data, the difference between the predicted

criterion scores forecasted by the subgroup regression models is computed, and the difference in prediction for each predictor score is weighted by the probability density associated with that score so that the differences in prediction over the full range of predictor scores can be averaged. Equations for the integration-based approach were presented earlier for d_{Mod_Signed} (Equation 8), $d_{Mod\_Unsigned}$ (Equation 9), d_{Mod_Under} (Equation 10), and d_{Mod_Over} (Equation 11) effect sizes. The integration-based approach for computing d_{Mod} estimates works quite well, but the operations involved in the computations are complex and this complexity gives rise to several limitations. The limitations outlined below provided the impetus to derive simpler formulas for computing d_{Mod} estimates.

First, in terms of practical usage, the integration-based formulas cannot be accommodated by all analysis software programs commonly used by I-O psychologists, such as spreadsheet-based programs like Microsoft Excel. Nye and Sackett (2017) released a MATLAB program that can compute d_{Mod} in a Microsoft Windows environment and Dahlke and Sackett (2018) released functions to compute d_{Mod} in the *R* programming language; these programs can be used by a sizable audience, but they still leave those who do not use Windows and/or *R* without a workable option for computing d_{Mod} effect sizes. Dahlke and Sackett (2018) did offer non-parametric methods for computing d_{Mod} estimates that can be implemented in any data-analysis software, but only for those interested in computing d_{Mod} from a raw data set; these methods are not usable by those computing d_{Mod} from secondary data (e.g., meta-analysts or those computing d_{Mod} from artifact-corrected descriptive statistics). Thus, for increased accessibility and

easier adoption of d_{Mod}, I view it as quite important that researchers have access to simpler formulas that can be computed algebraically without need for numeric integration.

Second, the derivation of algebraic formulas is important so that d_{Mod} can be more similar to other effect sizes used by psychologists, such that the computations require less sophisticated math knowledge to implement. Few other effect sizes, if any, use integration and not all psychologists who might be interested in using the d_{Mod} effect sizes are sufficiently comfortable with calculus concepts to use these formulas. The computation of effect sizes should ideally be as simple as possible to encourage widespread adoption of good statistical practices that include presenting indices of effect magnitude in addition to statistical significance tests. One's knowledge of calculus (or lack thereof) should not represent a barrier to the usage of effect-size formulas, particularly if calculus concepts are not strictly necessary to compute the effect sizes in question. Thus, beyond facilitating the implementation of d_{Mod} in more software environments, deriving algebraic versions of the d_{Mod} formulas will make d_{Mod} accessible to researchers with a broader range of math backgrounds.

Third, given that d_{Mod} summarizes the difference between two linear formulas, the difference should also be expressible as a linear function. Operational equations for effect sizes should be simplified as much as possible in the interest of streamlined computations. Thus, beyond the practical reasons for simplifying the equations that were outlined above, the mathematical parsimony of an effect-size formula should be a goal in and of itself.

Fourth and finally, if linear algebraic versions of the d_{Mod} formulas can be derived, these simplified formulas should permit analytic algebraic estimations of the standard error of d_{Mod_Signed} effect sizes. Closed-form analytic standard-error estimators will be much simpler to compute and yield more reliable estimates of statistical uncertainty than the bootstrapping methods recommended in prior research (cf. Nye & Sackett, 2017). Whereas bootstrapping procedures are computationally costly and are most applicable to analyses of primary data, analytic standard-error estimators for d_{Mod_Signed} would be efficient and broadly applicable to analyses of both primary and secondary data.

Algebraic Formulas for d_{Mod_Signed} Effect Sizes

The d_{Mod} formulas described by Nye and Sackett (2017) and Dahlke and Sackett (2018) require the regression formulas for both the referent and focal subgroups to be known, but it is possible to compute d_{Mod_Signed} when one only knows both subgroups' means and the referent group's validity and standard deviations. Additionally, it is possible to compute parametric d_{Mod} effect sizes using simple linear algebraic formulas that do not require integration. Research Question 1 asked how d_{Mod} effect sizes can be estimated algebraically and I present my derivations of algebraic d_{Mod} formulas below.

The principal objective of the d_{Mod} formulas is to quantify the average difference in predicted criterion scores between two groups' regression lines. If one considers that the mean difference between two linear functions is equal to the difference between the means of the functions, one can use the latter definition to greatly simply the computation of the mean difference. In other words, instead of integrating over a distribution of

differences between predictions made by the referent and focal groups' regression equations, one can simply compute the difference between the mean outputs of the regression equations, which is simply the difference between the levels of predicted performance associated with the focal group's mean predictor score. This observation is the guiding idea for the derivations below.

As a starting point for simplifying d_{Mod}, let $\hat{Y}_1^*$ represent the vector of predicted criterion scores for the focal group based on the referent regression equation (the asterisk indicates that this referent vector of predicted scores is based on the focal predictor distribution, not the referent predictor distribution), as shown in Equation 28.

$$\hat{Y}_1^* = b_{01} + b_{11} X_2 \quad (28)$$

Additionally, let $\hat{Y}_2$ represent the vector of predicted criterion scores for the focal group based on the focal regression equation, as shown in Equation 29.

$$\hat{Y}_2 = b_{02} + b_{12} X_2 \quad (29)$$

The means of $\hat{Y}_1^*$ and $\hat{Y}_2$ can be computed directly based on the focal group's mean predictor score, as shown in Equations 30 and 31, respectively

$$\bar{\hat{Y}}_1^* = b_{01} + b_{11} \bar{X}_2 \quad (30)$$

$$\bar{\hat{Y}}_2 = b_{02} + b_{12} \bar{X}_2 \quad (31)$$

When computing the difference between two linear transformations of the same vector, the mean of the differences is equal to the difference of the means. Thus, the algebraic d_{Mod_Signed} formula can be simplified accordingly by standardizing the difference between the mean predicted criterion values, as shown in Equation 32.

$$d_{Mod\_Signed} = \frac{\overline{\hat{Y}_1^* - \hat{Y}_2}}{SD_{Y_1}} \quad (32)$$

$$= \frac{\overline{\hat{Y}_1^*} - \overline{\hat{Y}_2}}{SD_{Y_1}}$$

After substituting the mean predicted criterion scores for their definitions from Equations 30 and 31, one now has the makings of a computational d_{Mod_Signed} formula, as shown in Equation 33.

$$d_{Mod\_Signed} = \frac{(b_{01} + b_{11}\,\bar{X}_2) - (b_{02} + b_{12}\,\bar{X}_2)}{SD_{Y_1}} \quad (33)$$

The b_{01} and b_{02} intercepts in Equation 33 can be re-expressed in terms of the subgroups' slopes using the definition of the intercept in linear regression, as shown in Equations 34 and 35, respectively.

$$b_{01} = \bar{Y}_1 - b_{11}\,\bar{X}_1 \quad (34)$$

$$b_{02} = \bar{Y}_2 - b_{12}\,\bar{X}_2 \quad (35)$$

After substituting b_{01} and b_{02} for their slope-based definitions, the formula for d_{Mod_Signed} is as shown in Equation 36.

$$d_{Mod\_Signed} = \frac{(\bar{Y}_1 - b_{11}\,\bar{X}_1 + b_{11}\,\bar{X}_2) - (\bar{Y}_2 - b_{12}\,\bar{X}_2 + b_{12}\,\bar{X}_2)}{SD_{Y_1}} \quad (36)$$

Next, one can group similar terms for the sake of clarity, simplify the expressions, and make appropriate adjustments to operators contained within the parentheses to arrive at the formula in Equation 37.

$$d_{Mod\_Signed} = \frac{(\bar{Y}_1 - b_{11}\,\bar{X}_1 + b_{11}\,\bar{X}_2) + (-\bar{Y}_2 + b_{12}\,\bar{X}_2 - b_{12}\,\bar{X}_2)}{SD_{Y_1}} \quad (37)$$

$$= \frac{\bar{Y}_1 - b_{11}\,\bar{X}_1 + b_{11}\,\bar{X}_2 - \bar{Y}_2 + b_{12}\,\bar{X}_2 - b_{12}\,\bar{X}_2}{SD_{Y_1}}$$

$$= \frac{\bar{Y}_1 - \bar{Y}_2 - b_{11}\,\bar{X}_1 + b_{11}\,\bar{X}_2}{SD_{Y_1}}$$

$$= \frac{(\bar{Y}_1 - \bar{Y}_2) - (b_{11}\,\bar{X}_1 - b_{11}\,\bar{X}_2)}{SD_{Y_1}}$$

$$= \frac{(\bar{Y}_1 - \bar{Y}_2) - b_{11}\,(\bar{X}_1 - \bar{X}_2)}{SD_{Y_1}}$$

Thus, d_{Mod_Signed} is simply a function of the difference in subgroup criterion means, the difference in subgroup predictor means, the referent group's regression slope, and the referent group's criterion *SD*.

Finally, to make the computational formula easier to use, one can substitute b_{11} with its definition, $r_{XY_1}\,\frac{SD_{Y_1}}{SD_{X_1}}$ (where r_{XY_1} is the validity within the referent group and SD_{X_1} is the referent predictor standard deviation), the formula is based entirely on means, standard deviations, and correlations. The final computational formula for d_{Mod_Signed} is given in Equation 38.

$$d_{Mod\_Signed} = \frac{(\bar{Y}_1 - \bar{Y}_2) \;-\; r_{XY_1}\,\frac{SD_{Y_1}}{SD_{X_1}}(\bar{X}_1 - \bar{X}_2)}{SD_{Y_1}} \quad (38)$$

Equation 38 allows one to compute d_{Mod_Signed} from differences in unstandardized subgroup means, but it is also possible to closely approximate d_{Mod_Signed} from standardized mean differences. If one assumes that subgroups have equal criterion

standard deviations and equal predictor standard deviations, d_{Mod_Signed} can be estimated using Equation 39,

$$d_{Mod\_Signed} \cong d_Y - r_{XY_1} d_X \qquad 39$$

where d_Y and d_X represent standardized mean differences on the criterion and predictor, respectively. Thus, d_{Mod_Signed} will be zero and there will be no net differences in prediction when $d_Y = r_{XY_1} d_X$. Equation 39 allows d_{Mod_Signed} to be estimated with reasonable accuracy from secondary data when subgroup mean differences are only expressed in standardized form; it also allows one to use meta-analytic estimates of mean differences and referent-group validities to estimate the meta-analytic mean magnitudes of d_{Mod_Signed} effects.

It is interesting to note that Equation 39 is very closely related to the formula for Berry and Zhao's (2015) unbiased test of intercept differences (i.e., $\Delta b_0 = d_Y - r\, d_X$; see their Equation 4, p. 165). The key difference between the d_{Mod_Signed} formula I offer in Equation 39 and Berry and Zhao's intercept-difference formula is that the d_{Mod_Signed} formula calls specifically for the referent group's validity to be used whereas Berry and Zhao's method simply states that "*r* is the correlation coefficient between the cognitive ability test and job performance" (p. 41); this implies that the validity coefficient in Berry and Zhao's formula is the total validity when the referent and focal groups are analyzed together, which will be larger than either subgroup's correlation because it includes between-group variance. Berry and Zhao's intercept-difference test was formulated under the assumption that subgroup slopes are equal, but Equation 39 is evidence that a

substantially similar formula can produce useful effect-size estimates regardless of whether subgroups have equal or differing slopes.

Correcting d_{Mod_Signed} Effect Sizes for Measurement Error

Given that d_{Mod_Signed} can be computed from only means, standard deviations, and a validity coefficient, and measurement error has well-known impacts on all of these input statistics, it is possible to compute measurement error-corrected estimates of d_{Mod_Signed}. The expected values of subgroup means are unaffected by measurement error, as measurement error only decreases the precision with which means are estimated and does not bias the parameter values of means. Correlations can be corrected for measurement error by the inverse of the attenuation process (depicted earlier in Equation 16), such that the true-score correlation is estimated by simply dividing the observed correlation by the square root of the products of the reliability coefficients for *X* and *Y*, as shown in Equation 40.

$$r_{TP} = \frac{r_{XY}}{\sqrt{r_{XX'} r_{YY'}}} \quad (40)$$

The formula to correct standard deviations for measurement error is also the inverse of the attenuation process (depicted earlier in Equations 14 and 15), such that the true-score standard deviation of a variable is estimated as the product of the variable's observed standard deviation and the square root of the variable's reliability coefficient; the formulas for the measurement error-corrected standard deviations of *X* and *Y* are show in Equations 41 and 42, respectively.

$$SD_T = SD_X \sqrt{r_{XX'}} \quad (41)$$

$$SD_P = SD_Y \sqrt{r_{YY'}} \quad (42)$$

When correcting for measurement error, d_{Mod_Signed} can be computed using Equation 43, where the corrections are implemented into the formula.

$$d_{Mod\_Signed} = \frac{(\bar{Y}_1 - \bar{Y}_2) - r_{TP_1}\frac{SD_{P_1}}{SD_{T_1}}(\bar{X}_1 - \bar{X}_2)}{SD_{P_1}} \quad (43)$$

$$= \frac{(\bar{Y}_1 - \bar{Y}_2) - \frac{r_{XY_1}}{\sqrt{r_{XX_1'}r_{YY_1'}}}\frac{SD_{Y_1}\sqrt{r_{YY_1'}}}{SD_{X_1}\sqrt{r_{XX_1'}}}(\bar{X}_1 - \bar{X}_2)}{SD_{Y_1}\sqrt{r_{YY_1'}}}$$

$$= \frac{(\bar{Y}_1 - \bar{Y}_2) - r_{XY_1}\frac{SD_{Y_1}}{SD_{X_1}r_{XX_1'}}(\bar{X}_1 - \bar{X}_2)}{SD_{Y_1}\sqrt{r_{YY_1'}}}$$

Note that for the types of operational analyses typically of interest in selection research, one should only correct for measurement error in *Y* and set the reliability of *X* to 1; although criterion measurement error cannot affect regression slopes, it does affect the standard deviation of criterion scores and it therefore important to account for this in estimates of artifact-corrected d_{Mod_Signed} effects. Although correcting for predictor measurement error is not advisable in selection research, a correction for predictor unreliability is included in Equation 43 for cases in which construct-level inferences would be of interest (e.g., in moderated regression analyses performed outside of the selection context).

Similar adjustments can be made to Equation 39 to correct standardized effect-size inputs for measurement error. The measurement error-correction for a criterion *d* value entails dividing the observed *d* value by the square root of the pooled subgroup reliability coefficient, as shown in Equation 44.

$$d_P = \frac{d_Y}{\sqrt{r_{YY'_{Pooled}}}} \tag{44}$$

$$= \frac{d_Y}{\sqrt{\frac{(n_1 - 1)\, r_{YY'_1} + (n_2 - 1)\, r_{YY'_2}}{n_1 + n_2 - 2}}}$$

The same procedure applies to corrections of predictor *d* values. However, given that Equation 39 already requires one to assume that subgroups have equal criterion standard deviations, it does not seem unreasonable for one to also assume that subgroups have equal criterion reliabilities. After assuming equal that measurements are equally reliable between subgroups, the formulas for estimating the corrected criterion and predictor *d* values are expressed in Equations 45 and 46, respectively.

$$d_P \cong \frac{d_Y}{\sqrt{r_{YY'_1}}} \tag{45}$$

$$d_T \cong \frac{d_X}{\sqrt{r_{XX'_1}}} \tag{46}$$

One can substitute observed correlations and observed *d* values from Equation 39 for their correction formulas to arrive at Equation 47, which allows one to correct the standardized-input version of d_{Mod_Signed} for measurement error.

$$d_{Mod\_Signed} = \frac{d_Y}{\sqrt{r_{YY'_1}}} - \frac{r_{XY_1}}{\sqrt{r_{XX'_1} r_{YY'_1}}} \frac{d_X}{\sqrt{r_{XX'_1}}} \tag{47}$$

$$= \frac{d_Y}{\sqrt{r_{YY'_1}}} - \frac{r_{XY_1} d_X}{r_{XX'_1} \sqrt{r_{YY'_1}}}$$

Algebraic Formulas for d_{Mod_Under}, d_{Mod_Over}, and $d_{Mod\_Unsigned}$ Effect Sizes

The derivations presented above showed how to compute d_{Mod_Under} algebraically, but it is also possible to algebraically estimate d_{Mod_Under} and d_{Mod_Over}, which separately

express magnitudes of under- and over-prediction, respectively, as well as $d_{Mod\_Unsigned}$. However, estimates for the directional and unsigned d_{Mod} effect sizes technically still require some form of integration, but only for the purpose of computing cumulative densities of the normal distribution. The normal cumulative density function is pre-programmed into commonly available programs such as Microsoft Excel and it is much simpler to use than the original d_{Mod} formulas; the cumulative densities associated with particular quantiles of a distribution can also be found in tables that are commonly included as appendices in statistics textbooks.

The process for estimating d_{Mod_Under}, d_{Mod_Over}, and $d_{Mod\_Unsigned}$ requires one to (1) determine the point in the predictor-score distribution at which the referent and focal regression lines intersect, (2) estimate the mean scores of those focal-group members whose scores are above the point of intersection and those are scores were below that point, and (3) estimate the proportion of the focal predictor distribution that exists on either side of the point of intersection. When the slopes of the referent and focal regression lines are not equal, the subgroup lines will intersect and the intersection point indicates where overprediction ends and underprediction begins. The X coordinate at which the referent and focal regression lines intersect is computed using Equation 48.

$$X_{Intersect} = \frac{b_{01} - b_{02}}{b_{12} - b_{11}} \qquad 48$$

$$= \frac{\left(\bar{Y}_1 - r_{XY_1}\frac{SD_{Y_1}}{SD_{X_1}r_{XX_1'}}\bar{X}_1\right) - \left(\bar{Y}_2 - r_{XY_2}\frac{SD_{Y_2}}{SD_{X_2}r_{XX_2'}}\bar{X}_2\right)}{r_{XY_2}\frac{SD_{Y_2}}{SD_{X_2}r_{XX_2'}} - r_{XY_1}\frac{SD_{Y_1}}{SD_{X_1}r_{XX_1'}}}$$

After computing $X_{Intersect}$, that value can be used to estimate the mean scores on

either side of the point at which $X_{Intersect}$ bisects the focal group's predictor distribution. To do this, one can use the general formula for the mean of a doubly truncated distribution (that is, a distribution in which cut scores have been used to censor scores above and below certain points), which is shown in Equation 49,

$$\mu_{Truncated} = \mu + \sigma \frac{f\left(\frac{a-\mu}{\sigma}\right) - f\left(\frac{b-\mu}{\sigma}\right)}{F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right)} \quad (49)$$

where μ is the unrestricted mean, σ is the unrestricted standard deviation, *a* is the cut score imposed below the mean, *b* is the cut score imposed above the mean, *f* is the normal probability density function, and *F* is the normal cumulative density function. Equation 49 can be used to obtain two special-case formulas necessary for computing the directional d_{Mod} effect sizes. To estimate the focal group mean below the point of intersection, let *b* equal $X_{Intersect}$ and let *a* equal positive infinity, which simplifies to Equation 50.

$$\bar{X}_{2\_Below} = \bar{X}_2 - SD_{X_2}\sqrt{r_{XX'_2}} \frac{f\left(\frac{X_{Intersect} - \bar{X}_2}{SD_{X_2}\sqrt{r_{XX'_2}}}\right)}{F\left(\frac{X_{Intersect} - \bar{X}_2}{SD_{X_2}\sqrt{r_{XX'_2}}}\right)} \quad (50)$$

Similarly, to estimate the focal group mean above the point of intersection, let *b* equal negative infinity and let *a* equal $X_{Intersect}$, which simplifies to Equation 51.

$$\bar{X}_{2\_Above} = \bar{X}_2 + SD_{X_2}\sqrt{r_{XX'_2}} \frac{f\left(\frac{X_{Intersect} - \bar{X}_2}{SD_{X_2}\sqrt{r_{XX'_2}}}\right)}{1 - F\left(\frac{X_{Intersect} - \bar{X}_2}{SD_{X_2}\sqrt{r_{XX'_2}}}\right)} \quad (51)$$

The final set of new values that need to be determined before estimating d_{Mod_Under}, d_{Mod_Over}, and $d_{Mod\_Unsigned}$ includes the proportions of cases in the focal group's

predictor distribution that fall on either side of $X_{Intersect}$. These proportions are necessary to account for the fact that less than 100% of the focal predictor distribution exists on either side of the intersection point and to estimate directional effect sizes that factor in the prevalence rates of underprediction and overprediction. The proportion of the focal distribution below the intersection point can be computed using Equation 52.

$$p_{2\_Below} = F\left(\frac{X_{Intersect} - \bar{X}_2}{SD_{X_2}\sqrt{r_{XX'_2}}}\right) \tag{52}$$

Likewise, the proportion above the intersection point can be computed using Equation 53.

$$p_{2\_Above} = 1 - F\left(\frac{X_{Intersect} - \bar{X}_2}{SD_{X_2}\sqrt{r_{XX'_2}}}\right) \tag{53}$$

With $\bar{X}_{2\_Below}$, $\bar{X}_{2\_Above}$, p_{2_Below}, p_{2_Above} determined, d_{Mod_Under} and d_{Mod_Over}, can finally be computed. To accomplish this, it will be helpful to first establish a generic version of the d_{Mod} formula in which any random value Z can be entered into the prediction formulas to determine the magnitude of differential prediction associated with a score of Z; the generic formula for d_{Mod} is given in Equation 54.

$$\begin{aligned} d_{Mod\_Generic} &= \frac{\left(\bar{Y}_1 - \frac{r_{XY_1}SD_{Y_1}}{SD_{X_1}r_{XX'_1}}\bar{X}_1 + \frac{r_{XY_1}SD_{Y_1}}{SD_{X_1}r_{XX'_1}}Z\right) - \left(\bar{Y}_2 - \frac{r_{XY_2}SD_{Y_2}}{SD_{X_2}r_{XX'_2}}\bar{X}_2 + \frac{r_{XY_2}SD_{Y_2}}{SD_{X_2}r_{XX'_2}}Z\right)}{SD_{Y_1}\sqrt{r_{YY'_1}}} \\ &= \frac{\left[\bar{Y}_1 - \left(\frac{r_{XY_1}SD_{Y_1}}{SD_{X_1}r_{XX'_1}}\bar{X}_1 - \frac{r_{XY_1}SD_{Y_1}}{SD_{X_1}r_{XX'_1}}Z\right)\right] - \left[\bar{Y}_2 - \left(\frac{r_{XY_2}SD_{Y_2}}{SD_{X_2}r_{XX'_2}}\bar{X}_2 - \frac{r_{XY_2}SD_{Y_2}}{SD_{X_2}r_{XX'_2}}Z\right)\right]}{SD_{Y_1}\sqrt{r_{YY'_1}}} \\ &= \frac{\left[\bar{Y}_1 - \frac{r_{XY_1}SD_{Y_1}}{SD_{X_1}r_{XX'_1}}(\bar{X}_1 - Z)\right] - \left[\bar{Y}_2 - \frac{r_{XY_2}SD_{Y_2}}{SD_{X_2}r_{XX'_2}}(\bar{X}_2 - Z)\right]}{SD_{Y_1}\sqrt{r_{YY'_1}}} \end{aligned} \tag{54}$$

As a generic formula, Equation 54 can be used with any value Z of interest; d_{Mod_Signed} is a special case of this formula when $Z = \bar{X}_2$. When rescaled according to the proportion of cases on one side of the intersection point defined earlier, Equation 54 provides the basis of the formulas for computing d_{Mod_Under} and d_{Mod_Over}. By defining Z as the mean predictor score of focal group members whose performance is underpredicted, d_{Mod_Under} can be computed using Equation 55,

$$d_{Mod\_Under} = p_{2\_Under} \frac{\left[\bar{Y}_1 - r_{XY_1} \frac{SD_{Y_1}(\bar{X}_1 - \bar{X}_{2\_Under})}{SD_{X_1} r_{XX'_1}}\right] - \left[\bar{Y}_2 - r_{XY_2} \frac{SD_{Y_2}(\bar{X}_2 - \bar{X}_{2\_Under})}{SD_{X_2} r_{XX'_2}}\right]}{SD_{Y_1}\sqrt{r_{YY'_1}}} \quad (55)$$

where $\bar{X}_{2\_Under}$ is $\bar{X}_{2\_Below}$ (and p_{2_Under} is p_{2_Below}) if predictor scores below the line-intersection point show underprediction or $\bar{X}_{2\_Above}$ (and p_{2_Under} is p_{2_Above}) if predictor scores above the line-intersection point show underprediction. Similarly, by defining Z as the mean predictor score of focal group members whose performance is overpredicted, d_{Mod_Over} can be computed using Equation 56,

$$d_{Mod\_Over} = p_{2\_Over} \frac{\left[\bar{Y}_1 - r_{XY_1} \frac{SD_{Y_1}(\bar{X}_1 - \bar{X}_{2\_Over})}{SD_{X_1} r_{XX'_1}}\right] - \left[\bar{Y}_2 - r_{XY_2} \frac{SD_{Y_2}(\bar{X}_2 - \bar{X}_{2\_Over})}{SD_{X_2} r_{XX'_2}}\right]}{SD_{Y_1}\sqrt{r_{YY'_1}}} \quad (56)$$

where $\bar{X}_{2\_Over}$ is $\bar{X}_{2\_Below}$ (and p_{2_Over} is p_{2_Below}) if predictor scores below the line-intersection point show overprediction or $\bar{X}_{2\_Above}$ (and p_{2_Over} is p_{2_Above}) if predictor scores above the line-intersection point show overprediction.

As d_{Mod_Under} and d_{Mod_Over} represent differences in prediction in non-overlapping ranges of predictor scores, the sum of their absolute values gives $d_{Mod\_Unsigned}$, as shown in Equation 57.

$$d_{Mod\_Unsigned} = \left|d_{Mod\_Under}\right| + d_{Mod\_Over} \quad (57)$$

The raw sum of d_{Mod_Under} and d_{Mod_Over} can also be used to compute d_{Mod_Signed}, as shown in Equation 58.

$$d_{Mod\_Signed} = d_{Mod\_Under} + d_{Mod\_Over} \quad (58)$$

However, if d_{Mod_Signed} is the main effect size of interest, it is more efficient to compute it directly using Equation 43 than with Equation 58.

Demonstration of Agreement Between Algebraic and Integral Formulas

To supplement the mathematical proofs provided above and to offer further evidence that my algebraic formulas are equivalent to the integration-based formulas provided by Nye and Sackett (2017) and Dahlke and Sackett (2018), I applied both sets of formulas to simulated data sets representing a wide variety of differential prediction scenarios. As d_{Mod} effect sizes are applicable in any case where a binary variable moderates a relationship and are not restricted to use in differential prediction studies, my simulation parameters were chosen to explore situations beyond those that one is likely to encounter in differential prediction analyses. For example, subgroup slopes in the simulation ranged from being equal to having opposite signs with comparable magnitudes. Subgroup means and *SD*s were also varied to produce a wide range of mean-difference scenarios with different metrics. The simulation parameters are listed in Table 2 and all possible combinations of the parameters produced a total of 54,000 conditions.

The results of the simulation are plotted in Figure 4, where the perfect associations between d_{Mod} estimates produced by the algebraic and integral formulas show an exact correspondence between the methods all four types of d_{Mod} statistics: d_{Mod_Signed}, d_{Mod_Under}, d_{Mod_Over}, and $d_{Mod\_Unsigned}$. The algebraic formulas for d_{Mod_Signed}

(Equation 43), d_{Mod_Under} (Equation 55), d_{Mod_Over} (Equation 56), and $d_{Mod\_Unsigned}$ (Equation 57) can be used in place of the more complex integration-based formulas without any loss of precision.

An Alternative Scaling for d_{Mod} Based on Conditional Criterion Variances

The d_{Mod} methods described above standardize the differences in referent and focal groups' predicted criterion values by the referent group's observed criterion standard deviation, as recommended by Nye and Sackett (2017). However, an alternative scaling method is possible that can be of value for certain types of inferences. Rather than standardizing d_{Mod} with respect to SD_{Y_1}, which expresses differences in prediction in terms of the referent group's overall criterion *SD*, one can standardize d_{Mod} with respect to the standard deviation of the residuals from the referent group's regression model. This alternative scaling puts d_{Mod} in the scale of the conditional standard deviation of referent criterion scores, which allows d_{Mod} to be interpreted as the average number of standard deviations by which the referent and focal groups differ, *conditional on their predictor score*.

With this alternative scaling, d_{Mod} effect sizes would be standardized with respect to the standard error of the referent group's regression model (i.e., the standard deviation of residuals). The referent group's conditional standard deviation of *Y* can be easily estimated using Equation 59.

$$SD_{Residuals} = SD_{Y_1}\sqrt{1 - r_{XY_1}^2} \tag{59}$$

This scaling works because one of the assumptions of linear regression is homoscedasticity, which means one must assume that the variance of criterion variable,

conditional on the value of the predictor variable, is constant across the entire distribution of predictor scores. When the assumption of homoscedasticity is satisfied, the conditional variance of *Y* should be equal across all levels of *X* and this conditional variance should be equal to the variance of residuals. Thus, one can use the *SD* of residuals to standardize differences in prediction in a scale that is conditional on predictor scores. Compared to effect sizes computed using the overall *SD* scaling, effect sizes computed using this conditional scaling will always be $\left(1 - r_{XY_1}^2\right)^{-.5}$ times larger in magnitude, which means that it is very simple to convert effect sizes from one metric to the other. The choice of scaling will generally have a small impact on effect sizes, but it is nonetheless important to choose the scaling method that supports the types of inferences one would like to draw from one's data.

For example, consider a hypothetical scenario in which the referent and focal groups have equal slopes, the referent group's intercept is .20 units higher than the focal group's intercept, the referent group's criterion standard deviation is 1.0, and the validity of the predictor in the referent group is $r = .50$. If one were to use the referent group's criterion *SD* to standardize the effect, the d_{Mod_Signed} effect size would be .20 / 1.0 = .20. That effect size indicates that, on average, the focal group's performance is overpredicted by .20 *SD* units when the referent group's regression line is used to forecast performance, relative to the referent group considered as a whole. Alternatively, d_{Mod} could be standardized with respect to the *SD* of residuals, which in this case would be $1 \times \sqrt{1 - .5^2} = 0.866$. Whereas 1.0 is the overall standard deviation of criterion scores (i.e., the *SD* irrespective of predictor score), 0.866 is the *SD* of criterion scores among

those with a particular predictor score. Standardizing by this conditional standard deviation produces a d_{Mod_Signed} estimate of 0.20 / 0.866 = 0.23. In this metric, d_{Mod_Signed} means that, on average, focal group members' performance is overpredicted by .23 standard deviations relative to referent group members with the same score (rather than relative to referent group members, considered as a whole).

Assuming that $SD_{Residuals}$ provides an intuitive metric for one's purpose, an advantage of scaling d_{Mod} by $SD_{Residuals}$ rather than SD_{Y_1} is that, like the subgroup regression lines, $SD_{Residuals}$ is invariant to direct selection (i.e., direct range restriction) on X (see Mulaik, 2010, pp. 408–412). This means that, if the referent and focal groups have equal slopes but different intercepts and applicants are selected into an organization on a top-down basis, d_{Mod_Signed} will have the same expected value in the applicant population and the population of selectees when it is scaled by $SD_{Residuals}$, regardless of the proportion of applicants selected. Unlike other effect sizes, d_{Mod_Signed} scaled by $SD_{Residuals}$ is not necessarily biased by top-down selection on the basis of predictor scores.

Standard Errors for d_{Mod_Signed} Effect Sizes

With algebraic versions of the d_{Mod} formulas established and Research Question 1 answered, I can now begin to address Research Question 2, which asked how the standard error of d_{Mod_Signed} could be estimated. A key advantage of the linear algebraic versions of d_{Mod_Signed} defined above is that the outputs of linear functions have well-defined standard errors. Specifically, the sampling variances and covariances of the inputs to a linear function can be combined into a composite variance to estimate the

sampling error of the output using the delta method, also known as a Taylor series approximation. The delta method is a method for expressing the propagation of error that occurs when variables are transformed and/or combined by mathematical functions. Simply put, the delta method is a way to analytically determine the linear regression weights that would result from regressing the output of a function on each of its inputs. The regression weights are found by taking the partial derivative of the function with respect to each of its inputs; these derivatives quantify the linear change in the output for each unit change in the inputs, exactly like any other linear regression coefficient. These weights can be used to combine the variances and covariances of input values to estimate the composite variance of the function's output values.

As an example of how the delta method could be applied, imagine that *X* and *Y* are statistics with known sampling distributions and that $Z = X + Y$. In this case, the partial derivatives of *Z* with respect to *X* and *Y* would both be 1; with both partial derivatives being equal to 1, the sampling variance of *Z* would be equal to the sum of the sampling variances of *X* and *Y* and two times the covariance between their sampling distributions (i.e., $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}$), just like any other unweighted composite variable. If, however, $Z = X \times Y$, the partial derivatives of *Z* with respect to *X* and *Y* would be *Y* and *X*, respectively, and the sampling variance of *Z* would be equal to $Y^2\sigma_X^2 + X^2\sigma_Y^2 + 2XY\sigma_{XY}$. When applying the delta method, the mean values of the inputs, together with the functional form of the formula, determine the ways in which the inputs' variances contribute to the output's variance.

The simplest implementation of the delta method involves computing a linear combination of variances with all covariances among terms assumed to be zero. These types of composite variance estimates can function quite well in some cases (e.g., artifact-distribution meta-analyses in which distributions of reliability and range-restriction statistics are assumed to be independent), but can produce highly inaccurate estimates if non-trivial associations exist among the distributions. As a simple example of this, consider that the variance of an equally weighted composite of two standardized variables that are correlated $r = .5$ is equal to 3 (i.e., $1 + 1 + 2 \times .5 = 3$), but the naïve estimate of this variance that assumes the variables do not covary would be 2 (i.e., $1 + 1 + 2 \times 0 = 2$), a 33% underestimate. The degree of misestimation from ignoring covariance in this example is serious and similar degrees of misestimation are possible when estimating sampling variances via the delta method.

The sampling distributions of inputs to the d_{Mod_Signed} formulas *do* covary and these covariances must be taken into account to accurately estimate the sampling variance of d_{Mod_Signed}. In addition to showing how the sampling variances of d_{Mod_Signed}'s input statistics can be combined via the delta method, I also describe methods for analytically determining the covariances among the inputs. In the following sections, I describe methods for estimating the standard error of d_{Mod_Signed} when computed using Equation 43 (with unstandardized statistics as inputs) or Equation 47 (with standardized effect sizes as inputs), as well as when either method for estimating d_{Mod_Signed} uses $SD_{Residuals}$ to standardize the effect size. After presenting these methods, I then briefly present simulation evidence supporting the accuracy of the analytic standard-error estimates.

Unfortunately, as the algebraic formulas for d_{Mod_under} and d_{Mod_over} rely on complex functions that introduce non-linearity (namely, the probability density and cumulative density functions of the normal distribution), they are not as well-suited to error-variance estimation via the delta method. As linear analytic error-variance approximations for d_{Mod_under} and d_{Mod_over} do not produce reliable estimates, bootstrapping and Monte Carlo estimation methods are the best options for estimating error variance because they do not assume linearity. The "compute_dmod" function in the *psychmeta* package for *R* (Dahlke & Wiernik, 2018, 2017/2019) can automate the process of bootstrapping d_{Mod} estimates.

Standard error for d_{Mod_Signed} computed using unstandardized inputs. The standard error of d_{Mod_Signed} can be estimated as a linear function of the multivariate sampling variances of the referent and focal groups' means, standard deviations, and validity coefficients; the sampling covariance matrix for these statistics can be used to compute a composite variance that represents the sampling variance of d_{Mod_Signed}.

The first step when estimating the sampling variance of d_{Mod_Signed} via the delta method is to obtain estimates of the variances and covariances of the input statistics' sampling distributions. The sampling variances of subgroup means can be estimated using Equation 60.

$$SE^2_{\bar{X}_1} = \frac{SD^2_{X_1}}{n_1} \quad (60a)$$

$$SE^2_{\bar{X}_2} = \frac{SD^2_{X_2}}{n_2} \quad (60b)$$

$$SE^2_{\bar{Y}_1} = \frac{SD^2_{Y_1}}{n_1} \quad (60c)$$

$$SE^2_{\bar{Y}_2} = \frac{SD^2_{Y_2}}{n_2} \quad (60d)$$

The corresponding sampling variances of subgroup *SD*s can be estimated using 61.

$$SE^2_{SD_{X_1}} = \frac{SD^2_{X_1}}{2(n_1 - 1)} \quad (61a)$$

$$SE^2_{SD_{X_2}} = \frac{SD^2_{X_2}}{2(n_2 - 1)} \quad (61b)$$

$$SE^2_{SD_{Y_1}} = \frac{SD^2_{Y_1}}{2(n_1 - 1)} \quad (61c)$$

$$SE^2_{SD_{Y_2}} = \frac{SD^2_{Y_2}}{2(n_2 - 1)} \quad (61d)$$

The sampling variance of the correlation between *X* and *Y* can be estimated using Equation 62 (Schmidt & Hunter, 2015, p. 101 Equation 3.7).

$$SE^2_{r_{XY_1}} = \frac{\left(1 - r^2_{XY_1}\right)^2}{n_1 - 1} \quad (62a)$$

$$SE^2_{r_{XY_2}} = \frac{\left(1 - r^2_{XY_2}\right)^2}{n_2 - 1} \quad (62b)$$

The sampling distributions of variables' means within a group are correlated by the same magnitude as are the variables themselves. The covariances among these sampling distributions can therefore be estimated using the product of the variables' correlations and the standard errors of the means as shown in Equation 63.

$$s_{\bar{X}\bar{Y}_1} = r_{XY_1} SE_{\bar{X}_1} SE_{\bar{Y}_1} \quad (63a)$$

$$s_{\bar{X}\bar{Y}_2} = r_{XY_2} SE_{\bar{X}_2} SE_{\bar{Y}_2} \quad (63b)$$

Correlations and standard deviations are expressions of variance and therefore also have correlated sampling distributions. The covariance between the sampling distributions of standard deviations is shown in Equation 64.

$$s_{SD_{X_1}SD_{Y_1}} = r^2_{XY_1} SE_{SD_{X_1}} SE_{SD_{Y_1}} \quad 64a$$

$$s_{SD_{X_2}SD_{Y_2}} = r^2_{XY_2} SE_{SD_{X_2}} SE_{SD_{Y_2}} \quad 64b$$

Whereas the correlation between the sampling distributions of means is equal to the correlation between *X* and *Y*, the correlation between the sampling distributions of standard deviations is equal to the squared correlation between *X* and *Y*. This is because standard deviations are computed as the square root of the mean of squared deviation scores and these squared deviation scores have a diminished degree of correlation compared to the original variables, as squaring linearly associated variables weakens the linear relationship. The expected value of the correlation between normally distributed variables whose values have been squared (i.e., chi-square distributed variables) is equal to the squared correlation between the original normally distributed variables; this association also applies to the standard deviations. Equation 64 produces estimates equal to those that can be obtained via Cheung and Chan's (2004) method for estimating the covariances among the sampling distributions of the elements of a covariance matrix.

The correlation between the sampling distributions of correlation coefficients and standard deviations can be estimated by making use of determinantal algebra. Specifically, the determinant of the correlation parameter between *X* and *Y* is equal to the determinant of the matrix of correlations among the sampling distributions of r_{XY}, SD_X, and SD_Y. Given that the expected values of the determinants for these two matrices are

equal, the determinant of the *X*-*Y* correlation matrix is known, and the correlation between the sampling distributions of SD_X and SD_Y is analytically estimable, the correlation between r_{XY} and SD_X can be easily solved. The correlation between r_{XY} and SD_Y can also be solved because it is equal to the correlation between r_{XY} and SD_X.

To solve for the correlation between r_{XY} and SD_X, one must first define the determinantal equivalence described above. The determinant of the 2x2 $\mathbf{R}_{XY}$ matrix containing the correlation between *X* and *Y* is shown in Equation 65.

$$|\mathbf{R}_{XY}| = 1 - r_{XY}^2 \tag{65}$$

The determinant of the 3x3 $\mathbf{R}_{SE}$ (where *SE* means sampling error) correlation matrix of relationships among the sampling distributions of r_{XY}, SD_X, and SD_Y is given in Equation 66.

$$\begin{aligned} |\mathbf{R}_{SE}| &= 1 - r_{r_{XY}SD_X}^2\left(2 - 2\, r_{SD_XSD_Y}\right) - r_{SD_XSD_Y}^2 \\ &= 1 - r_{r_{XY}SD_X}^2(2 - 2\, r_{XY}^2) - r_{XY}^4 \end{aligned} \tag{66}$$

By setting $|\mathbf{R}_{XY}|$ equal to $|\mathbf{R}_{SE}|$, one can solve for the unknown $r_{r_{XY}SD_X}$ value (which is also equal to the $r_{r_{XY}SD_Y}$ value). The process begins with defining the equivalence shown in Equation 67.

$$1 - r_{XY}^2 = 1 - r_{r_{XY}SD_X}^2(2 - 2\, r_{XY}^2) - r_{XY}^4 \tag{67}$$

Next, the term containing $r_{r_{XY}SD_X}$ can be isolated, as shown in Equation 68.

$$r_{r_{XY}SD_X}^2(2\, r_{XY}^2 - 2) = 1 - r_{XY}^4 - (1 - r_{XY}^2) \tag{68}$$

Further isolating the $r_{r_{XY}SD_X}$ coefficient and simplifying the result gives Equation 69.

$$r^2_{r_{XY}SD_X} = \frac{1 - r^4_{XY} - (1 - r^2_{XY})}{2 - 2\,r^2_{XY}} \quad 69$$

$$= \frac{r^2_{XY} - r^4_{XY}}{2 - 2\,r^2_{XY}}$$

$$= \frac{r^2_{XY}(1 - r^2_{XY})}{2(1 - r^2_{XY})}$$

$$= \frac{r^2_{XY}}{2}$$

The correlation between r_{XY} and SD_X (and between r_{XY} and SD_Y) is therefore $\sqrt{r^2_{XY}\,/2}$, as shown in Equation 70.

$$r_{r_{XY}SD_X} = r_{r_{XY}SD_Y} = \sqrt{\frac{r^2_{XY}}{2}} \quad 70$$

The result of Equation 70 is equal to the correlation between the sampling distributions of a correlation and the standard deviations of the covariates estimated via the methods introduced by Cheung and Chan (2004). The subgroup-specific sampling covariances for the associations between r_{XY} and SD_X and between r_{XY} and SD_Y can be computed as shown in Equations 71 and 72, respectively.

$$s_{r_{XY_1}SD_{X_1}} = \sqrt{\frac{r^2_{XY_1}}{2}}\,SE_{r_{XY_1}}SE_{SD_{X_1}} \quad 71a$$

$$s_{r_{XY_2}SD_{X_2}} = \sqrt{\frac{r^2_{XY_2}}{2}}\,SE_{r_{XY_2}}SE_{SD_{X_2}} \quad 71b$$

$$s_{r_{XY_1}SD_{Y_1}} = \sqrt{\frac{r_{XY_1}^2}{2}} SE_{r_{XY_1}} SE_{SD_{Y_1}} \quad 72a$$

$$s_{r_{XY_2}SD_{Y_2}} = \sqrt{\frac{r_{XY_2}^2}{2}} SE_{r_{XY_2}} SE_{SD_{Y_2}} \quad 72b$$

Unlike other statistics such as means, *SD*s, and correlations that are estimated the same way in all circumstances, there are many different methods for estimating reliability coefficients and the estimation method determines how the sampling variance should be computed. The sampling variance of a test-retest reliability coefficient, parallel-forms reliability coefficient, or other reliability coefficient that is computed as a Pearson correlation can be estimated using Equation 73.

$$SE_{r_{XX'}}^2 = \frac{\left(1 - r_{XX'}^2\right)^2}{N - 1} \quad 73$$

The sampling variance of coefficient alpha, which was derived by Duhachek and Iacobucci (2004), is given in Equation 74,

$$SE_{r_{XX'}}^2 = \frac{2k^2(\mathbf{1}^T\mathbf{S1}[\text{tr}(\mathbf{SS}) + \text{tr}(\mathbf{S})^2] - 2\text{tr}(\mathbf{S})\mathbf{1}^T\mathbf{SS1})}{N(k-1)^2(\mathbf{1}^T\mathbf{S1})^3} \quad 74$$

where k is the number of items in the scale, $\mathbf{S}$ is the inter-item covariance matrix, $\mathbf{1}$ is a column vector with as many entries as $\mathbf{S}$ has variables that consists entirely of 1s, and "tr" is the trace function for computing the sum of diagonal elements in a square matrix. Finally, the sampling variance of a generic reliability coefficient that does not have a dedicated sampling variance formula can be estimated using Equation 75,

$$SE_{r_{XX'}}^2 = \frac{4r_{XX}(1 - r_{XX'})^2}{N - 1} \quad 75$$

which is based on the definition of reliability as the squared correlation between observed scores and true scores.

Given the variety of ways in which reliability coefficients can be computed, it is not feasible to concretely estimate the association between the sampling distributions of reliability coefficients and other statistics. My simulations presented later indicate that modeling the multivariate sampling covariances involving reliabilities would contribute little to the estimates of corrected d_{Mod_Signed} sampling variances and can be rather safely constrained to zero. There is a strong precedent for not modeling the covariance between reliability coefficients and other statistics, as individual-correction psychometric meta-analyses involves correcting correlations for measurement error under the assumption that correlations and reliabilities coefficients have uncorrelated sampling distributions (cf. Hunter, Schmidt, & Le, 2006; Schmidt & Hunter, 2015).

The weights used to combine the sampling distributions estimated above into d_{Mod_Signed}'s composite sampling variance are the partial derivatives of the d_{Mod_Signed} formula (see Equation 47) with respect to each of its input values. The partial derivatives for when d_{Mod_Signed} is standardized using the referent group's overall criterion *SD* are given in Equation 76.

$$b_{\bar{Y}_1} = \frac{\partial d_{Mod}}{\partial \bar{Y}_1} = \frac{1}{SD_{Y_1}\sqrt{r_{YY_1'}}} \tag{76a}$$

$$b_{\bar{Y}_2} = \frac{\partial d_{Mod}}{\partial \bar{Y}_2} = -\frac{1}{SD_{Y_1}\sqrt{r_{YY_1'}}} \tag{76b}$$

$$b_{\bar{X}_1} = \frac{\partial d_{Mod}}{\partial \bar{X}_1} = -\frac{r_{XY_1}}{SD_{X_1} r_{XX_1'}\sqrt{r_{YY_1'}}} \tag{76c}$$

$$b_{\bar{X}_2} = \frac{\partial d_{Mod}}{\partial \bar{X}_2} = \frac{r_{XY_1}}{SD_{X_1} r_{XX_1'} \sqrt{r_{YY_1'}}} \quad 76d$$

$$b_{r_{XY_1}} = \frac{\partial d_{Mod}}{\partial r_{XY_1}} = \frac{\bar{X}_2 - \bar{X}_1}{SD_{X_1} r_{XX_1'} \sqrt{r_{YY_1'}}} \quad 76e$$

$$b_{SD_{Y_1}} = \frac{\partial d_{Mod}}{\partial SD_{Y_1}} = \frac{\bar{Y}_2 - \bar{Y}_1}{SD_{Y_1}^2 \sqrt{r_{YY_1'}}} \quad 76f$$

$$b_{SD_{X_1}} = \frac{\partial d_{Mod}}{\partial SD_{X_1}} = \frac{r_{XY_1}(\bar{X}_1 - \bar{X}_2)}{SD_{X_1}^2 r_{XX_1'} \sqrt{r_{YY_1'}}} \quad 76g$$

$$b_{r_{XX_1'}} = \frac{\partial d_{Mod}}{\partial r_{XX_1'}} = \frac{r_{XY_1}(\bar{X}_1 - \bar{X}_2)}{SD_{X_1} r_{XX_1'}^2 \sqrt{r_{YY_1'}}} \quad 76h$$

$$b_{r_{YY_1}} = \frac{\partial d_{Mod}}{\partial r_{YY_1'}} = -\frac{\left([Y_1 - \bar{Y}_2] - \frac{r_{XY_1} SD_{Y_1} [\bar{X}_1 - \bar{X}_2]}{SD_{X_1} r_{XX_1'}}\right)}{2 SD_{Y_1} r_{YY_1'}^{\frac{3}{2}}} \quad 76i$$

The partial derivatives of d_{Mod_Signed} for when the effect size is standardized using the referent group's residual criterion *SD* are given in Equation 77.

$$b_{\bar{Y}_1} = \frac{\partial d_{Mod}}{\partial \bar{Y}_1} = \frac{1}{SD_{Y_1} \sqrt{r_{YY_1'}} \sqrt{1 - \frac{r_{XY_1}^2}{r_{XX_1'} r_{YY_1'}}}} \quad 77a$$

$$b_{\bar{Y}_2} = \frac{\partial d_{Mod}}{\partial \bar{Y}_2} = -\frac{1}{SD_{Y_1} \sqrt{r_{YY_1'}} \sqrt{1 - \frac{r_{XY_1}^2}{r_{XX_1'} r_{YY_1'}}}} \quad 77b$$

$$b_{\bar{X}_1} = \frac{\partial d_{Mod}}{\partial \bar{X}_1} = -\frac{r_{XY_1}}{r_{XX_1'} SD_{X_1} \sqrt{r_{YY_1'}} \sqrt{1 - \frac{r_{XY_1}^2}{r_{XX_1'} r_{YY_1'}}}} \quad 77c$$

$$b_{\bar{X}_2} = \frac{\partial d_{Mod}}{\partial \bar{X}_2} = \frac{r_{XY_1}}{r_{XX_1'} SD_{X_1} \sqrt{r_{YY_1'}} \sqrt{1 - \frac{r_{XY_1}^2}{r_{XX_1'} r_{YY_1'}}}} \quad 77d$$

$$b_{r_{XY_1}} = \frac{\partial d_{Mod}}{\partial r_{XY_1}} = -\frac{\bar{X}_1 SD_{Y_1} r_{YY_1'} + \bar{Y}_2 r_{XY_1} SD_{X_1} - \bar{X}_2 SD_{Y_1} r_{YY_1'} - r_{XY_1} \bar{Y}_2 SD_{X_1}}{SD_{Y_1} r_{XX_1'} r_{YY_1'}^{\frac{3}{2}} SD_{X_1} \left(-\frac{r_{XY_1}^2 - r_{XX_1'} r_{YY_1'}}{r_{XX_1'} r_{YY_1'}} \right)^{\frac{3}{2}}} \quad 77e$$

$$b_{SD_{Y_1}} = \frac{\partial d_{Mod}}{\partial SD_{Y_1}} = \frac{\bar{Y}_2 - \bar{Y}_1}{SD_{Y_1} \sqrt{r_{YY_1'}} \sqrt{1 - \frac{r_{XY_1}^2}{r_{XX_1'} r_{YY_1'}}}} \quad 77f$$

$$b_{SD_{X_1}} = \frac{\partial d_{Mod}}{\partial SD_{X_1}} = \frac{r_{XY_1}(\bar{X}_1 - \bar{X}_2)}{r_{XX_1'} \sqrt{r_{YY_1'}} SD_{X_1}^2 \sqrt{1 - \frac{r_{XY_1}^2}{r_{XX_1'} r_{YY_1'}}}} \quad 77g$$

$$b_{r_{XX_1}} = \frac{\partial d_{Mod}}{\partial r_{XX_1'}} = \left(r_{XY_1} \left[-\bar{X}_1 SD_{Y_1} r_{XY_1}^2 + 2\bar{X}_1 SD_{Y_1} r_{XX_1'} r_{YY_1'} + \bar{Y}_2 r_{XY_1} r_{XX_1'} SD_{X_1} \right. \right.$$
$$\left. + SD_{Y_1} r_{XY_1}^2 \bar{X}_2 - 2SD_{Y_1} \bar{X}_2 r_{XX_1'} r_{YY_1'} - r_{XY_1} r_{XX_1'} \bar{Y}_1 SD_{X_1} \right] \Big) \quad 77h$$
$$\Big/ \left(2SD_{Y_1} r_{XX_1'}^3 r_{YY_1'}^{\frac{3}{2}} SD_{X_1} \left[\frac{r_{XX_1'} r_{YY_1'} - r_{XY_1}^2}{r_{XX_1'} r_{YY_1'}} \right]^{\frac{3}{2}} \right)$$

$$b_{r_{YY_1}} = \frac{\partial d_{Mod}}{\partial r_{YY_1'}} = \frac{\bar{X}_1 SD_{Y_1} r_{XY_1} + \bar{Y}_2 r_{XX_1'} SD_{X_1} - SD_{Y_1} r_{XY_1} \bar{X}_2 - r_{XX_1'} \bar{Y}_1 SD_{X_1}}{2SD_{Y_1} r_{XX_1'} r_{YY_1'}^{\frac{3}{2}} SD_{X_1} \left(\frac{r_{XX_1'} r_{YY_1'} - r_{XY_1}^2}{r_{XX_1'} r_{YY_1'}} \right)^{\frac{3}{2}}} \quad 77i$$

The sampling variances and covariances described above can be combined to represent the multivariate sampling distributions of all of the input statistics used to compute d_{Mod_Signed}, as shown in Equation 78 where a covariance matrix **S** is constructed.

$$\mathbf{S} = \begin{bmatrix} SE^2_{\bar{Y}_1} & 0 & 0 & 0 & 0 & s_{\bar{X}\bar{Y}_1} & 0 & 0 & 0 \\ 0 & SE^2_{\bar{Y}_2} & 0 & 0 & 0 & 0 & s_{\bar{X}\bar{Y}_2} & 0 & 0 \\ 0 & 0 & SE^2_{r_{XY_1}} & s_{r_{XY_1}SD_{Y_1}} & s_{r_{XY_1}SD_{Y_1}} & 0 & 0 & 0 & 0 \\ 0 & 0 & s_{r_{XY_1}SD_{X_1}} & SE^2_{SD_{X_1}} & s_{SD_{X_1}SD_{Y_1}} & 0 & 0 & 0 & 0 \\ 0 & 0 & s_{r_{XY_1}SD_{Y_1}} & s_{SD_{X_1}SD_{Y_1}} & SE^2_{SD_{Y_1}} & 0 & 0 & 0 & 0 \\ s_{\bar{X}\bar{Y}_1} & 0 & 0 & 0 & 0 & SE^2_{\bar{X}_1} & 0 & 0 & 0 \\ 0 & s_{\bar{X}\bar{Y}_2} & 0 & 0 & 0 & 0 & SE^2_{\bar{X}_2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & SE^2_{r_{XX'_1}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & SE^2_{r_{YY'_1}} \end{bmatrix} \quad (78)$$

The partial derivative weights computed using Equation 76 or Equation 77 can be organized into a weight vector denoted as **b** that is conformable with **S**, as shown in Equation 79.

$$\mathbf{b} = \begin{bmatrix} b_{\bar{Y}_1} & b_{\bar{Y}_2} & b_{r_{XY_1}} & b_{SD_{X_1}} & b_{SD_{Y_1}} & b_{\bar{X}_1} & b_{\bar{X}_2} & b_{r_{XX'_1}} & b_{r_{YY'_1}} \end{bmatrix}^T \quad (79)$$

Together, the matrix **S** and vector **b** can be used estimate the sampling variance of d_{Mod_Signed} via the quadratic-form equation for the variance of a composite, given in Equation 80.

$$SE^2_{d_{Mod\_signed}} = \mathbf{b}^T\,\mathbf{S}\,\mathbf{b} \quad (80)$$

The matrix-multiplication operation from Equation 80 can also be expressed as a scalar algebra equation, as shown in Equation 81.

$$\begin{aligned} SE^2_{d_{Mod\_signed}} = {} & b^2_{\bar{Y}_1} SE^2_{\bar{Y}_1} + b^2_{\bar{Y}_2} SE^2_{\bar{Y}_2} + b^2_{r_{XY_1}} SE^2_{r_{XY_1}} + b^2_{SD_{X_1}} SE^2_{SD_{X_1}} + b^2_{SD_{Y_1}} SE^2_{SD_{Y_1}} \\ & + b^2_{\bar{X}_1} SE^2_{\bar{X}_1} + b^2_{\bar{X}_2} SE^2_{\bar{X}_2} + b^2_{r_{XX'_1}} SE^2_{r_{XX'_1}} + b^2_{r_{YY'_1}} SE^2_{r_{YY'_1}} \\ & + 2\left(b_{r_{XY_1}} b_{SD_{X_1}} s_{r_{XY_1}SD_{X_1}} + b_{r_{XY_1}} b_{SD_{Y_1}} 2_{r_{XY_1}SD_{Y_1}}\right. \\ & \left. + b_{SD_{X_1}} b_{SD_{Y_1}} s_{SD_{X_1}SD_{Y_1}} + b_{\bar{X}_1} b_{\bar{Y}_1} r_{\bar{X}\bar{Y}_1} SE_{\bar{X}_1} SE_{Y_1} + b_{\bar{X}_2} b_{\bar{Y}_2} s_{\bar{X}\bar{Y}_2}\right) \end{aligned} \quad (81)$$

Standard error for d_{Mod_Signed} computed using standardized inputs. When d_{Mod_Signed} is computed from correlations and standardized *d* values, the standard error of d_{Mod_Signed} can be estimated as a linear function of the sampling variances and covariances of the predictor and criterion *d* values and the referent group's validity and reliability coefficients. The formulas for the sampling variances of correlations and reliabilities were presented in the previous section (see Equations 62, 73, 74, and 75) and the sampling variance of a *d* value can be estimated as shown in Equation 82.

$$SE_d^2 = \frac{n_1 + n_2}{n_1 \times n_2} + \frac{d^2}{2(n_1 + n_2)} \tag{82}$$

If the standard deviations of all variables in both groups are equal, as is assumed when using *d* values, the correlation between the sampling distributions of *d* values representing the mean differences of *X* and the mean differences of *Y* can be approximated via the delta method, as shown in Equation 83.

$$r_{d_X d_Y} = \frac{b_{\bar{X}_1} b_{\bar{Y}_1} r_{XY_1} SE_{\bar{X}_1} SE_{\bar{Y}_1} + b_{\bar{X}_2} b_{\bar{Y}_2} r_{XY_2} SE_{\bar{X}_2} SE_{\bar{Y}_2}}{\sqrt{\left(b_{\bar{X}_1}^2 SE_{\bar{X}_1}^2 + b_{\bar{X}_2}^2 SE_{\bar{X}_2}^2\right)\left(b_{\bar{Y}_1}^2 SE_{\bar{Y}_1}^2 + b_{\bar{Y}_2}^2 SE_{\bar{Y}_2}^2\right)}} \tag{83}$$

where $b_{\bar{X}_1} = b_{\bar{Y}_1} = 1$, $b_{\bar{Y}_1} = b_{\bar{Y}_1} = -1$, $SE_{\bar{X}_1} = SE_{\bar{Y}_1} = 1/n_1$, and $SE_{\bar{X}_2} = SE_{\bar{Y}_2} = 1/n_2$. After substituting replacing the weights and sampling variances in the equation with their definitions, $r_{d_X d_Y}$ can be estimated as shown in Equation 84.

$$r_{d_X d_Y} = \frac{(1)(1)r_{XY_1}\sqrt{1/n_1}\sqrt{1/n_1} + (-1)(-1)r_{XY_2}\sqrt{1/n_2}\sqrt{1/n_2}}{\sqrt{[(1)^2(1/n_1) + (-1)^2(1/n_2)][(1)^2(1/n_1) + (-1)^2(1/n_2)]}} \quad 84$$

$$= \frac{\frac{r_{XY_1}}{n_1} + \frac{r_{XY_2}}{n_2}}{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= \frac{\frac{r_{XY_1}n_2}{n_1n_2} + \frac{r_{XY_2}n_1}{n_1n_2}}{\frac{n_2}{n_1n_2} + \frac{n_1}{n_1n_2}}$$

$$= \frac{r_{XY_1}n_2 + r_{XY_2}n_1}{n_1 + n_2}$$

The covariance among sampling distributions of *d* values is then simply the product of $r_{d_X d_Y}$, SE_{d_X}, and SE_{d_Y}, as shown in Equation 85.

$$s_{d_X d_Y} = r_{d_X d_Y} SE_{d_X} SE_{d_Y} \quad 85$$

The covariances between r_{XY_1} and d_X and between r_{XY_1} and d_Y are assumed to be zero, as distributions of means are uncorrelated with distributions of correlations.

The weights used to combine the sampling distributions estimated above into d_{Mod_Signed}'s composite sampling variance are the partial derivatives of d_{Mod_Signed} with respect to each of its input values. The partial derivatives for when d_{Mod_Signed} is standardized using the referent group's overall criterion *SD* are given in Equation 86.

$$b_{d_Y} = \frac{\partial d_{Mod}}{\partial d_Y} = \frac{1}{\sqrt{r_{YY_1'}}} \quad 86a$$

$$b_{r_{XY_1}} = \frac{\partial d_{Mod}}{\partial r_{XY_1}} = -\frac{d_X}{r_{XX_1'}\sqrt{r_{YY_1'}}} \quad 86b$$

$$b_{d_X} = \frac{\partial d_{Mod}}{\partial d_X} = -\frac{r_{XY_1}}{r_{XX_1'}\sqrt{r_{YY_1'}}} \quad 86c$$

$$b_{r_{XX'_1}} = \frac{\partial d_{Mod}}{\partial r_{XX'_1}} = \frac{d_X r_{XY_1}}{r_{XX'_1}^2 \sqrt{r_{YY'_1}}} \quad (86d)$$

$$b_{r_{YY'_1}} = \frac{\partial d_{Mod}}{\partial r_{YY'_1}} = \frac{d_X r_{XY_1} - d_Y r_{XX'_1}}{2 r_{XX'_1} r_{YY'_1}^{\frac{3}{2}}} \quad (86e)$$

The partial derivatives for when d_{Mod_Signed} is standardized using the referent group's residual criterion *SD* are given in Equation 87.

$$b_{d_Y} = \frac{\partial d_{Mod}}{\partial d_Y} = \frac{1}{r_{YY'_1}\sqrt{1 - \frac{r_{XY_1}^2}{r_{XX'_1} r_{YY'_1}}}} \quad (87a)$$

$$b_{r_{XY_1}} = \frac{\partial d_{Mod}}{\partial r_{XY_1}} = \frac{d_Y r_{XY_1} - d_X r_{YY'_1}}{r_{XX'_1} r_{YY'_1}^2 \left(\frac{r_{XX'_1} r_{YY'_1} - r_{XY_1}^2}{r_{XX'_1} r_{YY'_1}}\right)^{\frac{3}{2}}} \quad (87b)$$

$$b_{d_X} = \frac{\partial d_{Mod}}{\partial d_X} = -\frac{r_{XY_1}}{r_{XX'_1} r_{YY'_1}\sqrt{1 - \frac{r_{XY_1}^2}{r_{XX'_1} r_{YY'_1}}}} \quad (87c)$$

$$b_{r_{XX'_1}} = \frac{\partial d_{Mod}}{\partial r_{XX'_1}} = -\frac{d_Y r_{XY_1}^2 r_{XX'_1} + r_{XY_1} d_X \left(r_{XY_1}^2 - 2 r_{XX'_1} r_{YY'_1}\right)}{2 r_{XX'_1}^3 r_{YY'_1}^2 \left(\frac{r_{XX'_1} r_{YY'_1} - r_{XY_1}^2}{r_{XX'_1} r_{YY'_1}}\right)^{\frac{3}{2}}} \quad (87d)$$

$$b_{r_{YY'_1}} = \frac{\partial d_{Mod}}{\partial r_{YY'_1}} = \frac{\left(r_{XY_1}^2 - 2 r_{XX'_1} r_{YY'_1}\right)\left(d_Y r_{XX'_1} - d_X r_{XY_1}\right)}{2 r_{XX'_1}^2 r_{YY'_1}^3 \left(\frac{r_{XX'_1} r_{YY'_1} - r_{XY_1}^2}{r_{XX'_1} r_{YY'_1}}\right)^{\frac{3}{2}}} \quad (87e)$$

As with the method for computing the sampling variance of d_{Mod_Signed} estimates based on unstandardized input statistics, the sampling variances and covariances can be organized into a matrix called **S** (see Equation 88) and the partial derivatives can be organized into a conformable weight vector called **b** (see Equation 89).

$$\mathbf{S} = \begin{bmatrix} SE_{d_Y}^2 & 0 & s_{d_X d_Y} & 0 & 0 \\ 0 & SE_{r_{XY_1}}^2 & 0 & 0 & 0 \\ s_{d_X d_Y} & 0 & SE_{d_X}^2 & 0 & 0 \\ 0 & 0 & 0 & SE_{r_{XX'_1}}^2 & 0 \\ 0 & 0 & 0 & 0 & SE_{r_{YY'_1}}^2 \end{bmatrix} \quad (88)$$

$$\mathbf{b} = \begin{bmatrix} b_{d_Y} & b_{r_{XY_1}} & b_{d_X} & b_{r_{XX'_1}} & b_{r_{YY'_1}} \end{bmatrix}^T \quad (89)$$

S and **b** can be used estimate the sampling variance of d_{Mod_Signed} via the quadratic-form formula from Equation 80. Alternatively, $E_{d_{Mod\_signed}}^2$ can also be expressed as a scalar algebra equation, as shown in Equation 90.

$$SE_{d_{Mod\_signed}}^2 = b_{d_Y}^2 SE_{d_Y}^2 + b_{r_{XY_1}}^2 SE_{r_{XY_1}}^2 + b_{d_X}^2 SE_{d_X}^2 + b_{r_{XX'_1}}^2 SE_{r_{XX'_1}}^2 + b_{r_{YY'_1}}^2 SE_{r_{YY'_1}}^2 + 2\, b_{d_X}\, b_{d_Y}\, s_{d_X d_Y} \quad (90)$$

Simulation examining the accuracy of analytically estimated correlations among sampling distributions. The correlations among sampling distributions of d_{Mod_Signed}'s input statistics described above during the course of presented the standard-error estimators are very reliable estimates of the associations observed among statistical distributions. As an illustration of this, I simulated 10,000 Monte Carlo samples for each of four different sample sizes (50, 100, 500, and 1,000 cases) and each of 10 different correlations between *X* and *Y* (ranging from 0 to .9 in increments of .1). As the *X*-*Y* correlation is the only statistic necessary to estimate the correlations among sampling distributions, the parameters of all other values were constrained.

In each iteration of the simulation, two samples were generated: A primary sample (used to obtain estimates of correlations, *SD*s, and means) and a secondary

sample (whose means were contrasted with the means of the primary sample to obtain estimates of d_X and d_Y). The primary and secondary samples within each condition had equal sample sizes and their data were generated from the same correlation parameter and the same *SD* parameters for *X* and *Y* (all *SD* parameters were set to 1). The mean parameters for *X* and *Y* were set to 0 in all primary samples and were set to 1 in all secondary samples; the magnitude of mean differences between samples is irrelevant to the correlation between the sampling distributions of *d* values, but I chose to simulate a 1 *SD* difference between samples for the sake of having a non-trivial difference.

After computing the correlations, *SD*s, means, and *d* values from the simulated samples, I computed the correlations among the statistics observed in each condition. I also computed the determinant of the correlation matrix including correlations and *SD*s in each condition so that the correspondence between $|\mathbf{R}_{XY}|$ and $|\mathbf{R}_{SE}|$ could be evaluated. As shown in Figure 5, the relationships between the analytically estimated values and the observed simulated values closely follow a line depicting a perfect association. Any deviations from a perfect correspondence are well within the margins of what one would expect from Monte Carlo simulation error; similar to sampling error, simulated sampling distributions asymptotically approximate the infinite sampling distributions they are meant to represent and all simulations with a finite number of iterations will produce parameter estimates depart at least trivially from the true parameters of the infinite sampling distributions.

As the correlations among sampling distributions play an important role in estimating the sampling variance of d_{Mod_Signed}, the evidence in Figure 5 that these

correlations can be estimated accurately provides indirect support for the accuracy of my delta-method procedures for computing $SE^2_{d_{Mod\_signed}}$. In the next section, I directly test the accuracy of my procedures by comparing analytic standard error estimates to standard errors generated via Monte Carlo simulation.

Simulation examining the accuracy of standard error estimates for d_{Mod_Signed}. Having demonstrated the accuracy of the estimates of correlations among sampling distributions used in the formulas for the standard error of d_{Mod_Signed}, I designed a simulation to demonstrate the convergence between analytically estimated d_{Mod_Signed} standard errors and Monte Carlo estimates. This simulation evaluates the accuracy of standard errors computed using observed data, data corrected for measurement error in *X* and *Y*, and data corrected for measurement error in *Y* only; the simulation also examines the standard errors of d_{Mod_Signed} statistics computed using unstandardized inputs and standardized effect-size inputs.

In the simulation, I varied (1) sample size, (2) the proportion of the sample belonging to the referent group, (3) the referent group's *X*-*Y* correlation, (4) the ratio of the focal group's *X*-*Y* correlation to the referent group's *X*-*Y* correlation, (5) the referent-focal standardized mean difference on *X*, (6) the referent-focal standardized mean difference on *Y*, and (7) the reliability of *Y*. I held the reliability of *X* constant at .9 to represent the high level of reliability expected of professionally developed predictor variables used to make operational selection decisions. All parameter values used in the simulation are shown in Table 3. The parameters were fully crossed for a total of 288 conditions and 10,000 Monte Carlo samples were generated per condition. In each

condition, d_{Mod_Signed} estimates were computed for observed scores, operational scores (the predictor was treated as an observed variable and the criterion was corrected for measurement error), and true scores (both the predictor and criterion were corrected for measurement error). In each condition, each type of d_{Mod_Signed} estimate just described was computed using unstandardized inputs as well as standardized effect-size inputs.

The results of the simulation are depicted in Figure 6. The results for observed-score estimates show a very strong correspondence between analytically estimated standard errors and the observed Monte Carlo standard deviations. The association between analytic and Monte Carlo estimates were close to perfect using both unstandardized and standardized input values. The results for true-score and operational-score estimates also show a strong correspondence between analytic standard-error estimates and Monte Carlo standard deviations, but with analytic estimates demonstrating a small negative bias. This negative bias is due to the assumption that reliability coefficients' sampling distributions are uncorrelated with the sampling distributions of other input statistics. However, in light of the complexity of modeling how reliability sampling distributions relate to other sampling distributions and the fact that it is a common practice to correct the standard errors of effect sizes for measurement error without accounting for the covariances among sampling distributions (cf. Hunter et al., 2006; Schmidt & Hunter, 2015), the analytically estimated standard errors for true-score and operational-score d_{Mod_Signed} values are substantially similar to the true values and would support useful statistical inferences.

Discussion

The d_{Mod} effect sizes derived by Nye and Sackett (2017) and Dahlke and Sackett (2018) represent an important advancement in the quantification of differential prediction, as they summarize the magnitude of differences between subgroup's regression formulas in a standardized metric. These statistics offer a much-needed index of effect size to complement the significance-testing approach outlined by the Cleary model of predictive bias. However, the need for numeric integration in the original set of formulas represents a potential barrier for the adoption of d_{Mod} effect sizes because they require usage of software capable of performing such an operation. In this study, I removed this barrier by simplifying the d_{Mod} formulas into algebraic equations that can be computed using commonplace software. My new formula for d_{Mod_Signed} is purely algebraic and the formulas for d_{Mod_Under}, d_{Mod_Over}, and $d_{Mod\_Unsigned}$ can be computed in any program that can implement the cumulative density function of the normal distribution (including Microsoft Excel or even the free Apache OpenOffice "Calc" program). I also derived a closed-form procedure for accurately estimating the standard error of d_{Mod_Signed}, which not only allows researchers to construct confidence intervals around d_{Mod_Signed} estimates in primary research, but will also allow differential prediction effect sizes to be meta-analyzed across studies that use different measures of predictor and performance constructs.

In summary, it is now possible to compute d_{Mod} effect sizes algebraically and estimate closed-form standard errors for d_{Mod_Signed}, which makes it possible to meta-analyze differential prediction effects. The potential for meta-analyzing d_{Mod_Signed}

estimates is a particularly exciting prospect, as differential prediction effects (e.g., regression coefficients indicating intercept and slope differences) have previously only been meta-analyzable when the predictor and criterion were in the same metric across all studies. By using my standard-error estimation procedures along with my algebraic formulas for d_{Mod_Signed}, future researchers will be able to pool differential prediction effects across contexts that examine the same predictor and criterion constructs, even if different measures of those constructs are reported.

Similar to how Cleary model analyses can be performed on composite predictors to examine differential prediction in a complete selection system, d_{Mod_Signed} effect sizes can be computed for composite predictors the same way they are computed for individual predictor variables. By estimating d_{Mod_Signed} for composite predictors, one can express the magnitude of system-wide differential prediction for a selection program. No prior study has detailed the effect of forming composites of magnitudes of d_{Mod_Signed} values, but such information would be useful so that psychologists using d_{Mod} can anticipate the effects of composites on d_{Mod_Signed}, just as they can anticipate the effects on other effect sizes such as correlations and *d* values. The following study fills this research gap and describes how d_{Mod_Signed} estimates are affected by the formation of composites.

Study 2: Effects of Forming Composite Predictors on Magnitudes of Differential Prediction

Performance is determined by a multitude of factors (Campbell, 1990) and thoughtfully developed selection systems account for this by gathering data on several predictor variables that each make a unique contribution to forecasted levels of performance. Given that multiple predictors are relevant to most, if not all, forms of performance and that real-world selection systems seldom rely on a single predictor, it is generally recommended that differential prediction be evaluated at the level of the selection system rather than at the level of individual predictors (Sackett et al., 2003; SIOP, 2018). Testing several predictors separately for differential prediction when they are collectively used to make selection decisions fails to capture how data are operationally used and can provide misleading indications of the differential prediction associated with a composite of the predictors. Focusing differential prediction analyses on composite predictors will not only do a better job of reflecting the implications of operational data usage for differential prediction, it will also natural avoid the interpretative issues associated with the omitted variables problem (Sackett et al., 2003) and multicollinearity. As composite predictors are the recommended focus of differential prediction analyses according to SIOP's (2018) *Principles* and the effect of forming composites on d_{Mod} effect sizes (see Study 1) has not previously been explored, the present study fills this gap by describing the implications of composites for d_{Mod} computations. This study also illustrates the effects of composites on d_{Mod} using three

popular methods for forming composites: Unit weighting, regression weighting, and Pareto-optimal weighting.

The effect of forming composites on the magnitudes of d_{Mod} effect sizes can be best understood by considering the formula for computing d_{Mod_Signed} from correlations and *d* values (see Equation 39 from Study 1), which was defined as:

$$d_{Mod\_Signed} \cong d_Y - r_{XY_1} d_X.$$

From this equation, it is clear that the direction and magnitude of differential prediction are determined by two key things: The magnitude of subgroup mean differences on the criterion, as indexed by d_Y, and the magnitude of the product of r_{XY_1} and d_X, which indicates the magnitude of subgroup mean differences on the predicted criterion scores when the referent group's regression equation is used to make predictions. With all else being equal, higher d_Y parameter values increase the probability of observing overprediction and higher r_{XY_1} and/or d_X parameter values increase the probability of observing underprediction. Below, I describe how this equation sheds light on Research Question 3 ("How are d_{Mod_Signed} effect sizes affected by the formation of composite predictors?") and Research Question 4 ("How does the usage of Pareto-optimal weighting solutions affect d_{Mod_Signed} effect sizes?").

Unless predictors are completely redundant with each other, their intercorrelations will be less than 1.00 in absolute value, such that a composite of predictors will have a validity coefficient larger than the average validity coefficient of the individual predictors. Similarly, as was shown by Sackett and Ellingson (1997), a composite predictor will also tend to have a *d* value larger than the average *d* value of its component

predictors because of incomplete overlap in the predictors' variance. All else equal, the smaller the average intercorrelation among predictors, the greater will be both the validity and the *d* value of the composite predictor. Given that White-minority mean differences tend to be positive and that predictors are generally analyzed in such a way that validity coefficients are positive, this means that predictor-combination practices that increase the validity of a composite predictor will also tend to increase mean differences on the composite, which increases the $r_{XY_1} \times d_X$ product. This implies that if one combines two predictors, each of which exhibits overprediction of focal group performance when analyzed separately, the composite of the two will exhibit a smaller magnitude of overprediction than either of its composites because the $r_{XY_1} \times d_X$ product increases in magnitude while d_Y remains unchanged. In fact, depending on the magnitude of mean differences on the criterion, it is theoretically possible that the increase of the $r_{XY_1} \times d_X$ product from combining two predictors that each individually overpredict performance could even result in a composite predictor that *underpredicts* performance.

When applied to a given data set, composite-formation practices such as unit weighting (i.e., giving equal weight to all variables) and regression weighting each yield one set of weights per set of predictor variables; there is only one set of unit weights that can be used and there is a single set of linear-regression weights that is optimal for explaining variance in the criterion variable using a given set of predictors. When using these weighting strategies, one can only change the expected degree of differential prediction associated with a composite by adding or removing predictors from the composite. However, Pareto-optimal weighting (De Corte et al., 2007) can yield infinite

sets of optimal weights for a given set of predictors, with each set representing a compromise between the objectives of maximizing validity and minimizing adverse-impact potential; each of these compromises is necessarily associated with a particular degree of differential prediction.

Whereas use of unit weighting or regression weighting with a given set of predictors produces a single d_{Mod_Signed} estimate (such that, after choosing a weighting strategy, the degree of differential prediction is dependent entirely upon which predictors are used), use of Pareto weighting means that there are infinite possible d_{Mod_Signed} values that could result from a single set of predictors (such that the degree of differential prediction is a result of one's preferred validity-diversity tradeoff in addition to the predictors one has chosen to use). If one prefers a tradeoff that favors minimizing adverse impact over maximizing validity, the resulting d_{Mod_Signed} value will be more likely to indicate overprediction, as both the validity estimate and the *d* value associated with the composite predictor will be smaller than would have occurred had one chosen a tradeoff that placed greater importance on maximizing validity. However, placing greater emphasis on validity will result in less overprediction because the $r_{XY_1} \times d_X$ product will be larger. The average amount of differential prediction in a selection system is a function of how the predictors are combined and it happens that Pareto weighting optimizes the same two predictor attributes (validity and mean differences) that are most critical for determining the magnitude of differential prediction.

The remainder of this study is dedicated to illustrating the principles described above by computing validity coefficients, White-Black *d* values, and d_{Mod_Signed} values for

composite predictors computed from meta-analytic correlations and *d* values. To convincingly demonstrate the effects of composite-formation methods on d_{Mod_Signed} estimates, the predictor variables to be combined into composites that represent varying degrees of validity for predicting job performance and varying degrees of adverse-impact potential for Black job applicants. Note that the value of d_Y will be unaffected by one's choice of predictor(s) or how one choses to compute composites, as this value can only be changed by altering the criterion variable; thus, I focus on the implications of r_{XY_1} and d_X in my demonstrations.

Method

For my demonstrations, I used the meta-analytic correlation matrix and vector of meta-analytic White-Black *d* values compiled by Song, Wee, and Newman (2017) to compute validities, standardized mean differences, and d_{Mod_Signed} effect sizes for composites consisting of varying sets of predictors using varying methods of assigning weight to predictors. The correlations and *d* values from Song et al. (2017) are shown in Table 4; these meta-analytic effect sizes were already corrected for artifacts, so no reliability corrections were included in my d_{Mod_Signed} computations. Note that Table 4's meta-analytic validities include between-group variance in addition to within-group variance, but Equations 39 and 47 call for one to use referent-group coefficients to compute d_{Mod_Signed}; the combined-group validities in Table 4 are therefore unsuitable for use in the d_{Mod_Signed} formula because they overestimate within-group validity. To account for this in my analyses, I partialed between-group variance out of the combined-group validity estimates to obtain estimates of within-group validity that could be used with

Equations 39 and 47. I note that my use of these pooled within-group validity estimates assumes equal slopes for White and Black samples; this would not be ideal for operational predictive bias analyses, but will be sufficient for achieving the illustrative goals of the present study. I computed unit-weighted and regression-weighted composites for all possible combinations of the predictors show in Table 4, with predictor sets ranging in size from two to five variables. I used the *ParetoR* package for *R* by Q. Chelsea Song (2017/2018) to compute Pareto-optimal solutions for the five predictors.

Results

Table 5 shows the effect sizes for composites consisting of different sets of predictors. The results for both unit- and regression-weighted composites show that composites consisting of more predictors, on average, exhibited larger validities and larger *d* values, which gave rise to larger products of those values and correspondingly smaller d_{Mod_Signed} estimates. The mean d_{Mod_Signed} effect size for individual predictors was .330 and smaller mean effect sizes were observed for composites. The mean d_{Mod_Signed} effect sizes for unit-weighted composites were .275 with two predictors (regression = .227), .237 with three predictors (regression = .146), .211 with four predictors (regression = .084), .193 with five predictors (regression = .038). As the mean d_{Mod_Signed} estimate for individual predictors was .330, these analyses offer a clear answer to Research Question 3: There is a distinct trend that larger composites are less likely to demonstrate overprediction of minority performance. Not only did larger composites tend to exhibit less overprediction, regression-weighted composites exhibited less overprediction than

their unit-weighted counterparts because regression composites had larger validities and *d* values.

Table 6 shows the effect sizes and adverse-impact ratios for 21 Pareto-optimal predictor composites. The Pareto solutions support the trends observed in Table 5: Solutions that had larger *d* values (i.e., greater adverse-impact potential) and larger validities also had smaller d_{Mod_Signed} effect sizes. The association between composite validity coefficients and d_{Mod_Signed} values is depicted in Figure 7 and the association between composite validity coefficients and d_{Mod_Signed} values is depicted in Figure 8. To answer Research Question 4, these Pareto solutions show that overprediction of minority performance is more likely when one uses Pareto-optimal weights that give greater emphasis to minimizing adverse-impact potential and less emphasis to maximizing validity.

Discussion

This study was designed to investigate the effect that forming composite predictors has on d_{Mod} effect sizes. The results of my illustrative analyses using meta-analytic data supported the trends that I anticipated based on my breakdown of the algebraic d_{Mod_Signed} formula that I derived in Study 1. A composite predictor produces smaller d_{Mod_Signed} effect sizes than does its average component predictor because, compared to the average component, a composite has a larger validity coefficient and a larger *d* value, and both of these factors lead to a lower d_{Mod_Signed} estimate (Research Question 3). Furthermore, when Pareto-optimal weights are used, solutions that give greater emphasis to validity will produce lower d_{Mod_Signed} estimates (i.e., a reduced

magnitude of overprediction) than will solutions that give greater emphasis to adverse-impact mitigation (Research Question 4). These insights offer much-needed insight into how d_{Mod_Signed} estimates are affected by combining multiple predictors into a single composite. These findings also nicely complement past work on the omitted variables problem and coincide closely in time with the release of SIOP's (2018) updated *Principles* in which researchers are advised to analyze composites for differential prediction rather than limiting their analyses to individual predictor variables.

My finding that composite predictors show less differential prediction is noteworthy not only because of the substantive implication that more inclusive sets of predictors are likely to help organizations predict performance more consistently across groups by avoiding the omitted variables problem, but also because of the technical implication that the d_{Mod_Signed} effect size functions quite differently from other effect sizes. Effect sizes such as correlations and *d* values are well-known to increase in magnitude as more indicators are added to a composite, but this trend does not necessarily apply to d_{Mod_Signed}. The direction of the change in d_{Mod_Signed} as indicators are added to a composite is predictable, but the absolute magnitude of a composite d_{Mod_Signed} effect size is not as systematic because d_{Mod_Signed} is affected by a larger number of factors. Assuming that validity is positive, d_{Mod_Signed} values will *decrease* in value as indicators are added to a composite if there are positive mean differences on predictors (i.e., the referent means are higher), but the d_{Mod_Signed} values will *increase* in value if there are negative mean differences on predictors.

Although the analyses presented in this study clearly show the practical impact of composites on differential prediction effects sizes, they suffer from several limitations. First, as I noted in the Methods section, I used estimates of pooled within-group validity to compute d_{Mod_Signed} rather than using validity coefficients that were specific to the referent group. This analysis choice was prudent given that meta-analytic estimates of subgroup validity coefficients are not available for all of the predictors examined here and I determined that within-group validities were sufficient to demonstrate a methodological phenomenon. However, this also obviously means that the d_{Mod_Signed} estimates reported here are only approximations of the real differential prediction effects and do not represent "true" estimates of differential prediction. Another limitation of this study is that my d_{Mod_Signed} estimates represent approximations of the average degree of differential prediction associated with different predictors, but do not provide insight into the random-effects variability of d_{Mod_Signed}.

This study demonstrated that a composite predictor tends to exhibit a smaller magnitude of differential prediction than its average component. In Study 3, I conceptually replicate the effects of forming composites on differential prediction analyses. I also test whether differences in prediction generalize across settings by meta-analyzing differential prediction effects computed from a large post-secondary education database.

Study 3: Testing the Generalizability of Differential Prediction in Post-Secondary Admissions Settings

Study 2 demonstrated that the magnitudes of differential prediction observed for composite predictors can be quite different from the magnitudes observed for the composites' components. The present study builds upon those findings by examining how statistical artifacts impact differential prediction analyses performed on real post-secondary admissions data featuring both individual predictors (i.e., high school GPAs and SAT subtest scores for Critical Reading, Mathematics, and Writing) and composite predictors (i.e., SAT composite scores and combinations of SAT scores and high school GPAs). Beyond expanding upon and replicating the findings of Study 2, Study 3 also tests the generalizability of differential prediction for White-Black, White-Hispanic, and male-female contrasts, with differential prediction quantified using d_{Mod_Signed}, subgroup intercept-difference regression coefficients, and subgroup slope-difference regression coefficients. By analyzing individual predictor variables along with composite predictors, I aim to provide a clearer indication of the generalizability of differential prediction of operational selection systems in post-secondary education than has previously been available (cf. Aguinis et al., 2016, who tested multiple predictors' differential prediction simultaneously in regression models heavily influenced by multicollinearity).

The focus of this study is on differential prediction in operational selection systems, which, as described in the Introduction chapter, means that all artifacts save for predictor measurement error should be accounted for. Operational estimates of differential prediction indicate how different subgroups' regression equations differ when

computed using applicant predictor data and incumbent criterion data that have been properly corrected for range restriction. Additionally, I correct for criterion measurement error in effect sizes to obtain fully operational estimates of differential prediction. Although criterion measurement error does not bias the expected values of regression coefficients, it does inflate the error variance of observed coefficients and accounting for it meta-analyses may help to account for additional artifactual variance and therefore arrive at better conclusions regarding generalizability. Correcting for criterion measurement error also allows for more accurate estimates of d_{Mod_Signed} effect sizes because, like other effect sizes, d_{Mod_Signed} is attenuated by measurement error in the dependent variable.

Method

Participants. Participants were 1,319,998 students enrolled at 270 U.S. colleges and universities that contributed to a data-collection initiative led by the College Board. The College Board also provided covariance matrices and means for the predictor scores of 8,969,539 applicants to my sample of institutions.

Measures. The College Board provided students' scores on each of the three SAT subtests: Critical Reading, Mathematics, and Writing. Students provided their self-reported high school GPAs (HSGPAs) in a survey administered by the College Board and schools reported students' first-year college GPAs for inclusion in the College Board's database.

Procedure. My analysis procedures consisted of several steps, including computing observed subgroup descriptive statistics, correcting for criterion measurement

error, correcting for range restriction, computing composite predictors, computing d_{Mod_signed} effect sizes, applying the Cleary model of bias to observed and artifact-corrected data, and meta-analyzing d_{Mod_Signed} effect sizes and regression coefficients. Each of these steps is detailed below.

Computation of observed subgroup means and covariances. I computed covariance matrices and vectors of means for each subgroup's first-year GPAs, SAT scores, and self-reported HSGPAs at each school in the database. For each school, I organized subgroup-specific applicant norms provided by the College Board; these norms represented the unrestricted means, variances, and covariances of all predictor variables. Subgroup norms were reported at the level of entering cohorts, but my interest was focused on institution-level differential prediction trends. Thus, for each subgroup at each school, I merged the multivariate predictor distributions from all cohorts into a single applicant mixture distribution that represented the combined within-cohort and between-cohort variance in predictor scores. Mixture distributions were constructed using the "mix_matrix" function from the *psychmeta* R package (Dahlke & Wiernik, 2018, 2017/2019).

In addition to my analyses of subgroup-specific data, I also performed the above operations on the complete dataset from each school that contained data from all subgroups. I then pooled the school-wise distributions of overall predictor information to obtain a distribution that represented the sample-size weighted average means, variances,

and covariances of predictors across all schools my analysis. These overall estimates of range-restricted and applicant data were necessary to support later steps in my procedure.

Corrections for criterion measurement error. I corrected observed fist-year GPA criteria for measurement error using subgroup-specific reliability estimates based on the internal consistency of students' first-year grades at each school. Reliability corrections were performed separately on the data from White, Black, and Hispanic samples, as well as male and female samples. I estimated reliability coefficients by computing the intraclass correlation coefficient (ICC_1) for first-year course grades (which indicates the average intercorrelation among individual students' individual course grades). After this, I used the Spearman-Brown formula to step-up each ICC_1 estimate by the average number of courses the students took to arrive at an estimate of the internal consistency of the grades that contributed to students' first-year GPAs. Summaries of the distributions of subgroup reliability coefficients are displayed in Table 7**Error! Reference source not found.**.

Range-restriction corrections. I used the Aitken-Lawley multivariate range-correction procedure (Aitken, 1934; Lawley, 1943) to correct first-year GPAs for range restriction using applicant norms from the College Board. Specifically, I used the "correct_matrix_mvrr" function from the *psychmeta R* package (Dahlke & Wiernik, 2018, 2017/2019) to correct covariance matrices for range restriction and I used *psychmeta*'s "correct_means_mvrr" to correct the criterion mean for range-restriction. When I corrected for both criterion measurement error and range restriction, I made corrections for criterion measurement error first because the criterion reliability

coefficients were based on range-restricted data; according to Hunter, Schmidt, and Le (2006), the measurement error of range-restricted criteria should always be corrected before making corrections for predictor range restriction.

Composites. To evaluate the differential prediction associated with different methods of using predictor scores to make holistic evaluations, I created four different composite predictors to include in my analyses. I created (1) a traditional two-test SAT composite consisting of equally weighted Critical Reading and Mathematics scores, (2) a three-test SAT composite consisting of equally weighted Critical Reading, Mathematics, and Writing scores, (3) an equally weighted composite of the two-test SAT composite and HSGPA, and (4) an equally weighted composite of the three-test SAT composite and HSGPA. The standardized weights assigned to predictors in each composite are displayed in Table 8.

It was during this process of computing composite predictors that it was necessary to use the estimates of overall predictor variances pooled across schools mentioned earlier. Regression coefficients are only meta-analyzable when the variables from all samples share the same metric, so it was necessary that comparable weights be used in forming composite predictors across all schools. When unit weights are used to form a composite, each variable only gets equal weight if its variance is taken into account. Without either standardizing the variables or converting the weights to an unstandardized metric, predictors will end up being weighted by their standard deviations; predictors with greater variance will unintentionally receive more weight. Thus, I used the pooled overall predictor variances to scale the unstandardized weights used in each sample to

ensure that the variances of the resulting composites would be on the same metric across schools. If not for this re-scaling procedure, each school's composite predictors would have had idiosyncratic variances because unit-weights would have been defined without regard for the predictor variances observed at the school. Applying consistent unstandardized metrics to all weights supports meaningful cross-school comparisons.

After all composites were formed, all subgroups' data were re-scaled relative to the pooled total-sample standard deviations of all variables across schools. This re-scaling process represented a pseudo-standardization procedure: All variables were standardized with respect to the pooled standard deviations, which kept between-sample variation in variance intact but ensured that, on average, the means and standard deviations of all variables were 0 and 1, respectively. I used this re-scaling process so that all predictors could be expressed in comparable and interpretable metrics when analyzed in regression models.

d_{Mod_Signed} analyses. I computed d_{Mod_Signed} using the formula in Equation 38, which calls for correlations and unstandardized descriptive statistics (note that corrections for measurement error were applied to statistics and sampling-variance estimates prior to computing d_{Mod_Signed}). I estimated the sampling variance of each d_{Mod_Signed} effect size using the procedure described in Study 1. The sampling variance of each input statistic was estimated based on the sample-size weighted mean statistic across all samples, as the mean observed value of a statistic is generally a better estimate of the parameter value than any sample statistic considered in isolation and thus permits more reliable estimates of sampling variance (Schmidt & Hunter, 2015).

Cleary model bias analyses. I computed Cleary-model regression tests for intercept and slope differences using the "lm_mat" function from the *psychmeta R* package (Dahlke & Wiernik, 2018, 2017/2019), which performs linear regression analyses based on covariance matrices, vectors of means, and sample sizes rather than on databases of random variates. I prepared the input information for these analyses by using the procedure described in the Appendix to combine subgroup matrices and means into mixture distributions that included group dummy variables and group-by-predictor interaction terms.[5] The group dummy variables were constructed such that the referent groups (i.e., White individuals in White-Black and White-Hispanic contrasts and males in female-male contrasts) were coded as "0" and the focal groups (i.e., Black individuals in White-Black contrasts, Hispanic individuals in White-Hispanic contrasts, and females in female-male contrasts) were coded as "1." This coding means that the intercept and slope coefficients represent the referent group's intercept and slope, while the coefficients for

---

[5] This procedure combines the within-group variance-covariance matrices with the between-group variance indicated by the subgroup means to give the overall variance-covariance matrix that describes predictor-criterion relationships when the groups are analyzed together. The procedure for combining subgroup distributions also adds to the multivariate distribution (1) a dummy variable indicating the association between group membership and scores on continuous variables, which is necessary to estimate intercept differences between groups, and (2) a product variable that represents the group-by-predictor interaction, which is necessary to estimate slope differences between groups. The vector of means and the variance-covariance matrix describing the two-group mixture distribution can be used to compute the regression models specified by the Cleary model of bias.

the group-membership dummy variable's main effect and interaction with the predictor indicate how the focal group's intercept and slope differ from those of the referent group.

In analyses of artifact-corrected data, the effective sample sizes associated with a given sample's predictor variance, criterion variance, and predictor-criterion covariance could differ because of (1) differences in sample size between the applicant and incumbent samples that provided the predictor and criterion data, respectively, in range-restriction corrected analyses and (2) differences in the extent to which corrections for artifacts impacted the adjusted sample size associated with the incumbent data.[6] The sample size used in each regression analysis was determined by computing the harmonic mean of the sample sizes associated with the elements of each sample's predictor-criterion covariance matrix; this approach accounts for the overall precision of regression estimates computed using matrices in which sample size varies across cells (Viswesvaran & Ones, 1995).

I computed a full array of Cleary-model regression analyses for each predictor in each sample to determine whether subgroups exhibited slope differences, intercept differences, or no differences in prediction. I extracted the significance test results from

---

[6] Corrections for artifacts impact the effective sample size associated with the corrected statistic because of the effect of the corrections on sampling variances. For example, correcting a correlation for criterion measurement error requires that the sample size be downwardly adjusted to account for the fact that $r_C = r/\sqrt{r'_{YY}}$ and $var_{e_C} = var_e/r'_{YY}$, so that $N_C = N \times r'_{YY}$. A more general formula for estimating the adjusted sample size for a corrected statistic is $N_C = N \times (stat_{observed}/stat_{corrected})^2$ (see Schmidt & Hunter, 2015, pp. 143–149).

each application of the Cleary model, recorded the nature of the differences, and then, if subgroups differed, recorded the direction of the difference (i.e., If there were slope differences, was the focal group's slope flatter or steeper than the referent group's slope? If there were intercept differences, was the focal group's intercept lower or higher than the referent group's intercept?). I also extracted all estimated coefficients from the regression models, as well as the estimates of their sampling variances so that the regression coefficients could be meta-analyzed.

Of the regression coefficients extracted from the models, two were of elevated importance: Those representing intercept differences and those representing slope differences. From the regression model in which GPA was regressed on the predictor variable and the group-membership dummy variable (i.e., "Model 2" in the Cleary framework), I extracted the group main-effect (i.e., intercept-difference) coefficient and its standard error. From the regression model in which GPA was regressed on the predictor variable, the group-membership dummy variable, and the group-by-predictor product term (i.e., "Model 3" in the Cleary framework), I extracted the group-by-predictor interaction (i.e., slope-difference) coefficient and its standard error. These regression effects were of interest because they are the Cleary model's indicators of the direction and magnitudes of intercept and slope differences, respectively. The group main-effect coefficient from Model 2 was the focus for intercept differences instead of the corresponding coefficient from Model 3 because intercept differences are arbitrary and uninterpretable in the presence of slope differences.

Meta-analyses. I used Schmidt and Hunter's (2015) random-effects meta-analytic methods to compute meta-analyses of d_{Mod_Signed} effect sizes and both intercept-difference and slope-difference regression coefficients. I used the "ma_generic" function from the *psychmeta R* package (Dahlke & Wiernik, 2018, 2017/2019) to compute meta-analyses with inverse sampling-variance weights. For meta-analyses of intercept-difference regression coefficients, I excluded coefficients from samples that exhibited significant slope differences, as these main effects were not interpretable.

Results

Table 9 shows the results for all meta-analyses of observed-score d_{Mod_Signed} effect sizes. Consistent with my findings from Study 2, these results show that d_{Mod_Signed} for composite predictors became smaller in magnitude as more indicators were added. Individual predictors reliably overpredicted the performance of Black and Hispanic students relative to White students and reliably underpredicted the performance of female students relative to male students. However, the magnitudes of overprediction and underprediction shrank as the composites became more inclusive and approached a fully specified model. Similar trends also occurred for d_{Mod_Signed} effect sizes corrected for criterion unreliability (see Table 10), range restriction (see Table 11), and both range restriction and criterion unreliability (see Table 12). In general, corrections for statistical artifacts had rather small impacts on the magnitudes of mean d_{Mod_Signed} estimates, but these impacts were large enough to cause some credibility intervals of range-restriction-corrected d_{Mod_Signed} estimates to overlap with zero. Specifically, the HSGPA and SAT composites (with and without writing) had credibility intervals that included zero for

White-Hispanic contrasts when range restriction was corrected (with and without applying corrections for criterion unreliability; see Table 11 and Table 12) and the three-test SAT composite's credibility interval also included zero for White-Hispanic contrasts when all artifacts were corrected (see Table 12).

To help explain the d_{Mod_Signed} trends depicted in Table 9, Table 10, Table 11, and Table 12 (particularly the White-Hispanic d_{Mod_Signed} effects that did not cleanly generalize), the values arrayed in Table 13 provide a detailed look at other standardized effect sizes implicated in the d_{Mod_Signed} calculations. Although the d_{Mod_Signed} meta-analyses described above were based on unstandardized slopes and means, examining patterns in validity coefficients and standardized mean differences can provide insights into the mechanisms driving the magnitudes and directions of d_{Mod_Signed} (cf. Equation 39 and Study 2). Table 13 shows meta-analytic means of all referent-group predictor validities, criterion and predictor mean differences, and the products of predictor mean differences and referent-group validities. The patterns of validity coefficients and *d* values support the findings from Study 2: Composite predictors have larger effect sizes than would be anticipated from the average effect sizes of their component predictors. Additionally, the differences between the composite effect sizes and the average effect sizes of components is larger for composites that include more predictors. These effect-size trends mean that composite predictors (particularly those with more components) tend to have *d* value-validity products that are larger than those associated with their average component. For White-Black and White-Hispanic contrasts, these larger product terms have the effect of driving down not only the mean d_{Mod_Signed} estimates, but also the

entire d_{Mod_Signed} parameter distributions. The negative effect of forming composites on d_{Mod_Signed} estimates associated with overprediction can cause the low-end of the parameter distribution to drop slightly below zero, which results in some credibility intervals that do not support strict generalizability inferences.

With the d_{Mod_Signed} meta-analyses showing strong support for the generalizability of historic trends regarding minority overprediction and female underprediction, it is important to consider whether the forms of the differences in prediction also agree with prior research. Table 14 summarizes the rates at which intercept differences and slope differences were detected in analyses of observed data. These results indicate that intercept and slope differences were detected at well-above chance rates across all three subgroup contrasts. Of the significant intercept differences, the overwhelming majority of differences occurred in the historic directions: Black and Hispanic individuals' performance was overpredicted by Whites' regression equations and females' performance was underpredicted by males' regression equations.

Of the significant slope differences, there was a very strong tendency for the differences to represent flatter slopes for Black and Hispanic individuals than for Whites. White-minority slope differences occurred at the greatest rates for HSGPA and composites that included HSGPA. In male-female contrasts, slopes tended to be steeper for females than males for pure SAT-based predictors, but tended to be flatter for females than males for HSGPA and composites that included HSGPA. These results provide evidence that HSGPA tends to exhibit more reliable slope differences than SAT tests (particularly in White-minority contrasts, where the rates of significant slope differences

for HSGPA were roughly double that of SAT-based predictors). Furthermore, HSGPA tends to correspond to steeper referent-group slopes than focal-group slopes, even in male-female contrasts.

As a complement to the rates of differential prediction detected for observed data, Table 15, Table 16, and Table 17 show rate of differential prediction for artifact-corrected data. The trends in the artifact-corrected summaries closely correspond to those discussed above with respect to Table 14, with both intercept and slope differences occurring at above-chance rates. Intercept differences were more common in corrected data than were slope differences, and the forms of these differences agreed with historic trends: minority performance was overwhelmingly overpredicted and female performance was overwhelmingly underpredicted. Slope differences manifested in patterns similar to those described for the observed data: Minority groups tended to have flatter slopes than Whites and females tended to exhibit steeper slopes than males for SAT-based predictors but flatter slopes for HSGPA. As with the observed data, White-minority differences were associated with HSGPA at a much higher rate than they were associated with SAT-based predictors.

The rates of significant differences in prediction summarized in Table 14, Table 15, Table 16, and Table 17 suggest generalizable differences in intercepts, as there were very few cases in which significant intercept differences occurred in the opposite directions of those detailed in the historic literature. To directly examine the generalizability of intercept differences, Table 18 shows meta-analyses of observed intercept-difference regression coefficients, Table 19 shows meta-analyses of intercept-

difference coefficients corrected for criterion unreliability, Table 20 shows meta-analyses of intercept-difference coefficients corrected for range restriction, and Table 21 shows meta-analyses of fully corrected intercept-difference coefficients. Indeed, the credibility intervals for all intercept differences across all methods of handling artifacts indicate generalizable overprediction of minority performance relative to Whites (that is, minority groups tended to have lower intercepts than Whites) and generalizable underprediction of female performance relative to males (that is, females tended to have higher intercepts than males).

Whereas the rates of significant difference in prediction detailed in Table 14, Table 15, Table 16, and Table 17 indicate consistency in the directions of intercept differences, trends regarding directions of subgroup slope differences were considerably less clear. Although slope differences tended to follow certain patterns, it was not clear whether these differences in prediction would generalize when subjected to meta-analysis. Table 22 presents meta-analyses of observed slope-difference coefficients and shows that the mean slope differences do indeed indicate that slopes were, on average, at least slightly flatter for minority groups than for Whites, slightly steeper for females than for males for SAT-based predictors, and flatter for females than for males for HSGPA. Of these trends, however, there were only two generalizable differences in subgroup slopes: The slopes of HSGPA and the composite of HSGPA with all three SAT subtests were reliably flatter for Black samples than White samples. All other 80% credibility intervals indicated that the random-effects distributions of slope difference coefficients included zero. The meta-analyses of artifact-corrected slope differences shown in Table 23, Table

24, and Table 25 are quite similar to the meta-analyses of observed slope differences, but only White-Black slope differences for HSGPA were found to generalize. The consistency of White-Black differences exhibited by HSGPA in rates of significant slope differences and meta-analyses of slope-difference coefficients provides a clear indication that HSGPA may be a problematic predictor by virtue of its high rates of slope differences. SAT-based predictors, however, did not exhibit generalizable slope differences.

In light of the mixture of significant slope differences and intercept differences observed in this study, it is instructive to consider how conditional differences in prediction are distributed across the operational range of predictor scores. The meta-analyses of d_{Mod_Signed} effects described earlier indicated that, on average, minority overprediction and female underprediction generalized across settings; however, these average magnitudes of differences could obscure differences in prediction that occur in the opposing directions in segments of the predictor distribution. To explore conditional differences in prediction, I plotted referent-focal differences across the operational ranges of standardized predictor scores for all predictors and subgroup contrasts. Figure 9 depicts differences in prediction from observed-score analyses. To aid in interpretation and avoid becoming distracted by wildly discrepant patterns of predictions in very small samples, I used darker lines to plot results from samples that had larger numbers of focal group members, as focal-group sample size is a strong indicator of the stability of a differential prediction regression analysis. The red lines represent the mean differences in predicted performance based on meta-analytic mean regression coefficients. As shown in

Figure 9, although the mean magnitudes of differences varied across predictor scores in some analyses (namely those involving HSGPA), the direction of the differences tended to be consistent, such that minority performance was overpredicted and female performance was underpredicted. The clustering of individual samples' lines tends to support the robustness of these trends. The trends depicted for observed data in Figure 9 were also supported by the range-restriction corrected data plotted in Figure 10; I did not generate plots for data corrected for criterion unreliability, as those plots would be identical to Figure 9 and Figure 10. Although HSGPA does exhibit problematic White-minority differential prediction trends, it does not appear that these differences in prediction translate into reliable degrees of underprediction in the operational range of scores when analyzed alone or as part of a composite that includes SAT scores.

Discussion

Like Study 2, Study 3 addressed Research Question 3 and showed that composite predictors, on average, have smaller magnitudes of overall differences in prediction than their individual components, as indexed by d_{Mod_Signed}. Study 3 also answered Research Question 5 and showed that d_{Mod_Signed} effects and intercept differences tend to demonstrate generalizable differences in prediction (with the exception of White-Hispanic d_{Mod_Signed} signed estimates for composites of HSGPA and SAT scores that were corrected for range restriction, which had credibility intervals that included zero), whereas slope differences do not. Meta-analyses of d_{Mod_Signed} effects and intercept differences showed that, regardless of artifact corrections, the performance of Black and Hispanic individuals is overpredicted relative to Whites while the performance of females

is consistently underpredicted relative to males. The only consistent generalizable difference in slopes detected in this study pertained to self-reported HSGPA, such that HSGPA had a flatter slope for Black samples than White samples, on average. With respect to d_{Mod_Signed} estimates, the lower bounds of White-Hispanic credibility intervals that overlapped with zero were of very small magnitudes (e.g., lower credibility bounds of -.01 and -.02), such that the lack of generalizable differential prediction corresponded to magnitudes of underprediction so small that they would not be considered practically meaningful by traditional effect-size interpretation practices.

Contrary to Aguinis et al.'s (2016) claims that differential prediction does not generalize across settings, I found that differential prediction overwhelmingly generalizes across samples in White-Black, White-Hispanic, and female-male comparisons. All intercept differences generalized in the historically expected directions across all contrasts and, after correcting for range restriction, only HSGPA had credibility intervals for slopes that excluded zero. It is important to note that the method for determining generalizability differed between Aguinis et al.'s study and mine: Aguinis et al. used *Q* tests to determine whether significant variance in effects remained after accounting for artifactual variance, whereas I used credibility intervals based on residual random-effects standard deviations. Rather than conveying generalizability in the Hunter-Schmidt sense (i.e., answering the question, "is zero contained within the middle 80% of the estimated parameter distribution?"), Aguinis et al.'s analysis addressed *heterogeneity* of effects (i.e., they answered the question "is there a significant amount of parameter variance?"). The *Q* test is affected by the number of samples in an analysis and can signal significant

parameter variance in large meta-analyses, even when the amount of parameter variance is not practically meaningful. Credibility intervals, however, provide a practical test of generalizability and avoid interpretation issues associated with significance testing. The *Q* test is commonly used to determine whether sufficient parameter variance exists that moderator variables could be contributing to between-study variation but, if the credibility interval for the overall effect does not include zero, such moderators will tend to simply correspond to different magnitudes of effects occurring in the same direction. Thus, I argue that my results offer a clearer indication about the generalizability of differential prediction effects than Aguinis et al.'s results.

An issue upon which my findings converge with Aguinis et al.'s (2016) is that slope differences do not generalize. Although the mean slope-difference effects were near zero for many predictors, the lack of generalizability combined with non-trivial amounts of estimated parameter variance means that slopes can differ in both directions across settings for reasons that are not explainable by statistical artifacts. Relatedly, I detected higher rates of slope differences than were reported in past reviews (cf. Bartlett et al., 1978; Hartigan & Wigdor, 1989; Schmidt et al., 1980) and my rates of significant slope differences were similar to those produced by Aguinis et al.'s analyses. Thus, slope differences may be more commonplace in the post-secondary education context than was previously believed. I suspect that this may be due to of the rather large average sample size used in my analyses. For example, the mean sample size in my analyses of White-Black differences in prediction was 4,135. This very large average sample size comes with substantial statistical power and means that even small slope differences could

trigger a significant result in the Cleary model. Such large samples were not common several decades ago and this could account for the high rates of slope differences observed in studies that analyze large databases from the College Board.

This study was notable for being the first meta-analysis of d_{Mod_Signed} effect sizes and the first application of my newly derived standard-error estimator for d_{Mod_Signed} effects. I recommend that future meta-analyses of differential prediction effects examine d_{Mod_Signed} estimates, as d_{Mod_Signed} can be easily meta-analyzed and it permits the examination of differences in prediction across studies that need not use the same predictor and criterion measures. By allowing various predictor and criterion measures to be meta-analyzed together, d_{Mod_Signed} will be a particularly useful effect size for research on personnel selection systems, as organizations vary widely in how they operationalize criterion and predictor constructs. Future research that uses d_{Mod_Signed} and applies high-fidelity artifact corrections to personnel selection data would be useful to empirically test whether differences in prediction generalize in in the work context.

Limitations. Despite the large database used in this study and the care with which statistical artifacts were corrected, several limitations must be noted. First, the data analyzed in this study came from a non-random sampling of both schools and students, which means that my results may not generalize to all U.S. post-secondary institutions. In particular, given that all schools analyzed here used the SAT as their preferred standardized admissions test, the trends associated with SAT scores may not correspond to other standardized tests such as the ACT.

Second, the only predictor variables available for analysis in this study were SAT scores and HSGPAs, but these are not the only predictors that colleges consider when making their admissions decisions. Predictors such as admissions officers' ratings of applicants' personal statements, letters of recommendation, extracurricular activities, and rigor of high school course choices would be necessary to establish a fully specified prediction model.

Third, given that differential prediction analyses require an unbiased criterion, it is possible that some form of criterion bias could have influenced my results, particularly magnitudes of slope differences. Dahlke, Sackett, and Kuncel (2019) recently investigated whether criterion contamination in the form of individual differences in course-taking choices could explain White-minority differential validity of the SAT. After controlling for differences in students' course-taking choices, Dahlke et al. found that White-Black and White-Hispanic differential validity disappeared whereas male-female differential validity increased in magnitude (with females having larger SAT validity coefficients than males). This type of criterion contamination may also influence estimates of slope and intercept differences; additional research is necessary to explore this possibility.

Fourth and finally, first-year GPA is not the only criterion that colleges care about; just as researchers should account for all relevant predictors when assessing differential prediction, it is important to also acknowledge that multiple criterion dimensions may be of interest. First-year GPA is arguably the most commonly studied performance criterion in the post-secondary academic domain, but other criteria such as

second-year retention, fourth-year cumulative GPA, and degree completion could also be of interest in operational differential prediction analyses.

Conclusion. This study demonstrated that findings of minority overprediction and female underprediction overwhelmingly generalize across settings. Specifically, d_{Mod_Signed} effects and intercept differences representing minority overprediction and female underprediction generalized and, when they did not, magnitudes of underprediction were trivial in magnitude; however, slope differences tended not to generalize (yet slope differences were detected at above-chance rates). Additionally, patterns of slope and intercept differences were not strongly affected by corrections for statistical artifacts. Given that statistical artifacts had rather small effects on slope and intercept differences in the post-secondary admissions context, the next logical question is "when *could* these artifacts meaningfully bias the results of differential prediction analyses?" This question is examined in the following study, where I present a simulation that demonstrates the effects of range restriction and criterion measurement error on differential prediction parameters.

Study 4: Impact of Measurement Error and Range Restriction on Differential Prediction Inferences

As indicated in the Introduction chapter, statistical artifacts can impact the standard errors and significance tests of coefficients used in tests of the Cleary model, but only a few artifacts (namely, indirect range restriction and predictor measurement error) can actually bias the expected values of the regression coefficients themselves. Effect sizes such as d_{Mod_Signed}, however, can be influenced by any type of statistical artifact because range restriction and measurement error impact the input values used to compute the effect size. Although range-restriction artifacts can theoretically bias differential prediction effects, the results of Study 3 show that artifacts did not seriously impact the inferences made about the magnitudes and prevalence of differential prediction trends in real-world post-secondary education data. This study aims to clarify the effects of statistical artifacts on differential prediction analyses by examining how direct range restriction, indirect range restriction, and criterion measurement error influence d_{Mod_Signed} statistics and the results of Cleary-model analyses of intercept and slope differences.

Given that making corrections for artifacts did not meaningfully alter conclusions about differential prediction in academic selection settings, a logical next step toward understanding the influence of artifacts in differential prediction analyses is to systematically simulate hypothetical selection scenarios to identify which aspects of selection systems can lead to biased statistical results. Simulations are useful for studying methodological phenomena because they (1) allow each of a set of parameters to be independently manipulated to isolate parameters' main effects and interactions, (2)

permit the exploration of extreme scenarios that may occur in certain settings but are seldom described in published research, and (3) they remove ambiguity surrounding how data are operationally used (e.g., organizations and the individuals employed by them may use idiosyncratic and inconsistent methods for making selection decisions based on data, but simulations avoid this inconsistency by allowing the researcher to specify how the data are to be used across all conditions). The systematic nature of simulations means that, as long as input parameters are chosen carefully and selection is modeled in a sensible manner, simulated results can supplement other empirical research findings and facilitate a more comprehensive understanding of a phenomenon. With the goal of modeling operational selection scenarios in a high-fidelity way, this study improves upon Aguinis et al.'s (2010) simulation design by treating measurement error as an unavoidable characteristic of predictor data and focusing on how statistical artifacts impact the data used in operational selection decisions.

To date, the largest differential prediction simulation examining the influence of statistical artifacts was reported by Aguinis et al. (2010); however, those authors compared observed statistical results impacted by artifacts to true-score parameters rather than operational parameters, ignoring the fact that differential prediction analyses should be computed using operational predictor scores. Aguinis et al. also only examined the effects of direct range restriction (DRR) on differential prediction analyses, despite the fact that DRR does not bias operational regression parameters and most organizational data are affected by indirect range restriction (IRR). As a consequence, Aguinis et al. failed to shed light on how operational differential prediction estimates are affected by

realistic types of range restriction. I address this oversight in the present study by exploring the impacts of range restriction and criterion measurement error on four key types of differential prediction indicators: d_{Mod_Signed} effect sizes, regression coefficients indicating intercept differences, regression coefficients indicating slope differences, and the test statistics that allow researchers to make comparisons among regression models in the Cleary framework (i.e., *F* ratios representing incremental model fit).

This study addresses two broad research questions that were posed in the Introduction chapter. The first research question relevant to this study is Research Question 6, which asked, "Which parameters of applicant populations and selection systems are most predictive of biased estimation of (a) d_{Mod_Signed} effect sizes, (b) intercept differences, and (c) slope differences?" The parameters of applicant populations and selection systems of interest in my study are the proportions of members from different subgroups, the subgroup mean differences on predictors and criteria, the correlations among predictors and criteria (including differences in the correlations between groups), the reliability of the criterion variable, the selection ratios applied to the selection variables, and whether a selection process causes DRR or IRR. As I described earlier, any statistical artifact could influence d_{Mod_Signed} estimates, but only IRR should be able to have an effect on observed subgroup intercept- and slope-difference regression parameters. By inducing IRR artifacts in my simulated data, I show which simulation parameters have the strongest biasing effects on regression estimates. I also examine which parameters are associated with sign changes of d_{Mod_Signed} estimates and regression

coefficients, such that the observed data indicate differences in prediction that occur in the opposite direction of the operational differences.

The second research question relevant to this study is Research Question 7, which asked, "Which parameters of applicant populations and selection systems have the biggest impact on the ability of researchers to detect (a) intercept differences and (b) slope differences?" Whereas Research Question 6 was focused on how statistical artifacts bias the differential prediction statistics that researchers use to discern the magnitudes and directions of differences in prediction, Research Question 7 has to do with the adequacy of the Cleary model's multiple moderated regression procedure for correctly identifying the presence or absence of operational differences in prediction. To answer this question, I examine the impacts of statistical artifacts on the parameter values of the test statistics that are used to make comparisons among nested regression models in the Cleary framework. In this study, I compare the *F* ratio parameters from observed data affected by both range restriction and criterion measurement error against the *F* ratio parameters from an operational sample of equal size that is not affected by criterion unreliability or systematic selection effects. However, the magnitudes of *F* ratios are affected by sample size and I intend for my simulation to illustrate general principles that are not specific to samples of any particular size; thus, I prepared an analysis method that rescales the differences between artifact-impacted *F* ratios and their artifact-free counterparts from their arbitrary sample-size-dependent metrics into a metric that is independent of sample size and that supports more intuitive interpretations in the context of the present research. I describe my method for addressing Research Question 7 in the

following section where I explain how my approach can support clearer inferences than the Monte Carlo simulation methods employed in previous simulations of artifacts and differential prediction (cf. Aguinis et al., 2010).

Method for Examining the Cleary Model's Potential for Type I and Type II Statistical Errors

A common approach for simulating the effects of artifacts on statistical results is to use Monte Carlo methods to generate simulated datasets, each of which can be subjected to a particular analysis procedure in order to obtain a set of statistical results; the results from each simulated sample are then gathered into a dataset for analysis (e.g., analyses to determine the mean effect, the variation of effects, and/or the rates of Type I or Type II statistical errors). This Monte Carlo approach represents a completely brute-force method for understanding a statistical phenomenon and the brute-force nature of this approach has both positive and negative consequences. A salient positive consequence is that, by virtue of relying on sheer brute force, Monte Carlo methods can be implemented by anyone with the requisite programming knowledge. This means that this approach to computational modeling can be used to explore problems that are too complex for a researcher to solve analytically, whether because of the complexity of the problem itself or the level of mathematical/statistical skills possessed by the researcher.

Apart from any positive attributes of Monte Carlo methods, a distinct negative consequence of these methods is that, because they are so easily applied to a variety of problems, researchers tend to resort to designing Monte Carlo simulations to explore problems that could be addressed with greater clarity via more efficient and more elegant

simulation techniques. For example, Aguinis et al. (2010) touted the scale and scope of their Monte Carlo simulation ("3,185,000 unique combinations" of parameters, "15 billion 925 million individual samples," and "more than 8 trillion 662 million individual scores," p. 648); although these numbers are initially impressive because of the gargantuan simulation effort they represent, they are significantly less impressive when one considers that a Monte Carlo simulation was not necessary to answer Aguinis et al.'s research questions.

Differential prediction is a regression-based phenomenon and linear regression is a very well understood analysis in terms of how coefficients and their standard errors are estimated. If one knows how a given statistical artifact impacts a set of means and variance-covariance parameters (e.g., measurement error inflates a variable's variance without impacting its covariances), one can determine how that artifact affects both the expected values of regression coefficients and whatever test statistics are of interest (e.g., *t* statistics for individual regression coefficients or *F* ratios for comparisons among nested regression models). Only when a statistical artifact violates the assumptions of linear regression is it strictly necessary to study the artifact's effect via Monte Carlo methods. Given that neither measurement error nor direct range restriction in any way violate regression's assumptions, Aguinis et al.'s (2010) simulation could have been executed more efficiently using analytic simulation methods in which statistical artifacts were induced into parameter values and in which the effects of the artifacts on Type I and Type II error rates were evaluated by comparing the analytically expected error rates associated with the artifact-free and artifact-attenuated parameters for a variety of illustrative sample

sizes. An analytic simulation of this sort obviates the need for Monte Carlo data generation, which makes the analytic simulation more efficient and precise. The present study relies on analytic simulation methods, as all of the analyses involved have well-defined estimation techniques and the effects of artifacts on test statistics' parameter values indicate whether there are elevated or decreased likelihoods of detecting significant results in the presence of artifacts.

One of the challenges in designing a Monte Carlo simulation is determining which sample size(s) to use in the simulated samples, with the intention of demonstrating how an effect varies as a function of sample size (e.g., illustrating that a method is asymptotically unbiased) or providing an intuitive frame of reference for readers by framing a phenomenon in terms of a sample size researchers are likely to encounter in their own work. An advantage of applying analytic simulation methods is that the statistical functions involved in such a simulation are usually sufficiently well understood that explicitly modeling sample-size effects is unnecessary, as the simulation's manipulation has comparable impacts on distributions of test statistics regardless of sample size. For example, if one were interested in examining the effects of measurement error on the expected values of observed correlations and the power of statistical tests for those correlations, a Monte Carlo simulation would not be necessary because measurement error has the same type of effect on sampling distributions of correlations across all sample sizes. In situations such as this, sample size need not be explicitly modeled because it is an abstraction not directly relevant to the phenomenon of real interest.

In the simulation described momentarily, the total sample size of the combined referent and focal groups were held constant across all conditions and all statistics that are dependent upon sample size were normalized prior to analysis to nullify the influence of the arbitrary sample-size value on the simulation results. The sample-size dependent parameters of primary interest in this study are the *F* ratios associated with the difference between nested regression models, as an *F* ratio is the test statistic used in the Cleary model to determine whether regressions models in which subgroups have different slopes or intercepts fit the data better than a model in which subgroups have equal regression lines. A difference in the magnitude of *F* ratio parameters between two models that have the same degrees of freedom indicates a difference in statistical power or error rates between the two models. By computing differences between the *F* ratio parameters associated with observed and operational regression models with sample size held constant, one obtains an index of the relative difference in power or error rates between the effects.

The magnitudes of differences in *F* ratio parameters are arbitrary and depend on the sample size used to estimate the standard error, but these differences can be normalized by dividing all values by the largest absolute-value difference. Such normalized differences are computed as shown in Equation 91.

$$\text{Normalized Difference in } F \text{ ratios} = \frac{F_{Observed} - F_{Operational}}{\max\left(\left|F_{Observed} - F_{Operational}\right|\right)} \qquad 91$$

This normalized difference in *F* ratio parameters removes the influence of sample size on the magnitudes of the differences so that, holding all else constant, the normalized differences are identical across simulation conditions that have different sample sizes but

are equal in terms of all other parameters. By removing the influence of sample size, the resulting differences yield indicators of differences in statistical power or Type I error rates that can be interpreted relative to the magnitude of the largest *F* ratio difference in the simulation. For example, a normalized difference in *F* parameters of -.50 would indicate that the magnitude of the negative difference between the observed and operational *F* parameters is half as large as the largest-magnitude difference observed in the entire simulation. Normalization not only gets rid of the arbitrary sample-size scaling of raw differences in *F* ratios, it also contextualizes the differences in the simulation so that each effect is interpretable relative to the most extreme magnitude observed.

As an illustration of how normalized differences in *F* ratios support valuable inferences in my main simulation, I have prepared a miniature analytic simulation to show how these normalized differences indicate artifacts' impacts on statistical results. This simulation held constant the referent-group validity ($r = .5$), the standardized mean difference on X ($d = 1$), and the proportion of the sample consisting of referent group members ($p = .5$). The variables used in this simulation were sample size (which ranged from 100 to 1,000 in increments of 100), selection ratios (which were varied as .1, .5, and .9), focal-group validities (which were varied as .1, .3, and .5), mean differences on Y (which were varied as .3, .5, and .7), and reliability coefficients for Y (which were varied as .6, .8, and 1.0). All possible combinations of simulation parameters were used to compute observed regression models; for each artifact-attenuated condition, a condition with a selection ratio of 1.0 and a perfectly reliable criterion was run to obtain the corresponding operational regression models.

The simulation procedures I used in my illustrative simulation are much like the procedures that I used in my main simulation. Reliability artifacts were introduced to the criterion variable prior to truncating the predictor distribution at a cut score that satisfied the selection ratio assigned to a given condition. The truncation procedure introduced range restriction into the data and adjusted all parameter values describing variances, covariances, means, and subgroup proportions according to the effect of censoring data below the cut score. After introducing unreliability and selection artifacts, the subgroup's bivariate distributions were combined into a two-group mixture distribution that could be used as the input to the Cleary model's regression analyses; this procedure is described in the Appendix and was summarized earlier in Footnote 5. From the Cleary model results, I extracted the *F* ratios associated with the comparisons among regression models.

The simulation results for *F* ratio parameters comparing Model 3 and Model 1 (i.e., tests of overall differences in prediction) are shown in Figure 11, the results for *F* ratio parameters comparing Model 3 and Model 2 (i.e., tests of slope differences) are shown in Figure 12, and the results for *F* ratio parameters comparing Model 2 and Model 1 (i.e., tests of intercept differences) are shown in Figure 13. In each figure, panel A shows the operational *F* ratio parameters that are not attenuated by artifacts, panel B shows the observed *F* ratio parameters that are attenuated by selection and measurement-error artifacts, panel C shows the raw difference between the observed and operational *F* ratio parameters, and panel D shows the normalized differences between the observed and operational *F* ratio parameters. Panels A, B, and C are in sample-size dependent metrics, but panel D show that normalizing the differences transforms the data so that the

differences across all sample sizes are identical. The normalized-difference trends show that conditions in which the influence of artifacts was minimal (e.g., conditions in which criterion reliability was perfect and a selection ratio of .90 was imposed) have differences close to 0, which indicates that the levels of statistical power and Type I errors associated with the artifact-attenuated parameters from panel B are only slightly different from the levels associated with the operational parameters from panel A. In contrast, conditions with substantial statistical artifacts (e.g., conditions in which criterion reliability is .60 and the selection ratio is .10) contain cases in which the largest differences in power/error were observed (i.e., normalized differences of 1.0 in absolute value). By scaling differences in *F* ratio parameters with respect to the magnitude of the largest difference, the relative impact of artifacts on the statistical power and error of intercept- and slope-difference tests is more readily discerned.

Method

In this simulation, infinite multivariate normal distributions were used to represent unrestricted applicant populations and appropriately truncated multivariate normal distributions were used to represent range-restricted incumbent populations (i.e., subpopulations of the applicant populations). In contrast to a Monte Carlo simulation in which only the operational parameters (i.e., parameters describing unrestricted applicant data with a perfectly measured criterion) are defined and the influence of statistical artifacts on observed results (i.e., artifact-attenuated results obtained by analyzing range-restricted incumbent data with an imperfectly measured criterion) is only apparent by accumulating a database of simulated statistics, this simulation directly compared the

parameters that describe the operational and observed data to determine how the artifacts impacted the expected values of statistical results as well as the statistical power and Type I errors associated with the analyses.

I modeled the effects of range restriction and criterion measurement error on differential prediction parameters. Variables *X* (i.e., the predictor) and *Y* (i.e., the criterion) were analyzed in all conditions, but *X* was only be the variable used to induce selection artifacts in DRR conditions; in IRR conditions, a third variable called *Z* functioned as the selection variable. Variables *X* and *Y* were designed to resemble cognitive ability tests and overall job performance, respectively, but *Z* was allowed to take on a variety of parameter values. Most characteristics of *Z* were varied so that *Z* could represent many different plausible operational selection variables. For example, *Z* could be a composite that includes *X* or it could be some other predictor that correlates weakly with *X* (e.g., holistic clinical/judgmental evaluations of applicant suitability). To enhance the fidelity of *Z*, I imposed no constraint that *Z* must be more valid than *X* or that it must exhibit smaller standardized subgroup mean differences than *X*.

Parameter values. All simulation parameter values are shown in Table 26. Across all experimental conditions, I held constant the validity of predictor *X* for the referent group (ρ_{XY_Ref} = .5) and the standardized subgroup mean difference on predictor *X* (δ_X = 1; δ is the population parameter of subgroup mean differences expressed as a *d* value). The ρ_{XY_Ref} constant was chosen to resemble the meta-analytic validity of cognitive ability tests (Schmidt & Hunter, 1998) and the δ_X constant was chosen to approximate the meta-analytic White-Black mean difference on cognitive ability tests

(Roth, Bevier, Bobko, Switzer III, & Tyler, 2001). Population variances for all variables were constrained to be equal between referent and focal applicant groups; I therefore use standardized correlations and *d* values as parameter metrics for differences in subgroup slopes and means, respectively.

I manipulated (1) the intercorrelations among all variables (with the exception of the validity of *X* for the referent group), (2) the relative proportion of the applicant population from the referent group, (3) the subgroup mean differences on *Y* and *Z*, and (4) the selection ratios used to induce range restriction. All possible values for all parameters are indicated in Table 26. The proportion of the applicant population belonging to the referent group were varied from .5 (equal representation of the focal and referent groups) to .9 (such that the referent group represents a clear majority) in increments of .2. Selection ratios were varied as .1, .5, and .9, which chosen to represent high, moderate, and low selectivity, respectively.

The validities of *X* and *Z* were varied between subgroups to produce slope-based differential prediction for both predictors; magnitudes of slope differences on both variables are purely exploratory and represent more extreme differences than have been documented in real-world assessment and selection research. Whereas the validity of *X* was held constant at .5 for the referent group, the validity of *X* for the focal group was varied from .1 (i.e., dramatically different subgroup slopes) to .5 (i.e., equal subgroup slopes) in increments of .2. The validity of *Z* for the referent group was varied from .2 to .6 in increments of .2; these values were chosen to represent settings ranging from those in which *Z* is a less effective predictor than *X* to those in which *Z* resembles a composite

predictor that demonstrates meaningful incremental validity over *X*. Given that the referent-group validity of *Z* was varied, I defined the focal-group validity of *Z* as a ratio of the referent group's validity to facilitate interpretations of simulation trends when these parameters are independently manipulated. The focal-group validity of *Z* was defined as 0% of the referent-group validity of *Z* (i.e., a single-group validity situation), 50% of the referent-group validity of *Z* (i.e., substantial subgroup differences in validities and slopes), or 100% of the referent-group validity of *Z* (i.e., equal subgroup validities and slopes) as a way to examine extreme slope/validity-difference scenarios. The values for the correlation between *X* and *Z* cover possibilities ranging from *X* being a component of *Z* when *Z* is cognitive construct or a mechanical composite of predictors that includes *X* (i.e., $\rho_{ZX} = .8$, which resembles the magnitude of a part-whole correlation) to *Z* being a non-cognitive construct or a holistic clinical judgment in which *X* has a small influence (i.e., $\rho_{ZX} = .2$). Correlations between *X* and *Z* were set equal between the referent and focal subgroups, as differential intercorrelations among predictors across subgroups have not been of serious research interest to psychologists and there is no empirical basis for selecting parameters to govern subgroup differences in predictor intercorrelations.

Mean differences on *Z* ranged from equal subgroup means ($\delta_Z = 0$) to subgroup differences that were comparable to the typical magnitude of White-Black mean differences on measures of cognitive ability ($\delta_Z = 1$) in increments of .5. The mean-difference parameters for the overall job performance criterion (δ_Y) were informed by the range-restricted meta-analytic estimate provided by McKay and McDaniel (2006), who reported an average observed *d* of .35 and an average unreliability corrected *d* of .46.

However, as those meta-analytic means were range-restricted, I required some basis from which to estimate the magnitude of criterion mean differences in an unrestricted worker population. There are no empirical estimates of range restriction for job performance criteria, so Berry and Zhao (2015) applied a variety of range-restriction corrections to McKay and McDaniel's estimate based on varied subgroup differences on a selection variable and varied correlations between the selection variable and job performance. The right half of Berry and Zhao's Table 1 (p. 170) indicates that the majority of criterion mean differences corrected for both IRR and criterion measurement error were between .3 and .7 in magnitude. Thus, I varied the unrestricted true-score mean differences on simulated criteria as .3, .5, and .7 to capture a wide range of plausible mean differences.

The final set of simulation parameters define the reliability of the variables. The reliabilities of *X* and *Z* were constrained to 1.0 across all conditions. Rather than indicating that these variables are measured perfectly, the choice to constrain these reliabilities to unity is meant to represent the fact that *X* and *Z* are operational predictors and whatever measurement error affects these variables is already factored into their other parameters. The reliability of *Y*, however, was manipulated and ranged from .60 (i.e., a value that resembles the inter-rater reliability of job performance), to .80 (i.e., a value that represents a plausible internal-consistency reliability), to 1.00 (i.e., a hypothetical condition in which the criterion is measured perfectly that allows the effects of pure range restriction artifacts to be observed).

All possible parameter values in Table 26 were fully crossed to create 243 unique DRR conditions and 19,683 unique IRR conditions.

Simulation procedure. For each combination of parameter values, I began by constructing a covariance matrix and a vector of means for each of the two subgroups. Next, I attenuated the covariance matrices for criterion measurement-error artifacts to produce observed-score matrices. The observed mean parameters were set equal to the true-score mean parameters because measurement error only affects the variability of scores, not the expected value of scores. I then combined the subgroups' observed predictor means and variances into a mixture distribution that contained both groups' data in order to identify the predictor cutoff score that would satisfy the scenario's selection ratio. I used the cutoff score to compute the truncated post-selection mean and variance of the selection variable's distribution in each group and then used the multivariate selection theorem (Aitken, 1934; Lawley, 1943) to transfer this selection effect to all of the other variables' variances, covariances, and means.

After artifacts were introduced into the subgroup distributions, I used the algebraic computational formula for d_{Mod_Signed} derived in Study 1 (see Equation 43) to compute differential prediction effect sizes for the observed data and the operational data. Next, I applied the Cleary model of bias to the data. I used the "lm_mat" function from the *psychmeta R* package (Dahlke & Wiernik, 2018, 2017/2019) to compute the regression models needed to evaluate differential prediction. As my simulation is focused on how the expected values of observed statistics differ from operational parameter values, two factors needed to be surmounted to apply the Cleary model to my data. First, my simulation conditions did not have sample sizes associated with them and sample sizes are necessary to compute the *F* ratios used to compare regression models. I resolved

this issue by defining an arbitrary sample size of 1,000 to run regression models for all conditions, regardless of the selection ratio imposed. Although the total sample size was held constant, the proportions of members from each of the two subgroups were free to vary as a function of the proportions in the applicant population, the selection ratio, and the subgroup mean difference on the selection variable. Holding the sample size constant kept the sample-size dependent *F* statistics comparable across conditions so that the biasing effects of selection could be interpreted unambiguously.

The second issue that complicated my regression analyses was that my use of parameter covariance matrices and mean vectors as data (rather than using random variates) precluded the use of conventional product terms that rely on the multiplication of the random deviates of group-membership and predictor variables. I resolved this limitation by deriving an algebraic procedure for appending a dummy-variable and product terms involving the dummy-variable moderator to subgroup covariance matrices and mean vectors and then merging the subgroup distributions into a multivariate mixture distribution that could be analyzed with the Cleary model (see the Appendix for the details of this procedure and see Footnote 5 for a summary). The matrix-mixing procedure was performed after the covariance matrix, mean vectors, and subgroup sample sizes had already been manipulated to reflect the influences of artifacts. To enhance the interpretability of observed-operational differences in regression coefficients, the referent group was coded as 1 and the focal group was coded as 0 in my group-membership dummy variables; this allowed for more intuitive interpretations of the directions in which the observed and operational regression parameters differed.

The results from each simulation condition involving selection and/or measurement error artifacts were matched with the results of an artifact-free condition that represented the operational comparison point for evaluating the biasing effects of artifacts on differential prediction statistics.

Simulation summarization procedure. After running all simulated conditions, I organized the statistical parameters of interest (i.e., d_{Mod_Signed} effects, regression coefficients, and *F* ratios) into a data matrix for further analysis and I matched results from artifact-attenuated observed conditions with their corresponding artifact-free operational conditions. I computed the difference between all observed parameters and their operational comparison values and, for d_{Mod_Signed} values and regression coefficients, I also generated variables indicating whether statistical artifacts altered the signs of the parameters. With observed-operational difference variables and sign-change variables appended to my database of simulation results, I used linear models to explain variation in these dependent variables.

I used ANOVA models to explain variation in my dependent variables, with the independent variables in these models representing all simulation parameters, parameters' quadratic effects, two-way interactions among parameters and quadratic parameter effects, and three-way interactions among parameters and quadratic parameter effects. I placed a constraint on all interactions that parameters could not interact with their own quadratic effects; this limited the highest-order polynomials to quadratic effects and prevented incidentally examining cubic effects. From each ANOVA model, I computed the variance explained by each effect (represented as η^2 effect sizes) and, to simplify the

expression of variance explained by the effects, I consolidated the variance explained by linear and quadratic effects involving the same parameters into composite η^2 effects. As a hypothetical example, if the SR and ρ_{YY} parameters were included in a model together with their quadratic effects, the total effect of the SR parameter would be the sum of η^2 for the SR and SR^2 effects, the total effect of the ρ_{YY} parameter would be the sum of η^2 for the ρ_{YY} and ρ_{YY}^2 effects, and the total effect of the SR x ρ_{YY} interaction would be the sum of η^2 for the SR x ρ_{YY}, SR^2 x ρ_{YY}, SR x ρ_{YY}^2, and SR^2 x ρ_{YY}^2 effects. I considered any consolidated main effect or interaction that explains at least 1% of the variance in a dependent variable to have an effect worthy of examination; all effects explaining less than 1% of variance in a dependent variable were ignored for the sake of parsimony. Although combined linear and quadratic terms' η^2 effects, the functional form of each effect was still apparent because I plotted all effects with $\eta^2 \geq .01$.

To reflect the conditions under which each of the differential prediction parameters would be of interest to researchers, my linear models were based on different sets of conditions depending on which dependent variable was involved. The criteria that determined which conditions were included in each analysis are summarized in Table 27, along with indications of the number of conditions included in each analysis. Note that analyses involving changes to regression coefficients were only conducted for the IRR simulation, as the DRR simulation did not involve processes that could create such changes. These inclusion criteria ensured that the results of each summary model would be clearly interpretable (e.g., in analyses of intercept-difference coefficients, conditions

with operational slope differences were excluded because the intercept differences would not be meaningful in those conditions).

For each dependent variable, I created a table of key results from the linear model summaries; these tables included η^2 values of all effects with $\eta^2 \geq .01$. Additionally, when an interaction effect exceeded my η^2 cutoff, the lower-order effects associated with the interaction were also tabled and the total η^2 of the interaction and its lower-order effects was computed to provide an index of the overall importance of the interaction and its subordinate effects. Each effect that exceeded my η^2 cutoff was plotted to supplement my textual interpretation of the interaction.

Results Preamble

All analyses were meant to show which factors in the simulation predicted differences between the parameters of observed results (i.e., range-restricted results computed from incumbent data with an imperfectly measured criterion) and the parameters of operational results (i.e., unrestricted results computed from applicant data with a perfectly measured criterion). Before detailing the results of my analyses, it is important to note that some of the analyses identified in Table 27 were ultimately unnecessary. Specifically, analyses of differences in *F* ratios representing changes to Type I error potentials from the DRR simulation revealed that there was no variation in *F*-ratio differences; therefore, neither DRR nor criterion unreliability impacted Type I error rates of model comparisons analyzed in the Cleary framework. This is a finding that could have been anticipated from the fact that neither DRR nor criterion unreliability can bias regression parameters, which means that observed regression lines will not differ if

operational regression lines do not differ. Thus, these artifacts cannot turn a null difference in intercepts or slopes into a non-null difference and therefore cannot affect rates of Type I errors. In light of this, I focused my analyses of Type I error potentials on data from the IRR simulation. However, analyses of *F*-ratio differences representing changes to statistical power were relevant to both the DRR and IRR simulations.

The main objective of my analyses was to determine which simulation parameters had the most pronounced effects on my dependent variables; detailing the intricacies of interaction effects was therefore be secondary to my goal of identifying which (sets of) parameters were associated with the biggest differences between observed and operational results. Table 28 offers a summary of parameter's contributions to the explanation of dependent variables' variance. The values in Table 28 are meant to give a rough indication of each parameter's "importance" by conveying the total amount of variance explained in the dependent variable by all effects that involve the parameter in question (i.e., the sum of all η^2 values associated with a parameter's main effect and the interaction terms of which it was a part). However, not all models explained 100% of the variance in their respective dependent variables (due to non-linear effects more complex than what quadratic terms can capture) and it is useful to account for this when attributing importance to parameters. Table 29 re-expresses the data from Table 28 by dividing each value by the model's R^2 (i.e., the sum of all effects' η^2 values) to scale each effect's contribution relative to the summary model's overall fit. This rescaling did not meaningfully alter the magnitudes of tabled values and therefore variation in model fit did not affect attributions of parameter importance. I have bolded the total effects of

parameters in Table 28 and Table 29 that appear to be particularly influential. My choices regarding which effects to bold are admittedly subjective, but were made with the intention of simplifying interpretations of results by signaling differences between distinctly impactful parameters and those that had comparatively little influence (typically, non-influential parameters were involved in effects explaining less than a total of 20% of variance).

I offer summaries of all effects that met my $\eta^2 \geq .01$ cutoff in the following results sections, but I also use the importance indications from Table 28 (and Table 29) to guide my interpretations and discussions of key effects. In describing the effects of simulation parameters on each dependent variable, I (1) list the important contributors, (2) clarify the roles of those contributors by describing the nature of their effects, and, when feasible, (3) provide explanations of why the effects occurred as they did. Parameters' effects on each dependent variable are also be conveyed in detailed tables and figures. As I present these results, I include some amount of discussion material along the way; however, I save bigger-picture discussion points for the Discussion section.

As a final clarification about how results will be organized prior to delving into the findings, I note that I have taken liberties in re-expressing some interaction effects in an attempt to provide greater clarity regarding how sets of three-way interactions occurred. Although my linear summary models were limited to testing three-way interactions as the most complex type of effect, some three-way interactions suggested the existence of four-way interactions by virtue of the same set of four parameters recurring in interactions in various permutations. In cases such as these, clear

interpretations of trends can be facilitated by examining the more complex implied four-way interaction because that interaction frames multiple effects within a shared context and offers a cleaner depiction of the effects. Whereas four-way interactions can be difficult to interpret in analyses of random real-world data, the structured and more-or-less lawful patterns of results produced by simulations can lend themselves to clearer interpretations through more complex representations.

The results of the direct range restriction simulation and the indirect range restriction simulation are presented next, with each simulation receiving attention in its own dedicated results section.

Results of the Direct Range Restriction Simulation

Effects of simulation parameters on d_{Mod_Signed} estimates. Table 30 provides a summary of the variance explained in observed-operational differences in d_{Mod_Signed} estimates by each effect that met my $\eta^2 \geq .01$ cutoff.

***Effects of simulation parameters on the values of* $\mathbf{d_{Mod\_Signed}}$ *estimates*.** Table 28 indicates that *SR* and ρ_{XY_Foc} were the most important contributors to observed-operational differences in d_{Mod_Signed} values. However, these were not the only parameters to meet my η^2 threshold. I also found that ρ_{YY} had a very small positive main effect. This effect is depicted in Figure 14, which shows that the direction of difference between observed and operational d_{Mod_Signed} values tended to be positive (i.e., the observed values were higher in value than the operational values) and the differences were smaller for lower levels of criterion reliability.

The SR x ρ_{XY_Foc} interaction is depicted in Figure 15, which shows that there was no relationship between selectivity and observed-operational differences in d_{Mod_Signed} values when subgroups had equal operational slopes, but a negative relationship between *SR* and differences in d_{Mod_Signed} values emerged and became stronger as the difference in operational subgroup slopes increased. The largest differences in d_{Mod_Signed} values occurred in settings with low selection ratios (i.e., highly selective settings) in which there were large operational slope differences.

*SR*s and operational slope differences affect observed estimates of d_{Mod_Signed} because selection in the presence of mean differences on the predictor causes the subgroups' range-restricted predictor means to be closer together are than their unrestricted means. Additionally, the effects of range restriction on the subgroups' criterion means are transmitted through the subgroups' slopes, which means that the restricted criterion means also become closer together, but to a lesser extent than is true of the predictor means. In the context of the $d_{Mod\_Signed} = d_Y - r_{XY\_Ref}d_X$ formula, all three of the effect sizes are attenuated by range restriction, but r_{XY_Ref} and d_X are attenuated to a greater degree than is d_Y. The effect becomes stronger as operational differences in slopes increase because slope differences cause the range restriction to be transmitted to the criterion to a different degree in each group. This results in a $r_{XY\_Ref}d_X$ product that is attenuated to a greater degree than d_Y, which causes d_{Mod_Signed} to be overestimated in the presence of range restriction.

***Effects of simulation parameters on the signs of* $\mathbf{d_{Mod\_Signed}}$ *estimates*.** Table 28 indicates that *SR,* ρ_{XY_Foc}, and $\delta_Y$were the most important contributors to observed-

operational differences in the signs of d_{Mod_Signed} estimates. These parameters, along with P_{Ref}, affected the signs of d_{Mod_Signed} values through a set of 4 three-way interactions. All four of these interactions involved combinations of the same four parameters and, as noted earlier, a network of three-way interactions may be understood with greater clarity when plotted as a four-way interaction.

Figure 16 depicts the four-way interaction among SR, P_{Ref}, ρ_{XY_Foc}, and δ_Y and shows that sign changes were prevalent only in scenarios with operational slope differences, moderate-to-high levels of selectivity, and small mean differences on *Y*. These trends were more pronounced when there was a closer-to-equal representation of the focal and referent groups in the applicant population. When the applicant population consisted overwhelmingly of referent group members (P_{Ref} = .9), conditions with moderate operational slope differences (i.e., conditions in which ρ_{XY_Foc} = .3) that exhibited sign changes; sign changes did not occur in conditions where slope differences were more extreme (i.e., ρ_{XY_Foc} = .1). There was no effect on d_{Mod_Signed} sign changes when selectivity was low (i.e., *SR* = .9).

These results indicate that operational slope differences make it more likely that the sign of d_{Mod_Signed} will change because slope differences increase the likelihood that subgroup regression lines will cross somewhere within the operational range of scores. This, in turn, makes it more likely that settings with average operational underprediction effect will yield observed data that suggest an average overprediction effect if enough range restriction occurs (i.e., if the system is selective enough). Smaller operational mean differences on the criterion increase the likelihood that the sign of d_{Mod_Signed} will change.

This occurs because a d_Y of .3 in this case implies an average operational underprediction effect (i.e., .3 – 1 x .5 = -.2) and range restriction can attenuate the $r_{XY\_Ref}d_X$ product enough to result in a net overprediction effect when computed from observed data.

The effects observed for changes in the values and signs of d_{Mod_Signed} estimates imply that DRR can result in misestimation of differential prediction effects from observed data relative to operational data. However, given that pure DRR can seldom be expected in operational selection programs and operational slope differences are required for DRR to have a meaningful effect on interpretations, substantive changes in d_{Mod_Signed} due to DRR appear unlikely. Study 3 showed that slope differences between groups are quite small when they occur; this simulation builds on that observation by showing that DRR will have no real effect when slopes are equal (or close to equal).

Effects of simulation parameters on statistical power of Cleary analyses. Table 31 provides a summary of the variance explained in observed-operational differences in *F* ratios from non-null operational Cleary analyses by each effect that met my $\eta^2 \geq .01$ cutoff.

Effects of simulation parameters on statistical power of tests of overall differences in prediction. Table 28 indicates that P_{Ref}, *SR*, ρ_{XY_Foc}, and ρ_{YY} were the most important contributors to observed-operational differences in *F* ratios representing effects on power for tests of overall differences in prediction. These parameters interacted with each other, as well as with δ_Y, to explain variation in *F*-ratio differences.

The first interaction to consider is the four-way interaction among SR, P_{Ref}, ρ_{XY_Foc}, and δ_Y depicted in Figure 17. This interaction gives a summary of the mechanics of 4 three-way interactions that met my criteria for follow-up examination:

- SR x P_{Ref} x ρ_{XY_Foc}
- SR x P_{Ref} x δ_Y
- SR x ρ_{XY_Foc} x δ_Y
- P_{Ref} x ρ_{XY_Foc} x δ_Y

Figure 17 shows that the simulation parameters did not have a consistent direction of impact on *F*-ratio differences across all conditions: Differences were positive in some cases and negative in others. On average, positive differences appear to predominate, which is indicative of power being *enhanced* in the presence of these artifacts. However, the form of differences in prediction (slope differences vs. intercept differences) seems to determine whether power increases or decreases in the presence of statistical artifacts.

The largest increase in power in observed analyses relative to operational analyses occurred when selectivity was high and subgroups exhibited slope differences; this effect was stronger when subgroups were closer to equal representation in the applicant population and there were larger mean differences on the criterion. When subgroups exhibited intercept differences only (i.e., they had equal slopes, but criterion mean differences were not equal to .5), power was reduced to some degree in all settings; however, these decreases in power were quite small compared to the increases in power observed in settings with high selectivity, large slope differences, large mean differences on the criterion, and equal representation of subgroups in the applicant population.

The second interaction of interest is the four-way interaction among P_{Ref}, ρ_{XY_Foc}, δ_Y, and ρ_{YY} depicted in Figure 18. This interaction summarizes 3 three-way interactions:

- P_{Ref} x ρ_{XY_Foc} x δ_Y
- P_{Ref} x ρ_{XY_Foc} x ρ_{YY}
- ρ_{XY_Foc} x δ_Y x ρ_{YY}

Whereas Figure 17 showed the effects of DRR, Figure 18 shows the effects of criterion unreliability and clarifies trends depicted in Figure 17. The first effect to note is that low criterion reliability corresponded to an average reduction in power across all depicted scenarios, regardless of whatever power-enhancing effects of DRR might be operating in the slope-difference conditions. Second, as criterion reliability increased, so did the observed-operational differences in power. In slope-difference conditions, this reliability effect facilitated the DRR-related power-enhancing effects such that the increases in power, averaged across all levels of selectivity, were largest when the criterion was perfectly reliable and slope differences were large. This trend was strengthened as mean differences on *Y* increased and subgroups approached equal representation in the applicant population. However, consistent with the trends from Figure 17, power was decreased in intercept-difference only conditions.

The finding that DRR artifacts can *increase* power to detect overall differences in prediction in certain settings may seem quite peculiar at first, as statistical artifacts are generally regarded as factors than decrease statistical power. However, it is important to bear in mind that these analyses represent comparisons among nested regression models

rather than evaluations of the fit of individual regression models (i.e., comparisons to null models). Although artifacts worsen the fit of a model when analyzed in isolation and when compared to a null model, there is no guarantee that this worsened fit decreases the power to identify differences between nested models, as not all variables included in these models are impacted by artifacts to the same degree.

In the present case, it is worth noting that artifacts have a more direct impact on the fit of Model 1 in the Cleary framework than on Model 3: Model 1 includes the predictor on its own whereas Model 3 includes the predictor, a group main effect, and a group-by-predictor interaction. The predictor is directly range restricted, but the group-membership dummy variable is only indirectly range restricted via selection on the predictor. The expected impact of range restriction on model fit is always greater when it occurs as DRR than when it occurs as IRR because DRR is the most efficient mechanism for attenuating the variance of a variable. Thus, if Model 1 is affected by DRR and Model 3 includes additional variables affected by other types of range-restriction artifacts, the fit of both models is worsened but the difference between the models can actually increase if the fit of each is was worsened to a different degree. This is true when groups exhibit slope differences such that Model 3 represents the addition of two effects over Model 1 (i.e., slope and intercept terms) as opposed to the addition of only one non-null effect in the case of intercept differences. In fact, Figure 17 shows that power was slightly attenuated in all cases where there were intercept differences but no slope differences; it is the presence of slope differences that allows the power of observed analyses to exceed the power of operational analyses when detecting overall differences in prediction.

Effects of simulation parameters on statistical power of tests of slope differences. With the analyses of overall differences in prediction having demonstrated some unusual DRR-related power effects for slope-difference conditions, it is important to now consider the power to explicitly detect slope differences in the Cleary framework by comparing Model 3 to Model 2. Table 28 indicates that P_{Ref}, *SR*, and ρ_{XY_Foc} were the most important contributors to observed-operational differences in *F* ratios representing effects on power for tests of slope differences. These parameters interacted with each other and with ρ_{YY} to explain variation in *F*-ratio differences, although ρ_{YY} did not appear to have an important effect in the broader scheme of results.

Figure 19 depicts the two-way interaction between SR and ρ_{YY} and shows that, despite the power-enhancing effects of DRR observed in tests of overall differences in prediction, artifacts had a distinct power-reducing effect on analyses of slope differences. Although there was a positive association between *SR* and observed-operational differences in power (i.e., lower selectivity settings [high *SR*s] were associated with smaller decrements in power), power was, on average, still lower in observed models than in operational models. This interaction was such that the positive association between *SR* and observed-operational differences in *F* ratios was stronger under conditions of higher reliability. Power was reduced to the greatest extent when selectivity was high (i.e., the *SR* was low) and, under conditions of high selectivity, the effect of criterion reliability on power was virtually non-existent. However, as noted above, the effect of ρ_{YY} in this situation appears to be quite small.

Figure 20 depicts the effect of the three-way interaction among SR, P_{Ref}, and ρ_{XY_Foc} on observed-operational differences in *F* ratios. This interaction is such that there was a positive association between magnitudes of operational slope differences and observed-operational differences in power and this positive relationship was stronger when there were fewer referent group members in the applicant population and the selection ratio was lower. Relative to the operational models, *F* ratios of the observed models were reduced in magnitude to the greatest extent when slope differences were large, the referent and focal groups were equally represented in the applicant population, and the selection ratio was low.

The group-membership dummy variable in Models 2 and 3 is affected by IRR, but the product term added in Model 3 to represent the group-by-predictor interaction is affected by the joint influence of IRR on the dummy variable and DRR on the predictor variable. Whereas the differences between Model 3 and Model 1 could increase due to DRR in the presence of slope differences, the differences between Model 3 and Model 2 tend to decrease because artifacts will tend to have a bigger fit-worsening effect on Model 3 than Model 2.

Effects of simulation parameters on statistical power of tests of intercept differences. In the above sections, I have shown that DRR tends to increase the power of tests of overall differences in prediction while decreasing the power of tests of slope differences. I now turn my attention to the final test involved in the Cleary model: The test of intercept differences. Table 28 indicates that P_{Ref}, *SR*, and ρ_{YY} were the most important predictors of observed-operational differences in *F* ratios representing effects

on power for tests of intercept differences. These parameters interacted with each other in 3 two-way interactions to explain variance in *F*-ratio differences.

Figure 21 depicts the two-way interaction between SR and P_{Ref} and shows that the positive association between *SR*s and differences in *F* ratios varied as a function of P_{Ref}. Compared to the relationship when 70% of the applicant population came from the referent group, the association was slightly weaker (but with higher values, overall) when 90% of applicants were from the referent group and the relationship was disjointed when subgroups were represented equally, such that the largest difference occurred between low and moderate *SR*s and there was little difference between moderate and high *SR*s. The biggest reductions in *F* ratios occurred under conditions of high selectivity in which there were fewer members of the referent group represented in the applicant population. In Cleary model analyses, power is in large part a function of how equally groups are represented in an analysis. If a low selection ratio is applied to a predictor with mean differences, adverse impact will occur and the lower-scoring group will be less-well represented in the observed dataset than in the operational dataset, which reduces power. The effect of adverse impact on subgroup representation and statistical power appears to be greater when subgroup representation is more similar (and power is therefore closer to optimal) in the applicant population, as selection artifacts have a relatively larger detrimental impact on power in these settings.

Figure 22 depicts the two-way interaction between SR and ρ_{YY}, showing that the positive relationship between *SR*s and differences in *F* ratios was stronger for higher levels of reliability. The biggest reductions in *F* ratios occurred under conditions of high

selectivity when the criterion was measured with low reliability. This is perhaps the most intuitive interaction effect presented thus far, as lower reliability and a larger degree of range restriction are conventionally associated with reduced statistical power.

Figure 23 depicts the two-way interaction between P_{Ref} and ρ_{YY}, which is similar to the SR x P_{Ref} interaction except that there were no disjointed trends. The positive association between ρ_{YY} and differences in *F* ratios was steepest when P_{Ref} = .5, slightly flatter when P_{Ref} = .7, and flattest (but with the highest values) when P_{Ref} = .9. The biggest reductions in *F* ratios occurred when the criterion was measured with low reliability and when there were fewer members of the referent group in the applicant population. Similar to the SR x P_{Ref} interaction, the P_{Ref} x ρ_{YY} interaction shows that lower measurement quality and departures from equal operational subgroup proportions are associated with reduced statistical power.

These results regarding power in Cleary analyses collectively indicate that, although DRR can increase the power to detect overall differences in prediction, the power of subsequent analyses to detect slope and intercept differences tends to be attenuated by statistical artifacts. In the next section, I turn my focus to the effects of indirect range restriction, which is more likely than direct range restriction to occur in operational selection systems and has the potential for more complex effects on differential prediction findings due to IRR's effects on regression parameters.

Results of the Indirect Range Restriction Simulation

Effects of simulation parameters on d_{Mod_Signed} estimates. Table 32 provides a summary of the variance explained in observed-operational differences in d_{Mod_Signed} values by each effect that met my $\eta^2 \geq .01$ cutoff.

***Effects of simulation parameters on the values of* $\mathbf{d_{Mod\_Signed}}$ *estimates*.** Table 28 indicates that *SR*, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} were the most important contributors to observed-operational differences in d_{Mod_Signed} values and these parameters interacted to explain variation in observed-operational differences. Figure 24 shows that, under conditions of IRR, differences between observed and operational d_{Mod_Signed} estimates were explained by the three-way interaction among SR, ρ_{ZY_Ref}, and ρ_{ZY_Ratio}. There was effectively no relationship between *SR* and differences in d_{Mod_Signed} when the *Z*-*Y* relationship was equal between groups (i.e., $\rho_{ZY\_Ratio} = 1$), as all differences were zero or near zero under these circumstances. Additionally, differences in d_{Mod_Signed} averaged zero under conditions of low selectivity (i.e., *SR* = .9). However, a negative relationship between *SR* and differences in d_{Mod_Signed} emerged and strengthened as ρ_{ZY_Ratio} shrank so that the focal ρ_{ZY} value became smaller than the referent ρ_{ZY} value; these negative relationships were stronger when the referent ρ_{ZY} value was large. The largest observed-operational d_{Mod_Signed} differences occurred when the *SR* was low, ρ_{ZY_Ref} was large, and ρ_{ZY_Foc} was small (via ρ_{ZY_Ratio} being small).

These results indicate that, as long as there is a similar relationship between the selection variable and the criterion, the d_{Mod_Signed} estimates computed for a predictor affected by IRR should not be systematically affected by statistical artifacts. Note that the

SR x ρ_{XY_Foc} DRR effect from Figure 15 was actually a special case of this effect. In that effect, *SR* interacted with validity differences on the selection variable (*X*) to produce a pattern of results very similar to those seen in the panels of Figure 24. The mechanism of the IRR effect on d_{Mod_Signed} differences is effectively the same as the mechanism of the DRR effect that I described earlier. Differences in how the selection variable relates to the criterion across subgroups correspond to differences in how range restriction is transferred to the criterion from the selection variable. Differences between subgroups in how predictor range restriction translates into criterion range restriction consequently affect how the restricted d_Y value compares to the magnitude of the restricted $r_{XY\_Ref}d_X$ product.

***Effects of simulation parameters on the signs of* $\mathbf{d_{Mod\_Signed}}$ *estimates*.** Table 28 indicates that *SR*, ρ_{ZY_Ratio}, and δ_Y were the most important contributors to observed-operational differences in the signs of d_{Mod_Signed} values. These parameters also interacted with less important parameters (i.e., ρ_{XZ}, ρ_{ZY_Ref}, and δ_Z) to explain IRR-related sign changes in d_{Mod_Signed}. The factors associated with sign differences between observed and operational d_{Mod_Signed} values can be characterized by three interactions.

The three-way interaction among ρ_{ZY_Ref}, δ_Y, and δ_Z is depicted in Figure 25, which shows that, for these parameters, the highest rates of sign changes in d_{Mod_Signed} values (nearly 50%) occurred when mean differences on *Y* were small, mean differences on *Z* were small, and ρ_{ZY_Ref} was large. In settings where mean differences on the selection variable *Z* were moderate in magnitude (i.e., .5) and *Z* was at least moderately correlated with the criterion, the rate of sign changes could be expected to range from 0%

when the mean differences on *Y* were .5 or .7 to about 38% when mean differences on *Y* were smaller (i.e., .3). However, the key effect here seems to be the main effect of δ_Y; the moderating effects of ρ_{ZY_Ref} and δ_Z were comparatively much smaller in magnitude.

The three-way interaction among ρ_{ZY_Ratio}, ρ_{XZ}, and δ_Y is depicted in Figure 26, where a small mean difference on *Y* was the key predictor of elevated rates of sign changes in d_{Mod_Signed} in settings where the validity of *Z* was smaller for the focal group than for the referent group. However, there was little risk of a sign change when the validity of *Z* was equal for the referent and focal groups. This was particularly true in scenarios in which *Z* was moderately to strongly correlated with *X* (the predictor of interest) – such a scenario resembles what might happen if *X* (a cognitive ability measure) were a component of a composite predictor were used to make selection decisions.

The four-way interaction among SR, ρ_{ZY_Ref}, ρ_{ZY_Ratio}, and δ_Y is depicted in Figure 27 and is an amalgam of 4 three-way interactions:

- SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio}
- SR x ρ_{ZY_Ref} x δ_Y
- SR x ρ_{ZY_Ratio} x δ_Y
- ρ_{ZY_Ref} x ρ_{ZY_Ratio} x δ_Y

Once again, a small mean difference on *Y* was the key predictor of elevated rates of sign changes in d_{Mod_Signed}, but only when the validity of *Z* was smaller for the focal group than for the referent group. These effects were stronger when the *SR* was low and *Z* was highly valid for the referent group.

Consistent with how the effects that explained variation in observed-operational differences in indirectly range restricted d_{Mod_Signed} values were closely conceptually related to DRR effects, this set of *SR*, ρ_{ZY_Ratio}, and δ_Y parameters were conceptually related to the parameters that explained DRR-related changes in the signs of d_{Mod_Signed} values. Sign changes were most likely when operational mean differences on *Y* were small in magnitude, the selection ratio was low, and there were operational slope differences on the selection variable. Just as with the DRR trends presented earlier, this IRR-related set of circumstances increases the likelihood that the average effect of underprediction in operational data will be distorted to appear as an average effect of overprediction in observed data. However, in light of the lack of evidence showing large magnitudes of differential validity or slope differences for high-stakes assessments after accounting for relevant artifacts, I view sign changes in d_{Mod_Signed} to be an unlikely event in applied selection systems.

Effects of simulation parameters on intercept-difference regression coefficients. Table 31 provides a summary of the variance explained in observed-operational differences in intercept-difference coefficients by each effect that met my $\eta^2 \geq .01$ cutoff.

Effects of simulation parameters on the values of intercept-difference regression coefficients. Table 28 indicates that *SR*, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} were the most important contributors to observed-operational differences in intercept-difference coefficients. The effect of the *SR* x ρ_{ZY_Ref} x ρ_{ZY_Ratio} interaction on differences in intercept-difference regression coefficients is depicted in Figure 28. The effects of these

simulation parameters on intercept-difference coefficients were very similar to their effects on d_{Mod_Signed} values. There was no association between *SR*s and differences in intercept-difference coefficients when *Z* was equally valid for both subgroups, but a negative association between *SR*s and differences in intercept-difference coefficients emerged as *Z* exhibited greater differential validity. This effect was stronger as the validity of *Z* in the referent group increased. The largest positive changes in intercept-difference coefficients (i.e., differences that indicate observed intercept differences were higher in value than operational intercept differences, representing more overprediction or less underprediction than in the operational analyses) occurred when the selection ratio was low, ρ_{ZY_Ref} was high, and ρ_{ZY_Ratio} was 0 (i.e., *Z* was highly valid for the referent group, but not at all valid for the focal group). As I stated earlier regarding the d_{Mod_Signed} effects, I do not expect this pattern of effects to occur in professionally developed selection systems because there is little evidence of consistent large differences in validity for high-stakes assessments.

Effects of simulation parameters on the signs of intercept-difference regression coefficients. Table 28 indicates that *SR*, $\rho_{ZY_{Ratio}}$, and δ_Y were the most important contributors to observed-operational differences in the signs of intercept-difference coefficients. These parameters were also the most important for explaining observed-operational sign differences in d_{Mod_Signed} effects. Similar to the effects noted earlier for sign differences in d_{Mod_Signed} effects, the *SR*, ρ_{ZY_Ratio}, and δ_Y parameters interacted with each other and with less-important parameters (i.e., ρ_{ZY_Ref} and δ_Z) in two interactions.

The three-way interaction among ρ_{ZY_Ref}, δ_Y, and δ_Z is depicted in Figure 29. This interaction is such that the negative relationship between mean differences on *Y* and proportions of intercept-difference sign changes tended to become stronger as ρ_{ZY_Ref} increased; this moderated effect became more pronounced as mean differences on *Z* decreased. The highest rates of sign changes were observed when δ_Y was small, ρ_{ZY_Ref} was large, and subgroups had equal means on *Z*. Low rates of sign changes were observed when δ_Y was large.

The four-way interaction among SR, ρ_{ZY_Ref}, ρ_{ZY_Ratio}, and δ_Y depicted in Figure 30 is an amalgam of 3 three-way interactions:

- SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio}
- SR x ρ_{ZY_Ratio} x δ_Y
- ρ_{ZY_Ref} x ρ_{ZY_Ratio} x δ_Y

This interaction resembles the interaction discussed earlier involving the same set of variables for predicting sign changes of d_{Mod_Signed} values (see Figure 27). Figure 30 shows that there was no association between mean differences on *Y* and proportions of sign changes when the selection ratio was high (.9). However, small mean differences on *Y* were predictive of high rates of sign differences when *Z* was less valid for the focal group than for the referent group (i.e., $\rho_{ZY\_Ratio} < 1$), particularly in highly selective settings.

As with sign changes of d_{Mod_Signed} values, IRR-related sign changes in intercept-difference coefficients appear to occur as a function of circumstances that are not

commonly encountered in operational selection programs. For sign changes to occur, the unrestricted mean differences on the criterion have to be small (so that, on average, there is underprediction) and the selection variable must exhibit rather extreme levels of differential validity. Neither of these conditions is commonly reported in practice, but small mean differences on a criterion will be more likely if the criterion represents performance on less cognitively loaded tasks.

Effects of simulation parameters on slope-difference regression coefficients. Table 34 provides a summary of the variance explained in observed-operational differences in slope-difference coefficients by each effect that met my $\eta^2 \geq .01$ cutoff.

Effects of simulation parameters on the values of slope-difference regression coefficients. Table 28 indicates that *SR*, ρ_{XY_Foc}, ρ_{XZ}, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} were the most important contributors to observed-operational differences in slope-difference coefficients. These parameters interacted with each other in two interactions.

The three-way interaction among SR, ρ_{XY_Foc}, and ρ_{XZ} is depicted in Figure 31. This interaction shows that there was no association between ρ_{XY_Foc} and changes in slope differences when the selection ratio was high, but a negative association developed as the selection ratio became smaller; the moderating effect of the selection ratio grew stronger as ρ_{XZ} increased. When ρ_{XZ} was small (.2) or moderate (.5) in magnitude, mean slope-difference changes were zero or negative, but the mean changes spanned positive values (indicating that observed slope differences favoring the referent group exceeded the operational differences in magnitude) and negative values (indicating that observed slope differences were smaller than operational; or, in the case of equal operational

slopes, that the focal group's observed slope was steeper than the referent group's) when ρ_{XZ} was large (.8) and this effect was stronger in highly selective settings. More specifically, when ρ_{XZ} was large and operational slope differences were large (ρ_{XY_Foc} = .1), observed slope-difference coefficients were more likely to overestimate the magnitude of slope differences. However, when operational slope differences were non-existent (i.e., ρ_{XY_Foc} = .5), observed slope-difference coefficients were more likely to falsely indicate the presence of slope differences favoring the focal group.

The four-way interaction among SR, ρ_{ZY_Ref}, ρ_{ZY_Ratio}, and ρ_{XZ} is depicted in Figure 32. This interaction confirms the earlier observation that parameters of the selection system did not systematically affect slope-difference coefficients when the selection ratio was large. A positive association emerged between ρ_{ZY_Ratio} and changes in slope differences as the selection ratio became smaller and the validity coefficient for *Z* in the referent group became larger. These effects mean that magnitudes of selection-related changes in slope differences were more likely to be positive in more selective contexts where *Z* is equally valid for the referent and focal groups. When *Z* was less valid for the focal group, selection-related changes in slope differences tended to be negative.

As Figure 32 did not include ρ_{XY_Foc} as a parameter and therefore did not convey the effects of IRR on slope differences of varying magnitudes, I have created an additional figure to aid the interpretation of results. Figure 33 shows an interaction among ρ_{XY_Foc}, ρ_{XZ}, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} with the selection ratio set at .5. This figure makes it clear that when ρ_{ZY_Ratio} = 1, all non-null slope-difference coefficients are overestimated; furthermore, when ρ_{ZY_Ratio} = 1, IRR has no effect on slope-difference estimates in

settings where operational slopes are equal. However, as ρ_{ZY_Ratio} decreases, slope differences are more likely to be underestimated and operationally equal slopes are more likely to produce observed slope-difference coefficients that artifactually indicate that the focal group has a steeper slope. These effects become stronger as ρ_{XZ} and ρ_{ZY_Ref} increase.

From these effects and the fact that ρ_{XY_Foc}, ρ_{XZ}, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} were all important contributors, it is clear that the configuration of relationships among the variables within subgroups (and differences in these configurations between subgroups) are of paramount importance for determining how statistical artifacts will affect observed-operational differences in slope-difference coefficients. Underestimating operational slope-differences from observed data was most likely when *Z*'s referent-group validity was higher and *Z* exhibited differential validity; higher selectivity and higher correlations between *X* and *Z* enhanced this effect.

Effects of simulation parameters on the signs of slope-difference regression coefficients. Table 28 indicates that *SR*, ρ_{XY_Foc}, ρ_{XZ}, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} were the most important contributors to observed-operational differences in the signs of slope-difference coefficients; these are the same parameters implicated in the changes of the values of slope-difference coefficients. The effects of these five parameters on the signs of slope-difference regression coefficients can be characterized by a set of 5 four-way interactions:

- SR x ρ_{XY_Foc} x ρ_{ZY_Ref} x ρ_{ZY_Ratio} (shown in Figure 34)
- SR x ρ_{XY_Foc} x ρ_{ZY_Ref} x ρ_{XZ} (shown in Figure 35)

- SR x ρ_{XY_Foc} x ρ_{ZY_Ratio} x ρ_{XZ} (shown in Figure 36)
- SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio} x ρ_{XZ} (shown in Figure 37)
- ρ_{XY_Foc} x ρ_{ZY_Ref} x ρ_{ZY_Ratio} x ρ_{XZ} (shown in Figure 38)

These 5 four-way interactions seem to suggest a five-way interaction, but the essence of that effect can be derived from the four-way interactions. Sign changes were most likely when the selection ratio was small, *Z* was highly valid for the referent group (i.e., ρ_{ZY_Ref} was large), *Z* was not valid for the focal group (i.e., $\rho_{ZY\_Ratio} = 0$), *X* and *Z* were highly correlated (i.e., ρ_{XZ} was large), and operational slope differences were smaller in magnitude (i.e., ρ_{XY_Foc} was .3 as opposed to .1). There was no association between parameters and sign changes when the selection ratio was high (.9) and/or ρ_{ZY_Ref} was small (.2).

Effects of simulation parameters on statistical power. Table 35 provides a summary of the variance explained in observed-operational differences in *F* ratios from non-null operational Cleary analyses by each effect that met my $\eta^2 \geq .01$ cutoff.

Effects of simulation parameters on statistical power of tests of overall differences in prediction. Table 28 indicates that *SR*, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} were the most important contributors to observed-operational differences in *F* ratios representing effects on power for tests of overall differences in prediction. These effects were accompanied by (and were involved in interactions with) the effects of less important parameters that explained little variance in the dependent variable.

The first three effects were the smallest effects and they contributed relatively little to the explanation of observed-operational differences in *F* ratios. They did,

however, pass my $\eta^2 \geq .01$ threshold and therefore merit consideration. Figure 39 shows that ρ_{XY_Foc} had a small positive association with differences in *F* ratios and that slightly larger increases in *F* ratios occurred as subgroup validities of *X* approached parity. Figure 40 depicts the small positive main effect of ρ_{YY} and indicates that higher levels of reliability corresponded to slightly larger increases in *F* ratios; this effect is essentially showing how criterion reliability modulated the average effect of range restriction on power. Figure 41 shows that *SR* was negatively associated with differences in *F* ratios, with larger positive differences occurring in highly selective conditions; this relationship was stronger when mean differences on the criterion were larger.

Two other interaction effects emerged that had larger impacts on differences in *F* ratios. The first of these is the three-way interaction among *SR*, ρ_{ZY_Ref}, and δ_Z shown in Figure 42. A negative relationship developed between *SR* and differences in *F* ratios as the validity of *Z* increased for the referent group; this moderated effect was stronger as the mean differences on *Z* decreased. *F* ratios increased the most when the *SR* was low, *Z* was highly valid, and there were no mean differences on *Z*.

The other impactful effect on differences in *F* ratios was the four-way interaction among *SR*, P_{Ref}, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} shown in Figure 43. This effect was such that there were no observed-operational differences in *F* ratios when *Z* had equal validity for both subgroups, regardless of the level of validity in the referent group. However, a negative relationship between ρ_{ZY_Ratio} and differences in *F* ratios emerged as *SR* decreased, P_{Ref} decreased, and ρ_{ZY_Ref} increased. The largest differences in *F* ratios occurred when ρ_{ZY_Ref} was high, ρ_{ZY_Ratio} was low, *SR* was low, and P_{Ref} was low.

These large differences seem unlikely to happen in real-world selection systems because of the magnitude of differential validity required for them to manifest.

Considered together, these effects show that large positive differences in *F* ratios indicating higher power in observed analyses than in operational analyses were most likely when *SR* was low, δ_Z was small, ρ_{ZY_Ref} was high, ρ_{ZY_Ratio} was low, and P_{Ref} was low; ρ_{XY_Foc} and ρ_{YY} had negligible effects. These trends mean that differential validity of the selection variable (especially when it is highly valid for one group) can affect the subgroup regression lines enough to enhance the improvement of Model 3's fit over Model 1 and that high selectivity boosts IRR's effects on power. Furthermore, lower P_{Ref} and lower δ_Z values both have the effect of increasing the representation of the focal group in the selected subpopulation. This facilitates the development of positive observed-operational differences in *F* ratios by preserving a good deal of variability in group membership post-selection. By preserving variation in the group-membership dummy variable, lower P_{Ref} and lower δ_Z values enhance the effect of IRR on *F*-ratio differences and increase the chance of a significant overall test of differential prediction being estimated from observed data.

Effects of simulation parameters on statistical power of tests of slope differences. Table 28 indicates that *SR*, ρ_{XY_Foc}, δ_Z, and ρ_{YY} were the most important contributors to observed-operational differences in *F* ratios representing effects on power for tests of slope differences. These parameters were accompanied by (and were involved in interactions with) the effects of less important parameters. The smallest of these effects was the main effect of ρ_{XZ}, such that bigger reductions in *F* ratios occurred when ρ_{XZ}

was small (.2) or moderate (.5) than when it was large (.8; see Figure 44). As for larger effects, I will briefly summarize each of the six interactions below and then draw overall conclusions from the patterns of effects.

The P_{Ref} parameter moderated the positive association between ρ_{XY_Foc} and *F*-ratio differences, such that the effect of ρ_{XY_Foc} was weaker when P_{Ref} was .9 than when it was .5 or .7 (see Figure 45). The biggest negative differences occurred when ρ_{XY_Foc} and P_{Ref} were both low.

There was a positive relationship between ρ_{YY} and differences in *F* ratios and this relationship was stronger when there were fewer referent group members in the population (see Figure 46). The biggest negative differences occurred when ρ_{YY} and P_{Ref} were low.

There was a positive relationship between ρ_{XY_Foc} and differences in *F* ratios, and this relationship was stronger when reliability was lower (see Figure 47). The biggest negative differences occurred when ρ_{YY} and ρ_{XY_Foc} were low.

There was a positive relationship between ρ_{YY} and differences in *F* ratios (see Figure 47 and Figure 48), with a stronger relationship occurring when ρ_{ZY_Ratio} was higher (see Figure 48). The biggest negative differences occurred when ρ_{YY} was low and ρ_{ZY_Ratio} was high.

There was no effect of δ_Z on *F*-ratio differences when *SR* was .9, but larger mean differences on *Z* were associated with stronger positive relationships between ρ_{XY_Foc} and

differences in *F* ratios as selectivity increased (see Figure 49). The biggest negative differences occurred when ρ_{XY_Foc} was small, *SR* was low, and δ_Z was large.

The *SR* and δ_Z parameters also interacted with ρ_{ZY_Ratio} (see Figure 50) so that the negative relationship between δ_Z and *F*-ratio differences strengthened as *SR* decreased and ρ_{ZY_Ratio} increased. The biggest negative differences occurred when *SR* was low, δ_Z was large, and *Z* had equal validity in both subgroups

Considered together, the effects described above indicate that negative observed-operational differences in *F* ratios indicating lower power for observed analyses than operational analyses were most pronounced when *SR* was low, ρ_{YY} was low, P_{Ref} was small, δ_Z was large, ρ_{XY_Foc} was small, and ρ_{ZY_Ratio} was high. This result contrasts with the results reported above for the power of tests of overall differences in prediction in which power increased under conditions of IRR. Similar to the effects observed for DRR, IRR appears to enhance the differences between Models 3 and 1 in the Cleary framework, but reduces the differences between Models 3 and 2.

Effects of simulation parameters on statistical power of tests of intercept differences. Table 28 indicates that *SR*, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} were the most important contributors to observed-operational differences in *F* ratios representing effects on power for tests of intercept differences. These parameters interacted with each other and with the P_{Ref}, δ_Y, and δ_Z parameters to explain differences in *F* ratios; ρ_{YY} also had a small main effect.

On average, observed-operational differences in *F* ratios for intercept-difference tests were positive, indicating that IRR increased the power to detect intercept differences. There was a very small negative relationship between ρ_{YY} and differences in *F* ratios (see Figure 51) and no relationship between *SR* and differences in *F* ratios when *Z* had equal validity for both subgroups (see Figure 52, Figure 53, and Figure 55) or when ρ_{ZY_Ref} was low (see Figure 54). However, a negative monotonic (negatively decelerating) relationship emerged as ρ_{ZY_Ref} increased, ρ_{ZY_Ratio} decreased (i.e., subgroup validities of *Z* became discrepant), P_{Ref} decreased, δ_Z decreased, and δ_Y increased. The biggest positive differences occurred when *SR* was low, ρ_{ZY_Ref} was high, *Z* was only valid for the referent group, P_{Ref} was low, δ_Z was 0, and δ_Y was high. Given the important role of ρ_{ZY_Ratio} in these effects and the fact that large differences in validity are not common or expected, the effects of IRR on the power of intercept-difference tests are likely to be modest in most circumstances (or near zero if *Z* exhibits no differences in validity).

As a set, the effects of IRR on power for the three Cleary model comparisons paint an interesting picture of selection artifacts' effects on researchers' abilities to detect significant differences among models. Selection artifacts increased the power of tests of overall differences in prediction and tests of intercept differences, but diminished the power of analyses to detect slope differences. With the effects on power established, I examine how IRR and selection system parameters affect Type I errors in the following section.

Effects of simulation parameters on Type I errors. Table 36 provides a summary of the variance explained in observed-operational differences in *F* ratios from null operational Cleary analyses by each effect that met my $\eta^2 \geq .01$ cutoff.

***Effects of simulation parameters on Type I errors of tests of overall differences in prediction*.** Table 28 indicates that *SR*, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} were the most important contributors to observed-operational differences in *F* ratios representing effects on Type I errors for tests of overall differences in prediction. These parameters interacted with each other, as well as with ρ_{YY}, P_{Ref}, and δ_Z, to explain variation in *F*-ratio differences.

Criterion reliability moderated the effects of two other parameters: *SR* and ρ_{ZY_Ratio}. Both *SR* and ρ_{ZY_Ratio} had negative associations with differences in *F* ratios and these associations were stronger when ρ_{YY} was higher (see Figure 56 and Figure 57, respectively). The largest negative differences occurred when ρ_{YY}, *SR*, and ρ_{ZY_Ratio} were low. There was no systematic effect of reliability on differences in *F* ratios when selectivity was low or subgroup validities for *Z* were equal.

Figure 58 and Figure 59 show the effects of other moderators on the association between *SR* and differences in *F* ratios. These figures show that there was no effect of *SR* on *F*-ratio differences when subgroup validities of *Z* were equal and that *SR* had very little effect when ρ_{ZY_Ref} was low. However, a negative association between *SR* and *F*-ratio differences emerged as ρ_{ZY_Ref} increased, P_{Ref} decreased, and δ_Z decreased. The largest positive differences in *F* ratios occurred when *SR* was low, ρ_{ZY_Ref} was high, ρ_{ZY_Ratio} was low, P_{Ref} was low, and δ_Z was low.

The explanation for these trends is similar to the explanation offered earlier for why IRR increased the power of analyses of overall differences in prediction. Although no slope or intercept differences existed in the operational data for the IRR conditions examined here, larger differences in validity of the selection variable created larger differences between subgroup regression lines linking *X* to *Y* computed from observed data. Performing selection on *Z* when there were subgroup differences in the *Z*-*Y* relationship and gave rise to artifactual subgroup regression-line differences for the *X*-*Y* relationship. Holding the selection ratio constant, there will be more variation in group membership in the selection subpopulation when P_{Ref} and δ_Z parameters are both low, which increases the likelihood that IRR's effects on subgroup regressions will cause a false-positive result when testing the null hypothesis that the subgroup regression lines are equal.

On average, differences in *F* ratios were positive, which means that rates of Type I errors will tend to be elevated for tests of overall differences in prediction. In the following two sections, I examine parameters' effects on follow-up analyses of slope and intercept differences to determine how artifacts might affect overall conclusions about differential prediction.

Effects of simulation parameters on Type I errors of tests of slope differences. Table 28 indicates that *SR*, ρ_{XZ}, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} were the most important contributors to observed-operational differences in *F* ratios representing effects on Type I errors for tests of slope differences. The interactions among these parameters, as well as their interactions with P_{Ref}, ρ_{YY}, and δ_Z, are depicted in a set of four figures (see Figure

60, Figure 61, Figure 62, and Figure 63). Collectively, these figures show how the effect of $\rho_{ZY_{Ratio}}$ is moderated by the effects of six other parameters.

There was no effect of ρ_{ZY_Ratio} when ρ_{XZ} was low (i.e., .2; see Figure 60, Figure 61, Figure 62, and Figure 63). Negative effects of ρ_{ZY_Ratio} on *F*-ratio differences emerged as ρ_{XZ} increased (see Figure 60, Figure 61, Figure 62, and Figure 63), ρ_{YY} increased (see Figure 60), *SR* decreased (see Figure 61, Figure 62, and Figure 63), P_{Ref} decreased (see Figure 61), ρ_{ZY_Ref} increased (see Figure 62), and δ_Z decreased (see Figure 63). The biggest positive differences occurred when ρ_{ZY_Ratio} was 0, ρ_{XZ} was high, ρ_{YY} was high, *SR* was low, P_{Ref} was low, ρ_{ZY_Ref} was high, and δ_Z was 0.

Similar to the effects of simulation parameters on Type I error potential for tests of overall differences in prediction, IRR caused Type I errors to be elevated for tests of slope differences. The mechanisms involved in this were similar to those involved in the tests of overall differences: Larger discrepancies in the subgroup validity of *Z* (via ρ_{ZY_Ratio}) created larger artifactual differences in observed subgroup slopes for *X* in more selective conditions. Additionally, smaller P_{Ref} and δ_Z parameters increased the likelihood that the artifactual differences in slopes would emerge as significant. The ρ_{XZ} parameter also influences slope-difference tests, such that Type I error rates were higher when ρ_{XZ} was larger. These results strongly indicate that the configurations of *Z*'s relationships with *X* and *Y* are critical for determining how range restriction of *Z* will affect differential prediction analyses applied to predictor *X*.

Effects of simulation parameters on Type I errors of tests of intercept differences. Table 28 indicates that *SR*, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} were the most important contributors to observed-operational differences in *F* ratios representing effects on Type I errors for tests of intercept differences. These parameters interacted with each other and with P_{Ref}, δ_Z, and ρ_{YY} to explain variation in *F*-ratio differences. Depicted in a set of four figures (see Figure 64, Figure 65, Figure 66, and Figure 67), three of these effects closely resemble interactions described earlier with regard to Type I errors of tests of overall differences in prediction.

The interaction between *SR* and ρ_{YY} in Figure 64 is almost identical to the interaction depicted in Figure 56. There was a negative relationship between *SR* and differences in *F* ratios that became stronger as reliability increased. The largest positive differences in *F* ratios occurred when *SR* was low and ρ_{YY} was high; variation in ρ_{YY} had virtually no effect when the *SR* was high.

The interaction between ρ_{ZY_Ratio} and ρ_{YY} in Figure 65 is very much like the interaction from Figure 57. There was a negative relationship between ρ_{ZY_Ratio} and differences in *F* ratios that became stronger as reliability increased; the largest positive differences were observed when $\rho_{ZY_Ratio}$was 0 and $\rho_{YY}$was 1.

The three-way interaction among *SR*, ρ_{ZY_Ref}, and δ_Z in Figure 66 reveals that there was a small negative effect of *SR* on differences in *F* ratios when ρ_{ZY_Ref} was low; this negative effect became stronger as ρ_{ZY_Ref} increased and δ_Z decreased. The largest positive differences were observed when *SR* was low, ρ_{ZY_Ref} was high, and δ_Z was 0.

Finally, the four-way interaction among SR, P_{Ref}, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} in Figure 67 is nearly identical to the interaction among the same parameters shown in Figure 58. There was no effect of *SR* when subgroup validities of *Z* were equal, but a negative relationship emerged as P_{Ref} decreased, ρ_{ZY_Ref} increased, and ρ_{ZY_Ratio} decreased. The largest positive differences occurred when *SR* was low, P_{Ref} was low, ρ_{ZY_Ref} was large, and ρ_{ZY_Ratio} was 0.

The same set of parameters that were important for explaining Type I errors for overall tests of differential prediction were also important for explaining Type I errors for overall tests of intercept differences. Lower selection ratios, stronger referent-group *Z*-*Y* relationships, and larger disparities in subgroup *Z*-*Y* relationships were all linked to increased Type I errors. Consistent with the previous Type I error analyses, smaller P_{Ref} and δ_Z parameters increased the risk of a positive observed-operational difference in *F* ratios; these parameters help to preserve variability of the group-membership dummy variable in the selected subpopulation, which increases the risk of detecting an artifactual intercept difference.

Discussion

The I-O psychology literature has been lacking a systematic treatment of the effects of range restriction and criterion measurement error on the detection of differential prediction in operational selection programs. The simulation by Aguinis et al. (2010) compared artifact-attenuated results to true-score differential prediction parameters and therefore did not support the field's understanding of operational predictive bias analyses. To address this gap in the literature, I designed an analytic

simulation in which operational mean, variance, and covariance parameters were impacted by statistical artifacts prior to use in differential prediction analyses. My comparisons of observed parameters to operational parameters allowed me to determine how the artifacts impacted the expected values of statistical results, including whether the artifacts biased the estimates of subgroup differences in slopes or intercepts and how artifacts impacted the accuracy with which operational slope and intercept differences can be detected. My simulation's focus on how statistical artifacts affect the parameters involved in operational differential prediction analyses supports insights regarding which characteristics of a selection system have the biggest impacts on the accurate identification of differential prediction trends from observed data. My parameter-oriented approach allowed me to demonstrate effects of artifacts that generalize across all sample sizes without having to explicitly manipulate sample size parameters in my simulation.

This simulation was designed to answer two research questions. The first of these was Research Question 6: "Which parameters of applicant populations and selection systems are most predictive of biased estimation of (a) d_{Mod_Signed} effect sizes, (b) intercept differences, and (c) slope differences?" With regard to d_{Mod_Signed} values, the key parameters were the selection ratio and the magnitude of subgroup differences in the relationship between the criterion and the selection variable. In both the DRR and IRR simulations, d_{Mod_Signed} values were overestimated (in terms of raw value, not necessarily absolute magnitude) in conditions where the selection variable was more strongly related to the criterion for the referent group than for the focal group; this effect was stronger in more selective contexts. However, selectivity had no effect on d_{Mod_Signed} values when the

selection variable was equally predictive for both subgroups. The effects of IRR on intercept differences were essentially identical to the effects on range restriction on d_{Mod_Signed}. Given that subgroup validities and slopes do not tend to exhibit large differences even after accounting for artifacts (see Study 3), it appears unlikely that d_{Mod_Signed} effects will be seriously misleading when calculated from observed data as opposed to unrestricted operational data. The theoretical possibility of overlooking underprediction based on interpreting d_{Mod_Signed} effects or intercept-difference coefficients is best regarded as a statistical curiosity rather than a probable phenomenon.

In terms of IRR's effects on slopes, operational slope differences were only likely to be underestimated in magnitude when the selection variable exhibited a weaker association with the criterion in the focal group than in the referent group (this effect was stronger when the selection variable was more highly correlated with the predictor being examined for bias). However, when the selection variable had similar relationships with the criterion in both groups, slope differences were more likely to be overestimated in magnitude. Selectivity also played a key role in this effect: There was no systematic effect on slope differences in low-selectivity contexts, but higher selectivity was associated with stronger versions of the trends described above. These patterns of findings indicate that differential validity/prediction of the selection variable is a key determinant of whether IRR will lead to the underestimation or overestimation of slope-difference effects. If the selection variable has a similar relationship with the criterion across groups, slope differences are likely to be overestimated from observed data. Given that there is little evidence of high-stakes selection variables exhibiting dramatically

different levels of predictive efficacy across groups (again, see Study 3, where artifact-corrected slope differences were nonexistent or of a very small magnitude), I view it as probable that slope differences estimated from observed data will be overestimates rather than underestimates of the operational differences in subgroup slopes.

The second question answered by this simulation was Research Question 7: "Which parameters of applicant populations and selection systems have the biggest impact on the ability of researchers to detect (a) intercept differences and (b) slope differences?" My results regarding the power of the Cleary analysis under conditions of DRR run contrary to Aguinis et al.'s (2010) findings: I found that, on average, DRR caused increases in power for tests of overall differences, but decreases in power for follow-up analyses to determine slope differences or intercept differences. Additionally, Type I errors of the Cleary model were not affected by DRR because DRR cannot bias regression parameters and therefore cannot inflate the range at which false-positive differences are detected. When subgroups' regression lines do not differ, only the *F* ratio parameters for the fit of individual models compared to null models are affected by DRR and criterion measurement error; *F* ratio parameters for model comparisons in the Cleary framework are not affected.

The differences in findings between my DRR simulation and Aguinis et al.'s (2010) are further evidence of the importance of attending only to operational artifacts in predictive bias analyses rather than comparing observed results to true-score parameters. Aguinis et al.'s choice to use true-score parameters as the point of comparison in their simulation seriously distorted their findings and made those findings incongruent with

practice. Whereas Aguinis et al.'s simulation showed that artifacts reduce power to detect slope differences but increase Type I errors for detecting intercept differences, my simulation shows that power is reduced for both types of differential-prediction effects and that Type I errors for all comparisons conducted within the Cleary model should be unaffected by DRR and criterion measurement error.

The arguably more interesting simulation presented here was my simulation of indirect range restriction, as IRR is more likely than DRR to occur in practice and IRR has the potential to bias regression parameters. Similar to the DRR simulation's results, IRR tends to increase the power for tests of overall differences in prediction and decrease the power for tests of slope differences. However, whereas DRR tended to decrease the power for tests of intercept differences, IRR increased the power for these tests. Additionally, IRR appears to increase Type I errors of all Cleary tests, on average, but large differences in error rates were only especially pronounced when there were large subgroup differences in the selection variable's relationship with the criterion. These results indicate that IRR makes it more likely that researchers will detect differences in prediction using observed data that do not exist in the operational population. They also indicate that IRR may increase the chances that researchers will detect non-null intercept differences that exist in the operational population, but may reduce the chances of detecting non-null operational slope differences.

With these trends established, the obvious question that needs to addressed is whether IRR could explain historical patterns of differential prediction findings. My answer is a qualified "no." Based on the effects of simulation parameters on d_{Mod_Signed}

values and intercept differences, IRR is unlikely to alter conclusions about the directions of differences in prediction. Operational underprediction is only likely to appear as observed overprediction if the selection variable has its own serious issues with predictive equivalency across groups. In terms of the form that differential prediction takes, it is theoretically possible that IRR could contribute to the high rates at which intercept differences are observed and the low rates at which slope differences are observed. However, historically documented rates of intercept differences are dramatically higher than the rates of slope differences and I view it as highly unlikely that large operational slope differences are much more persistent than has been indicated by decades of analyses conducted using observed data. Small slope differences are plausible and, as Study 3 showed, may be more likely than previously believed, but they are unlikely to contribute to bias in prediction by virtue of overprediction remaining the decidedly more dominant differential prediction effect.

Across nearly all dependent variables, the two most consistently impactful simulation parameters in both the DRR simulation and the IRR simulation were selection ratios and validity/slope differences of the selection variable. Selection ratios certainly vary across real-world selection systems, as organizations with higher levels of prestige encounter larger numbers of applicants for a limited number of positions. However, as there is not much evidence of high-stakes predictors exhibiting large validity or slope differences, I argue that the most egregious effects that distorted differential prediction analyses in my simulation would not be likely to occur in professionally developed selection systems. Although I view differential predictive efficacy of the selection

variable as an unlikely occurrence, I can offer a couple of scenarios in which such a problem could possibly arise. It would be possible for *Z* to have differential relationship with criteria across groups if *Z* were based on highly culturally contaminated constructs/measures (i.e., an assessment developed for the referent group's culture for which members of the focal group may not have a relevant frame of reference) or a biased clinical/judgmental data-combination process. For example, if *Z* were a composite variable, it could exhibit subgroup differences in its relationship with *Y* if composite scores were formed by human judges applying different standards to members of different groups. If the judges allowed knowledge of subgroup membership to influence their composite judgments, they could end up diluting the predictor information available to them for focal-group members with idiosyncratic rater errors. If errors in judges' holistic ratings occurred for the focal group at a higher rate than for the referent group, it would attenuate *Z*'s relationship with *Y* in the focal group and give rise to differences in validity/prediction for *Z* that could bias estimates of differential prediction effects for predictor *X* via IRR. However, these are only possibilities and I do not view these issues as likely to occur in professionally developed selection systems.

My analyses of observed-operational differences in *F* ratios showed the directions of effects and indicated whether statistically significant results are more or less likely in different settings. However, I note that these analyses did not offer a clear indication regarding how potent the effects on statistical conclusions might be. The actual effects of artifacts on power and Type I error rates will be dependent upon sample size, but my simulation results only convey general effects by design. I recommend that researchers

use power calculators to determine the potency of artifacts' effects on power and Type I error rates for specific sample sizes when such detailed information is necessary.

Limitations. No simulation is perfect, and mine is no exception. Although my findings regarding power and Type I errors differ from the conclusions reached by Aguinis et al. (2010), my specific research questions and the fact that I relied on separate examinations of power and Type I error implications for each of the Cleary model contrasts leaves the question of overall rates of power/error unaddressed. Even after examining many dependent variables in detail, I can only make clear statements about the power and Type I error rates of pairwise model contrasts and can only speculate as to how these effects might affect the power and Type I error rates of the Cleary model as a complete set of analyses. Errors will be made randomly across the pairwise contrasts, but these errors may also be correlated, so the next step in researching the impacts of statistical artifacts on differential prediction should be to examine system-wide errors in conclusions from applications of the Cleary model.

In my IRR simulation, the effects of range restriction on slopes makes it difficult to unambiguously interpret changes in intercept-difference coefficients. Intercept differences are only interpretable in the absence of slope differences and, although I restricted my intercept-difference analyses to settings in which subgroups' operational slopes were equal, this is no guarantee that subgroup slopes remained similar in the observed data. My analyses of Type I errors for slope tests showed elevated rates of errors when IRR occurred, and the slope differences indicated by that finding could interfere with the interpretability of range-restricted intercept difference coefficients.

My simulation modeled range restriction as a single-stage top-down selection process, but range restriction can also result from other selection methods (e.g., multiple hurdles, ideal-point models in which moderate scores on a predictor are considered best). I acknowledge that other mechanisms of inducing range restriction may produce different patterns of results than what my simulation could illustrate. Top-down selection on a single variable may be regarded as a special case of multiple hurdles in which only one hurdle is used, so I expect that my findings will generalize to more complicated selection designs to some extent. However, selection systems involving multiple selection criteria have nuances that cannot be captured in a univariate selection model, so any generalizations of my simulation's results to multivariate selection contexts will need to be made with caution.

My simulation's parameters were chosen to span a variety of slope- and intercept-difference scenarios that will be of interest to differential prediction researchers. These scenarios include those in which the focal group's unrestricted intercepts indicate underprediction, overprediction, or equal intercepts and those in which the focal group's unrestricted slopes are equal to or flatter than the referent groups' slopes. By focusing my parameter-value choices on these scenarios of interest, I acknowledge that my simulation does not address all theoretically possible configurations of differences in prediction, such as scenarios in which the focal group has a steeper slope than the referent group. I chose to focus on slope differences favoring the referent group because there is generally greater concern that Cleary analyses will overlook settings in which a predictor does not relate as strongly to a criterion for a minority group as it does for a majority group.

The operational correlations between *X* and *Z* were constrained to be equal for the referent and focal groups in my IRR simulation because there was no precedent for studying differential predictor intercorrelations. However, varying this parameter between groups would likely contribute to variation in observed-operational statistical results in a fashion similar to the ρ_{ZY_Ratio} parameter. So, although I believe my simulation handled the correlation between *X* and *Z* in a way that is consistent with current knowledge about predictor intercorrelations, varying this parameter between groups would undoubtedly have an effect on statistical results that could be of theoretical interest.

Implications. The differences in findings between my simulation and the simulation published by Aguinis et al. (2010) indicate the criticality of maintaining a focus on operational selection systems when simulating differential prediction effects. Comparing observed data to true-score data leads to distorted patterns of findings that result in utterly misguided conclusions about differential prediction and predictive bias. Researchers and practitioners who conduct predictive bias analyses should attempt to account for range restriction and criterion measurement error artifacts, but should not account for predictor measurement error, as doing so would cease to provide operational estimates of differential prediction. Furthermore, implementing latent-variable approaches to analyzing differential prediction, such as the method recently proposed by Culpepper et al. (2019), would undermine the integrity of the analysis and would violate the recommendations given in SIOP's (2018) *Principles*.

Contrary to prior simulation evidence, statistical artifacts do not offer plausible explanations for the long-observed finding of minority overprediction. The simulation conditions in which the most serious interpretation problems emerged are unlikely to occur in practice, as the selection variable would need to exhibit extreme differences in subgroup slopes/validity to have a substantive impact on differential prediction findings. There is little evidence of such extreme differences in slopes/validity for high-stakes predictors. Additionally, as long as researchers and practitioners involved in the enterprise of conducting differential prediction analyses are focused on analyzing the operational selection variable (e.g., a composite predictor, as opposed to analyzing some other predictor) for differential prediction, errors in interpretation should be minimal.

Conclusion. Differential prediction analyses should ideally be based on operational data (i.e., data that includes predictor measurement error but is corrected for range restriction and criterion measurement error) to support the most valid inferences about predictive differences in selection systems (SIOP, 2018). However, personnel psychologists often only analyze range-restricted observed data when conducting differential prediction studies. My simulation demonstrated that statistical artifacts and other parameters in the applicant population may cause observed differential prediction analyses to produce misleading results, but the most egregious effects were reserved for simulation conditions that I view as unlikely to occur in practice. A key implication of this study is that biased estimation of differential prediction effects is most likely when the selection variable exhibits extreme validity/slope differences, but there is no evidence that such differences occur in professionally developed selection systems. The simulation

evidence presented here does not offer a basis for refuting decades of prior findings regarding differential prediction (cf. Aguinis et al., 2010). It does, however, offer a motivation for selection professionals to focus on using the operational selection variable as the predictor of interest in differential prediction analyses, as doing so would avoid the complicated issues associated with indirect range restriction and allow for better estimates of differential prediction effects (even from observed data).

General Discussion

I presented a series of four studies to demonstrate important issues regarding differential prediction analyses that have not been explored in the extant research literature. In these studies, I derived simplified formulas for differential prediction effect sizes, presented standard error estimates for d_{Mod_Signed} effect sizes so that confidence intervals can be constructed around estimated values and so that d_{Mod_Signed} can be meta-analyzed, illustrated how overall magnitudes of differential prediction of composite predictors change as a function of which predictors are included in the composite and how they are weighted, and showed that historically documented patterns of subgroup intercept differences on the SAT generalize across schools when artifact-corrected regression coefficients are meta-analyzed (however, slope differences may be more common than prior research indicated). I also presented simulation evidence that demonstrated how criterion measurement error, direct range restriction, and indirect range restriction impact the accuracy and statistical power of differential prediction analyses computed using observed data compared to operational data.

Implications

This research has three key implications for researchers and practitioners interested in studying differential prediction. Regarding my improvements to d_{Mod} effect-size formulas, my algebraic formulas are much simpler to use than Nye and Sackett's (2017) and Dahlke and Sackett's (2018) integration-based formulas and my formulas require input values that are quite easy to obtain; this ease of implementation supports broader usage of these effect sizes. In fact, the simplest formula only calls for a validity

estimate and two *d* values, which would allow d_{Mod_Signed} to be closely approximated via literal back-of-the-envelope calculations. My analytically derived methods for estimating the standard error of d_{Mod_Signed} values mean that it is no longer necessary to rely on bootstrapping procedures to quantify the statistical uncertainty of d_{Mod_Signed} values. Given that all of the inputs to the d_{Mod_Signed} formula have well-established sampling variance estimators and analytically estimable covariances among their sampling distributions, the sampling variance of d_{Mod_Signed} is merely a linear combination of the sampling variances of its component statistics.

Beyond providing a simplified method for computing d_{Mod} effect sizes, my algebraic d_{Mod} formulas revealed an important insight about directions and magnitudes of differences in prediction observed for individual predictors as opposed to those observed for selection systems comprised of multiple predictors. Given that d_{Mod_Signed} is effectively just the difference between the referent-focal mean difference on the criterion and the product of the referent validity coefficient with the referent-focal mean difference on the predictor, the fact that composite predictors tend to have larger validities and larger mean differences than the average of their components means that composite predictors will tend to exhibit differences in prediction that are less extreme than the average difference in prediction of their components. This is a critical observation, as it clearly indicates the inadequacy of testing individual predictors for differential prediction when evaluating the fairness of systems that rely on multiple pieces of information to make selection decisions. Testing for differential prediction in multifaceted selection systems is best done using composites to support operational interpretations of statistical trends (Sackett

et al., 2003; SIOP, 2018). In addition to the fact that testing composites for bias is more statistically sound than including multiple predictor variables in a single Cleary model analysis, composite predictors tend to exhibit smaller standardized differences in prediction than their components and may lead to different conclusions about differential prediction than a multi-predictor model would support. One's choice of predictors and predictor-weighting scheme therefore has predictable impacts on magnitudes of differential prediction. In fact, one can reduce the risk of predictive bias against a minority group by using Pareto-optimal weights that give greater emphasis to adverse-impact reduction than validity maximization; thus, one could view Pareto-optimization as a method for balancing validity concerns against both adverse impact and risk of predictive bias.

My simulation provided insights into the effects of statistical artifacts on operational differential prediction analyses. My results suggested that range-restriction and criterion unreliability artifacts can potentially have biasing effects on the inferences drawn from differential prediction analyses, but the most serious effects that lead to distorted interpretations typically occurred in unlikely scenarios. Specifically, the most extreme effects of artifacts on substantive interpretations tended to occur in settings in which the selection variable exhibited large differences in validity/slopes between groups; there is little evidence that such dramatic differences in relationships occur between groups with any degree of regularity. In most other simulated settings, artifacts did not represent a significant impediment to interpreting differences in prediction. Even so, it is worth noting that methods exist that can help researchers to overcome the effects

of artifacts in their analyses. Specifically, range restriction's effects on analyses can be confronted by implementing modern missing-data approaches in differential prediction analyses. Sophisticated procedures to account for the effects of missing data on statistical results are becoming more accessible to researchers and these procedures hold promise for supporting more accurate conclusions about differential prediction. Methods such as multiple imputation and full-information maximum-likelihood (FIML) estimation (see Newman, 2014 for an overview) can help researchers to correct for range restriction in regression models and therefore compute more accurate tests of the Cleary model. Future development of best practices for applying missing-data methods to differential prediction analyses would be highly valuable.

Conclusion

The methodological advancements and research findings I have reported here support the usage of operational analyses for differential prediction as recommended in SIOP's *Principles* (2018). I derived updated d_{Mod} effect-size methods that will support the quantification of differential prediction in both primary research and meta-analyses and I demonstrated statistical principles related to composites that influence how the magnitude and direction of d_{Mod_Signed} can be anticipated and accounted for in selection-system design. My large-scale analyses of post-secondary admissions data support historic findings regarding overprediction of minority performance and underprediction of female performance, even after accounting for statistical artifacts, and also provided evidence that small slope differences may be more prevalent than previously thought. Additionally, my simulation will help I-O psychologists to better understand the ways in which their

differential prediction findings drawn from observed data may differ from the true state of affairs in their operational applicant populations. It is my hope that the tools, principles, and research findings I have presented here will contribute to the foundations underlying the next generation of research on differential prediction.

References

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

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

Aitken, A. C. (1934). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society (Series 2)*, *4*(2), 106–110. https://doi.org/10.1017/S0013091500008063

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

Anastasi, A. (1968). *Psychological testing*. Macmillan.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology*, *31*(2), 233–241. https://doi.org/10.1111/j.1744-6570.1978.tb00442.x

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of*

Organizational Psychology and Organizational Behavior, *2*(1), 435–463. https://doi.org/10.1146/annurev-orgpsych-032414-111256

Berry, C. M., & Zhao, P. (2015). Addressing criticisms of existing predictive bias research: Cognitive ability test scores still overpredict African Americans' job performance. *Journal of Applied Psychology*, *100*(1), 162–179. https://doi.org/10.1037/a0037615

Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on Black-White mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology*, *66*(1), 91–126. https://doi.org/10.1111/peps.12007

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1). Palo Alto, CA: Consulting Psychologists Press, Inc.

Cheung, M. W.-L., & Chan, W. (2004). Testing dependent correlation coefficients via structural equation modeling. *Organizational Research Methods*, *7*(2), 206–223. https://doi.org/10.1177/1094428104264024

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

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement*, *10*(4), 237–255.

Culpepper, S. A., Aguinis, H., Kern, J. L., & Millsap, R. E. (2019). High-stakes testing case study: A latent variable approach for assessing measurement and prediction invariance. *Psychometrika*. Advance online publication.

Dahlke, J. A., & Sackett, P. R. (2018). Refinements to effect sizes for tests of categorical moderation and differential prediction. *Organizational Research Methods*, *21*(1), 226–234. https://doi.org/10.1177/1094428117736591

Dahlke, J. A., Sackett, P. R., & Kuncel, N. R. (2019). Effects of range restriction and criterion contamination on differential validity of the SAT by race/ethnicity and sex. *Journal of Applied Psychology*. Advance online publication. https://doi.org/10.1037/apl0000382

Dahlke, J. A., & Wiernik, B. M. (2018). psychmeta: An R package for psychometric meta-analysis. *Applied Psychological Measurement*. Advance online publication. https://doi.org/10.1177/0146621618795933

Dahlke, J. A., & Wiernik, B. M. (2019). psychmeta: Psychometric meta-analysis toolkit (Version 2.3.2) [R, R Package]. Retrieved from https://CRAN.R-project.org/package=psychmeta (Original work published 2017)

Darlington, R. B. (1971). Another look at "cultural fairness." *Journal of Educational Measurement*, *8*(2), 71–82. Retrieved from JSTOR.

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology*, *92*(5), 1380–1393. https://doi.org/10.1037/0021-9010.92.5.1380

De Corte, W., Sackett, P. R., & Lievens, F. (2011). Designing Pareto-optimal selection systems: Formalizing the decisions required for selection system development. *Journal of Applied Psychology*, *96*(5), 907–926. https://doi.org/10.1037/a0023298

Druart, C., & De Corte, W. (2012). Designing Pareto-optimal systems for complex selection decisions. *Organizational Research Methods*, *15*(3), 488–513. https://doi.org/10.1177/1094428112440328

Duhachek, A., & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*, *89*(5), 792–808. https://doi.org/10.1037/0021-9010.89.5.792

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*(3), 429.

Fischer, F. T., Schult, J., & Hell, B. (2013). Sex-specific differential prediction of college admission tests: A meta-analysis. *Journal of Educational Psychology*, *105*(2), 478–488. https://doi.org/10.1037/a0031956

Guion, R. M. (1966). Employment tests and discriminatory hiring. *Industrial Relations*, *5*(2), 20–37. https://doi.org/10.1111/j.1468-232X.1966.tb00449.x

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the general aptitude test battery*. https://doi.org/10.17226/1338

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

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*(3), 594–612. https://doi.org/10.1037/0021-9010.91.3.594

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. M. (1977). Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. *Journal of Applied Psychology*, *62*(3), 245–260. https://doi.org/10.1037/0021-9010.62.3.245

Jencks, C. (1998). Racial bias in testing. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (p. 84). Washington, D.C.: Brookings Institution.

Keiser, H. N., Sackett, P. R., Kuncel, N. R., & Brothen, T. (2016). Why women perform better in college than admission scores would predict: Exploring the roles of conscientiousness and course-taking patterns. *Journal of Applied Psychology*, *101*(4), 569–581. https://doi.org/10.1037/apl0000069

Kling, K. C., Noftle, E. E., & Robins, R. W. (2013). Why do standardized tests underpredict women's academic performance? The role of conscientiousness. *Social Psychological and Personality Science*, *4*(5), 600–606. https://doi.org/10.1177/1948550612469038

Lautenschlager, G. J., & Mendoza, J. L. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement*, *10*(2), 133–139. https://doi.org/10.1177/014662168601000202

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh. Section A. Mathematical and Physical Sciences*, *62*(1), 28–30. https://doi.org/10/ckc2

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, *43*(2), 139–161. https://doi.org/10.2307/1169933

Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology*, *63*(4), 507–512.

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

Mattern, K. D., & Patterson, B. F. (2013). Test of slope and intercept bias in college admissions: A response to Aguinis, Culpepper, and Pierce (2010). *Journal of Applied Psychology*, *98*(1), 134–147. https://doi.org/10.1037/a0030610

McKay, P. F., & McDaniel, M. A. (2006). A reexamination of Black-White mean differences in work performance: More data, more moderators. *Journal of Applied Psychology*, *91*(3), 538–554. https://doi.org/10.1037/0021-9010.91.3.538

Meehl, P. E. (1998). *The power of quantitative thinking (Speech delivered upon receipt of the James McKeen Cattrell Fellow award)*. Presented at the American Psychological Society, Washington, DC.

Mulaik, S. A. (2010). *Foundations of factor analysis*. Boca Raton, FL: CRC Press.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*(4), 372–411. https://doi.org/10.1177/1094428114548590

Nye, C. D., & Sackett, P. R. (2017). New effect sizes for tests of categorical moderation and differential prediction. *Organizational Research Methods*, *20*(4). https://doi.org/10.1177/1094428116644505

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (pp. i–41) [College Board Research Report No. 93-1]. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.1994.tb01600.x/abstract

Roth, P. L., Bevier, C. A., Bobko, P., Switzer III, F. S., & Tyler, P. (2001). Ethnic group difference in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*(2), 297–330.

Roth, P. L., Switzer, F. S., Van Iddekinge, C. H., & Oh, I.-S. (2011). Toward better meta-analytic matrices: How input values can affect research conclusions in human resource management simulations. *Personnel Psychology*, *64*(4), 899–935.

Rotundo, M., & Sackett, P. R. (1999). *Effect of rater race on conclusions regarding differential prediction in cognitive ability tests*. *84*(5), 815–822.

Saad, S., & Sackett, P. R. (2002). Investigating differential prediction by gender in employment-oriented personality measures. *Journal of Applied Psychology*, *87*(4), 667–674. https://doi.org/10.1037//0021-9010.87.4.667

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, *63*(4), 215–227. https://doi.org/10.1037/0003-066X.63.4.215

Sackett, P. R., Corte, W. D., & Lievens, F. (2008). Pareto-optimal predictor composite formation: A complementary approach to alleviating the selection quality/adverse impact dilemma. *International Journal of Selection and Assessment*, *16*(3), 206–209. https://doi.org/10.1111/j.1468-2389.2008.00426.x

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*(3), 707–721.

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

Sackett, P. R., Schmitt, N., Kabin, M. B., & Ellingson, J. E. (2001). High-stakes testing in employment, credentialing, and higher education. *American Psychologist*, *56*(4), 302–318. https://doi.org/10.1037/0003-066X.56.4.302

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*(1), 112–118. https://doi.org/10.1037//0021-9010.85.1.112

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). https://doi.org/10/b6mg

Schmidt, F. L., Pearlman, K. L., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology*, *33*(4), 705–724.

Society for Industrial and Organizational Psychology. (2018). Principles for the validation and use of personnel selection procedures. *Industrial and Organizational Psychology*, *11*(S1). https://doi.org/10.1017/iop.2018.195

Song, Q. C. (2018). *Pareto-optimization via normal boundary intersection method in diversity hiring* [R]. Retrieved from https://github.com/Diversity-ParetoOptimal/ParetoR (Original work published 2017)

Song, Q. C., Wee, S., & Newman, D. A. (2017). Diversity shrinkage: Cross-validating pareto-optimal weights to enhance diversity via hiring practices. *Journal of Applied Psychology*, *102*(12), 1636–1657. https://doi.org/10.1037/apl0000240

Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). Sex differences in predictions of college grades from scholastic aptitude test scores. *Journal of Educational Psychology*, *85*(4), 710–718. https://doi.org/10.1037/0022-0663.85.4.710

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, *8*(2), 63–70. Retrieved from JSTOR.

Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology*, *97*(3), 499–530. https://doi.org/10.1037/a0021196

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology*, *48*(4), 865–885.

Weisberg, Y. J., DeYoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the Big Five. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00178

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

Table 1

Correlations Among Regression Coefficients from White-Black Analyses Reported by Aguinis, Culpepper, and Pierce (2016)

	Regression coefficient	1	2	3	4	5	6	7	8	9
1	Intercept									
2	HSGPA	-.66								
3	SAT-CR	.21	-.31							
4	SAT-M	-.02	.09	-.05						
5	SAT-W	-.24	.16	-.53	-.19					
6	Group	-.03	-.03	.10	-.14	.02				
7	Group × HSGPA	.15	-.35	-.03	.02	.14	.04			
8	Group × SAT-CR	.02	.10	-.26	-.05	.25	.18	.06		
9	Group × SAT-M	-.06	-.03	.18	-.22	-.09	.41	-.15	-.27	
10	Group × SAT-W	-.07	.05	.10	.07	-.29	.02	-.32	-.63	.01

Note. HSGPA = high school GPA; SAT-CR = SAT Critical Reading; SAT-M = SAT Mathematics; SAT-W = SAT Writing; Group = dummy variable representing subgroup membership (0 = White, 1 = Black).

Table 2

Parameter Values for Simulation Demonstrating the Convergence of Algebraic and Integration-Based d_{Mod} Formulas

Parameter name	Values
Referent parameters	
X-*Y* correlation	.00, .25, .50
Mean of *X*	0.0, 0.5, 1.0
Mean of *Y*	0.0, 0.4, 0.8
SD of *X*	1, 5
SD of *Y*	1, 5
Focal parameters	
X-*Y* correlation	-.50, -.25, .00, .25, .50
Mean of *X*	-1.0, -0.5, 0.0, 0.5, 1.0
Mean of *Y*	-0.8, -0.4, 0.0, 0.4, 0.8
SD of *X*	1, 5
SD of *Y*	1, 5

Table 3

Parameter Values for Simulation Demonstrating the Convergence of Algebraic and Monte Carlo Standard Errors of $d_{Mod\ Signed}$

Parameter name	Values
Sample size	100, 500, 1000
Proportion of referent group in sample	.5, .7, .9
Referent-group *X*-*Y* correlation	.2, .5
Ratio of focal-group and referent-group *X*-*Y* correlations	0.5, 1.0
Standardized referent-focal mean difference on *X*	0.5, 1.0
Standardized referent-focal mean difference on *Y*	.3, .6
Reliability of *X*	.9
Reliability of *Y*	.6, .8

Table 4

Meta-Analytic Correlation Matrix and White-Black Mean Differences from Song et al. (2017) with Measurement-Error Corrected Criterion d *Value*

	Variable	White-Black *d*	Correlation				
			1	2	3	4	5
1	Biodata	.39					
2	Conscientiousness	-.09	.51				
3	General mental ability	.72	.37	.03			
4	Integrity	.04	.25	.34	.02		
5	Structured interview	.39	.16	.13	.31	-.02	
6	Job performance	.46	.32	.22	.52	.20	.48

Note.

Roth, Switzer, Van Iddekinge, and Oh (2011) provided the predictor intercorrelations and validities for biodata, cognitive ability, conscientiousness, and structured interviews.

Van Iddekinge, Roth, Raymark, and Odle-Dusseau (2012) provided the integrity validity estimate, which Song et al. (2017) corrected for range restriction.

Bobko and Roth (2013) provided predictor *d* values.

McKay & McDaniel (2006) provided the measurement-error corrected *d* value for job performance.

Table 5

Unit- and Regression-Weighted Predictor Sets from Meta-Analytic Data in Table 4

	Unit-Weighted Composites				Regression-Weighted Composites			
Predictors	r_{XY}	r_{XY_WG}	d_X	d_{Mod}	r_{XY}	r_{XY_WG}	d_X	d_{Mod}
Individual predictors *								
Biodata (BD)	.320	.305	.390	.341	.320	.305	.390	.341
Conscientiousness (C)	.220	.218	-.090	.480	.220	.218	-.090	.480
General mental ability (GMA)	.520	.502	.720	.099	.520	.502	.720	.099
Integrity (I)	.200	.200	.040	.452	.200	.200	.040	.452
Structured interview (SI)	.480	.468	.390	.277	.480	.468	.390	.277
Mean	*.348*	*.339*	*.290*	*.330*	*.348*	*.339*	*.290*	*.330*
Two-predictor composites								
BD + C	.311	.305	.171	.408	.327	.315	.313	.362
BD + GMA	.507	.489	.673	.130	.538	.520	.733	.078
BD + I	.329	.319	.271	.374	.343	.330	.342	.347
BD + SI	.525	.511	.516	.196	.540	.526	.502	.196
C + GMA	.516	.503	.428	.244	.559	.544	.623	.121
C + I	.257	.258	-.031	.468	.257	.258	-.038	.470
C + SI	.466	.461	.198	.369	.506	.497	.324	.299
GMA + I	.504	.489	.525	.203	.554	.537	.684	.093
GMA + SI	.618	.604	.689	.043	.619	.605	.704	.034
I + SI	.486	.477	.306	.314	.524	.513	.376	.267
Mean	*.452*	*.442*	*.375*	*.275*	*.476*	*.465*	*.456*	*.227*
Three-predictor composites								
BD + C + GMA	.483	.469	.459	.245	.560	.544	.637	.113
BD + C + I	.325	.321	.148	.413	.345	.334	.305	.358
BD + C + SI	.476	.466	.320	.311	.541	.528	.472	.211
BD + GMA + I	.503	.487	.554	.190	.561	.545	.698	.080
BD + GMA + SI	.610	.596	.700	.042	.630	.617	.718	.017
BD + I + SI	.514	.502	.423	.248	.560	.548	.472	.201
C + GMA + I	.483	.474	.335	.301	.573	.559	.623	.112
C + GMA + SI	.615	.603	.509	.153	.640	.628	.642	.057
C + I + SI	.456	.453	.171	.383	.532	.523	.341	.282
GMA + I + SI	.631	.618	.604	.086	.650	.638	.679	.027
Mean	*.509*	*.499*	*.422*	*.237*	*.559*	*.546*	*.559*	*.146*
Four-predictor composites								
BD + C + GMA + I	.475	.463	.393	.278	.574	.559	.631	.107
BD + C + GMA + SI	.581	.568	.530	.159	.641	.629	.657	.047
BD + C + I + SI	.470	.462	.280	.331	.560	.548	.470	.202
BD + GMA + I + SI	.611	.598	.622	.088	.654	.642	.689	.018
C + GMA + I + SI	.599	.589	.441	.200	.658	.646	.645	.043
Mean	*.547*	*.536*	*.453*	*.211*	*.617*	*.605*	*.618*	*.084*
Five-predictor composite								
BD + C + GMA + I + SI	.574	.562	.475	.193	.658	.647	.653	.038

Note. r_{XY} = overall validity for the combined population of Black and White applicants. r_{XY_WG} = average within-group (i.e., pooled) validity across subgroups computed by partialling between-group predictor and criterion variance out of the overall validity estimate. d_X = standardized White-Black mean difference on predictor. d_{Mod} = d_{Mod_Signed} standardized mean difference in predicted performance between groups ($d_{Mod} = d_Y - d_X \times r_{XY\_WG}$). Estimates are based on a population consisting of 85% White individuals and 15% Black individuals. In all analyses, the mean difference for performance was .46.

* Results of individual predictors are shown in both unit- and regression-weighted solutions for ease of interpretation even though these are not truly composites.

Table 6
Pareto-Optimal Composite Solutions from Data in Table 4

Pareto solution #	Predictor weight								
	Biodata	Conscientiousness	GMA	Integrity	Structured interview	r_{XY}	r_{XY_WG}	d_X	d_{Mod}
1*	0.000	1.000	0.000	0.000	0.000	0.220	0.218	-0.090	0.480
2	0.000	0.842	0.000	0.146	0.012	0.244	0.243	-0.072	0.478
3	0.000	0.796	0.000	0.152	0.053	0.266	0.267	-0.052	0.474
4	0.000	0.751	0.000	0.157	0.091	0.289	0.291	-0.031	0.469
5	0.000	0.709	0.000	0.163	0.128	0.311	0.315	-0.010	0.463
6	0.000	0.668	0.000	0.168	0.164	0.334	0.338	0.013	0.456
7	0.000	0.627	0.000	0.173	0.200	0.356	0.359	0.037	0.447
8	0.000	0.587	0.000	0.178	0.235	0.379	0.381	0.062	0.437
9	0.000	0.546	0.000	0.183	0.270	0.402	0.402	0.088	0.425
10	0.000	0.505	0.000	0.189	0.307	0.424	0.423	0.116	0.411
11	0.000	0.461	0.000	0.194	0.345	0.447	0.445	0.146	0.395
12	0.000	0.415	0.000	0.200	0.385	0.469	0.466	0.177	0.377
13	0.000	0.383	0.019	0.200	0.398	0.490	0.486	0.213	0.357
14	0.000	0.362	0.048	0.197	0.393	0.512	0.506	0.248	0.334
15	0.000	0.339	0.079	0.194	0.388	0.535	0.528	0.287	0.308
16	0.000	0.315	0.112	0.190	0.383	0.557	0.549	0.330	0.279
17	0.000	0.289	0.147	0.186	0.377	0.580	0.571	0.375	0.246
18	0.000	0.261	0.187	0.182	0.371	0.602	0.592	0.425	0.208
19	0.000	0.227	0.233	0.177	0.363	0.624	0.613	0.481	0.165
20	0.000	0.183	0.294	0.170	0.353	0.645	0.634	0.549	0.112
21	0.030	0.091	0.387	0.156	0.336	0.658	0.646	0.653	0.038

Note.
GMA = general mental ability. r_{XY} = overall validity for the combined population of Black and White applicants. r_{XY_WG} = average within-group (i.e., pooled) validity across subgroups computed by partialling between-group predictor and criterion variance out of the overall validity estimate. d_X = standardized White-Black mean difference on predictor. $d_{Mod} = d_{Mod\_Signed}$ standardized mean differences in predicted performance between groups ($d_{Mod} = d_Y - d_X \times r_{XY\_WG}$). Estimates are based on a population consisting of 80% White individuals and 20% Black individuals. In all analyses, the mean difference for performance was .38. Solution 21 is the regression-weighted solution and is identical to the five-predictor regression composite in Table 5.

Table 7
Meta-Analyses of Internal-Consistency Reliabilities for First-Year Grades by Group

Group	*k*	*N*	$\overline{r_{YY'}}$	$SD_{r_{YY'}}$	SD_{res}	95% CI	80% CV
White (in White-Black contrasts)	236	875,294	.85	.03	.03	(.85, .85)	(.81, .89)
White (in White-Hispanic contrasts)	240	875,296	.85	.03	.03	(.85, .85)	(.81, .89)
Black	236	100,362	.82	.05	.05	(.81, .82)	(.75, .88)
Hispanic	240	123,395	.82	.05	.05	(.81, .82)	(.76, .88)
Male	266	600,314	.85	.03	.03	(.85, .86)	(.82, .89)
Female	266	710,776	.84	.03	.03	(.83, .84)	(.80, .88)

Note: k = number of studies contributing to meta-analysis; N = total sample size; $\overline{r_{YY'}}$ = mean observed effect size ($r_{YY'}$); $SD_{r_{YY'}}$ = observed standard deviation of $r_{YY'}$; SD_{res} = residual standard deviation of $r_{YY'}$; $\sigma^2_{r_{YY'}}$ = observed variance of $r_{YY'}$; CI = confidence interval around $\overline{r_{YY'}}$; CV = credibility interval around $\overline{r_{YY'}}$.

Table 8

Standardized Weights Assigned to Post-Secondary Academic Performance Predictors in Composite Calculations

	Component variable			
Composite name	SAT-CR	SAT-M	SAT-W	HSGPA
SAT Composite	1/2	1/2	0	0
SAT Composite w/ Writing	1/3	1/3	1/3	0
HSGPA + SAT Composite.	1/4	1/4	0	1/2
HSGPA + SAT Composite w/ Writing	1/6	1/6	1/6	1/2

Note. SAT-CR = SAT Critical Reading test. SAT-M = SAT Mathematics test. SAT-W = SAT Writing test. HSGPA = self-reported high school GPA.

Table 9

Meta-Analyses of Observed d_{Mod_Signed} Effect Sizes

Referent	Focal	*k*	*N*	Predictor	$\overline{d_{Mod}}$	$SD_{d_{Mod}}$	SD_{res}	95% CI	80% CV
White	Black	236	975,966	HSGPA	.44	.17	.16	(.42, .46)	(.23, .64)
				SAT Mathematics	.38	.16	.14	(.36, .40)	(.19, .57)
				SAT Critical Reading	.40	.16	.15	(.38, .42)	(.21, .59)
				SAT Writing	.36	.15	.14	(.34, .37)	(.18, .53)
				SAT Comp.	.32	.14	.12	(.30, .34)	(.16, .47)
				SAT Comp. w/ Writing	.28	.13	.11	(.27, .30)	(.14, .43)
				HSGPA + SAT Comp.	.24	.13	.11	(.23, .26)	(.09, .39)
				HSGPA + SAT Comp. w/ Writing	.23	.13	.11	(.21, .25)	(.09, .37)
White	Hispanic	240	999,018	HSGPA	.30	.18	.17	(.28, .33)	(.08, .53)
				SAT Mathematics	.24	.14	.13	(.22, .26)	(.07, .41)
				SAT Critical Reading	.23	.14	.12	(.22, .25)	(.07, .39)
				SAT Writing	.21	.13	.12	(.19, .22)	(.06, .36)
				SAT Comp.	.18	.12	.10	(.16, .19)	(.05, .31)
				SAT Comp. w/ Writing	.15	.11	.09	(.14, .17)	(.04, .27)
				HSGPA + SAT Comp.	.16	.12	.11	(.14, .17)	(.02, .30)
				HSGPA + SAT Comp. w/ Writing	.15	.12	.11	(.13, .16)	(.01, .29)
Male	Female	266	1,311,531	HSGPA	−.18	.08	.07	(−.19, −.17)	(−.27, −.08)
				SAT Mathematics	−.39	.09	.08	(−.40, −.38)	(−.50, −.28)
				SAT Critical Reading	−.28	.10	.09	(−.29, −.27)	(−.40, −.16)
				SAT Writing	−.22	.09	.09	(−.23, −.21)	(−.33, −.11)
				SAT Comp.	−.36	.09	.08	(−.37, −.35)	(−.46, −.26)
				SAT Comp. w/ Writing	−.31	.09	.08	(−.32, −.30)	(−.42, −.20)
				HSGPA + SAT Comp.	−.27	.07	.07	(−.28, −.26)	(−.36, −.19)
				HSGPA + SAT Comp. w/ Writing	−.23	.08	.07	(−.24, −.22)	(−.32, −.14)

Note. k = number of studies contributing to meta-analysis; N = total sample size; $\overline{d_{Mod}}$ = mean observed effect size (d_{Mod}); $SD_{d_{Mod}}$ = observed standard deviation of d_{Mod}; SD_{res} = residual standard deviation of d_{Mod}; CI = confidence interval around $\overline{d_{Mod}}$; CV = credibility interval around $\overline{d_{Mod}}$.

Table 10

Meta-Analyses of d_{Mod_Signed} Effect Sizes Corrected for Criterion Unreliability

Referent	Focal	*k*	*N*	Predictor	$\overline{d_{Mod}}$	$SD_{d_{Mod}}$	SD_{res}	95% CI	80% CV
White	Black	236	975,966	HSGPA	.47	.19	.18	(.45, .50)	(.24, .71)
				SAT Mathematics	.41	.17	.16	(.39, .43)	(.20, .62)
				SAT Critical Reading	.43	.17	.17	(.41, .46)	(.22, .65)
				SAT Writing	.39	.17	.16	(.36, .41)	(.18, .59)
				SAT Comp.	.34	.15	.14	(.32, .36)	(.17, .52)
				SAT Comp. w/ Writing	.31	.14	.13	(.29, .32)	(.14, .47)
				HSGPA + SAT Comp.	.26	.14	.13	(.24, .28)	(.09, .43)
				HSGPA + SAT Comp. w/ Writing	.25	.14	.13	(.23, .27)	(.09, .41)
White	Hispanic	240	999,018	HSGPA	.33	.20	.19	(.31, .36)	(.08, .58)
				SAT Mathematics	.26	.16	.15	(.24, .28)	(.07, .45)
				SAT Critical Reading	.26	.15	.14	(.24, .28)	(.08, .44)
				SAT Writing	.23	.14	.13	(.21, .24)	(.06, .39)
				SAT Comp.	.19	.13	.11	(.18, .21)	(.05, .34)
				SAT Comp. w/ Writing	.17	.12	.10	(.15, .18)	(.03, .30)
				HSGPA + SAT Comp.	.17	.13	.12	(.16, .19)	(.01, .33)
				HSGPA + SAT Comp. w/ Writing	.16	.13	.12	(.15, .18)	(.01, .32)
Male	Female	266	1,311,531	HSGPA	−.19	.09	.08	(−.20, −.18)	(−.30, −.09)
				SAT Mathematics	−.42	.09	.09	(−.43, −.41)	(−.54, −.31)
				SAT Critical Reading	−.30	.10	.10	(−.32, −.29)	(−.43, −.18)
				SAT Writing	−.24	.10	.09	(−.25, −.22)	(−.36, −.11)
				SAT Comp.	−.39	.09	.09	(−.40, −.38)	(−.50, −.28)
				SAT Comp. w/ Writing	−.34	.09	.09	(−.35, −.32)	(−.45, −.22)
				HSGPA + SAT Comp.	−.29	.08	.07	(−.30, −.29)	(−.39, −.20)
				HSGPA + SAT Comp. w/ Writing	−.25	.08	.08	(−.26, −.24)	(−.35, −.14)

Note. k = number of studies contributing to meta-analysis; N = total sample size; $\overline{d_{Mod}}$ = mean observed effect size (d_{Mod}); $SD_{d_{Mod}}$ = observed standard deviation of d_{Mod}; SD_{res} = residual standard deviation of d_{Mod}; CI = confidence interval around $\overline{d_{Mod}}$; CV = credibility interval around $\overline{d_{Mod}}$.

Table 11

Meta-Analyses of d_{Mod_Signed} Effect Sizes Corrected for Range Restriction

Referent	Focal	*k*	*N*	Predictor	$\overline{d_{Mod}}$	$SD_{d_{Mod}}$	SD_{res}	95% CI	80% CV
White	Black	236	975,966	HSGPA	.42	.17	.16	(.40, .44)	(.22, .62)
				SAT Mathematics	.35	.16	.15	(.33, .37)	(.16, .54)
				SAT Critical Reading	.39	.15	.14	(.37, .41)	(.21, .57)
				SAT Writing	.33	.15	.14	(.32, .35)	(.15, .52)
				SAT Comp.	.28	.15	.13	(.27, .30)	(.11, .45)
				SAT Comp. w/ Writing	.24	.15	.13	(.22, .26)	(.07, .42)
				HSGPA + SAT Comp.	.19	.14	.13	(.18, .21)	(.03, .36)
				HSGPA + SAT Comp. w/ Writing	.18	.14	.13	(.17, .20)	(.02, .35)
White	Hispanic	240	999,018	HSGPA	.31	.17	.17	(.29, .34)	(.10, .53)
				SAT Mathematics	.21	.14	.13	(.19, .23)	(.04, .38)
				SAT Critical Reading	.22	.13	.12	(.20, .23)	(.06, .37)
				SAT Writing	.19	.13	.12	(.17, .21)	(.04, .34)
				SAT Comp.	.15	.12	.10	(.14, .17)	(.02, .29)
				SAT Comp. w/ Writing	.13	.11	.10	(.11, .14)	(.00, .26)
				HSGPA + SAT Comp.	.14	.13	.12	(.12, .16)	(−.01, .29)
				HSGPA + SAT Comp. w/ Writing	.13	.12	.11	(.12, .15)	(−.01, .28)
Male	Female	266	1,311,531	HSGPA	−.16	.08	.07	(−.17, −.15)	(−.26, −.07)
				SAT Mathematics	−.41	.08	.08	(−.42, −.40)	(−.51, −.31)
				SAT Critical Reading	−.29	.09	.09	(−.30, −.28)	(−.40, −.18)
				SAT Writing	−.21	.09	.08	(−.22, −.20)	(−.32, −.10)
				SAT Comp.	−.36	.08	.08	(−.37, −.35)	(−.46, −.26)
				SAT Comp. w/ Writing	−.31	.08	.08	(−.32, −.30)	(−.41, −.21)
				HSGPA + SAT Comp.	−.25	.07	.07	(−.26, −.24)	(−.34, −.16)
				HSGPA + SAT Comp. w/ Writing	−.21	.08	.08	(−.22, −.20)	(−.31, −.10)

Note. k = number of studies contributing to meta-analysis; N = total sample size; $\overline{d_{Mod}}$ = mean observed effect size (d_{Mod}); $SD_{d_{Mod}}$ = observed standard deviation of d_{Mod}; SD_{res} = residual standard deviation of d_{Mod}; CI = confidence interval around $\overline{d_{Mod}}$; CV = credibility interval around $\overline{d_{Mod}}$.

Table 12

Meta-Analyses of d_{Mod_Signed} Effect Sizes Corrected for Range Restriction and Criterion Unreliability

Referent	Focal	*k*	*N*	Predictor	$\overline{d_{Mod}}$	$SD_{d_{Mod}}$	SD_{res}	95% CI	80% CV
White	Black	236	975,966	HSGPA	.45	.18	.17	(.43, .48)	(.23, .67)
				SAT Mathematics	.38	.17	.16	(.36, .40)	(.17, .59)
				SAT Critical Reading	.42	.17	.16	(.40, .44)	(.22, .62)
				SAT Writing	.36	.17	.16	(.34, .38)	(.16, .56)
				SAT Comp.	.31	.16	.15	(.29, .33)	(.12, .49)
				SAT Comp. w/ Writing	.26	.16	.15	(.24, .28)	(.07, .45)
				HSGPA + SAT Comp.	.21	.15	.14	(.19, .23)	(.02, .39)
				HSGPA + SAT Comp. w/ Writing	.20	.15	.14	(.18, .22)	(.02, .38)
White	Hispanic	240	999,018	HSGPA	.34	.19	.18	(.31, .36)	(.10, .57)
				SAT Mathematics	.23	.16	.15	(.21, .25)	(.04, .42)
				SAT Critical Reading	.23	.14	.13	(.21, .25)	(.06, .40)
				SAT Writing	.21	.14	.13	(.19, .22)	(.04, .37)
				SAT Comp.	.17	.13	.12	(.15, .18)	(.02, .31)
				SAT Comp. w/ Writing	.14	.12	.11	(.12, .16)	(−.00, .28)
				HSGPA + SAT Comp.	.15	.14	.13	(.13, .17)	(−.01, .32)
				HSGPA + SAT Comp. w/ Writing	.14	.14	.13	(.13, .16)	(−.02, .31)
Male	Female	266	1,311,531	HSGPA	−.17	.08	.08	(−.18, −.16)	(−.28, −.07)
				SAT Mathematics	−.43	.09	.08	(−.45, −.42)	(−.54, −.33)
				SAT Critical Reading	−.31	.10	.09	(−.32, −.29)	(−.42, −.19)
				SAT Writing	−.23	.09	.09	(−.24, −.22)	(−.34, −.11)
				SAT Comp.	−.39	.09	.08	(−.40, −.38)	(−.49, −.28)
				SAT Comp. w/ Writing	−.33	.09	.08	(−.34, −.32)	(−.44, −.22)
				HSGPA + SAT Comp.	−.27	.08	.08	(−.28, −.26)	(−.37, −.17)
				HSGPA + SAT Comp. w/ Writing	−.22	.09	.09	(−.23, −.21)	(−.33, −.11)

Note. k = number of studies contributing to meta-analysis; N = total sample size; $\overline{d_{Mod}}$ = mean observed effect size (d_{Mod}); $SD_{d_{Mod}}$ = observed standard deviation of d_{Mod}; SD_{res} = residual standard deviation of d_{Mod}; CI = confidence interval around $\overline{d_{Mod}}$; CV = credibility interval around $\overline{d_{Mod}}$.

Table 13

Meta-Analytic Means of Referent-Group Validities and Referent-Focal d Values Corresponding to d_{Mod_Signed} Computations

		Referent Validity of Predictor				Referent-Focal *d* Value				Product of Validity and *d* Value			
Contrast	Variable	Obs.	Crit.	MVRR	Crit. + MVRR	Obs.	Crit.	MVRR	Crit. + MVRR	Obs.	Crit.	MVRR	Crit. + MVRR
W-B	First-Year College GPA	---	---	---	---	0.62	0.67	0.72	0.78	---	---	---	---
	HSGPA	.37	.41	.45	.49	0.42	---	0.65	---	.16	.17	.29	.32
	SAT Mathematics	.24	.26	.33	.36	0.98	---	1.15	---	.24	.25	.38	.41
	SAT Critical Reading	.27	.29	.35	.37	0.79	---	1.00	---	.21	.23	.35	.37
	SAT Writing	.32	.35	.40	.43	0.82	---	1.01	---	.26	.29	.40	.43
	SAT Comp.	.29	.32	.38	.41	1.02	---	1.20	---	.30	.33	.46	.49
	SAT Comp. w/ Writing	.33	.36	.41	.44	1.03	---	1.20	---	.34	.37	.49	.53
	HSGPA+SAT Comp.	.42	.46	.50	.54	0.88	---	1.09	---	.37	.40	.55	.59
	HSGPA+SAT Comp. w/Writing	.44	.48	.52	.55	0.87	---	1.08	---	.38	.42	.56	.59
W-H	First-Year College GPA	---	---	---	---	0.39	0.42	0.36	0.39	---	---	---	---
	HSGPA	.37	.41	.45	.49	0.18	---	0.29	---	.07	.07	.13	.14
	SAT Mathematics	.24	.26	.33	.36	0.60	---	0.73	---	.14	.16	.24	.26
	SAT Critical Reading	.27	.29	.35	.38	0.56	---	0.70	---	.15	.16	.25	.27
	SAT Writing	.32	.35	.40	.43	0.57	---	0.69	---	.18	.20	.28	.30
	SAT Comp.	.29	.32	.38	.41	0.67	---	0.80	---	.19	.21	.30	.33
	SAT Comp. w/ Writing	.33	.36	.41	.44	0.69	---	0.81	---	.23	.25	.33	.36
	HSGPA+SAT Comp.	.42	.46	.50	.54	0.53	---	0.64	---	.22	.24	.32	.35
	HSGPA+SAT Comp. w/Writing	.44	.48	.52	.56	0.53	---	0.64	---	.23	.25	.33	.36
M-F	First-Year College GPA	---	---	---	---	−0.28	−0.31	−0.29	−0.31	---	---	---	---
	HSGPA	.35	.38	.45	.48	−0.24	---	−0.25	---	−.08	−.09	−.11	−.12
	SAT Mathematics	.28	.31	.39	.42	0.47	---	0.35	---	.13	.15	.14	.15
	SAT Critical Reading	.26	.28	.36	.39	0.08	---	0.04	---	.02	.02	.01	.02
	SAT Writing	.30	.33	.40	.43	−0.14	---	−0.14	---	−.04	−.05	−.06	−.06
	SAT Comp.	.31	.34	.42	.45	0.32	---	0.21	---	.10	.11	.09	.09
	SAT Comp. w/ Writing	.33	.36	.43	.46	0.16	---	0.09	---	.05	.06	.04	.04
	HSGPA+SAT Comp.	.42	.45	.51	.55	0.03	---	−0.03	---	.01	.01	−.02	−.02
	HSGPA+SAT Comp. w/Writing	.43	.47	.52	.56	−0.07	---	−0.10	---	−.03	−.03	−.05	−.06

Note. W-B = White-Black contrast in which Whites are the referent group; W-H = White-Hispanic contrast in which Whites are the referent group; M-F = Male-Female contrast in which males are the referent group; Obs. = observed effects in measured and range-restricted data; Crit. =

range-restricted estimates corrected for criterion measurement error; MVRR = unrestricted or multivariate range-restriction-corrected estimates that include criterion measurement error; Crit. + MVRR = estimates corrected for both range restriction and criterion measurement error.

Table 14
Summary of Differences in Prediction Detected in Observed Data

			Type of Difference			Breakdown of Intercept Differences		Breakdown of Slope Differences	
Referent	Focal	Predictor	None	Intercept	Slope	Lower Focal Intercept	Higher Focal Intercept	Flatter Focal Slope	Steeper Focal Slope
White	Black	HSGPA	22 (9.3)	115 (48.7)	99 (41.9)	115 (100.0)	0 (0.0)	99 (100.0)	0 (0.0)
		SAT Mathematics	41 (17.4)	151 (64.0)	44 (18.6)	151 (100.0)	0 (0.0)	36 (81.8)	8 (18.2)
		SAT Critical Reading	32 (13.6)	156 (66.1)	48 (20.3)	156 (100.0)	0 (0.0)	35 (72.9)	13 (27.1)
		SAT Writing	41 (17.4)	149 (63.1)	46 (19.5)	149 (100.0)	0 (0.0)	37 (80.4)	9 (19.6)
		SAT Comp.	48 (20.3)	142 (60.2)	46 (19.5)	142 (100.0)	0 (0.0)	39 (84.8)	7 (15.2)
		SAT Comp. w/ Writing	54 (22.9)	131 (55.5)	51 (21.6)	131 (100.0)	0 (0.0)	43 (84.3)	8 (15.7)
		HSGPA+SAT Comp.	61 (25.8)	106 (44.9)	69 (29.2)	105 (99.1)	1 (0.9)	64 (92.8)	5 (7.2)
		HSGPA+SAT Comp. w/ Writing	67 (28.4)	94 (39.8)	75 (31.8)	93 (98.9)	1 (1.1)	70 (93.3)	5 (6.7)
White	Hispanic	HSGPA	69 (28.7)	110 (45.8)	61 (25.4)	110 (100.0)	0 (0.0)	59 (96.7)	2 (3.3)
		SAT Mathematics	94 (39.2)	113 (47.1)	33 (13.8)	113 (100.0)	0 (0.0)	20 (60.6)	13 (39.4)
		SAT Critical Reading	99 (41.2)	108 (45.0)	33 (13.8)	108 (100.0)	0 (0.0)	26 (78.8)	7 (21.2)
		SAT Writing	110 (45.8)	97 (40.4)	33 (13.8)	97 (100.0)	0 (0.0)	26 (78.8)	7 (21.2)
		SAT Comp.	114 (47.5)	83 (34.6)	43 (17.9)	83 (100.0)	0 (0.0)	32 (74.4)	11 (25.6)
		SAT Comp. w/ Writing	125 (52.1)	75 (31.2)	40 (16.7)	75 (100.0)	0 (0.0)	32 (80.0)	8 (20.0)
		HSGPA+SAT Comp.	119 (49.6)	77 (32.1)	44 (18.3)	77 (100.0)	0 (0.0)	39 (88.6)	5 (11.4)
		HSGPA+SAT Comp. w/ Writing	123 (51.2)	77 (32.1)	40 (16.7)	77 (100.0)	0 (0.0)	36 (90.0)	4 (10.0)
Male	Female	HSGPA	56 (21.1)	154 (57.9)	56 (21.1)	0 (0.0)	154 (100.0)	40 (71.4)	16 (28.6)
		SAT Mathematics	9 (3.4)	168 (63.2)	89 (33.5)	0 (0.0)	168 (100.0)	17 (19.1)	72 (80.9)
		SAT Critical Reading	29 (10.9)	162 (60.9)	75 (28.2)	0 (0.0)	162 (100.0)	6 (8.0)	69 (92.0)
		SAT Writing	47 (17.7)	167 (62.8)	52 (19.5)	0 (0.0)	167 (100.0)	10 (19.2)	42 (80.8)
		SAT Comp.	16 (6.0)	159 (59.8)	91 (34.2)	0 (0.0)	159 (100.0)	14 (15.4)	77 (84.6)
		SAT Comp. w/ Writing	20 (7.5)	172 (64.7)	74 (27.8)	0 (0.0)	172 (100.0)	10 (13.5)	64 (86.5)
		HSGPA+SAT Comp.	32 (12.0)	173 (65.0)	61 (22.9)	0 (0.0)	173 (100.0)	33 (54.1)	28 (45.9)
		HSGPA+SAT Comp. w/ Writing	40 (15.0)	169 (63.5)	57 (21.4)	0 (0.0)	169 (100.0)	34 (59.6)	23 (40.4)

Note. Values in parentheses are percentages. Percentages reported for breakdowns of intercept and slope differences are condition on the detection of a significant difference; these percentages represent rates of directions of subgroup differences in prediction among the significant differences.

Table 15
Summary of Differences in Prediction Detected in Data Corrected for Criterion Unreliability

			Type of Difference			Breakdown of Intercept Differences		Breakdown of Slope Differences	
Referent	Focal	Predictor	None	Intercept	Slope	Lower Focal Intercept	Higher Focal Intercept	Flatter Focal Slope	Steeper Focal Slope
White	Black	HSGPA	20 (8.5)	113 (47.9)	103 (43.6)	113 (100.0)	0 (0.0)	103 (100.0)	0 (0.0)
		SAT Mathematics	37 (15.7)	150 (63.6)	49 (20.8)	150 (100.0)	0 (0.0)	41 (83.7)	8 (16.3)
		SAT Critical Reading	32 (13.6)	156 (66.1)	48 (20.3)	156 (100.0)	0 (0.0)	35 (72.9)	13 (27.1)
		SAT Writing	40 (16.9)	148 (62.7)	48 (20.3)	148 (100.0)	0 (0.0)	38 (79.2)	10 (20.8)
		SAT Comp.	46 (19.5)	142 (60.2)	48 (20.3)	142 (100.0)	0 (0.0)	39 (81.2)	9 (18.8)
		SAT Comp. w/ Writing	50 (21.2)	131 (55.5)	55 (23.3)	131 (100.0)	0 (0.0)	46 (83.6)	9 (16.4)
		HSGPA+SAT Comp.	60 (25.4)	106 (44.9)	70 (29.7)	105 (99.1)	1 (0.9)	66 (94.3)	4 (5.7)
		HSGPA+SAT Comp. w/ Writing	62 (26.3)	97 (41.1)	77 (32.6)	96 (99.0)	1 (1.0)	73 (94.8)	4 (5.2)
White	Hispanic	HSGPA	64 (26.7)	110 (45.8)	66 (27.5)	110 (100.0)	0 (0.0)	61 (92.4)	5 (7.6)
		SAT Mathematics	93 (38.8)	111 (46.2)	36 (15.0)	111 (100.0)	0 (0.0)	23 (63.9)	13 (36.1)
		SAT Critical Reading	94 (39.2)	108 (45.0)	38 (15.8)	108 (100.0)	0 (0.0)	30 (78.9)	8 (21.1)
		SAT Writing	106 (44.2)	96 (40.0)	38 (15.8)	96 (100.0)	0 (0.0)	29 (76.3)	9 (23.7)
		SAT Comp.	109 (45.4)	80 (33.3)	51 (21.2)	80 (100.0)	0 (0.0)	38 (74.5)	13 (25.5)
		SAT Comp. w/ Writing	116 (48.3)	77 (32.1)	47 (19.6)	77 (100.0)	0 (0.0)	38 (80.9)	9 (19.1)
		HSGPA+SAT Comp.	116 (48.3)	78 (32.5)	46 (19.2)	78 (100.0)	0 (0.0)	41 (89.1)	5 (10.9)
		HSGPA+SAT Comp. w/ Writing	120 (50.0)	76 (31.7)	44 (18.3)	76 (100.0)	0 (0.0)	39 (88.6)	5 (11.4)
Male	Female	HSGPA	54 (20.3)	153 (57.5)	59 (22.2)	0 (0.0)	153 (100.0)	42 (71.2)	17 (28.8)
		SAT Mathematics	9 (3.4)	163 (61.3)	94 (35.3)	0 (0.0)	163 (100.0)	20 (21.3)	74 (78.7)
		SAT Critical Reading	28 (10.5)	162 (60.9)	76 (28.6)	0 (0.0)	162 (100.0)	7 (9.2)	69 (90.8)
		SAT Writing	42 (15.8)	168 (63.2)	56 (21.1)	0 (0.0)	168 (100.0)	11 (19.6)	45 (80.4)
		SAT Comp.	13 (4.9)	158 (59.4)	95 (35.7)	0 (0.0)	158 (100.0)	14 (14.7)	81 (85.3)
		SAT Comp. w/ Writing	18 (6.8)	169 (63.5)	79 (29.7)	0 (0.0)	169 (100.0)	10 (12.7)	69 (87.3)
		HSGPA+SAT Comp.	28 (10.5)	173 (65.0)	65 (24.4)	0 (0.0)	173 (100.0)	36 (55.4)	29 (44.6)
		HSGPA+SAT Comp. w/ Writing	37 (13.9)	165 (62.0)	64 (24.1)	0 (0.0)	165 (100.0)	37 (57.8)	27 (42.2)

Note. Values in parentheses are percentages. Percentages reported for breakdowns of intercept and slope differences are condition on the detection of a significant difference; these percentages represent rates of directions of subgroup differences in prediction among the significant differences.

Table 16
Summary of Differences in Prediction Detected in Data Corrected for Range Restriction

			Type of Difference			Breakdown of Intercept Differences		Breakdown of Slope Differences	
Referent	Focal	Predictor	None	Intercept	Slope	Lower Focal Intercept	Higher Focal Intercept	Flatter Focal Slope	Steeper Focal Slope
White	Black	HSGPA	34 (14.4)	114 (48.3)	88 (37.3)	114 (100.0)	0 (0.0)	87 (98.9)	1 (1.1)
		SAT Mathematics	69 (29.2)	135 (57.2)	32 (13.6)	135 (100.0)	0 (0.0)	22 (68.8)	10 (31.2)
		SAT Critical Reading	59 (25.0)	145 (61.4)	32 (13.6)	145 (100.0)	0 (0.0)	20 (62.5)	12 (37.5)
		SAT Writing	59 (25.0)	137 (58.1)	40 (16.9)	137 (100.0)	0 (0.0)	24 (60.0)	16 (40.0)
		SAT Comp.	73 (30.9)	120 (50.8)	43 (18.2)	120 (100.0)	0 (0.0)	34 (79.1)	9 (20.9)
		SAT Comp. w/ Writing	80 (33.9)	105 (44.5)	51 (21.6)	105 (100.0)	0 (0.0)	41 (80.4)	10 (19.6)
		HSGPA+SAT Comp.	68 (28.8)	82 (34.7)	86 (36.4)	82 (100.0)	0 (0.0)	82 (95.3)	4 (4.7)
		HSGPA+SAT Comp. w/ Writing	72 (30.5)	74 (31.4)	90 (38.1)	74 (100.0)	0 (0.0)	84 (93.3)	6 (6.7)
White	Hispanic	HSGPA	63 (26.2)	119 (49.6)	58 (24.2)	119 (100.0)	0 (0.0)	55 (94.8)	3 (5.2)
		SAT Mathematics	120 (50.0)	93 (38.8)	27 (11.2)	92 (98.9)	1 (1.1)	14 (51.9)	13 (48.1)
		SAT Critical Reading	114 (47.5)	97 (40.4)	29 (12.1)	96 (99.0)	1 (1.0)	16 (55.2)	13 (44.8)
		SAT Writing	123 (51.2)	89 (37.1)	28 (11.7)	89 (100.0)	0 (0.0)	17 (60.7)	11 (39.3)
		SAT Comp.	134 (55.8)	66 (27.5)	40 (16.7)	65 (98.5)	1 (1.5)	29 (72.5)	11 (27.5)
		SAT Comp. w/ Writing	137 (57.1)	61 (25.4)	42 (17.5)	60 (98.4)	1 (1.6)	31 (73.8)	11 (26.2)
		HSGPA+SAT Comp.	121 (50.4)	72 (30.0)	47 (19.6)	71 (98.6)	1 (1.4)	40 (85.1)	7 (14.9)
		HSGPA+SAT Comp. w/ Writing	124 (51.7)	64 (26.7)	52 (21.7)	62 (96.9)	2 (3.1)	44 (84.6)	8 (15.4)
Male	Female	HSGPA	68 (25.6)	147 (55.3)	51 (19.2)	1 (0.7)	146 (99.3)	38 (74.5)	13 (25.5)
		SAT Mathematics	16 (6.0)	177 (66.5)	73 (27.4)	0 (0.0)	177 (100.0)	21 (28.8)	52 (71.2)
		SAT Critical Reading	32 (12.0)	173 (65.0)	61 (22.9)	0 (0.0)	173 (100.0)	7 (11.5)	54 (88.5)
		SAT Writing	50 (18.8)	171 (64.3)	45 (16.9)	0 (0.0)	171 (100.0)	12 (26.7)	33 (73.3)
		SAT Comp.	16 (6.0)	176 (66.2)	74 (27.8)	0 (0.0)	176 (100.0)	15 (20.3)	59 (79.7)
		SAT Comp. w/ Writing	23 (8.6)	173 (65.0)	70 (26.3)	0 (0.0)	173 (100.0)	18 (25.7)	52 (74.3)
		HSGPA+SAT Comp.	35 (13.2)	161 (60.5)	70 (26.3)	0 (0.0)	161 (100.0)	34 (48.6)	36 (51.4)
		HSGPA+SAT Comp. w/ Writing	44 (16.5)	155 (58.3)	67 (25.2)	0 (0.0)	155 (100.0)	35 (52.2)	32 (47.8)

Note. Values in parentheses are percentages. Percentages reported for breakdowns of intercept and slope differences are condition on the detection of a significant difference; these percentages represent rates of directions of subgroup differences in prediction among the significant differences.

Table 17
Summary of Differences in Prediction Detected in Data Corrected for Range Restriction and Criterion Unreliability

			Type of Difference			Breakdown of Intercept Differences		Breakdown of Slope Differences	
Referent	Focal	Predictor	None	Intercept	Slope	Lower Focal Intercept	Higher Focal Intercept	Flatter Focal Slope	Steeper Focal Slope
White	Black	HSGPA	33 (14.0)	115 (48.7)	88 (37.3)	115 (100.0)	0 (0.0)	87 (98.9)	1 (1.1)
		SAT Mathematics	72 (30.5)	133 (56.4)	31 (13.1)	133 (100.0)	0 (0.0)	21 (67.7)	10 (32.3)
		SAT Critical Reading	58 (24.6)	146 (61.9)	32 (13.6)	146 (100.0)	0 (0.0)	20 (62.5)	12 (37.5)
		SAT Writing	59 (25.0)	136 (57.6)	41 (17.4)	136 (100.0)	0 (0.0)	24 (58.5)	17 (41.5)
		SAT Comp.	76 (32.2)	118 (50.0)	42 (17.8)	118 (100.0)	0 (0.0)	33 (78.6)	9 (21.4)
		SAT Comp. w/ Writing	78 (33.1)	110 (46.6)	48 (20.3)	110 (100.0)	0 (0.0)	39 (81.2)	9 (18.8)
		HSGPA+SAT Comp.	69 (29.2)	81 (34.3)	86 (36.4)	81 (100.0)	0 (0.0)	81 (94.2)	5 (5.8)
		HSGPA+SAT Comp. w/ Writing	72 (30.5)	76 (32.2)	88 (37.3)	76 (100.0)	0 (0.0)	83 (94.3)	5 (5.7)
White	Hispanic	HSGPA	63 (26.2)	119 (49.6)	58 (24.2)	119 (100.0)	0 (0.0)	55 (94.8)	3 (5.2)
		SAT Mathematics	119 (49.6)	94 (39.2)	27 (11.2)	93 (98.9)	1 (1.1)	14 (51.9)	13 (48.1)
		SAT Critical Reading	114 (47.5)	98 (40.8)	28 (11.7)	97 (99.0)	1 (1.0)	15 (53.6)	13 (46.4)
		SAT Writing	121 (50.4)	90 (37.5)	29 (12.1)	90 (100.0)	0 (0.0)	18 (62.1)	11 (37.9)
		SAT Comp.	135 (56.2)	65 (27.1)	40 (16.7)	64 (98.5)	1 (1.5)	29 (72.5)	11 (27.5)
		SAT Comp. w/ Writing	136 (56.7)	62 (25.8)	42 (17.5)	61 (98.4)	1 (1.6)	31 (73.8)	11 (26.2)
		HSGPA+SAT Comp.	122 (50.8)	70 (29.2)	48 (20.0)	69 (98.6)	1 (1.4)	39 (81.2)	9 (18.8)
		HSGPA+SAT Comp. w/ Writing	126 (52.5)	61 (25.4)	53 (22.1)	60 (98.4)	1 (1.6)	43 (81.1)	10 (18.9)
Male	Female	HSGPA	65 (24.4)	147 (55.3)	54 (20.3)	1 (0.7)	146 (99.3)	38 (70.4)	16 (29.6)
		SAT Mathematics	15 (5.6)	177 (66.5)	74 (27.8)	0 (0.0)	177 (100.0)	21 (28.4)	53 (71.6)
		SAT Critical Reading	32 (12.0)	173 (65.0)	61 (22.9)	0 (0.0)	173 (100.0)	8 (13.1)	53 (86.9)
		SAT Writing	50 (18.8)	170 (63.9)	46 (17.3)	0 (0.0)	170 (100.0)	12 (26.1)	34 (73.9)
		SAT Comp.	14 (5.3)	175 (65.8)	77 (28.9)	0 (0.0)	175 (100.0)	17 (22.1)	60 (77.9)
		SAT Comp. w/ Writing	23 (8.6)	170 (63.9)	73 (27.4)	0 (0.0)	170 (100.0)	20 (27.4)	53 (72.6)
		HSGPA+SAT Comp.	34 (12.8)	164 (61.7)	68 (25.6)	0 (0.0)	164 (100.0)	35 (51.5)	33 (48.5)
		HSGPA+SAT Comp. w/ Writing	44 (16.5)	156 (58.6)	66 (24.8)	0 (0.0)	156 (100.0)	35 (53.0)	31 (47.0)

Note. Values in parentheses are percentages. Percentages reported for breakdowns of intercept and slope differences are condition on the detection of a significant difference; these percentages represent rates of directions of subgroup differences in prediction among the significant differences.

Table 18

Meta-Analyses of Observed Intercept-Difference Regression Coefficients from Samples without Slope Differences

Referent	Focal	*k*	*N*	Predictor	$\overline{b_{Group}}$	$SD_{b_{Group}}$	SD_{res}	95% CI	80% CV
White	Black	137	301,952	HSGPA	−.43	.13	.11	(−.45, −.40)	(−.57, −.28)
		192	655,158	SAT Mathematics	−.40	.13	.12	(−.41, −.38)	(−.55, −.25)
		188	644,641	SAT Critical Reading	−.40	.12	.10	(−.41, −.38)	(−.53, −.26)
		190	640,861	SAT Writing	−.37	.12	.11	(−.38, −.35)	(−.51, −.23)
		190	692,197	SAT Comp.	−.33	.12	.10	(−.35, −.32)	(−.46, −.20)
		185	662,433	SAT Comp. w/ Writing	−.29	.12	.10	(−.31, −.28)	(−.42, −.16)
		167	435,285	HSGPA + SAT Comp.	−.26	.11	.10	(−.28, −.24)	(−.38, −.14)
		161	411,787	HSGPA + SAT Comp. w/ Writing	−.25	.11	.09	(−.26, −.23)	(−.36, −.13)
White	Hispanic	179	605,921	HSGPA	−.25	.13	.12	(−.27, −.23)	(−.40, −.10)
		207	774,624	SAT Mathematics	−.20	.11	.10	(−.22, −.19)	(−.34, −.07)
		207	657,385	SAT Critical Reading	−.22	.10	.09	(−.23, −.20)	(−.33, −.10)
		207	673,644	SAT Writing	−.19	.10	.08	(−.20, −.17)	(−.29, −.08)
		197	668,751	SAT Comp.	−.16	.10	.09	(−.17, −.14)	(−.27, −.04)
		200	657,585	SAT Comp. w/ Writing	−.14	.10	.08	(−.15, −.12)	(−.24, −.03)
		196	699,810	HSGPA + SAT Comp.	−.14	.09	.08	(−.16, −.13)	(−.25, −.04)
		200	700,293	HSGPA + SAT Comp. w/ Writing	−.14	.09	.08	(−.15, −.13)	(−.24, −.04)
Male	Female	210	822,888	HSGPA	.18	.10	.09	(.17, .20)	(.07, .30)
		177	529,966	SAT Mathematics	.35	.13	.13	(.33, .37)	(.19, .52)
		191	535,454	SAT Critical Reading	.24	.13	.12	(.23, .26)	(.09, .40)
		214	704,315	SAT Writing	.20	.11	.11	(.18, .21)	(.06, .34)
		175	491,552	SAT Comp.	.32	.12	.12	(.30, .34)	(.17, .47)
		192	585,316	SAT Comp. w/ Writing	.28	.11	.11	(.26, .29)	(.14, .42)
		205	735,975	HSGPA + SAT Comp.	.27	.10	.10	(.26, .28)	(.15, .39)
		209	766,392	HSGPA + SAT Comp. w/ Writing	.23	.10	.09	(.21, .24)	(.11, .34)

Note: k = number of studies contributing to meta-analysis; *N* = total sample size; $\overline{b_{Group}}$ = mean observed effect size (b_{Group}); $SD_{b_{Group}}$ = observed standard deviation of b_{Group}; SD_{res} = residual standard deviation of b_{Group}; CI = confidence interval around $\overline{b_{Group}}$; CV = credibility interval around $\overline{b_{Group}}$.

Table 19

Meta-Analyses of Intercept-Difference Regression Coefficients from Samples without Slope Difference Corrected for Criterion Unreliability

Referent	Focal	*k*	*N*	Predictor	$\overline{b_{Group}}$	$SD_{b_{Group}}$	SD_{res}	95% CI	80% CV
White	Black	133	298,095	HSGPA	−.46	.14	.12	(−.48, −.44)	(−.61, −.30)
		187	647,273	SAT Mathematics	−.43	.14	.12	(−.45, −.41)	(−.59, −.27)
		188	644,641	SAT Critical Reading	−.43	.13	.11	(−.44, −.41)	(−.57, −.28)
		188	639,053	SAT Writing	−.39	.13	.12	(−.41, −.38)	(−.55, −.24)
		188	674,045	SAT Comp.	−.35	.12	.11	(−.37, −.34)	(−.49, −.21)
		181	649,964	SAT Comp. w/ Writing	−.32	.12	.11	(−.33, −.30)	(−.46, −.17)
		166	421,087	HSGPA + SAT Comp.	−.28	.12	.10	(−.30, −.26)	(−.42, −.15)
		159	387,531	HSGPA + SAT Comp. w/ Writing	−.27	.12	.10	(−.29, −.25)	(−.40, −.15)
White	Hispanic	174	601,555	HSGPA	−.27	.14	.13	(−.29, −.25)	(−.43, −.11)
		204	764,737	SAT Mathematics	−.22	.12	.11	(−.24, −.20)	(−.36, −.08)
		202	641,020	SAT Critical Reading	−.23	.11	.10	(−.25, −.22)	(−.36, −.11)
		202	650,238	SAT Writing	−.20	.11	.09	(−.21, −.18)	(−.32, −.08)
		189	628,702	SAT Comp.	−.17	.12	.10	(−.19, −.15)	(−.30, −.04)
		193	623,192	SAT Comp. w/ Writing	−.16	.10	.09	(−.17, −.15)	(−.27, −.05)
		194	691,532	HSGPA + SAT Comp.	−.16	.10	.09	(−.17, −.14)	(−.27, −.04)
		196	684,506	HSGPA + SAT Comp. w/ Writing	−.15	.10	.09	(−.16, −.14)	(−.26, −.04)
Male	Female	207	817,621	HSGPA	.20	.10	.10	(.18, .21)	(.07, .32)
		172	514,886	SAT Mathematics	.39	.14	.13	(.37, .41)	(.22, .56)
		190	534,486	SAT Critical Reading	.26	.13	.13	(.24, .28)	(.09, .43)
		210	691,222	SAT Writing	.21	.12	.12	(.20, .23)	(.06, .36)
		171	476,833	SAT Comp.	.35	.13	.13	(.33, .37)	(.18, .51)
		187	572,339	SAT Comp. w/ Writing	.30	.12	.12	(.28, .32)	(.14, .45)
		201	731,297	HSGPA + SAT Comp.	.29	.11	.10	(.28, .31)	(.16, .42)
		202	734,660	HSGPA + SAT Comp. w/ Writing	.24	.10	.10	(.23, .26)	(.12, .37)

Note: k = number of studies contributing to meta-analysis; *N* = total sample size; $\overline{b_{Group}}$ = mean observed effect size (b_{Group}); $SD_{b_{Group}}$ = observed standard deviation of b_{Group}; SD_{res} = residual standard deviation of b_{Group}; CI = confidence interval around $\overline{b_{Group}}$; CV = credibility interval around $\overline{b_{Group}}$.

Table 20

Meta-Analyses of Intercept-Difference Regression Coefficients from Samples without Slope Difference Corrected for Range Restriction

Referent	Focal	*k*	*N*	Predictor	$\overline{b_{Group}}$	$SD_{b_{Group}}$	SD_{res}	95% CI	80% CV
White	Black	148	342,094	HSGPA	−.48	.13	.11	(−.50, −.46)	(−.63, −.34)
		204	719,045	SAT Mathematics	−.36	.13	.10	(−.38, −.34)	(−.49, −.23)
		204	802,447	SAT Critical Reading	−.39	.13	.11	(−.41, −.37)	(−.53, −.25)
		196	636,758	SAT Writing	−.34	.12	.10	(−.36, −.33)	(−.48, −.21)
		193	695,202	SAT Comp.	−.31	.12	.10	(−.33, −.29)	(−.45, −.18)
		185	616,386	SAT Comp. w/ Writing	−.29	.12	.10	(−.30, −.27)	(−.42, −.15)
		150	340,689	HSGPA + SAT Comp.	−.26	.12	.10	(−.28, −.24)	(−.38, −.13)
		146	322,927	HSGPA + SAT Comp. w/ Writing	−.24	.12	.10	(−.26, −.22)	(−.37, −.11)
White	Hispanic	182	666,211	HSGPA	−.31	.15	.13	(−.34, −.29)	(−.49, −.14)
		213	853,532	SAT Mathematics	−.19	.11	.09	(−.21, −.18)	(−.31, −.08)
		211	813,018	SAT Critical Reading	−.20	.11	.08	(−.21, −.18)	(−.31, −.09)
		212	817,635	SAT Writing	−.17	.10	.08	(−.19, −.16)	(−.28, −.07)
		200	794,027	SAT Comp.	−.15	.10	.08	(−.16, −.13)	(−.25, −.05)
		198	783,336	SAT Comp. w/ Writing	−.12	.10	.08	(−.14, −.11)	(−.22, −.03)
		193	684,920	HSGPA + SAT Comp.	−.14	.10	.08	(−.15, −.12)	(−.24, −.03)
		188	640,225	HSGPA + SAT Comp. w/ Writing	−.13	.10	.08	(−.15, −.12)	(−.24, −.03)
Male	Female	215	843,106	HSGPA	.18	.10	.09	(.17, .19)	(.06, .30)
		193	648,952	SAT Mathematics	.41	.12	.12	(.39, .43)	(.26, .56)
		205	610,590	SAT Critical Reading	.28	.13	.12	(.26, .30)	(.13, .43)
		221	796,200	SAT Writing	.21	.11	.10	(.19, .22)	(.08, .34)
		192	552,752	SAT Comp.	.36	.13	.12	(.34, .38)	(.20, .52)
		196	637,920	SAT Comp. w/ Writing	.31	.11	.11	(.29, .32)	(.17, .45)
		196	722,046	HSGPA + SAT Comp.	.26	.09	.09	(.25, .27)	(.15, .38)
		199	757,686	HSGPA + SAT Comp. w/ Writing	.22	.09	.09	(.21, .23)	(.11, .33)

Note: k = number of studies contributing to meta-analysis; *N* = total sample size; $\overline{b_{Group}}$ = mean observed effect size (b_{Group}); $SD_{b_{Group}}$ = observed standard deviation of b_{Group}; SD_{res} = residual standard deviation of b_{Group}; CI = confidence interval around $\overline{b_{Group}}$; CV = credibility interval around $\overline{b_{Group}}$.

Table 21

Meta-Analyses of Intercept-Difference Regression Coefficients from Samples without Slope Difference Corrected for Range Restriction and Criterion Unreliability

Referent	Focal	*k*	*N*	Predictor	$\overline{b_{Group}}$	$SD_{b_{Group}}$	SD_{res}	95% CI	80% CV
White	Black	148	343,110	HSGPA	−.52	.14	.12	(−.54, −.49)	(−.67, −.36)
		205	719,724	SAT Mathematics	−.39	.14	.11	(−.41, −.37)	(−.53, −.25)
		204	802,447	SAT Critical Reading	−.42	.14	.11	(−.44, −.40)	(−.57, −.27)
		195	636,351	SAT Writing	−.37	.13	.11	(−.39, −.35)	(−.51, −.23)
		194	696,805	SAT Comp.	−.33	.13	.11	(−.35, −.31)	(−.48, −.19)
		188	630,637	SAT Comp. w/ Writing	−.31	.13	.11	(−.33, −.29)	(−.45, −.17)
		150	348,542	HSGPA + SAT Comp.	−.27	.13	.10	(−.29, −.25)	(−.41, −.14)
		148	328,064	HSGPA + SAT Comp. w/ Writing	−.26	.13	.11	(−.28, −.24)	(−.39, −.12)
White	Hispanic	182	666,211	HSGPA	−.34	.16	.14	(−.36, −.31)	(−.52, −.15)
		213	853,532	SAT Mathematics	−.21	.12	.09	(−.22, −.19)	(−.33, −.09)
		212	813,548	SAT Critical Reading	−.21	.11	.09	(−.23, −.20)	(−.33, −.10)
		211	817,561	SAT Writing	−.19	.11	.09	(−.20, −.17)	(−.30, −.08)
		200	794,393	SAT Comp.	−.16	.11	.08	(−.17, −.14)	(−.27, −.05)
		198	783,336	SAT Comp. w/ Writing	−.13	.10	.08	(−.15, −.12)	(−.24, −.03)
		192	694,024	HSGPA + SAT Comp.	−.15	.11	.09	(−.16, −.13)	(−.26, −.03)
		187	639,657	HSGPA + SAT Comp. w/ Writing	−.14	.11	.09	(−.16, −.13)	(−.25, −.03)
Male	Female	212	839,697	HSGPA	.19	.10	.10	(.18, .21)	(.07, .32)
		192	647,541	SAT Mathematics	.44	.13	.13	(.42, .46)	(.28, .60)
		205	621,164	SAT Critical Reading	.30	.14	.13	(.28, .32)	(.13, .47)
		220	795,779	SAT Writing	.22	.11	.11	(.21, .24)	(.08, .36)
		189	549,929	SAT Comp.	.39	.14	.13	(.37, .41)	(.22, .55)
		193	632,647	SAT Comp. w/ Writing	.33	.12	.12	(.31, .35)	(.18, .48)
		198	755,563	HSGPA + SAT Comp.	.28	.10	.10	(.27, .30)	(.16, .40)
		200	782,354	HSGPA + SAT Comp. w/ Writing	.24	.10	.09	(.23, .25)	(.12, .36)

Note: k = number of studies contributing to meta-analysis; *N* = total sample size; $\overline{b_{Group}}$ = mean observed effect size (b_{Group}); $SD_{b_{Group}}$ = observed standard deviation of b_{Group}; SD_{res} = residual standard deviation of b_{Group}; CI = confidence interval around $\overline{b_{Group}}$; CV = credibility interval around $\overline{b_{Group}}$.

Table 22

Meta-Analyses of Observed Slope-Difference Regression Coefficients

Referent	Focal	*k*	*N*	Predictor	$\overline{b_{Int.}}$	$SD_{b_{Int.}}$	SD_{res}	95% CI	80% CV
White	Black	236	975,966	HSGPA	−.11	.08	.06	(−.12, −.10)	(−.20, −.03)
				SAT Mathematics	−.02	.10	.07	(−.03, −.00)	(−.11, .08)
				SAT Critical Reading	−.02	.09	.07	(−.03, −.00)	(−.11, .08)
				SAT Writing	−.02	.08	.06	(−.03, −.01)	(−.10, .05)
				SAT Comp.	−.02	.08	.06	(−.03, −.01)	(−.10, .06)
				SAT Comp. w/ Writing	−.03	.08	.06	(−.04, −.02)	(−.11, .04)
				HSGPA + SAT Comp.	−.06	.07	.06	(−.07, −.05)	(−.14, .01)
				HSGPA + SAT Comp. w/ Writing	−.07	.07	.05	(−.08, −.06)	(−.14, −.00)
White	Hispanic	240	999,018	HSGPA	−.07	.07	.05	(−.08, −.06)	(−.14, .00)
				SAT Mathematics	−.01	.09	.06	(−.02, .00)	(−.09, .07)
				SAT Critical Reading	−.01	.08	.06	(−.02, .00)	(−.08, .07)
				SAT Writing	−.01	.08	.06	(−.02, .00)	(−.08, .06)
				SAT Comp.	−.01	.07	.06	(−.02, −.00)	(−.08, .06)
				SAT Comp. w/ Writing	−.02	.07	.05	(−.03, −.01)	(−.09, .05)
				HSGPA + SAT Comp.	−.03	.07	.05	(−.04, −.02)	(−.10, .03)
				HSGPA + SAT Comp. w/ Writing	−.03	.07	.05	(−.04, −.02)	(−.10, .03)
Male	Female	266	1,311,531	HSGPA	−.02	.05	.04	(−.02, −.01)	(−.06, .03)
				SAT Mathematics	.03	.06	.06	(.02, .04)	(−.04, .10)
				SAT Critical Reading	.04	.05	.04	(.04, .05)	(−.01, .10)
				SAT Writing	.02	.05	.04	(.02, .03)	(−.02, .07)
				SAT Comp.	.03	.05	.05	(.03, .04)	(−.03, .09)
				SAT Comp. w/ Writing	.03	.05	.04	(.02, .03)	(−.03, .08)
				HSGPA + SAT Comp.	−.00	.04	.04	(−.01, .00)	(−.05, .05)
				HSGPA + SAT Comp. w/ Writing	−.00	.04	.03	(−.01, .00)	(−.05, .04)

Note: k = number of studies contributing to meta-analysis; N = total sample size; $b_{Int.}$ = group-by-score interaction coefficient from regression analysis; $\overline{b_{Int.}}$ = mean observed effect size ($b_{Int.}$); $SD_{b_{Int.}}$ = observed standard deviation of $b_{Int.}$; SD_{res} = residual standard deviation of $b_{Int.}$; CI = confidence interval around $\overline{b_{Int.}}$; CV = credibility interval around $\overline{b_{Int.}}$.

Table 23

Meta-Analyses of Slope-Difference Regression Coefficients Corrected for Criterion Unreliability

Referent	Focal	*k*	*N*	Predictor	$\overline{b_{Int.}}$	$SD_{b_{Int.}}$	SD_{res}	95% CI	80% CV
White	Black	236	975,966	HSGPA	−.11	.08	.06	(−.12, −.10)	(−.19, −.03)
				SAT Mathematics	−.01	.09	.07	(−.02, −.00)	(−.10, .07)
				SAT Critical Reading	−.01	.08	.07	(−.02, −.00)	(−.10, .07)
				SAT Writing	−.02	.08	.06	(−.03, −.01)	(−.09, .05)
				SAT Comp.	−.02	.08	.07	(−.03, −.01)	(−.11, .06)
				SAT Comp. w/ Writing	−.03	.08	.06	(−.04, −.02)	(−.11, .05)
				HSGPA + SAT Comp.	−.06	.08	.06	(−.07, −.05)	(−.14, .01)
				HSGPA + SAT Comp. w/ Writing	−.07	.07	.06	(−.08, −.06)	(−.14, .00)
White	Hispanic	240	999,018	HSGPA	−.06	.07	.05	(−.07, −.05)	(−.13, .00)
				SAT Mathematics	−.01	.08	.06	(−.02, .00)	(−.08, .07)
				SAT Critical Reading	−.01	.08	.06	(−.01, .00)	(−.08, .07)
				SAT Writing	−.01	.07	.05	(−.02, .00)	(−.08, .06)
				SAT Comp.	−.01	.08	.06	(−.02, −.00)	(−.09, .06)
				SAT Comp. w/ Writing	−.02	.07	.06	(−.03, −.01)	(−.09, .06)
				HSGPA + SAT Comp.	−.03	.07	.05	(−.04, −.02)	(−.10, .04)
				HSGPA + SAT Comp. w/ Writing	−.03	.07	.05	(−.04, −.02)	(−.10, .04)
Male	Female	266	1,311,531	HSGPA	−.02	.04	.03	(−.02, −.01)	(−.06, .03)
				SAT Mathematics	.03	.06	.05	(.02, .04)	(−.04, .09)
				SAT Critical Reading	.04	.05	.04	(.03, .04)	(−.01, .09)
				SAT Writing	.02	.04	.03	(.02, .03)	(−.02, .06)
				SAT Comp.	.04	.06	.05	(.03, .04)	(−.03, .10)
				SAT Comp. w/ Writing	.03	.05	.04	(.02, .03)	(−.03, .08)
				HSGPA + SAT Comp.	−.00	.05	.04	(−.01, .00)	(−.05, .05)
				HSGPA + SAT Comp. w/ Writing	−.00	.04	.04	(−.01, .00)	(−.05, .04)

Note: k = number of studies contributing to meta-analysis; N = total sample size; $b_{Int.}$ = group-by-score interaction coefficient from regression analysis; $\overline{b_{Int.}}$ = mean observed effect size ($b_{Int.}$); $SD_{b_{Int.}}$ = observed standard deviation of $b_{Int.}$; SD_{res} = residual standard deviation of $b_{Int.}$; CI = confidence interval around $\overline{b_{Int.}}$; CV = credibility interval around $\overline{b_{Int.}}$.

Table 24

Meta-Analyses of Slope-Difference Regression Coefficients Corrected for Range Restriction

Referent	Focal	*k*	*N*	Predictor	$\overline{b_{Int.}}$	$SD_{b_{Int.}}$	SD_{res}	95% CI	80% CV
White	Black	236	975,966	HSGPA	−.10	.08	.06	(−.11, −.09)	(−.18, −.02)
				SAT Mathematics	−.00	.09	.07	(−.02, .01)	(−.09, .08)
				SAT Critical Reading	−.00	.09	.06	(−.02, .01)	(−.09, .08)
				SAT Writing	−.01	.09	.06	(−.02, .00)	(−.09, .08)
				SAT Comp.	−.02	.09	.07	(−.04, −.01)	(−.12, .07)
				SAT Comp. w/ Writing	−.03	.09	.07	(−.04, −.02)	(−.13, .06)
				HSGPA + SAT Comp.	−.08	.09	.07	(−.09, −.07)	(−.17, .01)
				HSGPA + SAT Comp. w/ Writing	−.08	.09	.07	(−.10, −.07)	(−.18, .01)
White	Hispanic	240	999,018	HSGPA	−.05	.08	.06	(−.06, −.04)	(−.12, .02)
				SAT Mathematics	.01	.08	.06	(.00, .02)	(−.06, .08)
				SAT Critical Reading	.01	.08	.05	(−.00, .02)	(−.06, .08)
				SAT Writing	.00	.08	.06	(−.01, .02)	(−.07, .08)
				SAT Comp.	−.01	.08	.06	(−.02, .00)	(−.09, .07)
				SAT Comp. w/ Writing	−.01	.08	.06	(−.02, −.00)	(−.09, .07)
				HSGPA + SAT Comp.	−.04	.09	.07	(−.05, −.03)	(−.13, .05)
				HSGPA + SAT Comp. w/ Writing	−.04	.09	.07	(−.05, −.03)	(−.13, .05)
Male	Female	266	1,311,531	HSGPA	−.01	.04	.03	(−.02, −.01)	(−.06, .03)
				SAT Mathematics	.02	.06	.05	(.01, .02)	(−.05, .09)
				SAT Critical Reading	.03	.05	.04	(.02, .03)	(−.02, .08)
				SAT Writing	.01	.05	.03	(.01, .02)	(−.03, .06)
				SAT Comp.	.02	.06	.05	(.02, .03)	(−.04, .09)
				SAT Comp. w/ Writing	.02	.05	.05	(.01, .02)	(−.04, .08)
				HSGPA + SAT Comp.	.00	.05	.04	(−.01, .01)	(−.05, .05)
				HSGPA + SAT Comp. w/ Writing	−.00	.05	.04	(−.01, .00)	(−.05, .05)

Note: k = number of studies contributing to meta-analysis; N = total sample size; $b_{Int.}$ = group-by-score interaction coefficient from regression analysis; $\overline{b_{Int.}}$ = mean observed effect size ($b_{Int.}$); $SD_{b_{Int.}}$ = observed standard deviation of $b_{Int.}$; SD_{res} = residual standard deviation of $b_{Int.}$; CI = confidence interval around $\overline{b_{Int.}}$; CV = credibility interval around $\overline{b_{Int.}}$.

Table 25

Meta-Analyses of Slope-Difference Regression Coefficients Corrected for Range Restriction and Criterion Unreliability

Referent	Focal	*k*	*N*	Predictor	$\overline{b_{Int.}}$	$SD_{b_{Int.}}$	SD_{res}	95% CI	80% CV
White	Black	236	975,966	HSGPA	−.11	.08	.07	(−.12, −.10)	(−.19, −.02)
				SAT Mathematics	−.00	.10	.07	(−.02, .01)	(−.10, .09)
				SAT Critical Reading	−.00	.10	.07	(−.02, .01)	(−.09, .09)
				SAT Writing	−.01	.09	.07	(−.02, .00)	(−.10, .08)
				SAT Comp.	−.03	.10	.08	(−.04, −.01)	(−.12, .07)
				SAT Comp. w/ Writing	−.03	.10	.08	(−.05, −.02)	(−.14, .07)
				HSGPA + SAT Comp.	−.08	.09	.08	(−.10, −.07)	(−.18, .01)
				HSGPA + SAT Comp. w/ Writing	−.09	.09	.08	(−.10, −.08)	(−.19, .01)
White	Hispanic	240	999,018	HSGPA	−.06	.08	.06	(−.07, −.04)	(−.13, .02)
				SAT Mathematics	.01	.09	.06	(.00, .02)	(−.07, .09)
				SAT Critical Reading	.01	.08	.06	(.00, .02)	(−.06, .09)
				SAT Writing	.01	.09	.07	(−.01, .02)	(−.08, .09)
				SAT Comp.	−.01	.09	.07	(−.02, .00)	(−.10, .08)
				SAT Comp. w/ Writing	−.01	.09	.07	(−.03, −.00)	(−.10, .07)
				HSGPA + SAT Comp.	−.04	.10	.08	(−.05, −.03)	(−.15, .06)
				HSGPA + SAT Comp. w/ Writing	−.04	.09	.08	(−.06, −.03)	(−.15, .06)
Male	Female	266	1,311,531	HSGPA	−.02	.05	.04	(−.02, −.01)	(−.06, .03)
				SAT Mathematics	.02	.07	.06	(.01, .03)	(−.06, .09)
				SAT Critical Reading	.03	.05	.04	(.02, .04)	(−.03, .09)
				SAT Writing	.01	.05	.04	(.01, .02)	(−.04, .06)
				SAT Comp.	.03	.06	.05	(.02, .03)	(−.04, .09)
				SAT Comp. w/ Writing	.02	.06	.05	(.01, .03)	(−.04, .08)
				HSGPA + SAT Comp.	.00	.05	.05	(−.01, .01)	(−.06, .06)
				HSGPA + SAT Comp. w/ Writing	−.00	.05	.04	(−.01, .00)	(−.06, .06)

Note: k = number of studies contributing to meta-analysis; N = total sample size; $b_{Int.}$ = group-by-score interaction coefficient from regression analysis; $\overline{b_{Int.}}$ = mean observed effect size ($b_{Int.}$); $SD_{b_{Int.}}$ = observed standard deviation of $b_{Int.}$; SD_{res} = residual standard deviation of $b_{Int.}$; CI = confidence interval around $\overline{b_{Int.}}$; CV = credibility interval around $\overline{b_{Int.}}$.

Table 26
Parameters Used in Range-Restriction Simulations

Parameter Name (Abbreviation)	Constant	Variable Values			# of Levels
		Low	High	Increment	
Constants					
Referent Validity of X (ρ_{XY_Ref})	.5	---	---	---	1
Standardized Mean Difference on X (δ_X)	1	---	---	---	1
Reliability of X (ρ_{XX}) *	1	---	---	---	1
Reliability of Z (ρ_{ZZ}) *	1	---	---	---	1
Variables for all simulations					
Overall Selection Ratio (SR) **	---	.1	.9	.4	3
Referent Group Proportion (P_{Ref})	---	.5	.9	.2	3
Focal Validity of X (ρ_{XY_Foc})	---	.1	.5	.2	3
Standardized Mean Difference on Y (δ_Y)	---	.3	.7	.2	3
Reliability of Y (ρ_{YY}) *	---	0.6	1.0	.2	3
Variables for IRR simulation only					
Referent Validity of Z (ρ_{ZY_Ref})	---	.2	.6	.2	3
ρ_{ZY_Ref}/ ρ_{ZY_Foc} Ratio (ρ_{ZY_Ratio}) ***	---	0	1	.5	3
Correlation between X and Z (ρ_{ZX}) *	---	.2	.8	.3	3
Standardized Mean Difference on Z (δ_Z)	---	0.0	1.0	0.5	3

Note.
IRR = indirect range restriction; ρ is the parameter notation that corresponds to correlations and reliability coefficients; δ is the parameter notation that corresponds to *d* effect sizes.
* ρ_{ZX} , ρ_{XX}, ρ_{YY}, and ρ_{ZZ} parameters were constrained to be equal between subgroups.
** In addition to the tabled selection ratios, each combination of parameters was also simulated using a selection ratio of 1, meaning all applicants were selected or, equivalently, selection was performed at random. These random-selection scenarios were used as comparison conditions for the conditions in which systematic selection was performed.
*** Focal-group validities were determined by multiplying referent-group validities (ρ_{ZY_Ref}) by ρ_{ZY_Ratio} values.

Table 27

Criteria Determining which Simulation Conditions Were Included in Summary Analyses Involving Each Dependent Variable

Dependent variable	Inclusion criteria for simulation conditions	Summary of conditions included			
		DRR		IRR	
		Number	% of total	Number	% of total
Δd_{Mod_Signed}	No restrictions	243	100.0%	19,683	100.0%
ΔSign of d_{Mod_Signed}	No restrictions	243	100.0%	19,683	100.0%
$\Delta\beta_{\Delta Intercepts}$	Not DRR and Equal slopes	0	0%	6,561	33.3%
ΔSign of $\beta_{\Delta Intercepts}$	Not DRR, equal slopes, and unequal intercepts	0	0%	4,374	22.2%
$\Delta\beta_{\Delta Slopes}$	Not DRR	0	0%	19,683	100.0%
ΔSign of $\beta_{\Delta Slopes}$	Not DRR and unequal slopes	0	0%	13,122	66.7%
$\Delta F_{\Delta Overall}$ Power	Unequal slopes or intercepts	216	88.9%	17,496	88.9%
$\Delta F_{\Delta Slopes}$ Power	Unequal slopes	162	66.7%	13,122	66.7%
$\Delta F_{\Delta Intercepts}$ Power	Equal slopes and unequal intercepts	54	22.2%	4,374	22.2%
$\Delta F_{\Delta Overall}$ Type I	Equal slopes and intercepts	27	11.1%	2,187	11.1%
$\Delta F_{\Delta Slopes}$ Type I	Equal slopes	81	33.3%	6,561	33.3%
$\Delta F_{\Delta Intercepts}$ Type I	Equal slopes and intercepts	27	11.1%	2,187	11.1%

Note. Inclusion criteria involving comparisons of slopes and/or intercepts were based on the unrestricted parameter values of regression coefficients, not the observed range-restricted coefficients.

Δd_{Mod_Signed} = change in d_{Mod_Signed} effect sizes; ΔSign of d_{Mod_Signed} = change in the signs of d_{Mod_Signed} effect sizes; $\Delta\beta_{\Delta Intercepts}$ = change in intercept-difference regression coefficients; ΔSign of $\beta_{\Delta Intercepts}$ = change in the signs of intercept-difference regression coefficients; $\Delta\beta_{\Delta Slopes}$ = change in slope-difference regression coefficients; ΔSign of $\beta_{\Delta Slopes}$ = change in the signs of slope-difference regression coefficients; $\Delta F_{\Delta Overall}$ Power = change in power for tests of overall differences in prediction as indicated by normalized differences in F ratios; $\Delta F_{\Delta Slopes}$ Power = change in power for tests of slope differences as indicated by normalized differences in F ratios; $\Delta F_{\Delta Intercepts}$ Power = change in power for tests of intercept differences as indicated by normalized differences in F ratios; $\Delta F_{\Delta Overall}$ Type I = change in Type I errors for tests of overall differences in prediction as indicated by normalized differences in F ratios; $\Delta F_{\Delta Slopes}$ Type I = change in Type I errors for tests of slope differences as indicated by normalized differences in F ratios; $\Delta F_{\Delta Intercepts}$ Type I = change in Type I errors for tests of intercept differences as indicated by normalized differences in F ratios. Analyses of $\Delta\beta_{\Delta Slopes}$, ΔSign of $\beta_{\Delta Slopes}$, $\Delta\beta_{\Delta Intercepts}$, and ΔSign of $\beta_{\Delta Intercepts}$ were not conducted for the DRR simulation because the statistical artifacts induced in that simulation were not capable of altering the values of regression coefficients.

Table 28

Summary of Simulation Parameters' Total Contributions to Explaining Dependent Variables

	Direct range restriction					Indirect range restriction								
Dependent variable	P_{Ref}	SR	ρ_{XY_Foc}	δ_Y	ρ_{YY}	P_{Ref}	SR	ρ_{XY_Foc}	ρ_{XZ}	ρ_{ZY_Ref}	ρ_{ZY_Ratio}	δ_Y	δ_Z	ρ_{YY}
Δd_{Mod_Signed}	.00	**.66**	**.58**	.01	*.04*	.00	**.59**	.00	.02	**.27**	**.52**	.01	.02	.03
ΔSign of d_{Mod_Signed}	*.13*	**.36**	**.43**	**.74**	.00	.01	**.32**	.00	*.05*	*.12*	**.33**	**.58**	*.05*	.00
$\Delta\beta_{\Delta Intercepts}$	—	—	—	—	—	.00	**.58**	—	.02	**.24**	**.55**	.00	.04	.00
ΔSign of $\beta_{\Delta Intercepts}$	—	—	—	—	—	.02	**.37**	—	.05	*.14*	**.37**	**.49**	*.05*	.00
$\Delta\beta_{\Delta Slopes}$	—	—	—	—	—	.00	**.39**	**.29**	**.41**	**.19**	**.55**	.00	.00	.00
ΔSign of $\beta_{\Delta Slopes}$	—	—	—	—	—	.01	**.25**	**.29**	**.36**	**.26**	**.50**	.00	.01	.00
$\Delta F_{\Delta Overall}$ Power	**.37**	**.51**	**.47**	*.14*	**.30**	*.11*	**.49**	*.03*	.01	**.36**	**.43**	*.07*	*.13*	*.10*
$\Delta F_{\Delta Slopes}$ Power	**.37**	**.36**	**.49**	.00	*.05*	*.16*	**.20**	**.38**	*.07*	.03	*.16*	.00	**.21**	**.28**
$\Delta F_{\Delta Intercepts}$ Power	**.33**	**.51**	—	.00	**.31**	*.11*	**.50**	—	.01	**.36**	**.47**	*.13*	*.08*	.08
$\Delta F_{\Delta Overall}$ Type I	—	—	—	—	—	*.13*	**.53**	—	.02	**.39**	**.51**	—	*.08*	*.08*
$\Delta F_{\Delta Slopes}$ Type I	—	—	—	—	—	*.15*	**.34**	—	**.41**	**.25**	**.43**	.00	*.11*	*.07*
$\Delta F_{\Delta Intercepts}$ Type I	—	—	—	—	—	*.12*	**.55**	—	.01	**.39**	**.50**	—	*.08*	*.07*

Note. Tabled values represent the sums of all η^2 effects involving the parameter for explaining the dependent variable (i.e., each value is the sum of η^2 values for a parameter's main effect and all interactions in which that parameter was included). Dash = Parameter was not included in the model used to explain variation in the dependent variable. Bold = Parameter is considered to have an important effect. Italic = Parameter was included in a plotted effect but is not considered to have an important effect. Δd_{Mod_Signed} = change in d_{Mod_Signed} effect sizes; ΔSign of d_{Mod_Signed} = change in the signs of d_{Mod_Signed} effect sizes; $\Delta\beta_{\Delta Slopes}$ = change in slope-difference regression coefficients; ΔSign of $\beta_{\Delta Slopes}$ = change in the signs of slope-difference regression coefficients; $\Delta\beta_{\Delta Intercepts}$ = change in intercept-difference regression coefficients; ΔSign of $\beta_{\Delta Intercepts}$ = change in the signs of intercept-difference regression coefficients; $\Delta F_{\Delta Overall}$ Power = change in power for tests of overall differences in prediction as indicated by normalized differences in F ratios; $\Delta F_{\Delta Slopes}$ Power = change in power for tests of slope differences as indicated by normalized differences in F ratios; $\Delta F_{\Delta Intercepts}$ Power = change in power for tests of intercept differences as indicated by normalized differences in F ratios; $\Delta F_{\Delta Overall}$ Type I = change in Type I errors for tests of overall differences in prediction as indicated by normalized differences in F ratios; $\Delta F_{\Delta Slopes}$ Type I = change in Type I errors for tests of slope differences as indicated by normalized differences in F ratios; $\Delta F_{\Delta Intercepts}$ Type I = change in Type I errors for tests of intercept differences as indicated by normalized differences in F ratios. Analyses of $\Delta\beta_{\Delta Slopes}$, ΔSign of $\beta_{\Delta Slopes}$, $\Delta\beta_{\Delta Intercepts}$, and ΔSign of $\beta_{\Delta Intercepts}$ were not conducted for the DRR simulation because the statistical artifacts induced in that simulation were not capable of altering the values of regression coefficients. P_{Ref} = proportion of referent-group members in the applicant population; SR = overall selection ratio applied to Z; ρ_{XY_Foc} = operational validity of X for predicting Y in the focal group's applicant population; ρ_{XZ} = correlation between X and Z in the referent and focal groups' applicant populations; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on Y; δ_Z = standardized mean difference between the referent and focal groups on Z; ρ_{YY} = reliability of Y.

Table 29

Summary of Simulation Parameters' Total Relative Contributions to Explaining Dependent Variables

	Direct range restriction					Indirect range restriction								
Dependent variable	P_{Ref}	SR	ρ_{XY_Foc}	δ_Y	ρ_{YY}	P_{Ref}	SR	ρ_{XY_Foc}	ρ_{XZ}	ρ_{ZY_Ref}	ρ_{ZY_Ratio}	δ_Y	δ_Z	ρ_{YY}
Δd_{Mod_Signed}	.00	**.66**	**.58**	.01	.04	.00	**.59**	.00	.02	**.27**	**.52**	.01	.02	.03
ΔSign of d_{Mod_Signed}	.13	**.37**	**.44**	**.76**	.00	.01	**.40**	.00	.06	.15	**.40**	**.72**	.06	.00
$\Delta\beta_{\Delta Intercepts}$	—	—	—	—	—	.00	**.59**	—	.02	**.24**	**.55**	.00	.04	.00
ΔSign of $\beta_{\Delta Intercepts}$	—	—	—	—	—	.02	**.44**	—	.05	.17	**.44**	**.58**	.06	.00
$\Delta\beta_{\Delta Slopes}$	—	—	—	—	—	.00	**.39**	**.29**	**.42**	**.19**	**.56**	.00	.00	.00
ΔSign of $\beta_{\Delta Slopes}$	—	—	—	—	—	.01	**.32**	**.37**	**.46**	**.34**	**.64**	.00	.01	.00
$\Delta F_{\Delta Overall}$ Power	**.38**	**.53**	**.48**	.14	**.31**	.12	**.52**	.03	.01	**.38**	**.45**	.07	.13	.11
$\Delta F_{\Delta Slopes}$ Power	**.37**	**.36**	**.49**	.00	.05	.17	**.21**	**.42**	.08	.04	.17	.00	**.23**	**.30**
$\Delta F_{\Delta Intercepts}$ Power	**.33**	**.51**	—	.00	**.31**	.12	**.54**	—	.01	**.38**	**.50**	.14	.09	.09
$\Delta F_{\Delta Overall}$ Type I	—	—	—	—	—	.13	**.56**	—	.02	**.41**	**.53**	—	.08	.08
$\Delta F_{\Delta Slopes}$ Type I	—	—	—	—	—	.18	**.40**	—	**.48**	**.30**	**.51**	.00	.12	.09
$\Delta F_{\Delta Intercepts}$ Type I	—	—	—	—	—	.12	**.58**	—	.01	**.41**	**.53**	—	.08	.07

Note. Tabled values represent the sums of the percent contribution to model fit involving the parameter for explaining the dependent variable (i.e., each value is the sum of $\eta^2 / \sum \eta^2$ values for a parameter's main effect and all interactions in which that parameter was included). Dash = Parameter was not included in the model used to explain variation in the dependent variable. Bold = Parameter is considered to have an important effect. Italic = Parameter was included in a plotted effect but is not considered to have an important effect. Δd_{Mod_Signed} = change in d_{Mod_Signed} effect sizes; ΔSign of d_{Mod_Signed} = change in the signs of d_{Mod_Signed} effect sizes; $\Delta\beta_{\Delta Slopes}$ = change in slope-difference regression coefficients; ΔSign of $\beta_{\Delta Slopes}$ = change in the signs of slope-difference regression coefficients; $\Delta\beta_{\Delta Intercepts}$ = change in intercept-difference regression coefficients; ΔSign of $\beta_{\Delta Intercepts}$ = change in the signs of intercept-difference regression coefficients; $\Delta F_{\Delta Overall}$ Power = change in power for tests of overall differences in prediction as indicated by normalized differences in F ratios; $\Delta F_{\Delta Slopes}$ Power = change in power for tests of slope differences as indicated by normalized differences in F ratios; $\Delta F_{\Delta Intercepts}$ Power = change in power for tests of intercept differences as indicated by normalized differences in F ratios; $\Delta F_{\Delta Overall}$ Type I = change in Type I errors for tests of overall differences in prediction as indicated by normalized differences in F ratios; $\Delta F_{\Delta Slopes}$ Type I = change in Type I errors for tests of slope differences as indicated by normalized differences in F ratios; $\Delta F_{\Delta Intercepts}$ Type I = change in Type I errors for tests of intercept differences as indicated by normalized differences in F ratios. Analyses of $\Delta\beta_{\Delta Slopes}$, ΔSign of $\beta_{\Delta Slopes}$, $\Delta\beta_{\Delta Intercepts}$, and ΔSign of $\beta_{\Delta Intercepts}$ were not conducted for the DRR simulation because the statistical artifacts induced in that simulation were not capable of altering the values of regression coefficients. P_{Ref} = proportion of referent-group members in the applicant population; SR = overall selection ratio applied to Z; ρ_{XY_Foc} = operational validity of X for predicting Y in the focal group's applicant population; ρ_{XZ} = correlation between X and Z in the referent and focal groups' applicant populations; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on Y; δ_Z = standardized mean difference between the referent and focal groups on Z; ρ_{YY} = reliability of Y.

Table 30
Summary of Variance Explained in Directly Range Restricted Observed-Operational Differences in d_{Mod_Signed} *Values*

Parameter	Δd_{Mod_Signed}	ΔSign of d_{Mod_Signed}
SR	.39	.07
P_{Ref}	—	.01
ρ_{XY_Foc}	.31	.08
δ_Y	—	.28
ρ_{YY}	.01	—
SR x P_{Ref}	—	.01
SR x ρ_{XY_Foc}	.26	.04
SR x δ_Y	—	.14
P_{Ref} x ρ_{XY_Foc}	—	.02
P_{Ref} x δ_Y	—	.02
ρ_{XY_Foc} x δ_Y	—	.16
SR x P_{Ref} x ρ_{XY_Foc}	—	.01
SR x P_{Ref} x δ_Y	—	.01
SR x ρ_{XY_Foc} x δ_Y	—	.08
P_{Ref} x ρ_{XY_Foc} x δ_Y	—	.05
Main Effect of ρ_{YY}	.01	—
Overall Effect of SR x ρ_{XY_Foc}	.96	—
Overall Effect of SR x P_{Ref} x ρ_{XY_Foc}	—	.24
Overall Effect of SR x P_{Ref} x δ_Y	—	.54
Overall Effect of SR x ρ_{XY_Foc} x δ_Y	—	.85
Overall Effect of P_{Ref} x ρ_{XY_Foc} x δ_Y	—	.62
Total	1.00	.98
Restricted Total	.97	.98

Note. Overall Effect = total variance explained by an interaction effect and all lower-order effects subsumed by the interaction; Total = total variance explained by all effects in the summary model; Restricted Total = total variance explained by all effects that explained at least 1% of variance in the dependent variable, plus the variance explained by any lower-order effects implicated in interactions that met the 1% threshold. Δd_{Mod_Signed} = change in d_{Mod_Signed} effect sizes; ΔSign of d_{Mod_Signed} = change in the signs of d_{Mod_Signed} effect sizes; SR = overall selection ratio applied to X; P_{Ref} = proportion of referent-group members in the applicant population; ρ_{XY_Foc} = operational validity of X for predicting Y in the focal group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on Y; ρ_{YY} = reliability of Y.

Table 31

Summary of Variance Explained in Directly Range Restricted Observed-Operational Differences in Power as Indicated by Normalized F *Ratios*

Parameter	$\Delta F_{\Delta Overall}$	$\Delta F_{\Delta Slopes}$	$\Delta F_{\Delta Intercepts}$
SR	.11	.21	.42
P_{Ref}	.04	.22	.21
ρ_{XY_Foc}	.06	.32	—
δ_Y	.02	—	—
ρ_{YY}	.12	.02	.22
SR x P_{Ref}	.10	.05	.06
SR x ρ_{XY_Foc}	.14	.07	—
SR x δ_Y	.02	—	—
SR x ρ_{YY}	—	.01	.03
P_{Ref} x ρ_{XY_Foc}	.04	.07	—
P_{Ref} x δ_Y	.02	—	—
P_{Ref} x ρ_{YY}	.03	—	.06
ρ_{XY_Foc} x δ_Y	.01	—	—
ρ_{XY_Foc} x ρ_{YY}	.06	—	—
δ_Y x ρ_{YY}	.02	—	—
SR x P_{Ref} x ρ_{XY_Foc}	.09	.02	—
SR x P_{Ref} x δ_Y	.02	—	—
SR x ρ_{XY_Foc} x δ_Y	.01	—	—
P_{Ref} x ρ_{XY_Foc} x δ_Y	.01	—	—
P_{Ref} x ρ_{XY_Foc} x ρ_{YY}	.02	—	—
ρ_{XY_Foc} x δ_Y x ρ_{YY}	.01	—	—
Overall Effect of SR x P_{Ref}	—	—	.69
Overall Effect of SR x ρ_{YY}	—	.23	.67
Overall Effect of P_{Ref} x ρ_{YY}	—	—	.49
Overall Effect of SR x P_{Ref} x ρ_{XY_Foc}	.58	.95	—
Overall Effect of SR x P_{Ref} x δ_Y	.31	—	—
Overall Effect of SR x ρ_{XY_Foc} x δ_Y	.37	—	—
Overall Effect of P_{Ref} x ρ_{XY_Foc} x δ_Y	.19	—	—
Overall Effect of P_{Ref} x ρ_{XY_Foc} x ρ_{YY}	.38	—	—
Overall Effect of ρ_{XY_Foc} x δ_Y x ρ_{YY}	.30	—	—
Total	.98	1.00	1.00
Restricted Total	.95	.98	1.00

Note. Overall Effect = total variance explained by an interaction effect and all lower-order effects subsumed by the interaction; Total = total variance explained by all effects in the summary model; Restricted Total = total variance explained by all effects that explained at least 1% of variance in the dependent variable, plus the variance explained by any lower-order effects implicated in interactions that met the 1% threshold. $\Delta F_{\Delta Overall}$ = change in power for tests of overall differences in prediction as indicated by normalized differences in *F* ratios; $\Delta F_{\Delta Slopes}$ = change in power for tests of slope differences as indicated by normalized differences in *F* ratios; $\Delta F_{\Delta Intercepts}$ = change in power for tests of intercept differences as indicated by normalized differences in *F* ratios; SR = overall selection ratio applied to *X*; P_{Ref} = proportion of referent-group members in the applicant population; ρ_{XY_Foc} = operational validity of *X* for predicting *Y* in the focal group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on *Y*; ρ_{YY} = reliability of *Y*.

Table 32
Summary of Variance Explained in Indirectly Range Restricted Observed-Operational Differences in d_{Mod_Signed} *Values*

Parameter	Δd_{Mod_Signed}	ΔSign of d_{Mod_Signed}
SR	.26	.08
ρ_{ZY_Ref}	.09	.01
ρ_{ZY_Ratio}	.23	.05
ρ_{XZ}	—	.00
δ_Y	—	.19
δ_Z	—	.00
SR x ρ_{ZY_Ref}	.07	.01
SR x ρ_{ZY_Ratio}	.19	.03
SR x δ_Y	—	.11
ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.05	.01
ρ_{ZY_Ref} x δ_Y	—	.02
ρ_{ZY_Ref} x δ_Z	—	.00
ρ_{ZY_Ratio} x ρ_{XZ}	—	.00
ρ_{ZY_Ratio} x δ_Y	—	.12
ρ_{XZ} x δ_Y	—	.00
δ_Y x δ_Z	—	.00
SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.04	.01
SR x ρ_{ZY_Ref} x δ_Y	—	.01
SR x ρ_{ZY_Ratio} x δ_Y	—	.07
ρ_{ZY_Ref} x ρ_{ZY_Ratio} x δ_Y	—	.01
ρ_{ZY_Ref} x δ_Y x δ_Z	—	.02
ρ_{ZY_Ratio} x ρ_{XZ} x δ_Y	—	.01
Overall Effect of SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.93	.19
Overall Effect of SR x ρ_{ZY_Ref} x δ_Y	—	.43
Overall Effect of SR x ρ_{ZY_Ratio} x δ_Y	—	.63
Overall Effect of ρ_{ZY_Ref} x ρ_{ZY_Ratio} x δ_Y	—	.41
Overall Effect of ρ_{ZY_Ref} x δ_Y x δ_Z	—	.24
Overall Effect of ρ_{ZY_Ratio} x ρ_{XZ} x δ_Y	—	.38
Total	1.00	.81
Restricted Total	.93	.75

Note. Overall Effect = total variance explained by an interaction effect and all lower-order effects subsumed by the interaction; Total = total variance explained by all effects in the summary model; Restricted Total = total variance explained by all effects that explained at least 1% of variance in the dependent variable, plus the variance explained by any lower-order effects implicated in interactions that met the 1% threshold. Δd_{Mod_Signed} = change in d_{Mod_Signed} effect sizes; ΔSign of d_{Mod_Signed} = change in the signs of d_{Mod_Signed} effect sizes; SR = overall selection ratio applied to *Z*; ρ_{ZY_Ref} = operational validity of *Z* for predicting *Y* in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of *Z* for predicting *Y* in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between *X* and *Z* in the referent and focal groups' applicant populations; δ_Y = true-score standardized mean difference between the referent and focal groups on *Y*; δ_Z = standardized mean difference between the referent and focal groups on *Z*.

Table 33

Summary of Variance Explained in Indirectly Range Restricted Observed-Operational Differences in Intercept-Difference Regression Coefficients

Parameter	$\Delta\beta_{\Delta Intercepts}$	ΔSign of $\beta_{\Delta Intercepts}$
SR	.27	.12
ρ_{ZY_Ref}	.08	.02
ρ_{ZY_Ratio}	.26	.07
δ_Y	—	.16
δ_Z	—	.00
SR x ρ_{ZY_Ref}	.06	.01
SR x ρ_{ZY_Ratio}	.20	.04
SR x δ_Y	—	.09
ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.04	.01
ρ_{ZY_Ref} x δ_Y	—	.02
ρ_{ZY_Ref} x δ_Z	—	.00
ρ_{ZY_Ratio} x δ_Y	—	.11
δ_Y x δ_Z	—	.00
SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.03	.02
SR x ρ_{ZY_Ratio} x δ_Y	—	.06
ρ_{ZY_Ref} x ρ_{ZY_Ratio} x δ_Y	—	.01
ρ_{ZY_Ref} x δ_Y x δ_Z	—	.01
Overall Effect of SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.94	.29
Overall Effect of SR x ρ_{ZY_Ratio} x δ_Y	—	.65
Overall Effect of ρ_{ZY_Ref} x ρ_{ZY_Ratio} x δ_Y	—	.40
Overall Effect of ρ_{ZY_Ref} x δ_Y x δ_Z	—	.21
Total	1.00	.84
Restricted Total	.94	.75

Note. Overall Effect = total variance explained by an interaction effect and all lower-order effects subsumed by the interaction; Total = total variance explained by all effects in the summary model; Restricted Total = total variance explained by all effects that explained at least 1% of variance in the dependent variable, plus the variance explained by any lower-order effects implicated in interactions that met the 1% threshold. $\Delta\beta_{\Delta Intercepts}$ = change in intercept-difference regression coefficients from scenarios without slope differences; ΔSign of $\beta_{\Delta Intercepts}$ = change in the signs of intercept-difference regression coefficients from scenarios with intercept differences and without slope differences; SR = overall selection ratio applied to Z; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on Y; δ_Z = standardized mean difference between the referent and focal groups on Z.

Table 34

Summary of Variance Explained in Indirectly Range Restricted Observed-Operational Differences in Slope-Difference Regression Coefficients

Parameter	$\Delta\beta_{\Delta Slopes}$	ΔSign of $\beta_{\Delta Slopes}$
SR	.02	.02
ρ_{XY_Foc}	.09	.03
ρ_{ZY_Ref}	.05	.02
ρ_{ZY_Ratio}	.18	.07
ρ_{XZ}	.00	.04
SR x ρ_{XY_Foc}	.05	.01
SR x ρ_{XZ}	.00	.02
SR x ρ_{ZY_Ref}	.03	.02
SR x ρ_{ZY_Ratio}	.11	.04
ρ_{XY_Foc} x ρ_{XZ}	.09	.02
ρ_{XY_Foc} x ρ_{ZY_Ref}	—	.01
ρ_{XY_Foc} x ρ_{ZY_Ratio}	—	.05
ρ_{ZY_Ref} x ρ_{XZ}	.02	.02
ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.03	.04
ρ_{ZY_Ratio} x ρ_{XZ}	.12	.07
SR x ρ_{XY_Foc} x ρ_{XZ}	.06	.02
SR x ρ_{XY_Foc} x ρ_{ZY_Ratio}	—	.03
SR x ρ_{ZY_Ref} x ρ_{XZ}	.02	.01
SR x ρ_{ZY_Ratio} x ρ_{XZ}	.08	.04
SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.02	.03
ρ_{XY_Foc} x ρ_{ZY_Ref} x ρ_{XZ}	—	.02
ρ_{XY_Foc} x ρ_{ZY_Ratio} x ρ_{XZ}	—	.05
ρ_{XY_Foc} x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	—	.03
ρ_{ZY_Ref} x ρ_{ZY_Ratio} x ρ_{XZ}	.02	.04
Overall Effect of SR x ρ_{XY_Foc} x ρ_{XZ}	.31	.16
Overall Effect of SR x ρ_{XY_Foc} x ρ_{ZY_Ratio}	—	.25
Overall Effect of SR x ρ_{ZY_Ref} x ρ_{XZ}	.15	.15
Overall Effect of SR x ρ_{ZY_Ratio} x ρ_{XZ}	.50	.30
Overall Effect of SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.43	.24
Overall Effect of ρ_{XY_Foc} x ρ_{ZY_Ref} x ρ_{XZ}	—	.16
Overall Effect of ρ_{XY_Foc} x ρ_{ZY_Ratio} x ρ_{XZ}	—	.33
Overall Effect of ρ_{XY_Foc} x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	—	.26
Overall Effect of ρ_{ZY_Ref} x ρ_{ZY_Ratio} x ρ_{XZ}	.42	.30
Total	.99	.78
Restricted Total	.98	.76

Note. Overall Effect = total variance explained by an interaction effect and all lower-order effects subsumed by the interaction; Total = total variance explained by all effects in the summary model; Restricted Total = total variance explained by all effects that explained at least 1% of variance in the dependent variable, plus the variance explained by any lower-order effects implicated in interactions that met the 1% threshold. $\Delta\beta_{\Delta Slopes}$ = change in slope-difference regression coefficients; ΔSign of $\beta_{\Delta Slopes}$ = change in the signs of slope-difference regression coefficients from scenarios with slope differences; SR = overall selection ratio applied to Z; ρ_{XY_Foc} = operational validity of X for

predicting *Y* in the focal group's applicant population; ρ_{ZY_Ref} = operational validity of *Z* for predicting *Y* in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of *Z* for predicting *Y* in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between *X* and *Z* in the referent and focal groups' applicant populations.

Table 35

Summary of Variance Explained in Indirectly Range Restricted Observed-Operational Differences in Power as Indicated by Normalized F *Ratios*

Parameter	$\Delta F_{\Delta Overall}$	$\Delta F_{\Delta Slopes}$	$\Delta F_{\Delta Intercepts}$
SR	.09	.04	.10
P_{Ref}	.01	.05	.01
ρ_{XY_Foc}	.01	.20	—
ρ_{ZY_Ref}	.07	—	.07
ρ_{ZY_Ratio}	.10	.00	.10
ρ_{XZ}	—	.01	—
δ_Y	.01	—	.03
δ_Z	.03	.05	.01
ρ_{YY}	.04	.14	.02
SR x P_{Ref}	.02	—	.02
SR x ρ_{XY_Foc}	—	.01	—
SR x ρ_{ZY_Ref}	.07	—	.07
SR x ρ_{ZY_Ratio}	.10	.01	.11
SR x δ_Y	.01	—	.02
SR x δ_Z	.03	.04	.01
P_{Ref} x ρ_{XY_Foc}	—	.02	—
P_{Ref} x ρ_{ZY_Ref}	.01	—	—
P_{Ref} x ρ_{ZY_Ratio}	.01	—	.01
P_{Ref} x ρ_{YY}	—	.02	—
ρ_{XY_Foc} x δ_Z	—	.02	—
ρ_{XY_Foc} x ρ_{YY}	—	.06	—
ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.06	—	.06
ρ_{ZY_Ref} x δ_Z	.01	—	.01
ρ_{ZY_Ratio} x δ_Y	—	—	.02
ρ_{ZY_Ratio} x δ_Z	—	.02	—
ρ_{ZY_Ratio} x ρ_{YY}	—	.02	—
SR x P_{Ref} x ρ_{ZY_Ref}	.01	—	—
SR x P_{Ref} x ρ_{ZY_Ratio}	.02	—	.02
SR x ρ_{XY_Foc} x δ_Z	—	.02	—
SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.06	—	.07
SR x ρ_{ZY_Ref} x δ_Z	.01	—	.01
SR x ρ_{ZY_Ratio} x δ_Y	—	—	.01
SR x ρ_{ZY_Ratio} x δ_Z	—	.01	—

Table continues.

Table 35 (Continued)

Summary of Variance Explained in Indirectly Range Restricted Observed-Operational Differences in Power as Indicated by Normalized F *Ratios*

Parameter	$\Delta F_{\Delta Overall}$	$\Delta F_{\Delta Slopes}$	$\Delta F_{\Delta Intercepts}$
Main Effect of ρ_{XY_Foc}	.01	—	—
Main Effect of ρ_{XZ}	—	.01	—
Main Effect of ρ_{YY}	.04	—	.02
Overall Effect of SR x δ_Y	.11	—	—
Overall Effect of P_{Ref} x ρ_{XY_Foc}	—	.27	—
Overall Effect of P_{Ref} x ρ_{YY}	—	.21	—
Overall Effect of ρ_{XY_Foc} x ρ_{YY}	—	.39	—
Overall Effect of ρ_{ZY_Ratio} x ρ_{YY}	—	.16	—
Overall Effect of SR x P_{Ref} x ρ_{ZY_Ref}	.28	—	—
Overall Effect of SR x P_{Ref} x ρ_{ZY_Ratio}	.35	—	.37
Overall Effect of SR x ρ_{XY_Foc} x δ_Z	—	.38	—
Overall Effect of SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.56	—	.58
Overall Effect of SR x ρ_{ZY_Ref} x δ_Z	.31	—	.28
Overall Effect of SR x ρ_{ZY_Ratio} x δ_Y	—	—	.39
Overall Effect of SR x ρ_{ZY_Ratio} x δ_Z	—	.18	—
Total	.94	.92	.93
Restricted Total	.79	.75	.77

Note. Overall Effect = total variance explained by an interaction effect and all lower-order effects subsumed by the interaction; Total = total variance explained by all effects in the summary model; Restricted Total = total variance explained by all effects that explained at least 1% of variance in the dependent variable, plus the variance explained by any lower-order effects implicated in interactions that met the 1% threshold. $\Delta F_{\Delta Overall}$ = change in power for tests of overall differences in prediction as indicated by normalized differences in *F* ratios; $\Delta F_{\Delta Slopes}$ = change in power for tests of slope differences as indicated by normalized differences in *F* ratios; $\Delta F_{\Delta Intercepts}$ = change in power for tests of intercept differences as indicated by normalized differences in *F* ratios; SR = overall selection ratio applied to *Z*; P_{Ref} = proportion of referent-group members in the applicant population; ρ_{XY_Foc} = operational validity of *X* for predicting *Y* in the focal group's applicant population; ρ_{ZY_Ref} = operational validity of *Z* for predicting *Y* in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of *Z* for predicting *Y* in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between *X* and *Z* in the referent and focal groups' applicant populations; δ_Y = true-score standardized mean difference between the referent and focal groups on *Y*; δ_Z = standardized mean difference between the referent and focal groups on *Z*; ρ_{YY} = reliability of *Y*.

Table 36

Summary of Variance Explained in Indirectly Range Restricted Observed-Operational Differences in Type I Errors as Indicated by Normalized F *Ratios*

Parameter	$\Delta F_{\Delta Overall}$	$\Delta F_{\Delta Slopes}$	$\Delta F_{\Delta Intercepts}$
SR	.13	.04	.14
P_{Ref}	.02	.02	.02
ρ_{ZY_Ref}	.08	.03	.08
ρ_{ZY_Ratio}	.12	.06	.12
ρ_{XZ}	—	.05	—
δ_Z	.01	.01	.01
ρ_{YY}	.01	.01	.01
SR x P_{Ref}	.02	.01	.02
SR x ρ_{XZ}	—	.04	—
SR x ρ_{ZY_Ref}	.08	.02	.08
SR x ρ_{ZY_Ratio}	.12	.04	.13
SR x δ_Z	.01	.01	.01
SR x ρ_{YY}	.01	—	.01
P_{Ref} x ρ_{XZ}	—	.02	—
P_{Ref} x ρ_{ZY_Ref}	.01	—	.01
P_{Ref} x ρ_{ZY_Ratio}	.02	.02	.02
ρ_{ZY_Ref} x ρ_{XZ}	—	.03	—
ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.07	.03	.07
ρ_{ZY_Ref} x δ_Z	.01	—	.01
ρ_{ZY_Ratio} x ρ_{XZ}	—	.06	—
ρ_{ZY_Ratio} x δ_Z	.01	.01	—
ρ_{ZY_Ratio} x ρ_{YY}	.01	.01	.01
ρ_{XZ} x δ_Z	—	.01	—
ρ_{XZ} x ρ_{YY}	—	.01	—
SR x P_{Ref} x ρ_{XZ}	—	.01	—
SR x P_{Ref} x ρ_{ZY_Ref}	.01	—	.01
SR x P_{Ref} x ρ_{ZY_Ratio}	.02	.01	.02
SR x ρ_{ZY_Ref} x ρ_{XZ}	—	.02	—
SR x ρ_{ZY_Ratio} x ρ_{XZ}	—	.05	—
SR x ρ_{XZ} x δ_Z	—	.01	—
SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.08	.02	.08
SR x ρ_{ZY_Ref} x δ_Z	.01	—	.01
SR x ρ_{ZY_Ratio} x δ_Z	.01	.01	—
P_{Ref} x ρ_{ZY_Ratio} x ρ_{XZ}	—	.02	—
P_{Ref} x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.01	—	—
ρ_{ZY_Ref} x ρ_{ZY_Ratio} x ρ_{XZ}	—	.03	—
ρ_{ZY_Ratio} x ρ_{XZ} x δ_Z	—	.01	—
ρ_{ZY_Ratio} x ρ_{XZ} x ρ_{YY}	—	.01	—

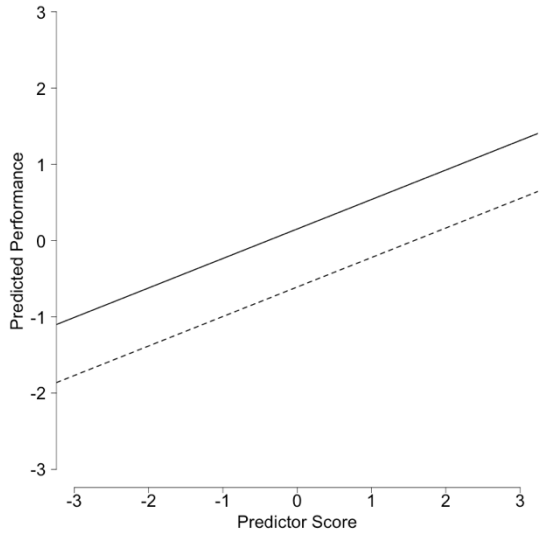
Table Continues.

Table 36 (Continued)

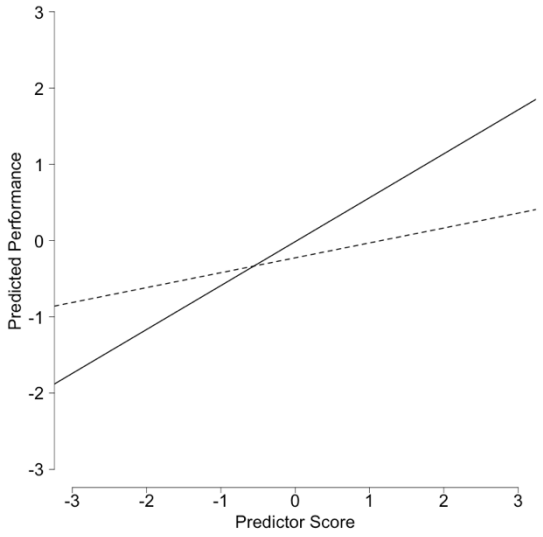
Summary of Variance Explained in Indirectly Range Restricted Observed-Operational Differences in Type I Errors as Indicated by Normalized F *Ratios*

Parameter	$\Delta F_{\Delta Overall}$	$\Delta F_{\Delta Slopes}$	$\Delta F_{\Delta Intercepts}$
Overall Effect of SR x ρ_{YY}	.15	—	.16
Overall Effect of ρ_{ZY_Ratio} x ρ_{YY}	.15	—	.14
Overall Effect of SR x P_{Ref} x ρ_{XZ}	—	.19	—
Overall Effect of SR x P_{Ref} x ρ_{ZY_Ref}	.34	—	.35
Overall Effect of SR x P_{Ref} x ρ_{ZY_Ratio}	.45	.20	.46
Overall Effect of SR x ρ_{ZY_Ref} x ρ_{XZ}	—	.24	—
Overall Effect of SR x ρ_{ZY_Ratio} x ρ_{XZ}	—	.35	—
Overall Effect of SR x ρ_{XZ} x δ_Z	—	.18	—
Overall Effect of SR x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.68	.25	.70
Overall Effect of SR x ρ_{ZY_Ref} x δ_Z	.32	—	.33
Overall Effect of SR x ρ_{ZY_Ratio} x δ_Z	.41	.19	—
Overall Effect of P_{Ref} x ρ_{ZY_Ratio} x ρ_{XZ}	—	.25	—
Overall Effect of P_{Ref} x ρ_{ZY_Ref} x ρ_{ZY_Ratio}	.33	—	—
Overall Effect of ρ_{ZY_Ref} x ρ_{ZY_Ratio} x ρ_{XZ}	—	.30	—
Overall Effect of ρ_{ZY_Ratio} x ρ_{XZ} x δ_Z	—	.22	—
Overall Effect of ρ_{ZY_Ratio} x ρ_{XZ} x ρ_{YY}	—	.22	—
Total	.95	.85	.95
Restricted Total	.88	.77	.86

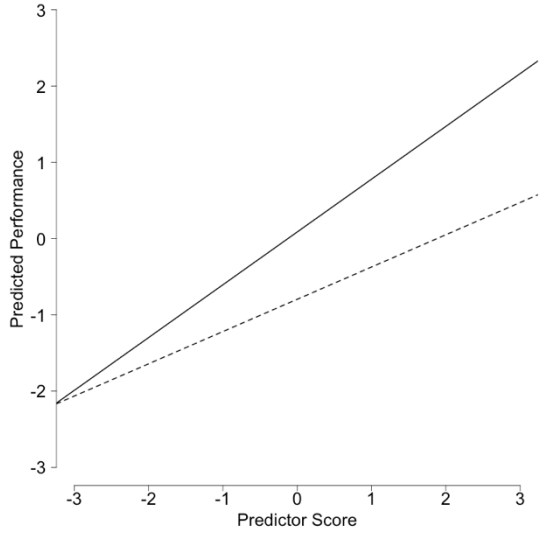
Note. Overall Effect = total variance explained by an interaction effect and all lower-order effects subsumed by the interaction; Total = total variance explained by all effects in the summary model; Restricted Total = total variance explained by all effects that explained at least 1% of variance in the dependent variable, plus the variance explained by any lower-order effects implicated in interactions that met the 1% threshold. $\Delta F_{\Delta Overall}$ = change in Type I errors for tests of overall differences in prediction as indicated by normalized differences in *F* ratios; $\Delta F_{\Delta Slopes}$ = change in Type I errors for tests of slope differences as indicated by normalized differences in *F* ratios; $\Delta F_{\Delta Intercepts}$ = change in Type I errors for tests of intercept differences as indicated by normalized differences in *F* ratios; SR = overall selection ratio applied to Z; P_{Ref} = proportion of referent-group members in the applicant population; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between X and Z in the referent and focal groups' applicant populations; δ_Z = standardized mean difference between the referent and focal groups on Z; ρ_{YY} = reliability of Y.



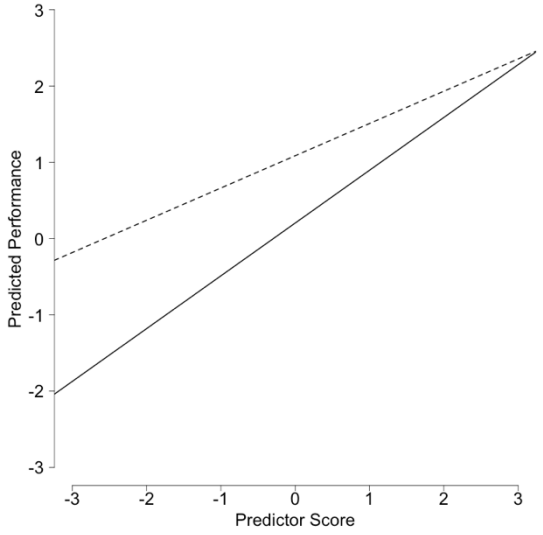
(A) Intercept bias: Minority overprediction



(B) Slope bias: Mixed patterns of over- and under-prediction



(C) Slope bias: Consistent minority overprediction



(D) Slope bias: Consistent minority underprediction

Figure 1

Examples of predictive bias scenarios for hypothetical predictors.

Solid lines represent the majority group's regression equations and dashed lines represent the minority group's regression equations.

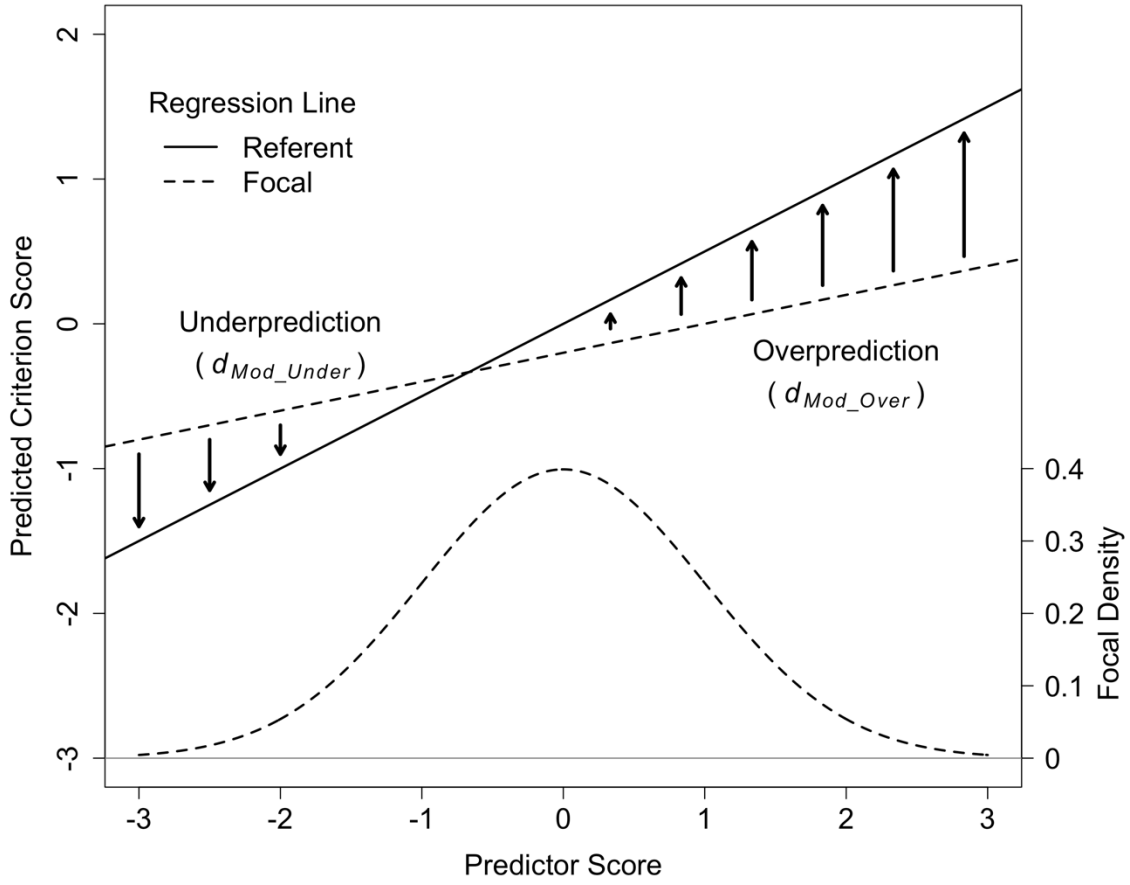


Figure 2
Conceptual illustration of d_{Mod} using hypothetical data with arbitrary parameters. Referent slope = .5; focal slope = .2, referent intercept = 0, focal intercept = -.2, focal predictor mean = 0; focal predictor *SD* = 1; referent criterion *SD* = 1; d_{Mod_Under} = -0.045; d_{Mod_Over} = 0.245; d_{Mod_Signed} = .200; proportion underpredicted = .25; proportion overpredicted = .75. Arrows indicate the magnitude and direction of differential prediction of focal group performance relative to the referent group line. The focal density function is plotted as a reference for interpreting the prevalence of different magnitudes of differential prediction across the range of predictor scores. To compute d_{Mod} effect sizes, the normal distribution is used to integrate the weighted average of the differences in prediction between the regression lines.

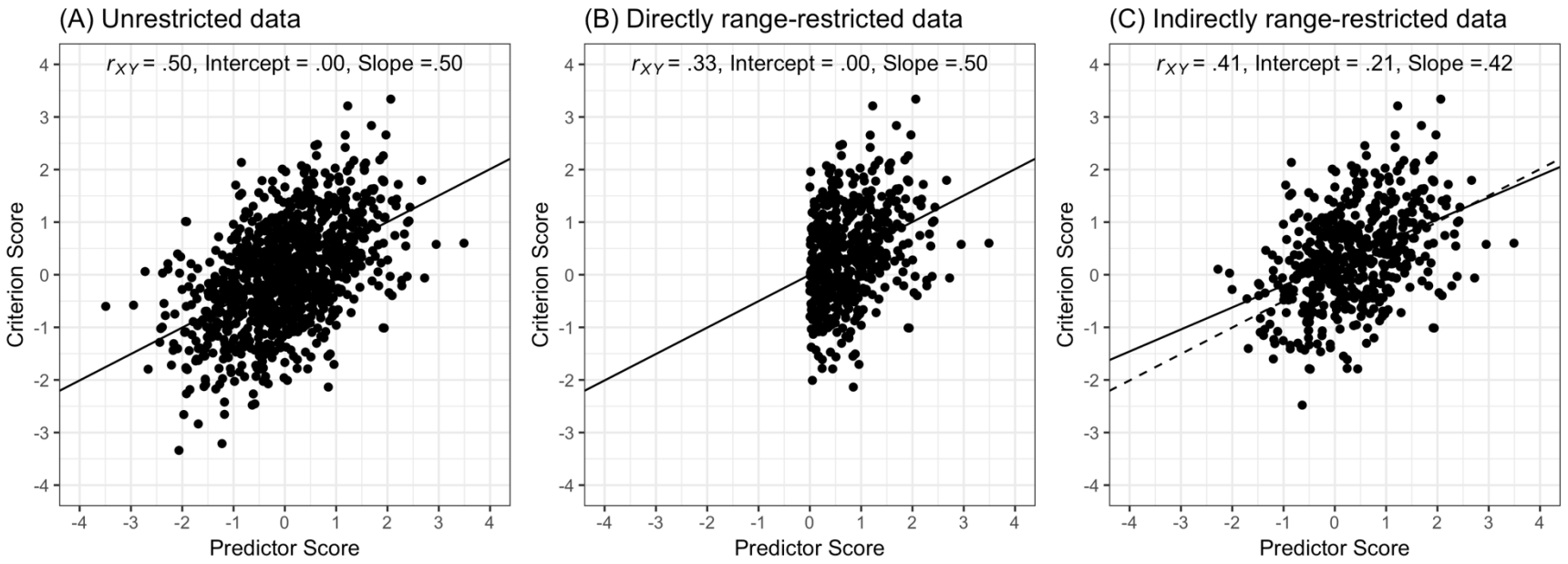


Figure 3

Demonstration of the effects of direct and indirect range restriction on validity and regression analyses.

Panel A shows unrestricted data in which the predictor and criterion correlate .5 and the regression line has an intercept of .00 and a slope of .50. Panel B shows directly range-restricted data in which only those cases with scores in the top 50% of the predictor distribution are selected; the regression coefficients are not biased by selection, but the validity estimate is attenuated. Panel C shows indirectly range-restricted data in which only those cases with scores in the top 50% of the distribution of a third variable are selected; this third variable correlates .5 with the predictor and the criterion in the unrestricted sample. Indirect range restriction attenuates the predictor's validity estimate and also biases the regression coefficients; the dashed line in Panel C shows the unrestricted regression line for reference.

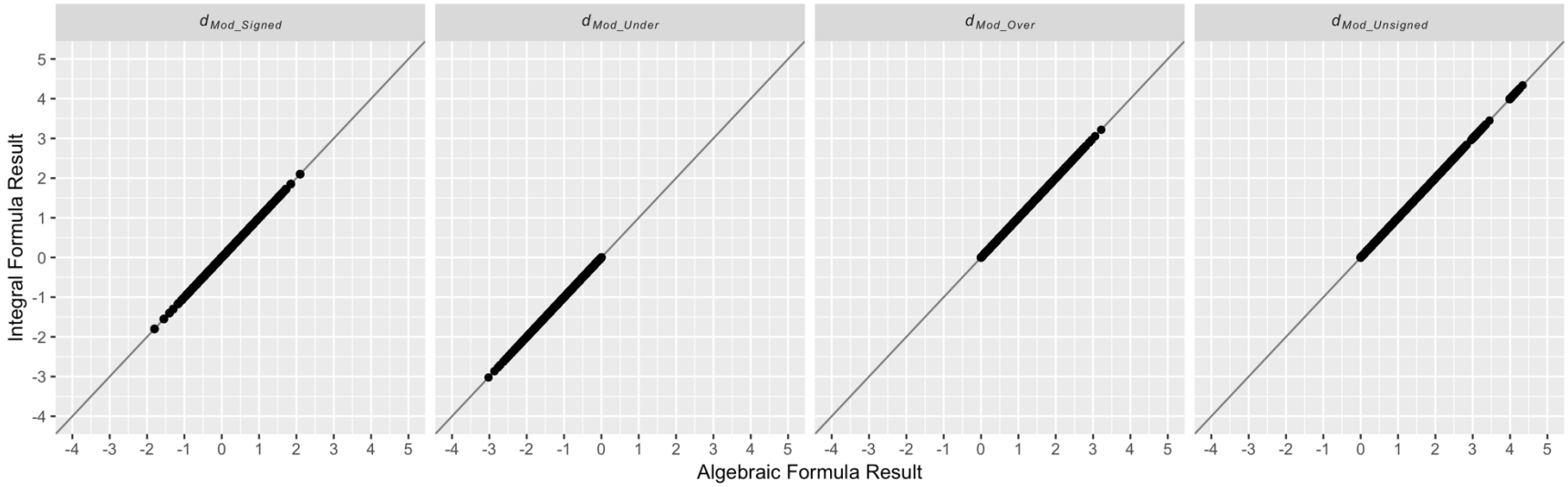


Figure 4
Correspondence between estimates of d_{Mod} effect sizes computed for simulated scenarios using new algebraic formulas derived in Study 1 and the integration-based formulas presented by Nye and Sackett (2017) and Dahlke and Sackett (2018).

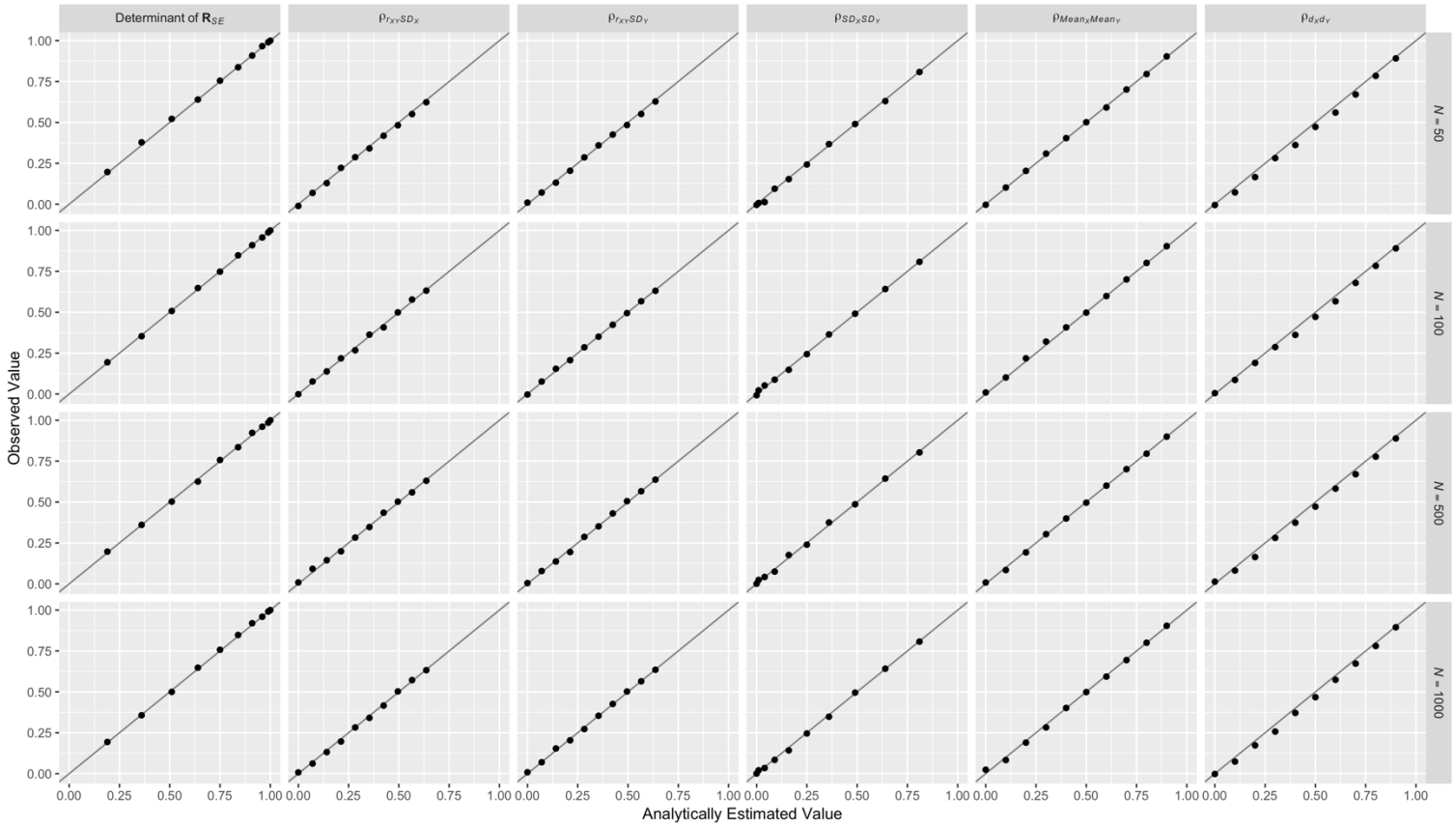


Figure 5

Correspondence between analytically estimated correlations between sampling distributions and mean observed Monte Carlo estimates. Columns of the plot grid differentiate the type of parameter being estimated and rows of the plot grid report results for different sample sizes. r_{XY} is the correlation between the observed scores for X and the observed scores for Y, SD_X and SD_Y are the standard deviations of observed scores for X and Y, respectively, $Mean_X$ and $Mean_Y$ are the means of observed scores for X and Y, respectively, d_X and d_Y are mean differences of observed scores between two groups for X and Y, respectively. "Determinant of $\mathbf{R}_{SE}$" is the determinant of the correlation matrix describing the sampling distributions of r_{XY}, SD_X, and SD_Y.

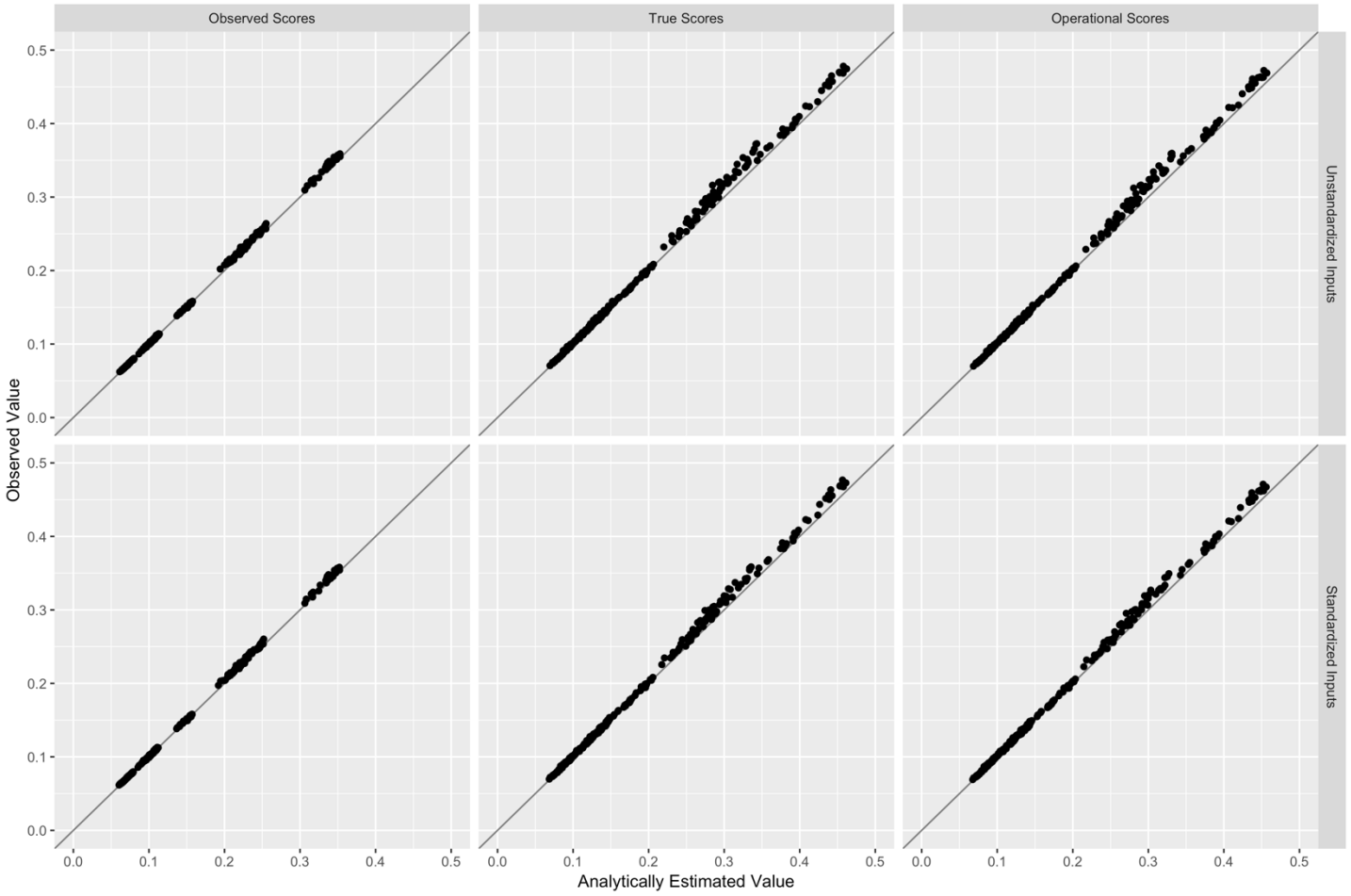


Figure 6

Correspondence between analytically estimated standard errors of d_{Mod_Signed} and standard deviations of Monte Carlo estimates. "Observed Scores" represent data in which the both the predictor and criterion are measured with error, "True Scores" represent data in which both the predictor and criterion are corrected for measurement error, and "Operational Scores" represent data in which the predictor is measured with error and the criterion is corrected for measurement error.

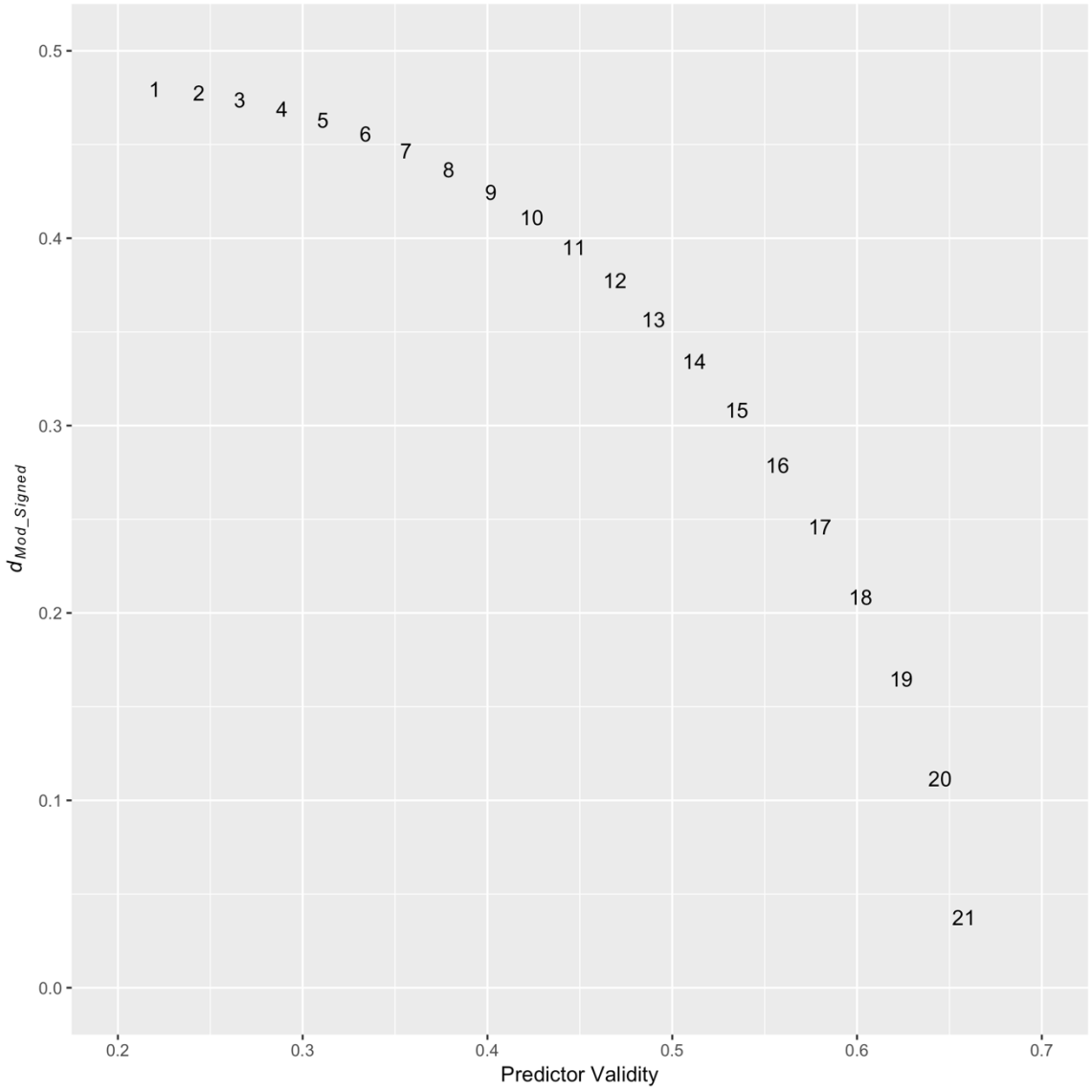


Figure 7
Association between validity coefficients and d_{Mod_Signed} effect sizes for Pareto-optimal composites shown in Table 6.
The numeric labels used as data points in the plot correspond to the "Pareto solution #" identifiers assigned to the Pareto-optimal solutions in Table 6.

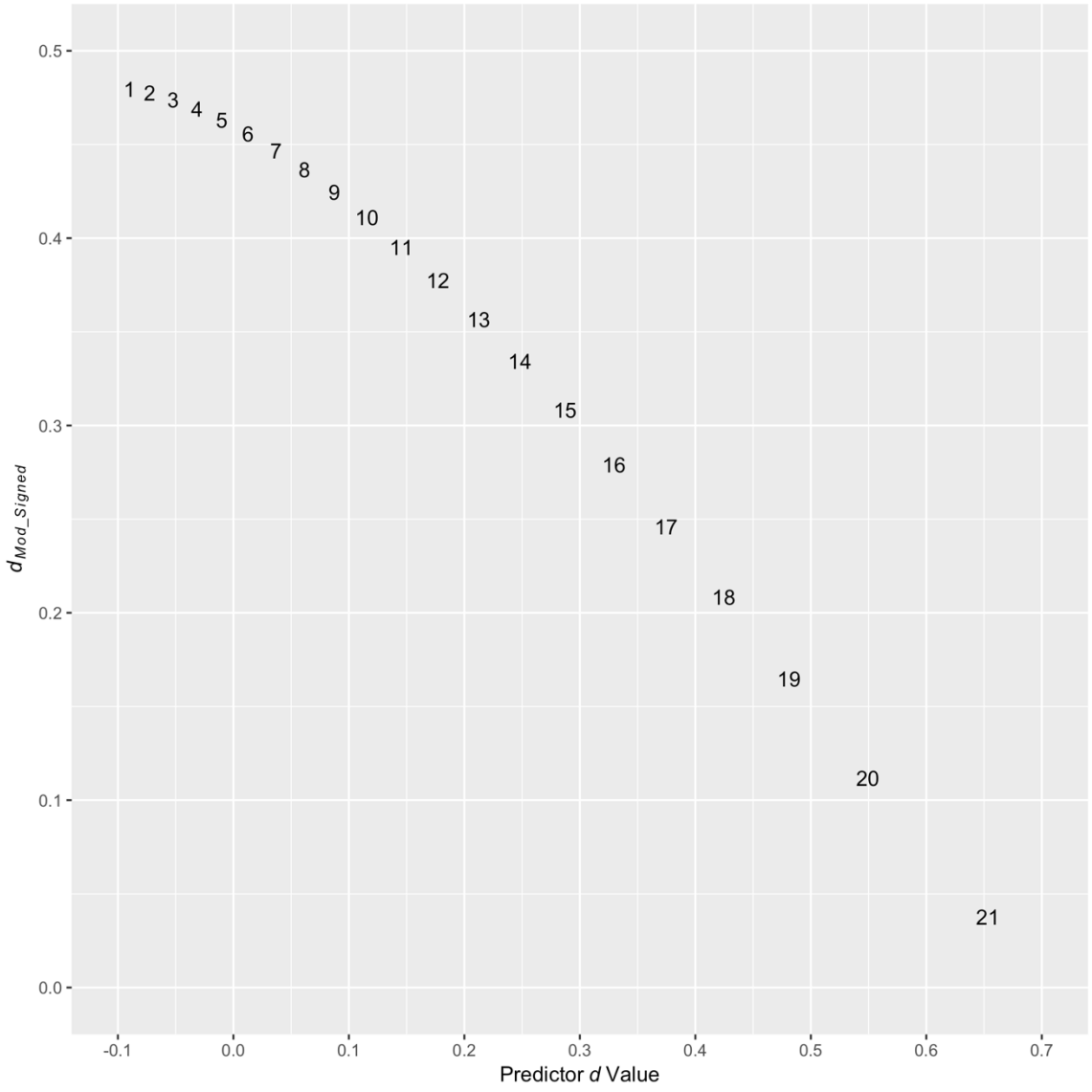


Figure 8
Association between predictor *d* values and d_{Mod_Signed} effect sizes for Pareto-optimal composites shown in Table 6.
The numeric labels used as data points in the plot correspond to the "Pareto solution #" identifiers assigned to the Pareto-optimal solutions in Table 6.

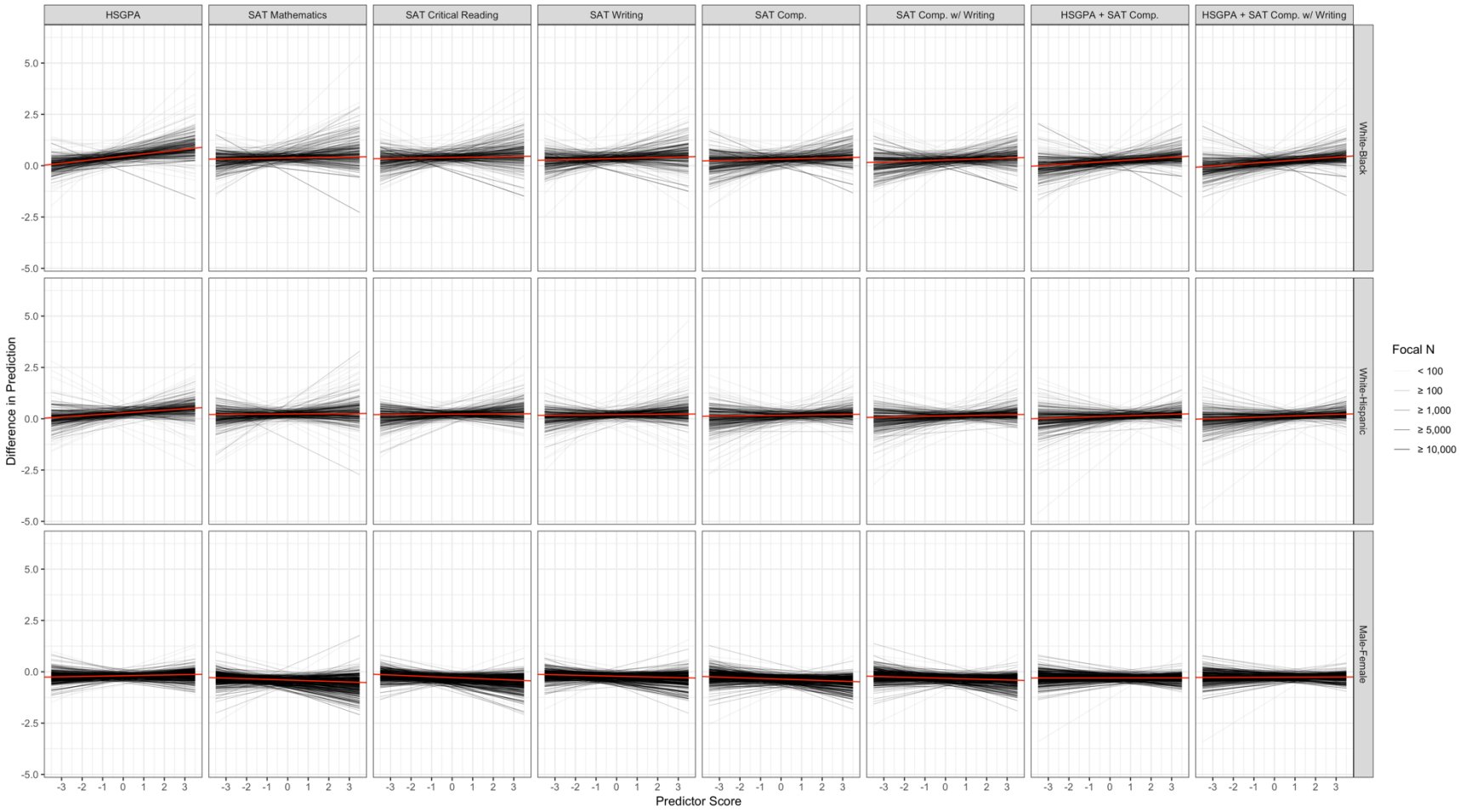


Figure 9

Plot of magnitudes in differences in prediction between subgroups over the operational range of predictor scores for observed data. Positive (negative) differences indicate overprediction (underprediction) of the focal group's performance when the referent group's regression formula is used to forecast performance. Samples with more focal group members are plotted with darker lines. Red lines indicate meta-analytic averages in differential prediction.

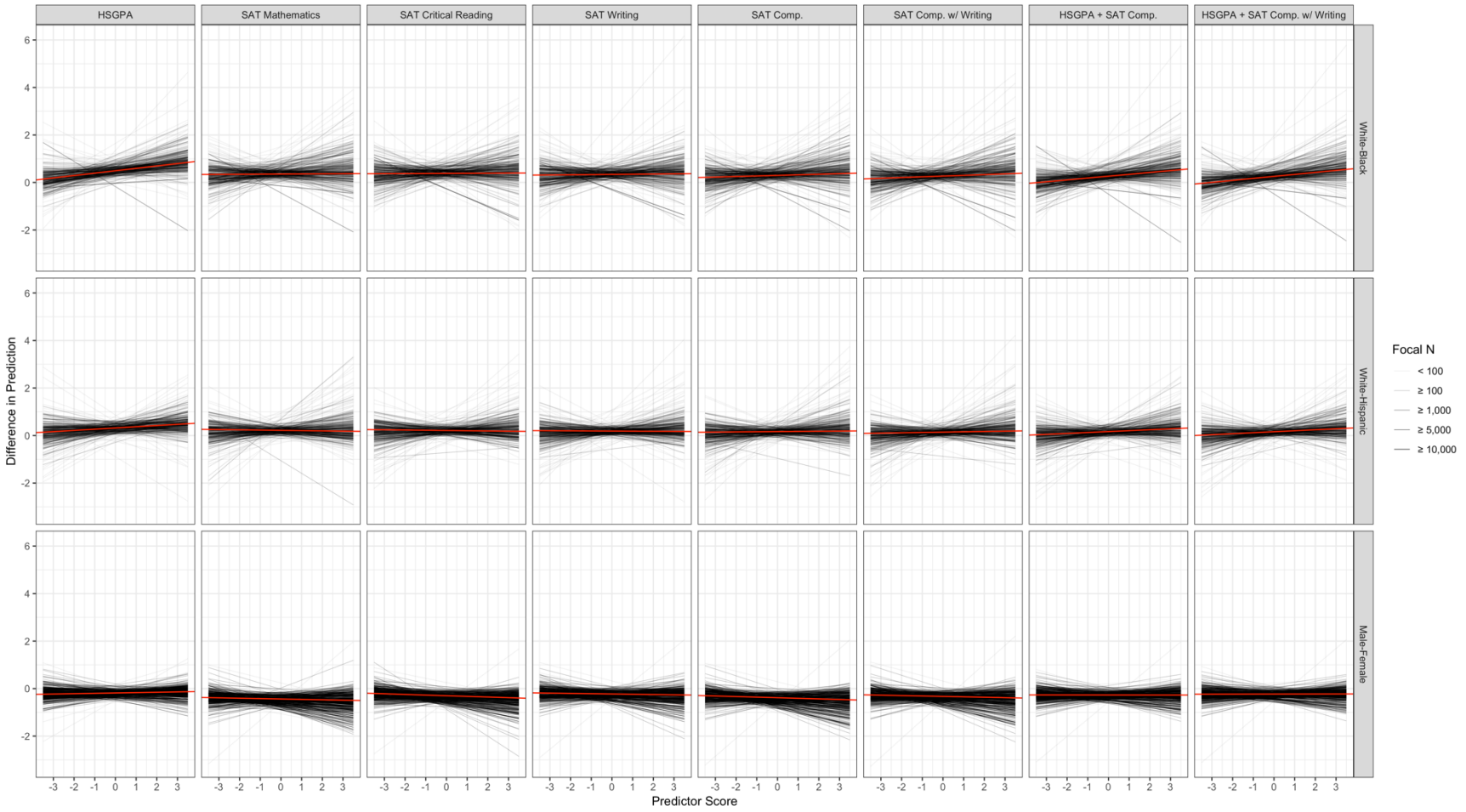


Figure 10

Plot of magnitudes in differences in prediction between subgroups over the operational range of predictor scores for range-restriction corrected data. Positive (negative) differences indicate overprediction (underprediction) of the focal group's performance when the referent group's regression formula is used to forecast performance. Samples with more focal group members are plotted with darker lines. Red lines indicate meta-analytic averages in differential prediction.

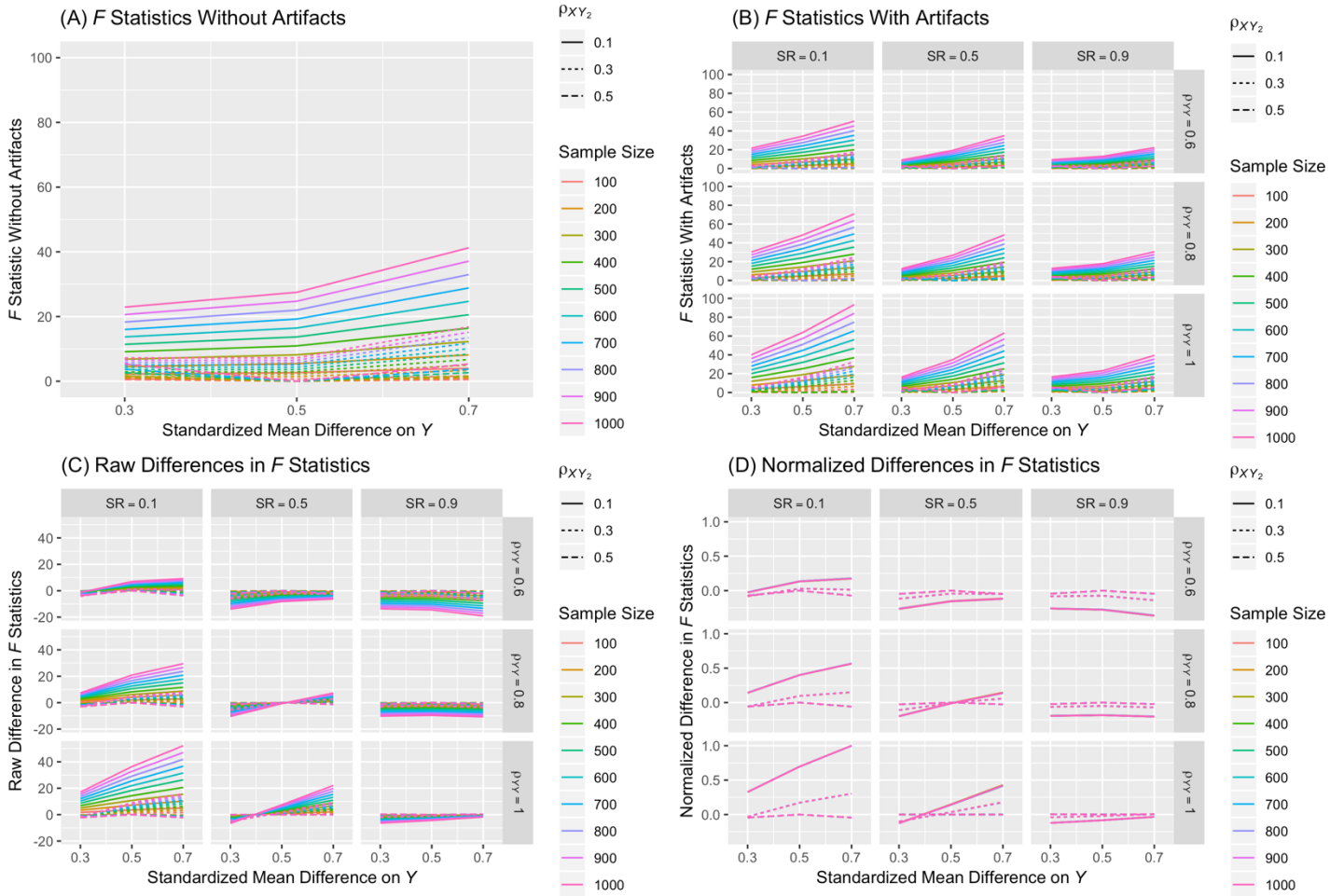


Figure 11

Demonstration of the equivalence of normalized differences between observed and operational *F* ratio parameters for comparisons of Model 3 and Model 1 in the Cleary framework across sample sizes (tests of overall differential prediction).

Panel A shows the operational *F* parameters regardless of artifacts, panel B shows observed *F* parameters with artifacts, panel C shows the raw differences between observed and operational *F* values, and panel D shows normalized differences (all lines overlap).

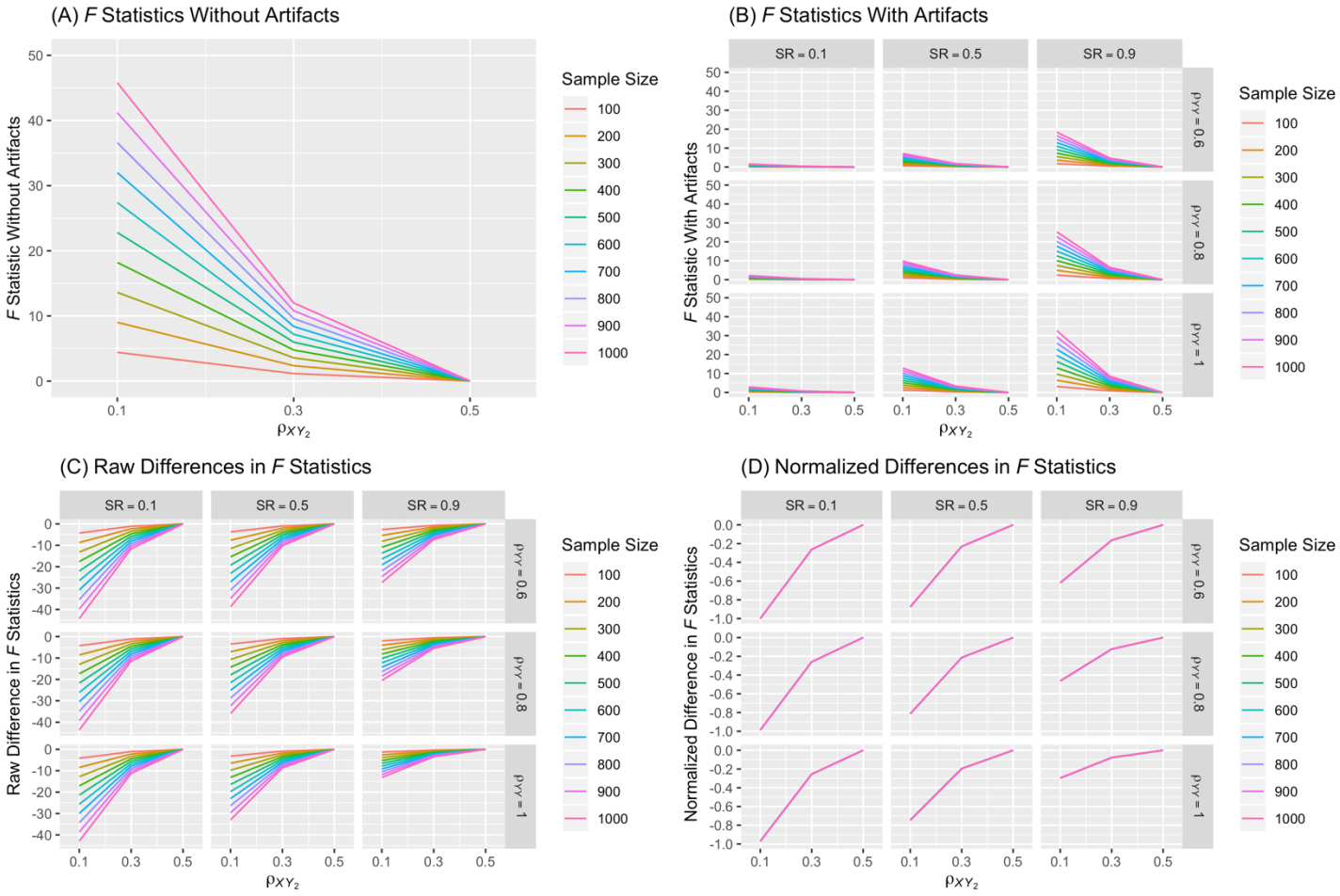


Figure 12

Demonstration of the equivalence of normalized differences between observed and operational *F* ratio parameters for comparisons of Model 3 and Model 2 in the Cleary framework across sample sizes (tests of slope differences).

Panel A shows the operational *F* parameters regardless of artifacts, panel B shows observed *F* parameters with artifacts, panel C shows the raw differences between observed and operational *F* values, and panel D shows normalized differences (all lines overlap).

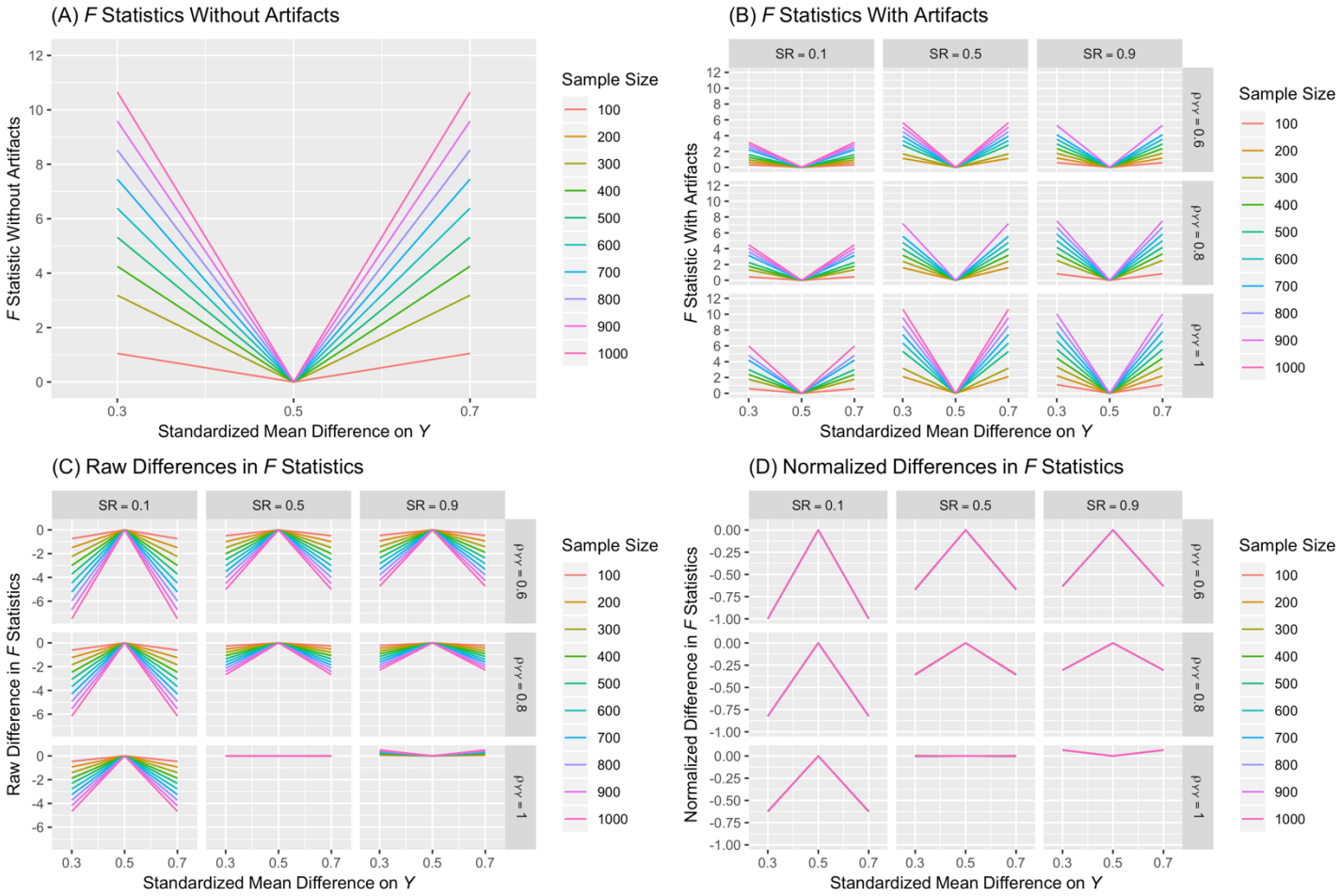


Figure 13

Demonstration of the equivalence of normalized differences between observed and operational *F* ratio parameters for comparisons of Model 2 and Model 1 in the Cleary framework across sample sizes (tests of intercept differences).

Panel A shows the operational *F* parameters regardless of artifacts, panel B shows observed *F* parameters with artifacts, panel C shows the raw differences between observed and operational *F* values, and panel D shows normalized differences (all lines overlap).

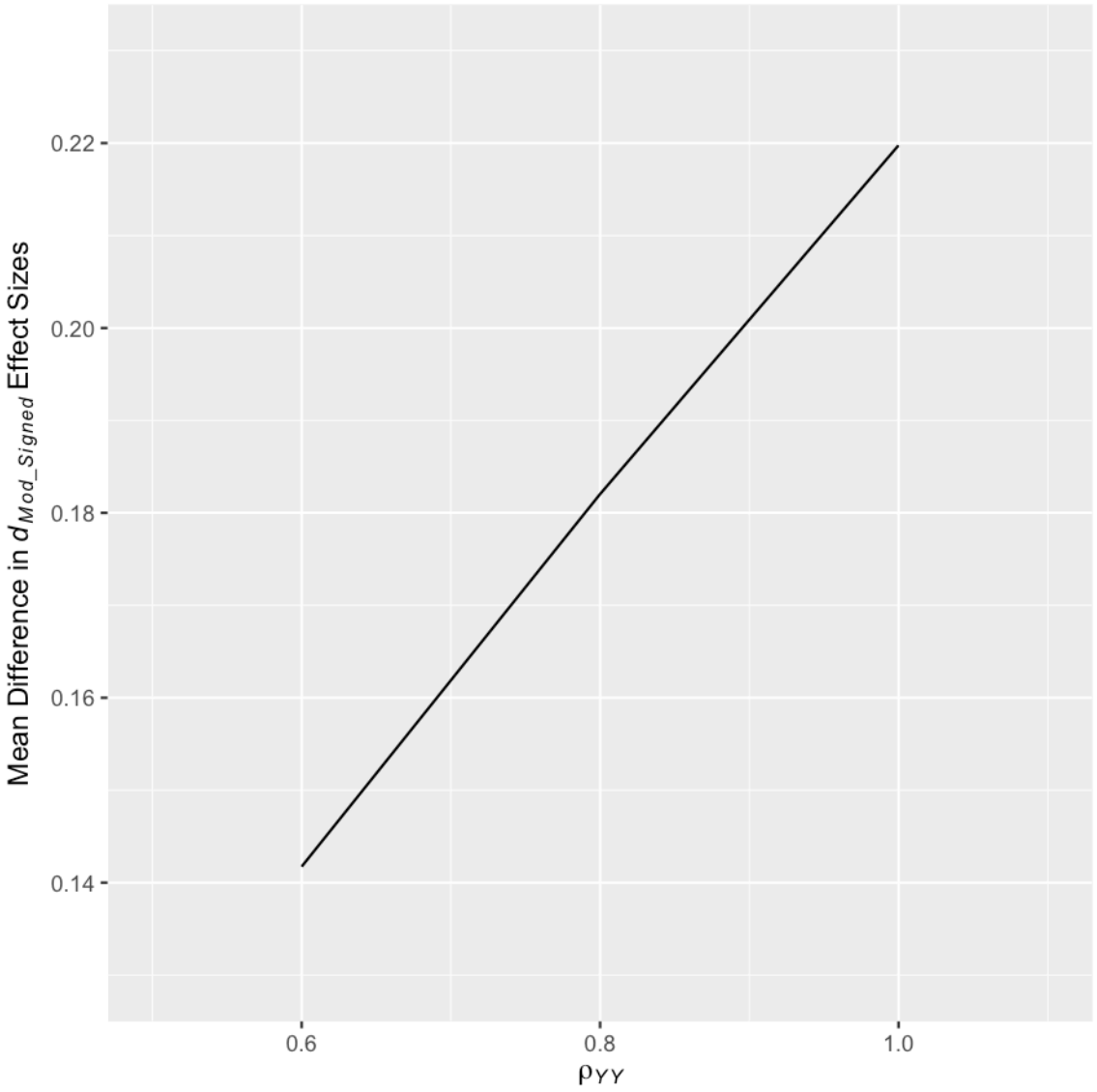


Figure 14
Main effect of ρ_{YY} on d_{Mod_Signed} effect sizes under conditions of direct range restriction.
Positive values indicate higher d_{Mod_Signed} effect sizes estimated from observed parameters than from operational parameters. ρ_{YY} = reliability of *Y*.
Figure is based on data from 243 (100.0%) direct range restriction conditions.
Total η^2 of plotted effects = .01.

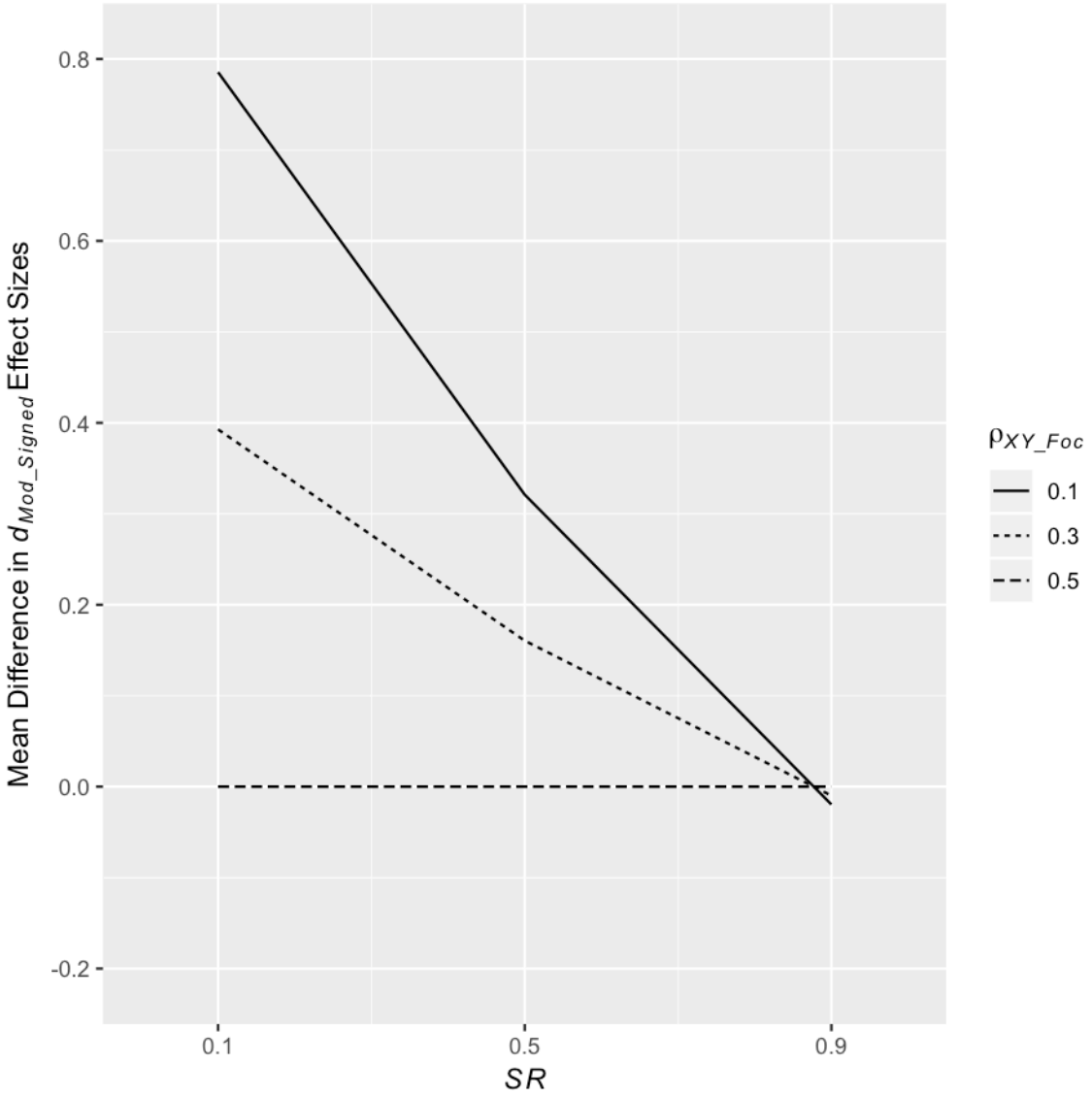


Figure 15

Effect of the two-way interaction between SR and ρ_{XY_Foc} on d_{Mod_Signed} effect sizes under conditions of direct range restriction.

Positive values indicate higher d_{Mod_Signed} effect sizes estimated from observed parameters than from operational parameters. SR = overall selection ratio applied to X; ρ_{XY_Foc} = operational validity of X for predicting Y in the focal group's applicant population.

Figure is based on data from 243 (100.0%) direct range restriction conditions.

Total η^2 of plotted effects = .96.

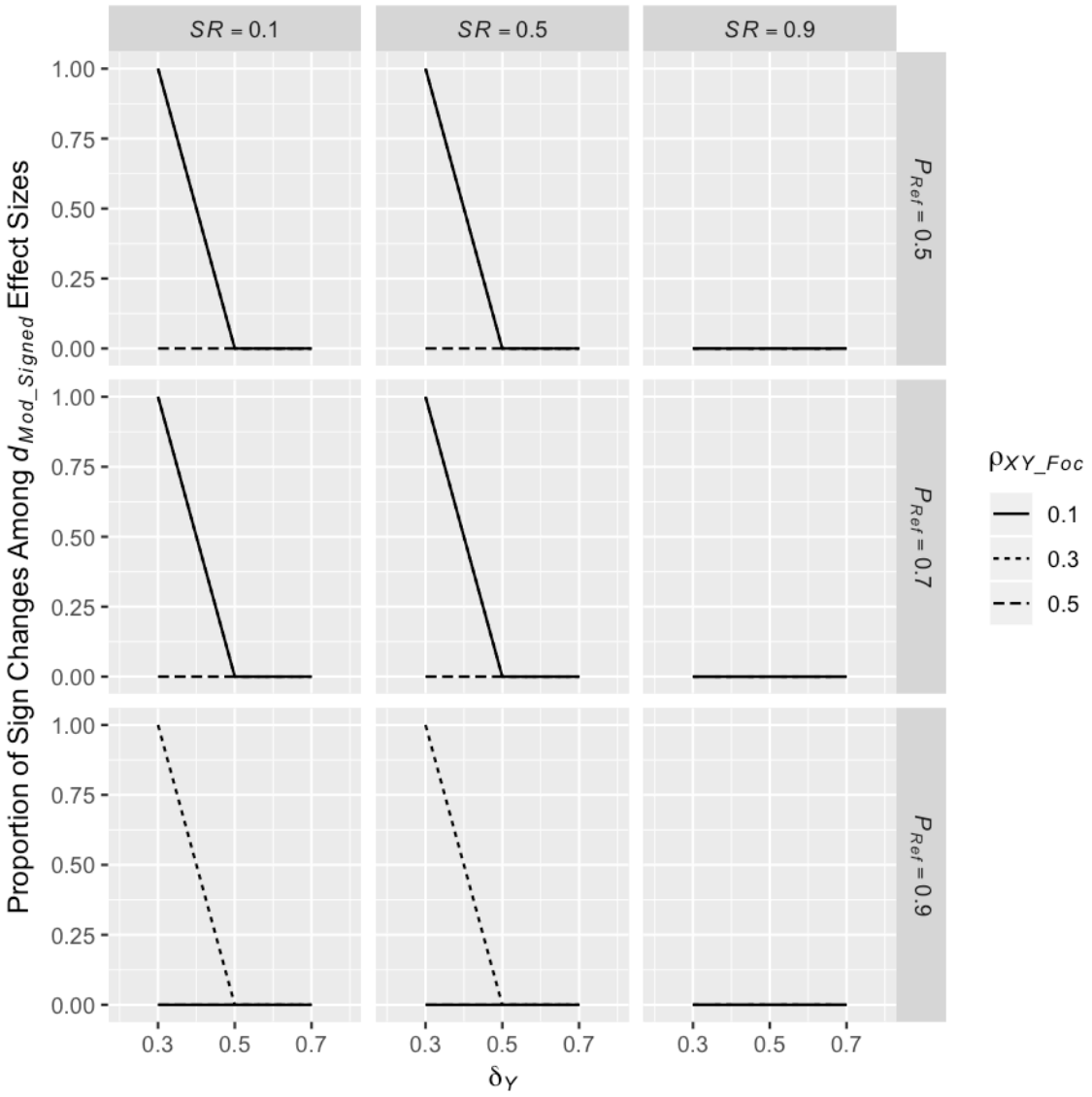


Figure 16

Effect of the four-way interaction among SR, P_{Ref}, ρ_{XY_Foc}, and δ_Y on the signs of d_{Mod_Signed} effect sizes under conditions of direct range restriction.

Larger values indicate higher proportions of sign differences between observed and operational parameters. SR = overall selection ratio applied to X; P_{Ref} = proportion of referent-group members in the applicant population; ρ_{XY_Foc} = operational validity of X for predicting Y in the focal group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on Y.

Figure is based on data from 243 (100.0%) direct range restriction conditions.

Total η^2 of plotted effects = .98.

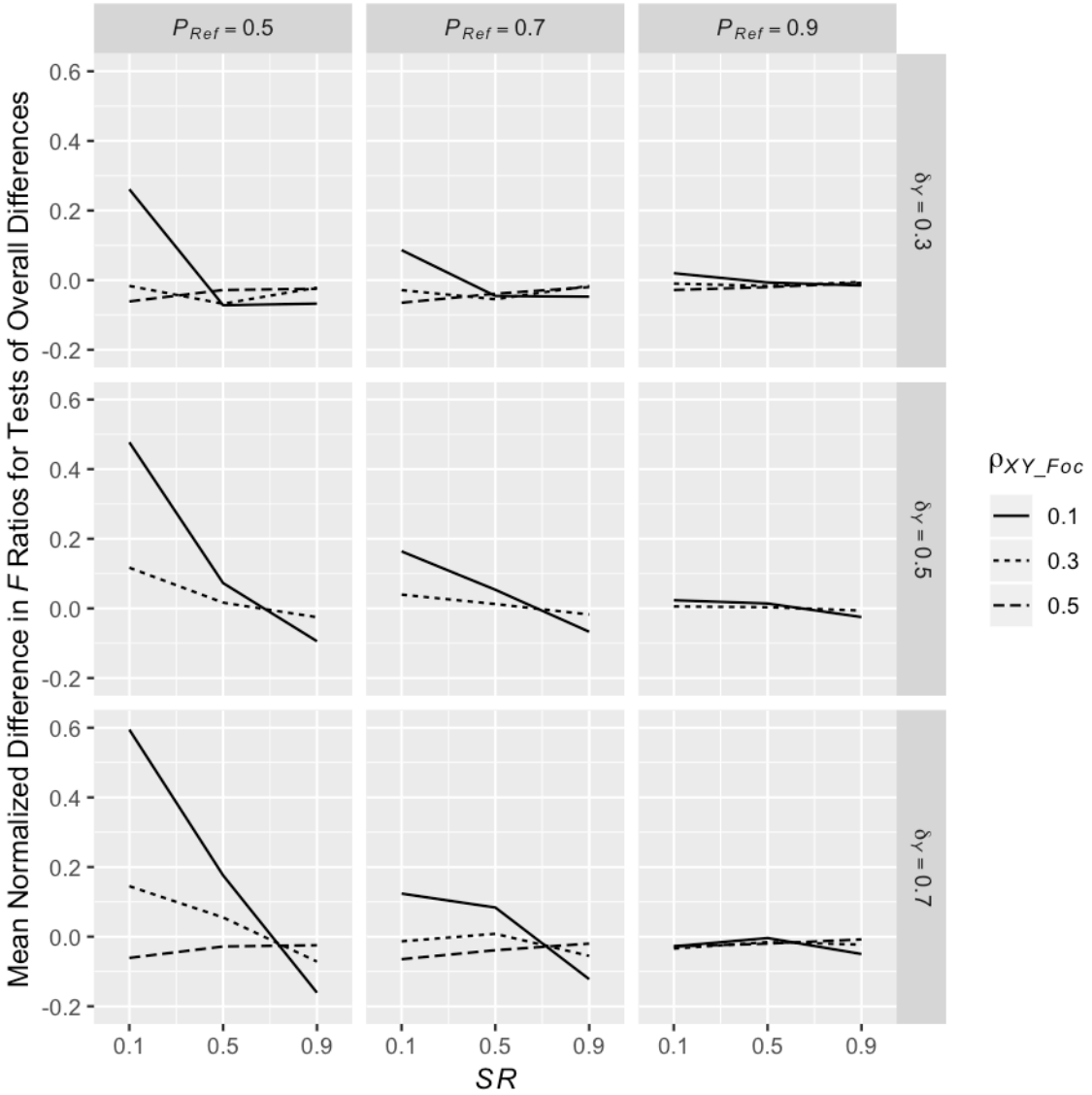


Figure 17

Effect of the four-way interaction among SR, P_{Ref}, ρ_{XY_Foc}, and δ_Y on the power of tests of overall differences in prediction (as indicated by normalized differences in *F* ratios) under conditions of direct range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to X; P_{Ref} = proportion of referent-group members in the applicant population; ρ_{XY_Foc} = operational validity of X for predicting Y in the focal group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on Y.

Figure is based on data from 216 (88.9%) direct range restriction conditions.

Total η^2 of plotted effects = .68.

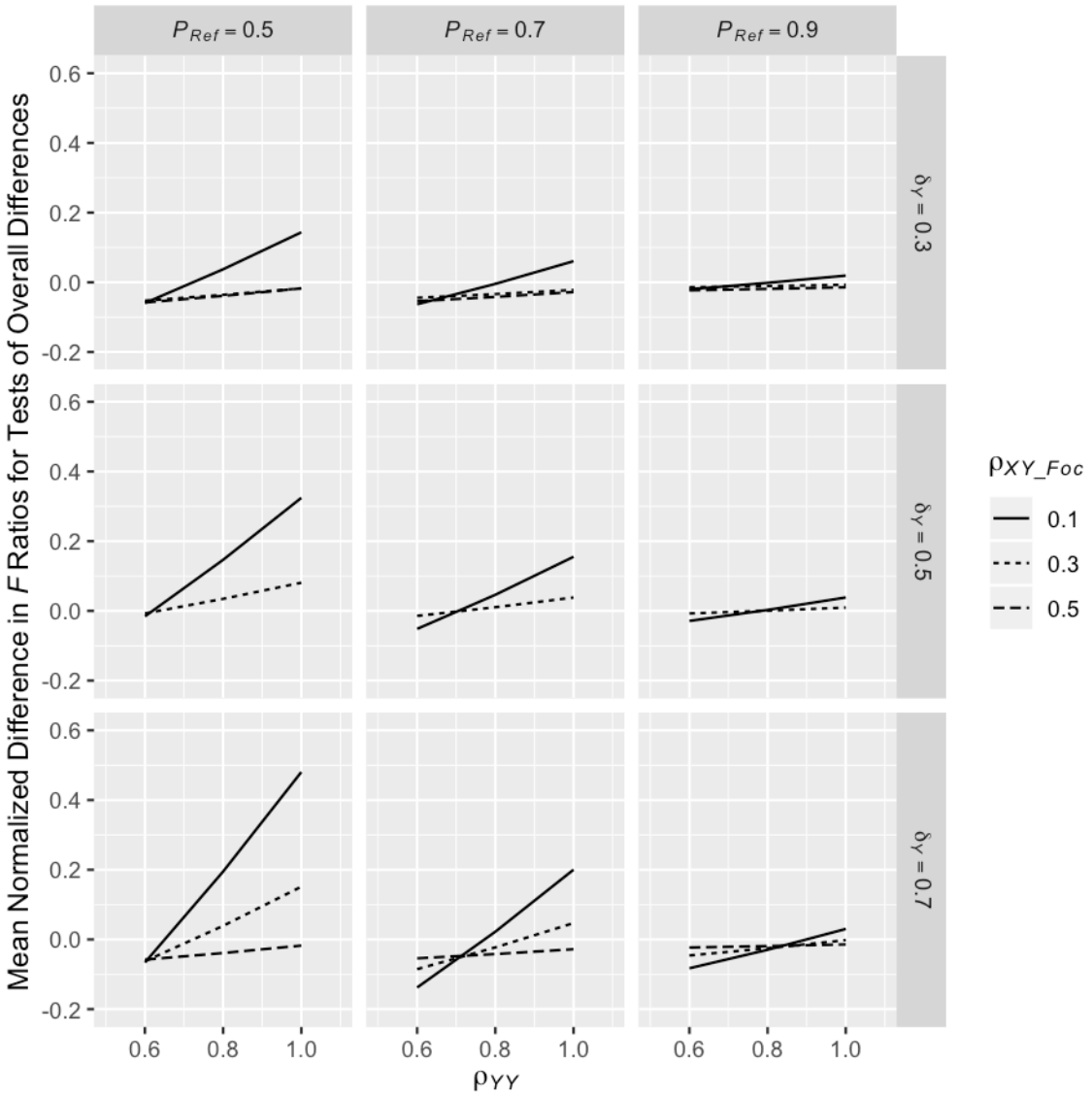


Figure 18

Effect of the four-way interaction among P_{Ref}, ρ_{XY_Foc}, δ_Y, and ρ_{YY} on the power of tests of overall differences in prediction (as indicated by normalized differences in *F* ratios) under conditions of direct range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. P_{Ref} = proportion of referent-group members in the applicant population; ρ_{XY_Foc} = operational validity of *X* for predicting *Y* in the focal group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on *Y*; ρ_{YY} = reliability of *Y*.

Figure is based on data from 216 (88.9%) direct range restriction conditions.

Total η^2 of plotted effects = .46.

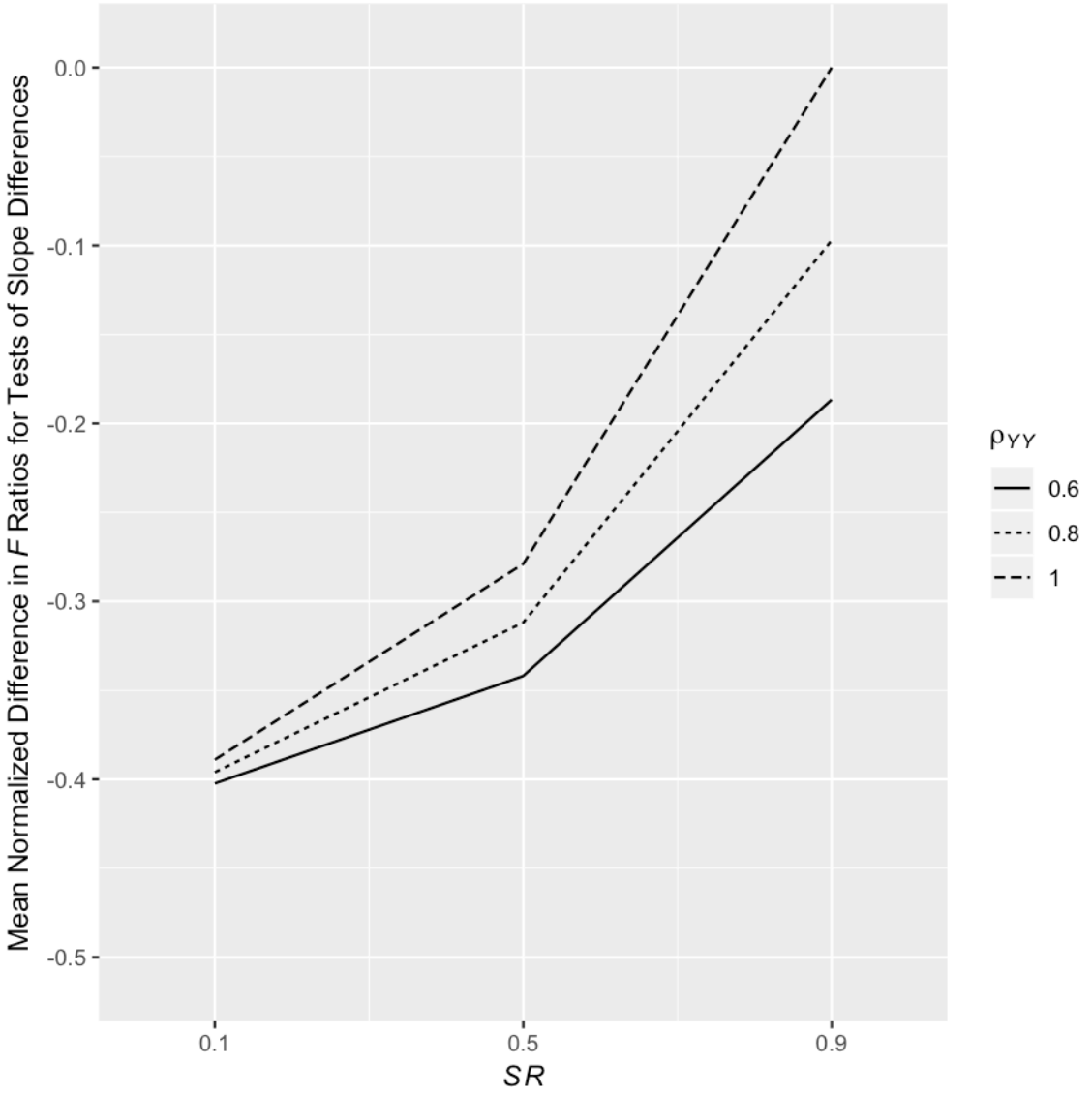


Figure 19

Effect of the two-way interaction between SR and ρ_{YY} on the power of tests of slope differences (as indicated by normalized differences in F ratios) under conditions of direct range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to X; ρ_{YY} = reliability of Y.

Figure is based on data from 162 (66.7%) direct range restriction conditions.

Total η^2 of plotted effects = .23.

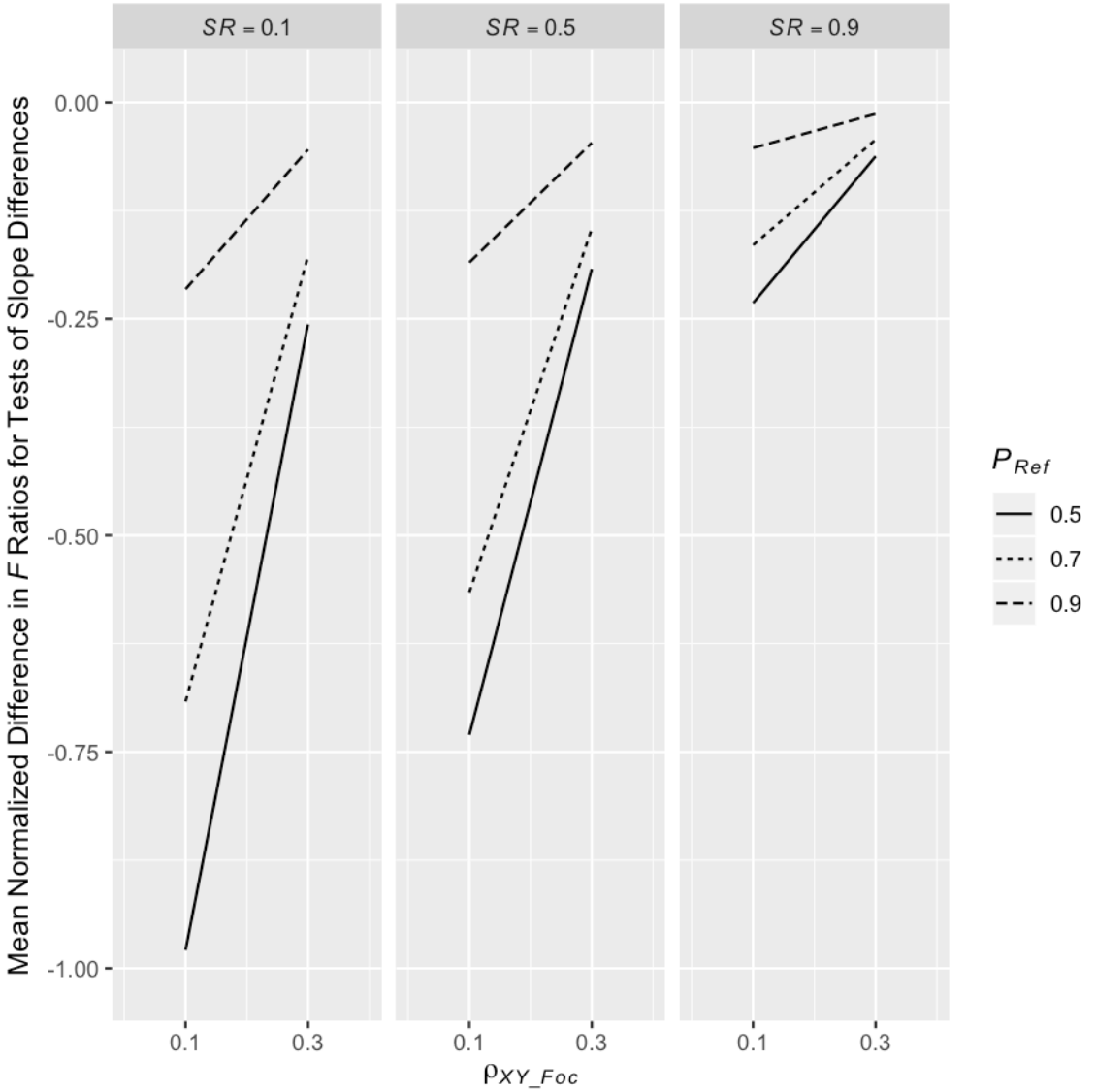


Figure 20

Effect of the three-way interaction among SR, P_{Ref}, and ρ_{XY_Foc} on the power of tests of slope differences (as indicated by normalized differences in F ratios) under conditions of direct range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to X; P_{Ref} = proportion of referent-group members in the applicant population; ρ_{XY_Foc} = operational validity of X for predicting Y in the focal group's applicant population.

Figure is based on data from 162 (66.7%) direct range restriction conditions.

Total η^2 of plotted effects = .95.

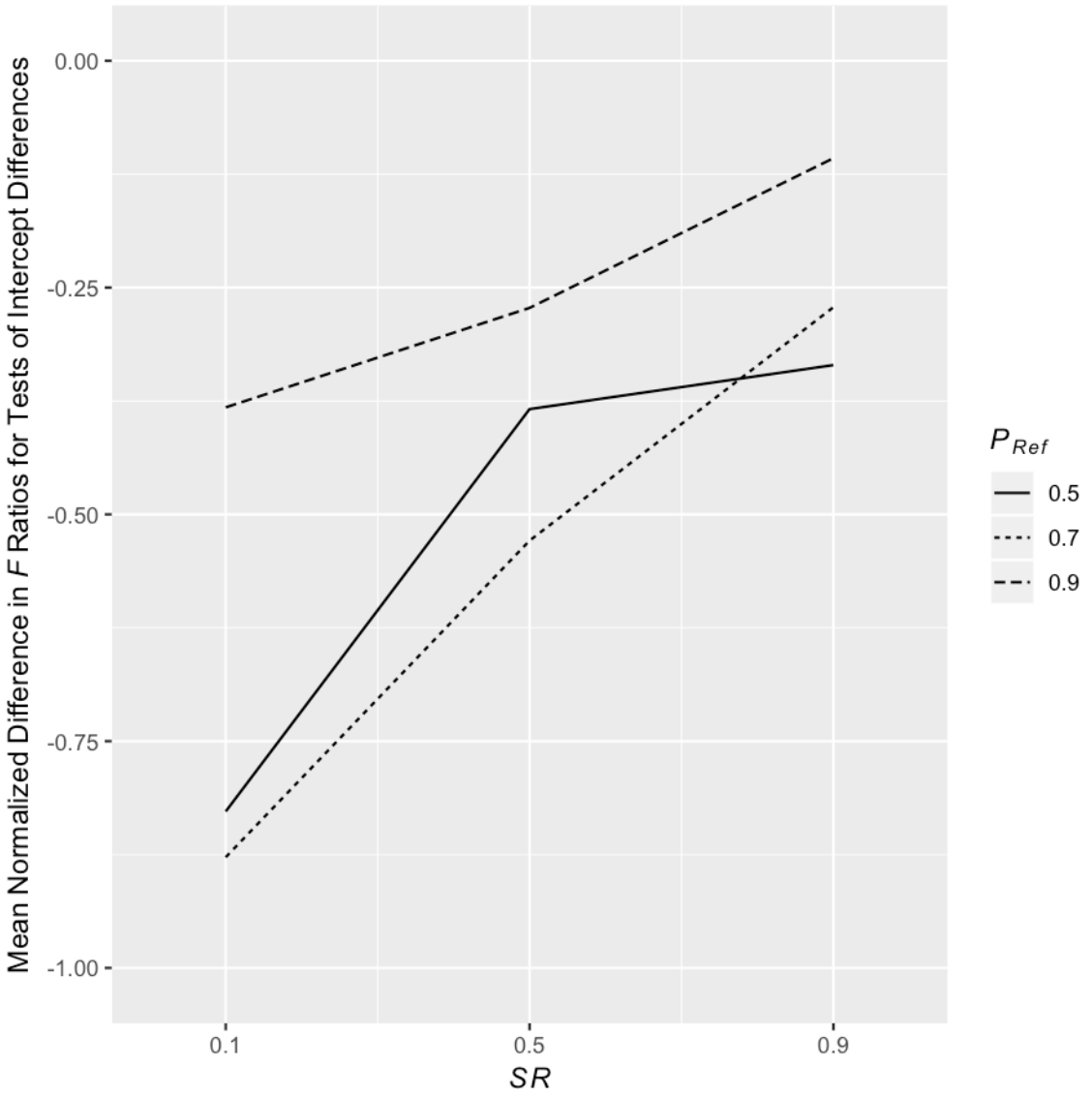


Figure 21

Effect of the two-way interaction between SR and P_{Ref} on the power of tests of intercept differences (as indicated by normalized differences in F ratios) under conditions of direct range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to X; P_{Ref} = proportion of referent-group members in the applicant population.

Figure is based on data from 54 (22.2%) direct range restriction conditions.

Total η^2 of plotted effects = .69.

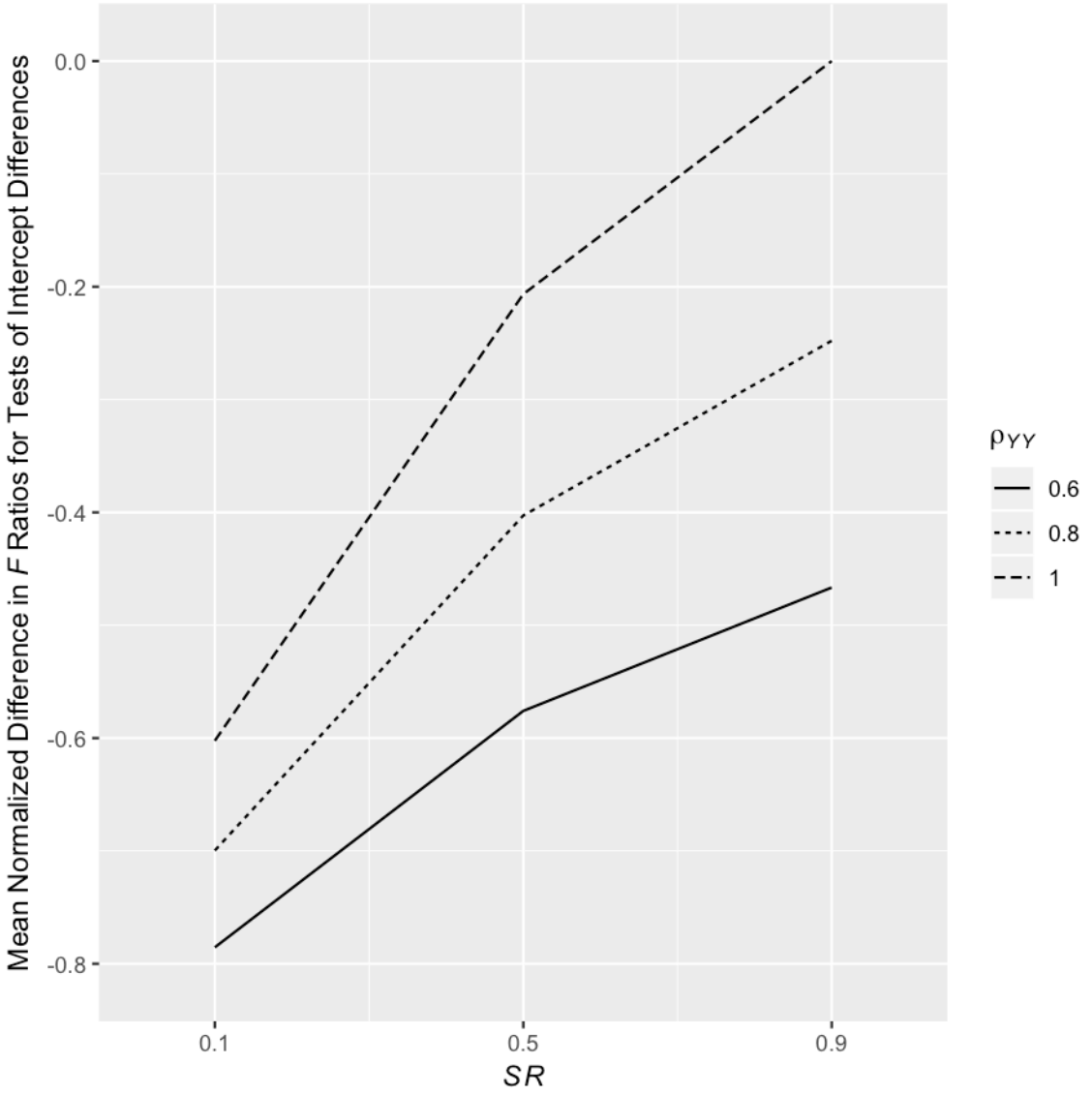


Figure 22

Effect of the two-way interaction between SR and ρ_{YY} on the power of tests of intercept differences (as indicated by normalized differences in F ratios) under conditions of direct range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to X; ρ_{YY} = reliability of Y.

Figure is based on data from 54 (22.2%) direct range restriction conditions.

Total η^2 of plotted effects = .67.

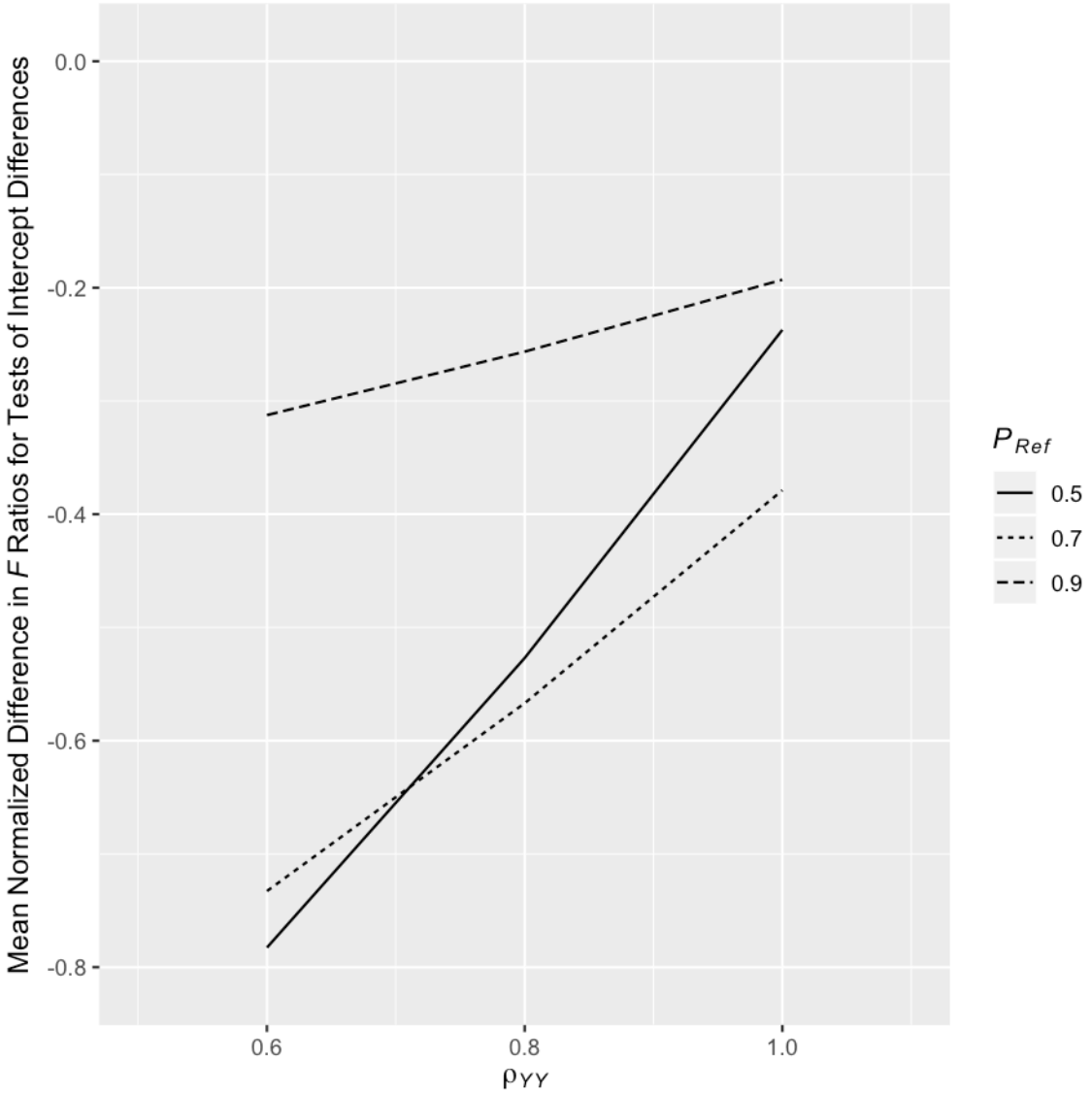


Figure 23

Effect of the two-way interaction between P_{Ref} and ρ_{YY} on the power of tests of intercept differences (as indicated by normalized differences in *F* ratios) under conditions of direct range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. P_{Ref} = proportion of referent-group members in the applicant population; ρ_{YY} = reliability of *Y*.

Figure is based on data from 54 (22.2%) direct range restriction conditions.

Total η^2 of plotted effects = .49.

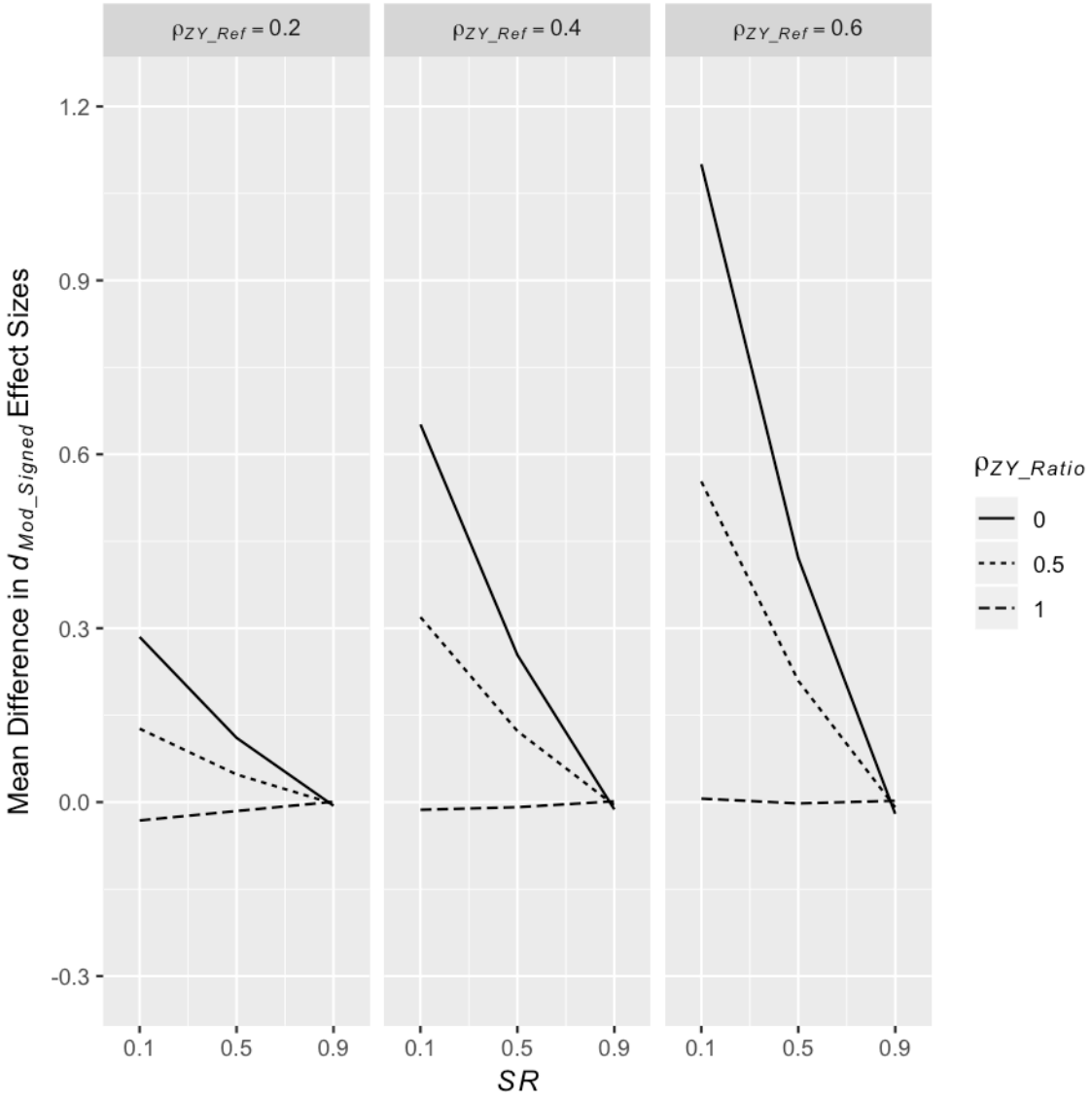


Figure 24

Effect of the three-way interaction among SR, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} on d_{Mod_Signed} effect sizes under conditions of indirect range restriction.

Positive values indicate higher d_{Mod_Signed} effect sizes estimated from observed parameters than from operational parameters. SR = overall selection ratio applied to Z; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population. Figure is based on data from 19,683 (100.0%) indirect range restriction conditions. Total η^2 of plotted effects = .93.

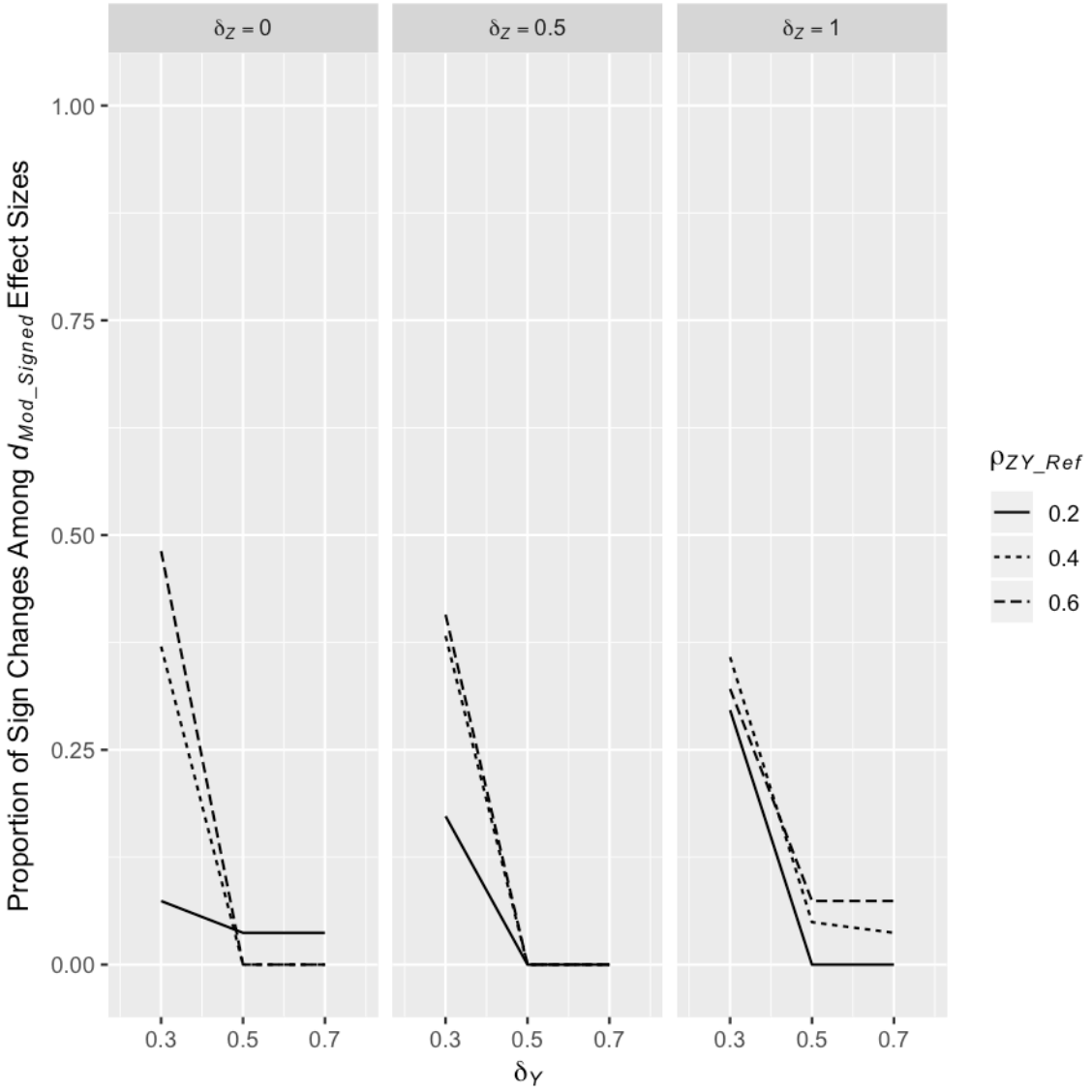


Figure 25

Effect of the three-way interaction among ρ_{ZY_Ref}, δ_Y, and δ_Z on the signs of d_{Mod_Signed} effect sizes under conditions of indirect range restriction.

Larger values indicate higher proportions of sign differences between observed and operational parameters. ρ_{ZY_Ref} = operational validity of *Z* for predicting *Y* in the referent group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on *Y*; δ_Z = standardized mean difference between the referent and focal groups on *Z*.

Figure is based on data from 19,683 (100.0%) indirect range restriction conditions.

Total η^2 of plotted effects = .24.

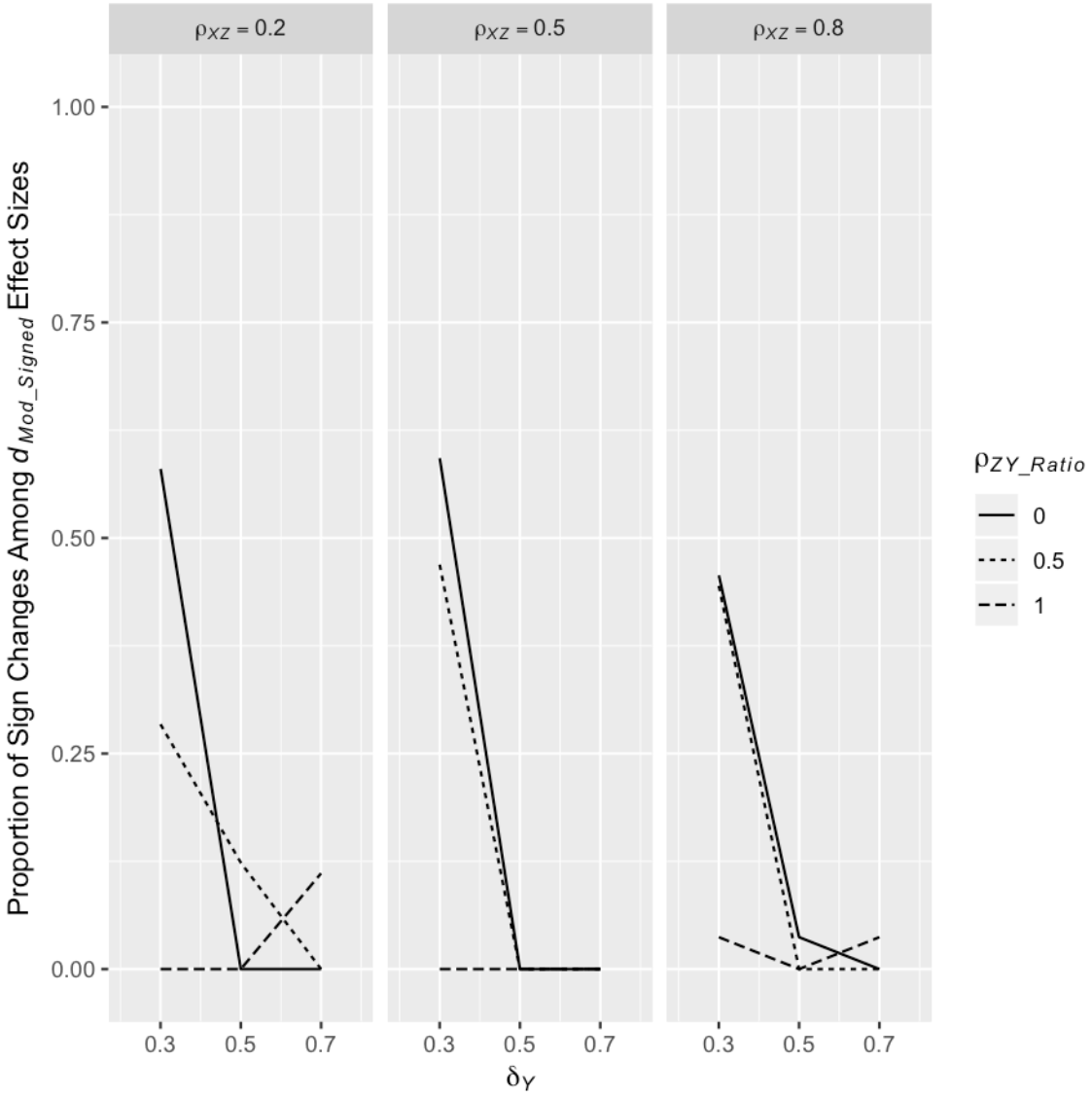


Figure 26

Effect of the three-way interaction among ρ_{ZY_Ratio}, ρ_{XZ}, and δ_Y on the signs of d_{Mod_Signed} effect sizes under conditions of indirect range restriction.

Larger values indicate higher proportions of sign differences between observed and operational parameters. ρ_{ZY_Ratio} = ratio of the validity of *Z* for predicting *Y* in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between *X* and *Z* in the referent and focal groups' applicant populations; δ_Y = true-score standardized mean difference between the referent and focal groups on *Y*.

Figure is based on data from 19,683 (100.0%) indirect range restriction conditions. Total η^2 of plotted effects = .38.

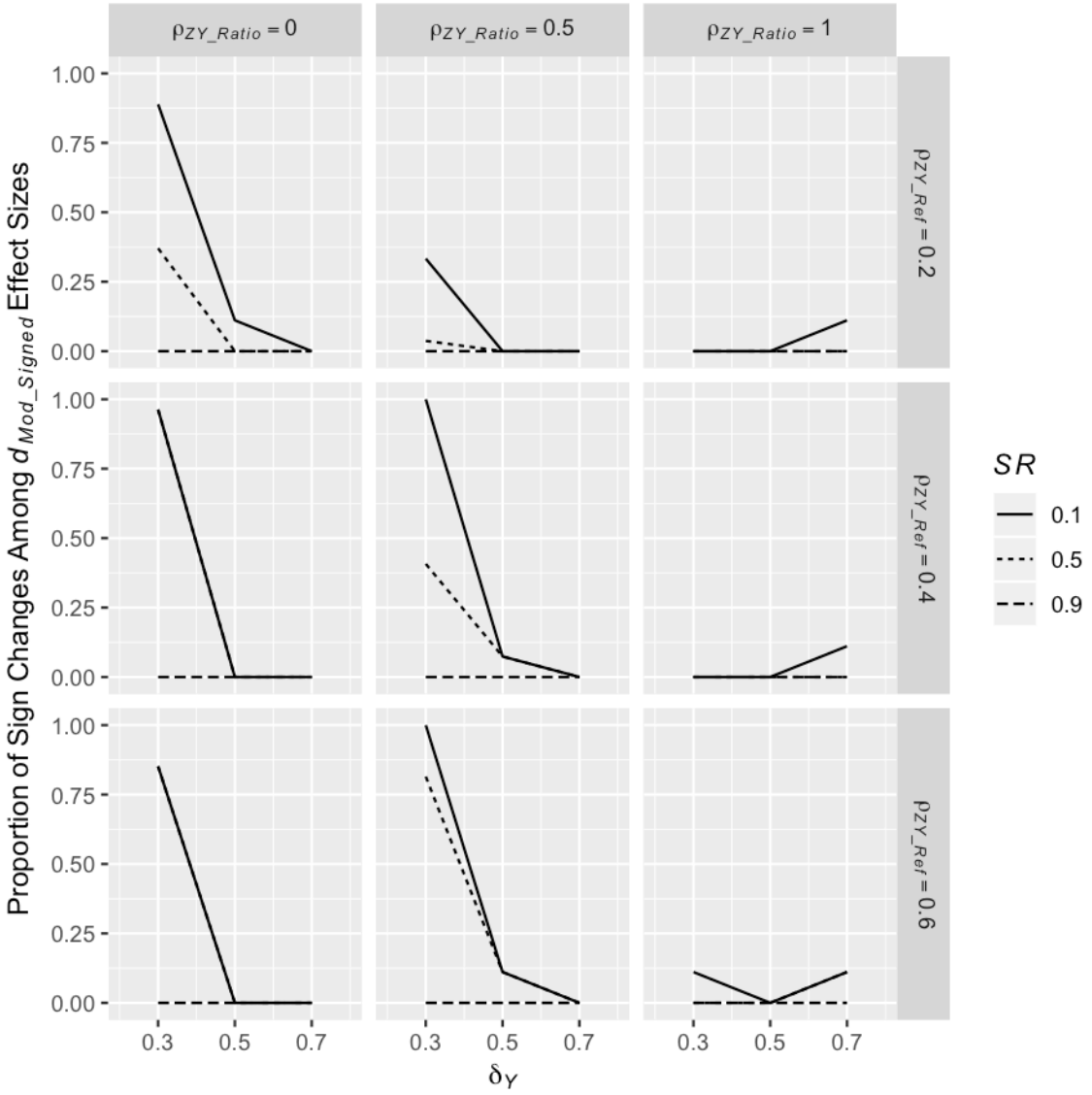


Figure 27

Effect of the four-way interaction among SR, ρ_{ZY_Ref}, ρ_{ZY_Ratio}, and δ_Y on the signs of d_{Mod_Signed} effect sizes under conditions of indirect range restriction.

Larger values indicate higher proportions of sign differences between observed and operational parameters. SR = overall selection ratio applied to Z; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on Y.

Figure is based on data from 19,683 (100.0%) indirect range restriction conditions. Total η^2 of plotted effects = .72.

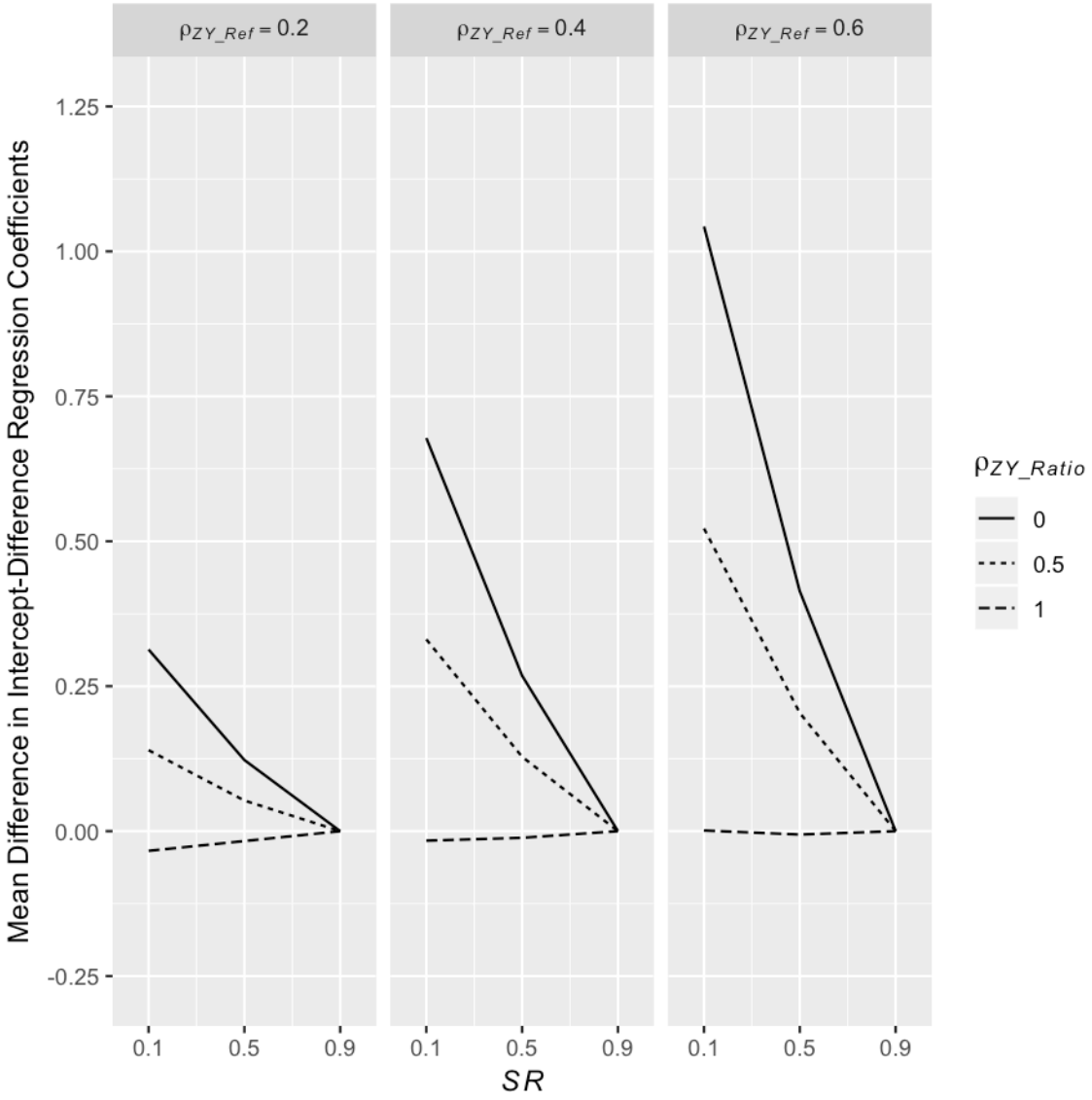


Figure 28

Effect of the three-way interaction among SR, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} on intercept-difference regression coefficients under conditions of indirect range restriction. Positive values indicate higher intercept-differences regression coefficients estimated from observed parameters than from operational parameters. SR = overall selection ratio applied to Z; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population. Figure is based on data from 6,561 (33.3%) indirect range restriction conditions. Total η^2 of plotted effects = .94.

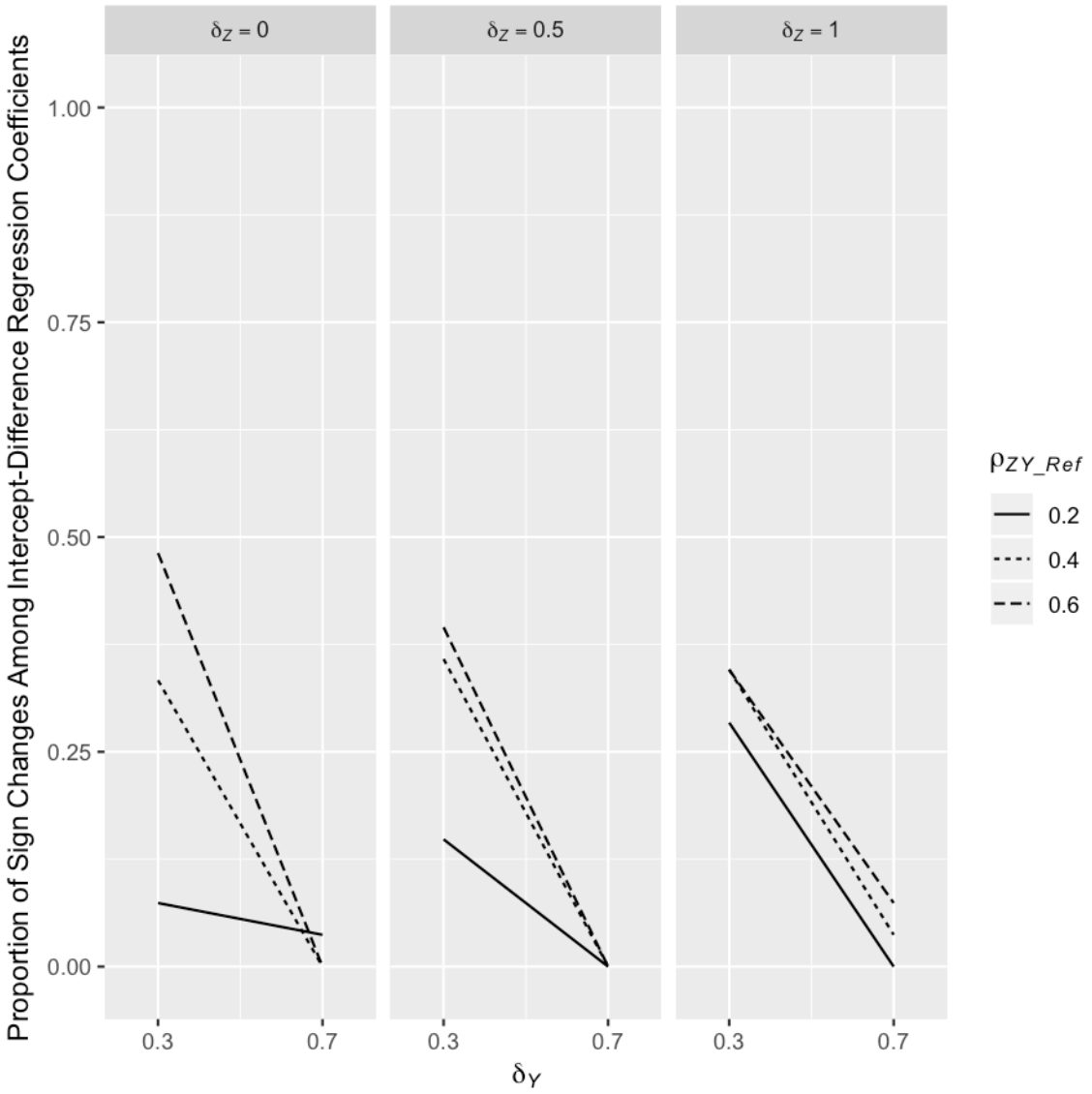


Figure 29

Effect of the three-way interaction among ρ_{ZY_Ref}, δ_Y, and δ_Z on the signs of intercept-difference regression coefficients from scenarios with intercept differences under conditions of indirect range restriction.

Larger values indicate higher proportions of sign differences between observed and operational parameters. ρ_{ZY_Ref} = operational validity of *Z* for predicting *Y* in the referent group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on *Y*; δ_Z = standardized mean difference between the referent and focal groups on *Z*.

Figure is based on data from 4,374 (22.2%) indirect range restriction conditions.

Total η^2 of plotted effects = .21.

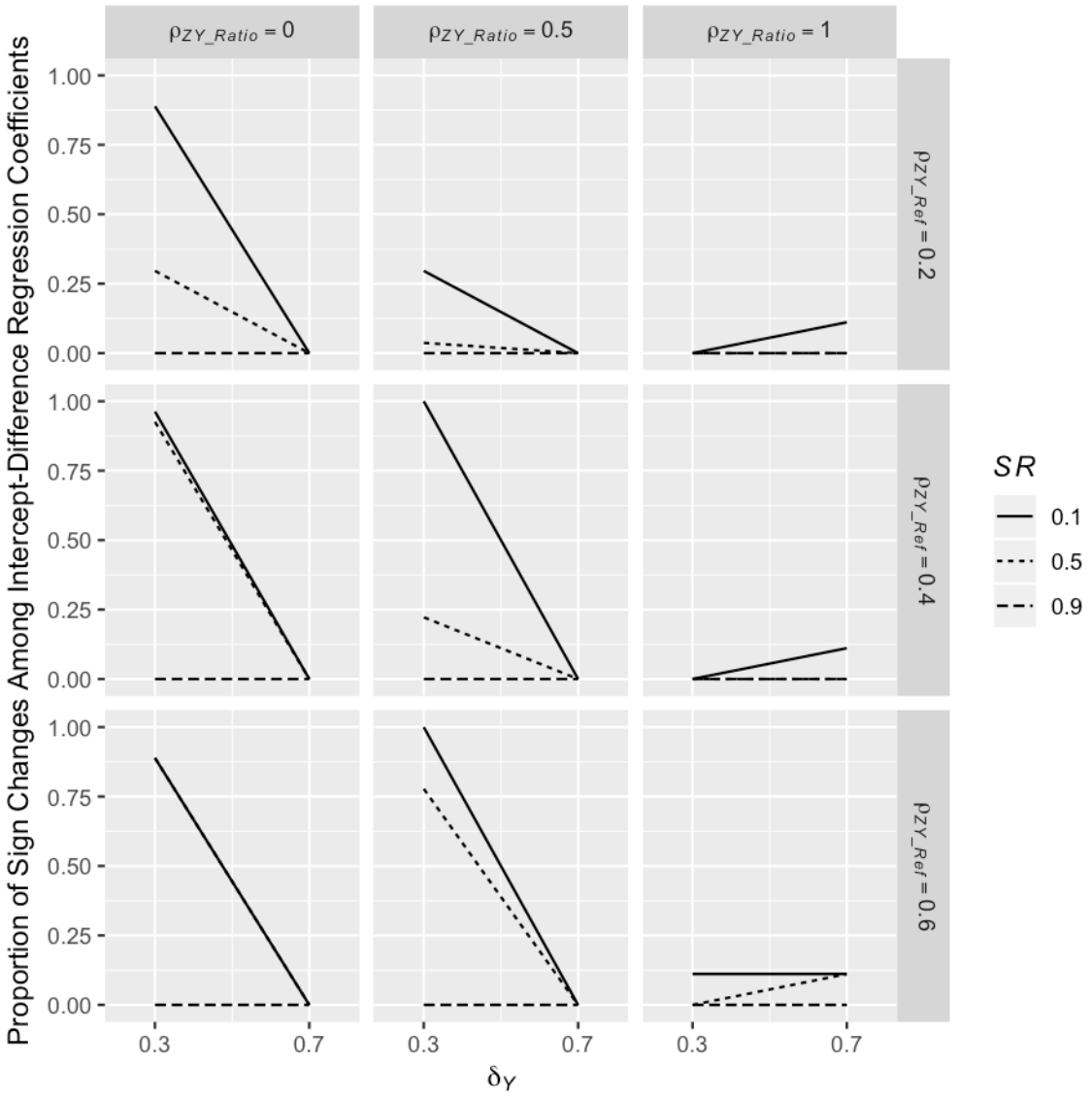


Figure 30

Effect of the four-way interaction among SR, ρ_{ZY_Ref}, ρ_{ZY_Ratio}, and δ_Y on the signs of intercept-difference regression coefficients from scenarios with intercept differences under conditions of indirect range restriction.

Larger values indicate higher proportions of sign differences between observed and operational parameters. SR = overall selection ratio applied to Z; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on Y.

Figure is based on data from 4,374 (22.2%) indirect range restriction conditions.

Total η^2 of plotted effects = .75.

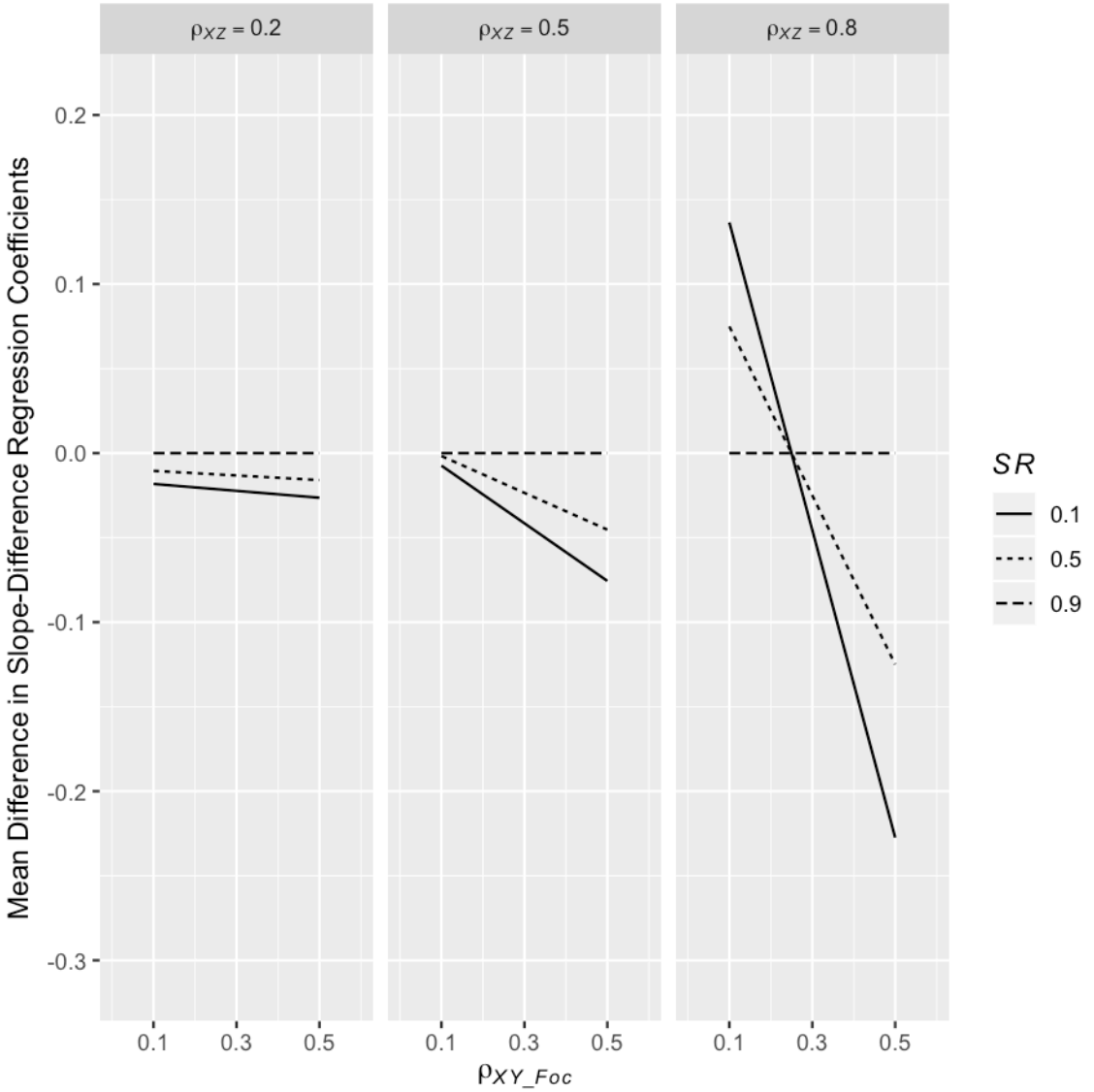


Figure 31

Effect of the three-way interaction among SR, ρ_{XY_Foc}, and ρ_{XZ} on slope-difference regression coefficients under conditions of indirect range restriction.

Positive values indicate higher slope-differences regression coefficients estimated from observed parameters than from operational parameters. SR = overall selection ratio applied to Z; ρ_{XY_Foc} = operational validity of X for predicting Y in the focal group's applicant population; ρ_{XZ} = correlation between X and Z in the referent and focal groups' applicant populations.

Figure is based on data from 19,683 (100.0%) indirect range restriction conditions. Total η^2 of plotted effects = .31.

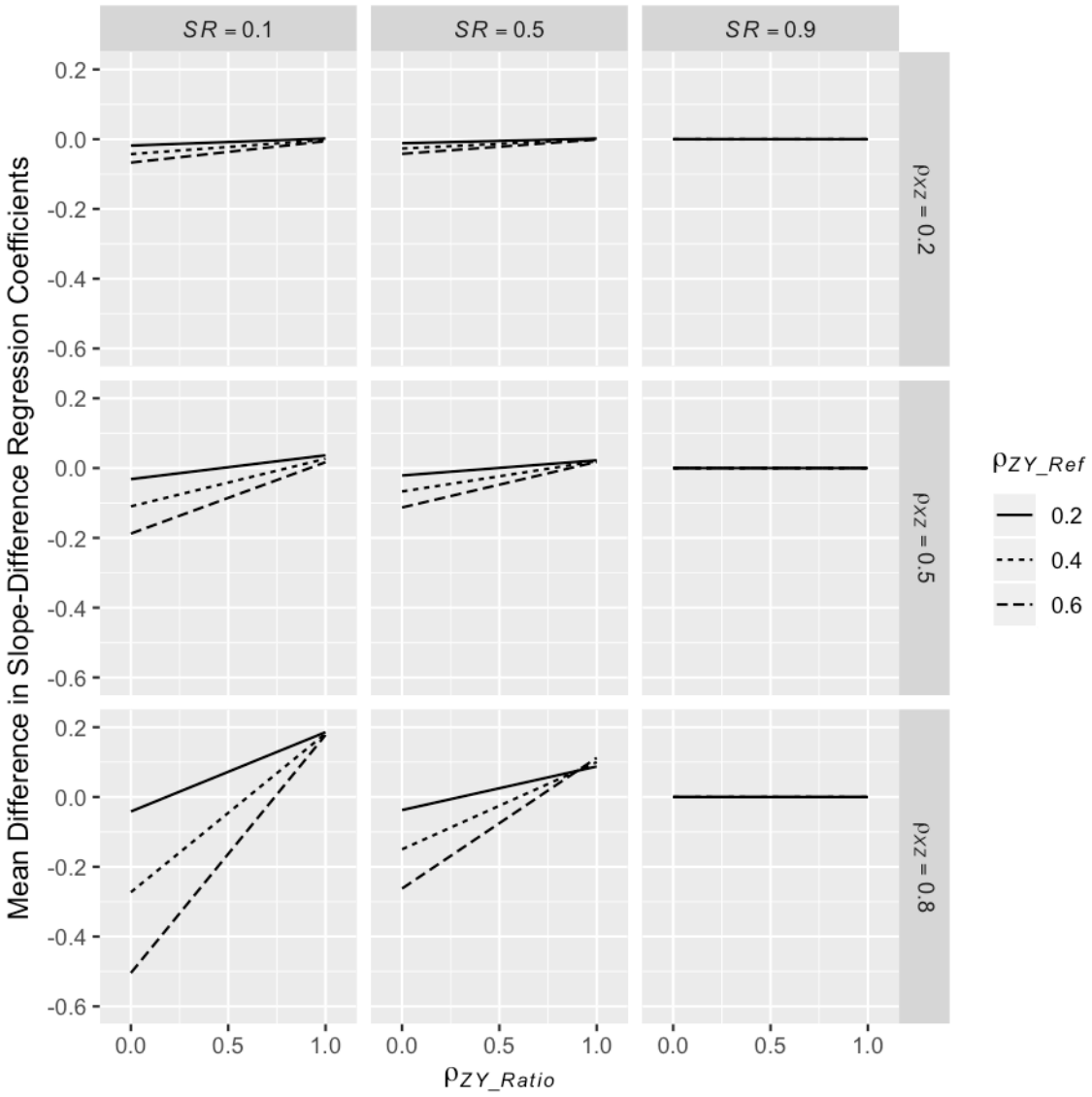


Figure 32

Effect of the four-way interaction among SR, ρ_{ZY_Ref}, ρ_{ZY_Ratio}, and ρ_{XZ} on slope-difference regression coefficients under conditions of indirect range restriction. Positive values indicate higher slope-differences regression coefficients estimated from observed parameters than from operational parameters. SR = overall selection ratio applied to Z; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between X and Z in the referent and focal groups' applicant populations.

Figure is based on data from 19,683 (100.0%) indirect range restriction conditions. Total η^2 of plotted effects = .70.

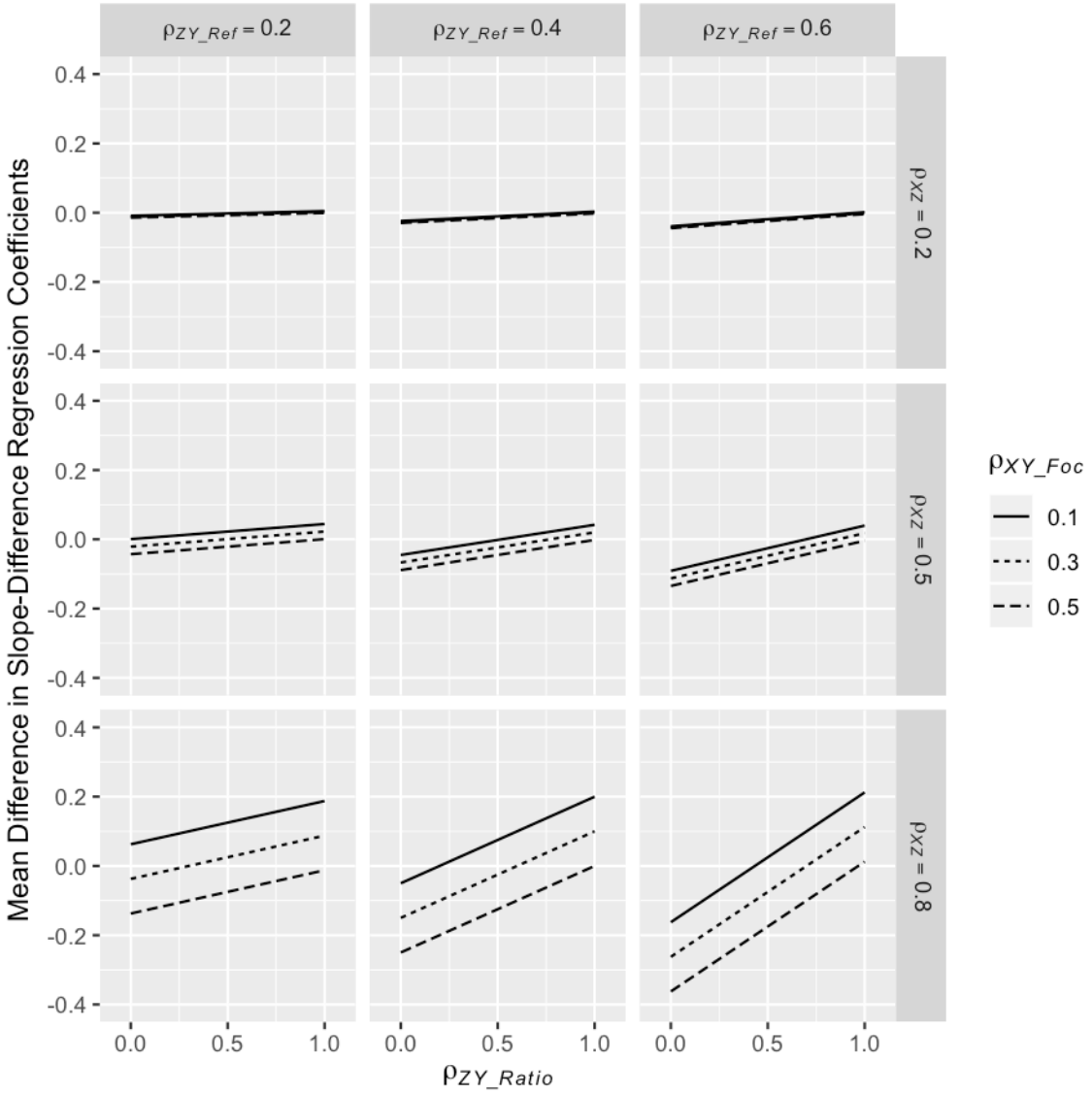


Figure 33

Effect of the four-way interaction among ρ_{XY_Foc}, ρ_{ZY_Ref}, ρ_{ZY_Ratio}, and ρ_{XZ} on slope-difference regression coefficients under conditions of indirect range restriction when the selection ratio is .50.

Positive values indicate higher slope-differences regression coefficients estimated from observed parameters than from operational parameters. ρ_{XY_Foc} = operational validity of *X* for predicting *Y* in the focal group's applicant population; ρ_{ZY_Ref} = operational validity of *Z* for predicting *Y* in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of *Z* for predicting *Y* in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between *X* and *Z* in the referent and focal groups' applicant populations.

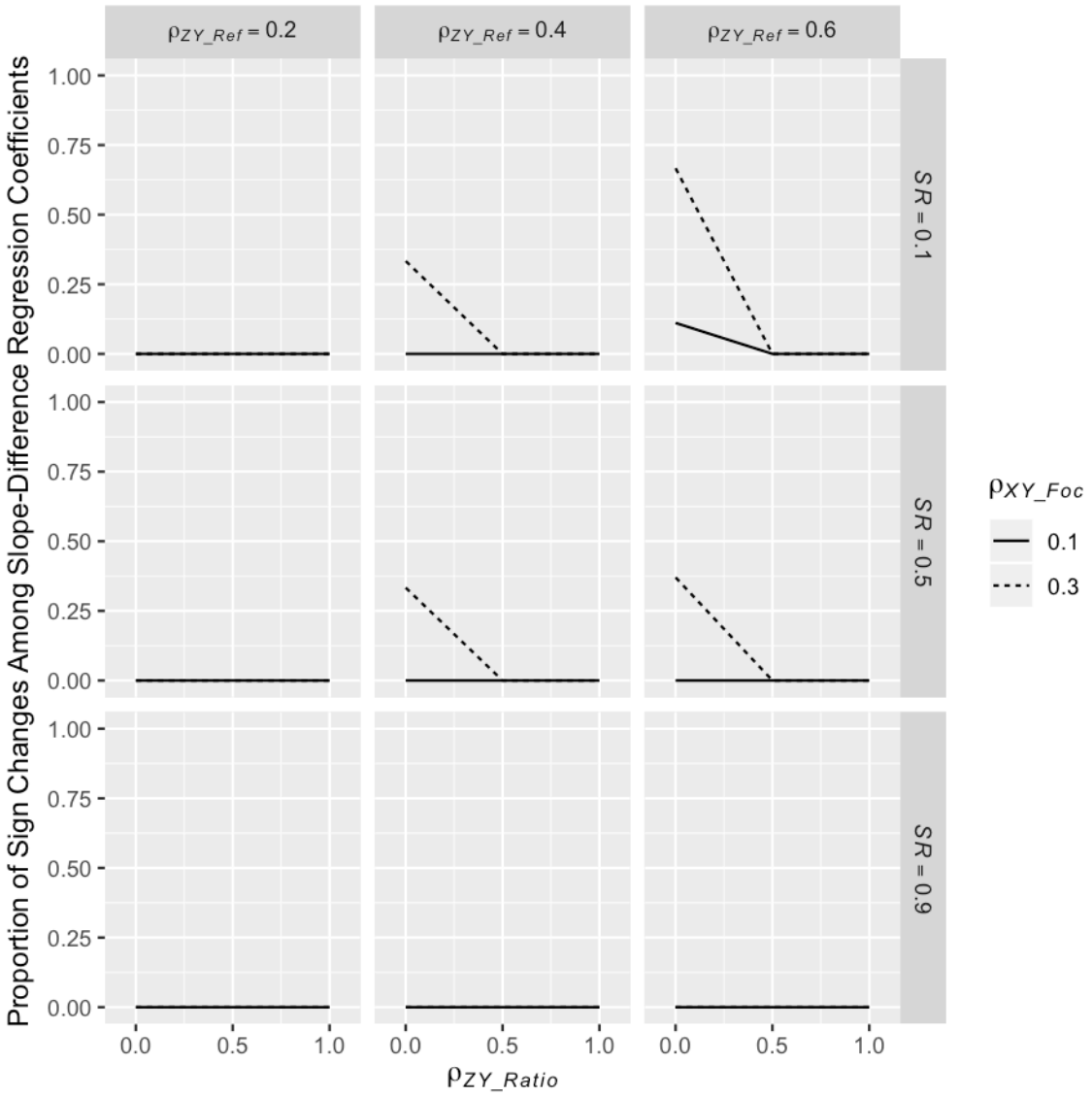


Figure 34

Effect of the four-way interaction among SR, ρ_{XY_Foc}, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} on the signs of slope-difference regression coefficients from scenarios with slope differences under conditions of indirect range restriction.

Larger values indicate higher proportions of sign differences between observed and operational parameters. SR = overall selection ratio applied to Z; ρ_{XY_Foc} = operational validity of X for predicting Y in the focal group's applicant population; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population.

Figure is based on data from 13,122 (66.7%) indirect range restriction conditions. Total η^2 of plotted effects = .41.

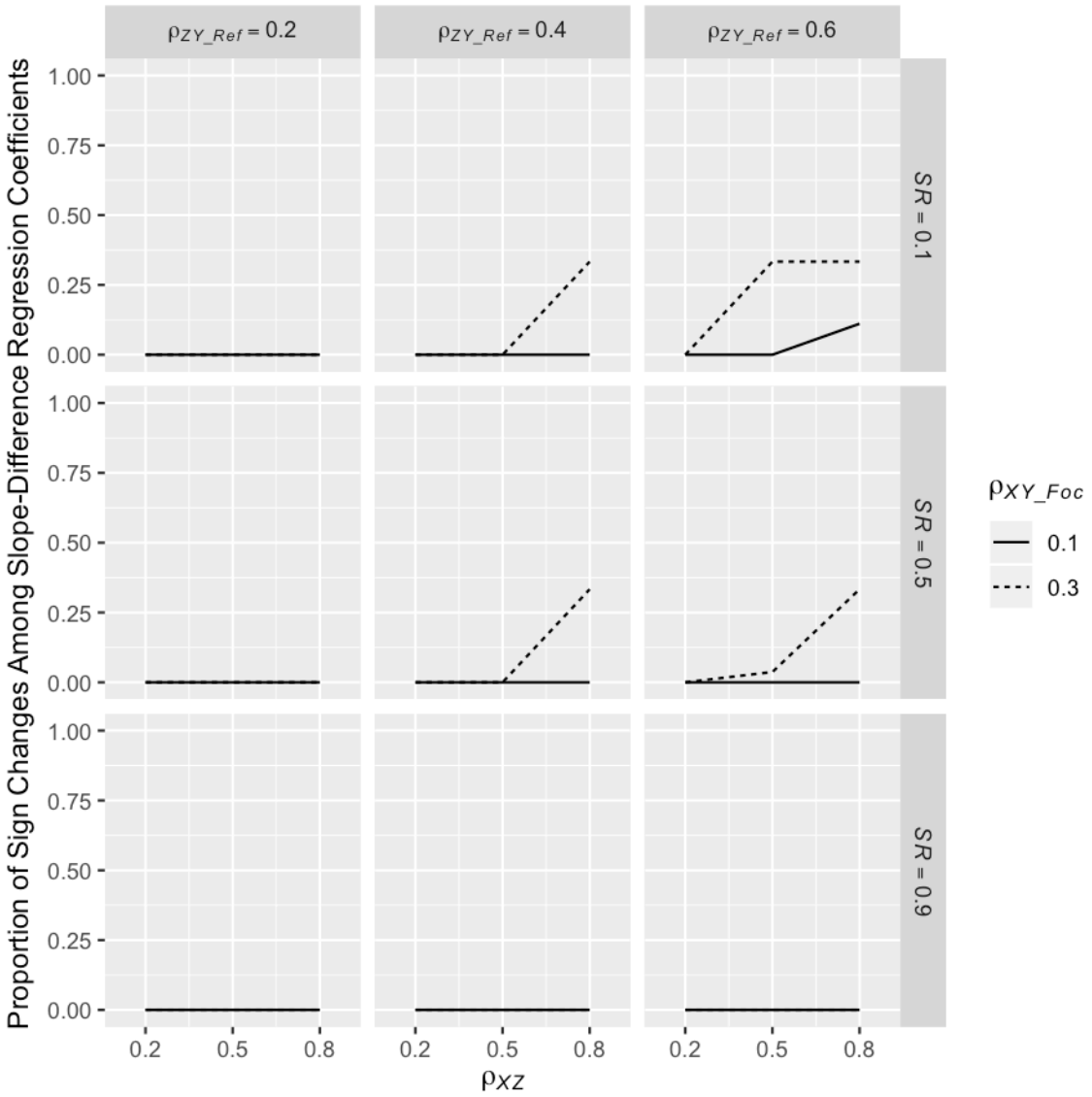


Figure 35

Effect of the four-way interaction among SR, ρ_{XY_Foc}, ρ_{ZY_Ref}, and ρ_{XZ} on the signs of slope-difference regression coefficients from scenarios with slope differences under conditions of indirect range restriction.

Larger values indicate higher proportions of sign differences between observed and operational parameters. SR = overall selection ratio applied to Z; ρ_{XY_Foc} = operational validity of X for predicting Y in the focal group's applicant population; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{XZ} = correlation between X and Z in the referent and focal groups' applicant populations. Figure is based on data from 13,122 (66.7%) indirect range restriction conditions. Total η^2 of plotted effects = .27.

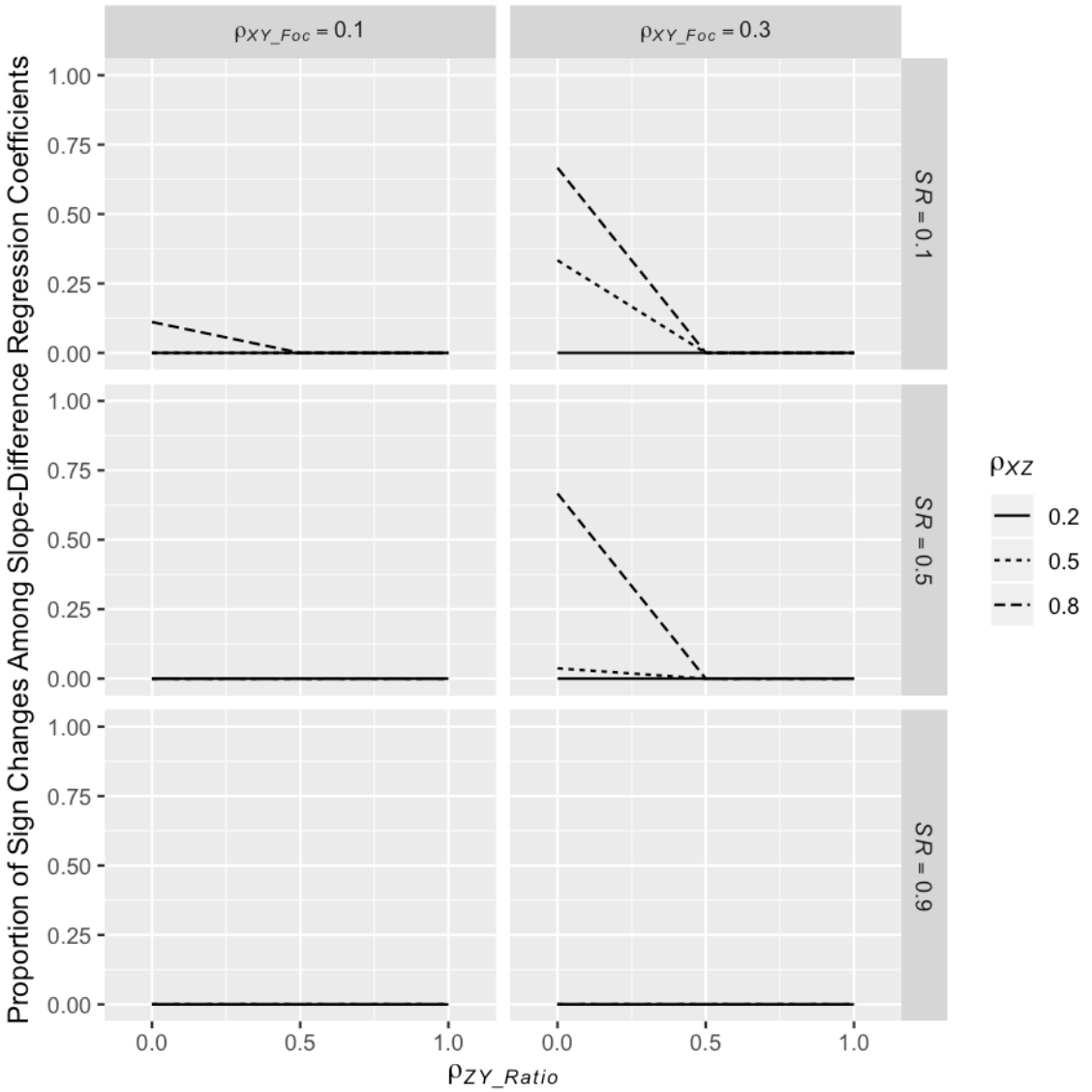


Figure 36

Effect of the four-way interaction among SR, ρ_{XY_Foc}, ρ_{ZY_Ratio}, and ρ_{XZ} on the signs of slope-difference regression coefficients from scenarios with slope differences under conditions of indirect range restriction.

Larger values indicate higher proportions of sign differences between observed and operational parameters. SR = overall selection ratio applied to Z; ρ_{XY_Foc} = operational validity of X for predicting Y in the focal group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between X and Z in the referent and focal groups' applicant populations.

Figure is based on data from 13,122 (66.7%) indirect range restriction conditions. Total η^2 of plotted effects = .51.

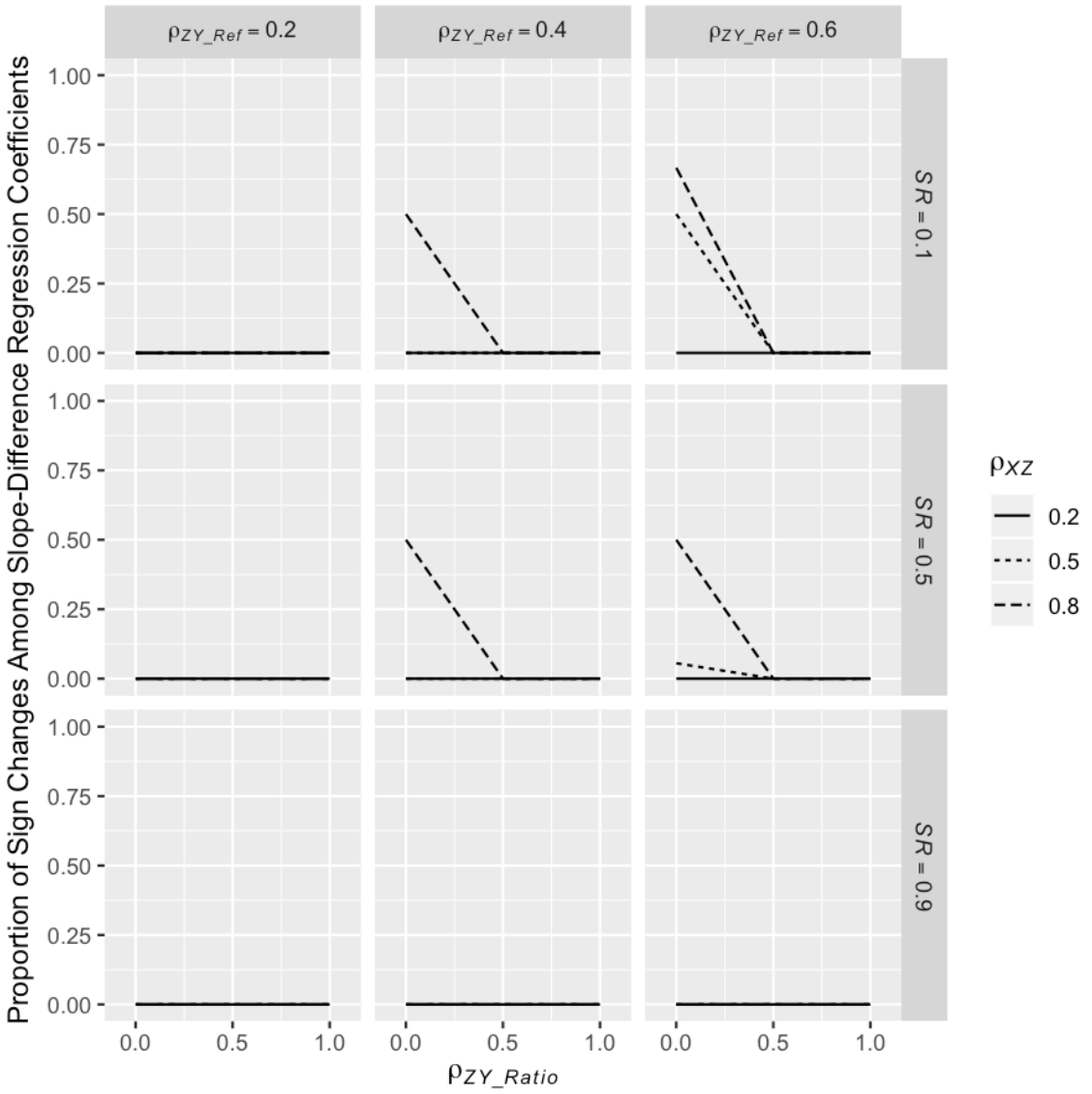


Figure 37

Effect of the four-way interaction among SR, ρ_{ZY_Ref}, ρ_{ZY_Ratio}, and ρ_{XZ} on the signs of slope-difference regression coefficients from scenarios with slope differences under conditions of indirect range restriction.

Larger values indicate higher proportions of sign differences between observed and operational parameters. SR = overall selection ratio applied to Z; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between X and Z in the referent and focal groups' applicant populations.

Figure is based on data from 13,122 (66.7%) indirect range restriction conditions.

Total η^2 of plotted effects = .49.

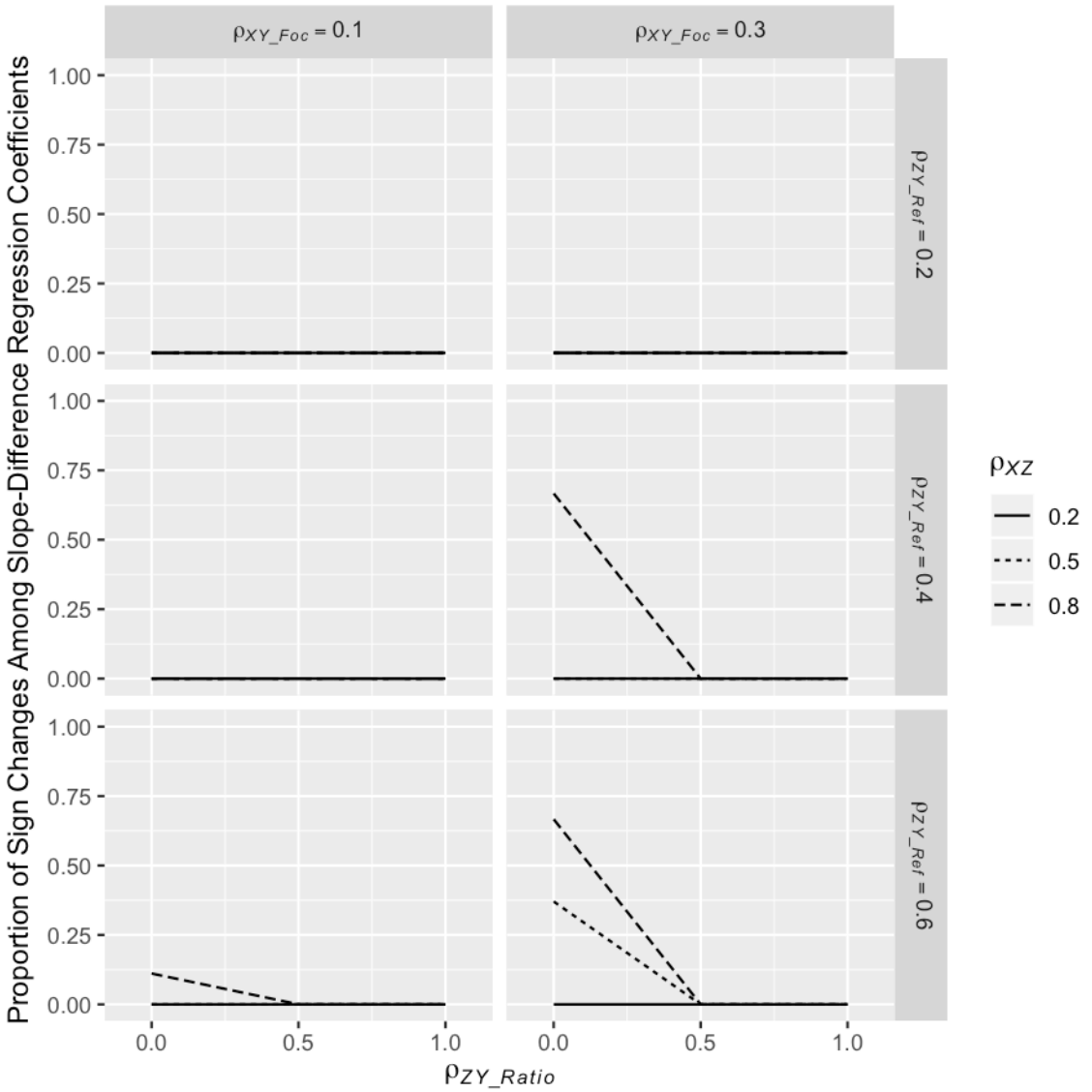


Figure 38

Effect of the four-way interaction among ρ_{XY_Foc}, ρ_{ZY_Ref}, ρ_{ZY_Ratio}, and ρ_{XZ} on the signs of slope-difference regression coefficients from scenarios with slope differences under conditions of indirect range restriction.

Larger values indicate higher proportions of sign differences between observed and operational parameters. ρ_{XY_Foc} = operational validity of *X* for predicting *Y* in the focal group's applicant population; ρ_{ZY_Ref} = operational validity of *Z* for predicting *Y* in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of *Z* for predicting *Y* in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between *X* and *Z* in the referent and focal groups' applicant populations.

Figure is based on data from 13,122 (66.7%) indirect range restriction conditions.

Total η^2 of plotted effects = .52.

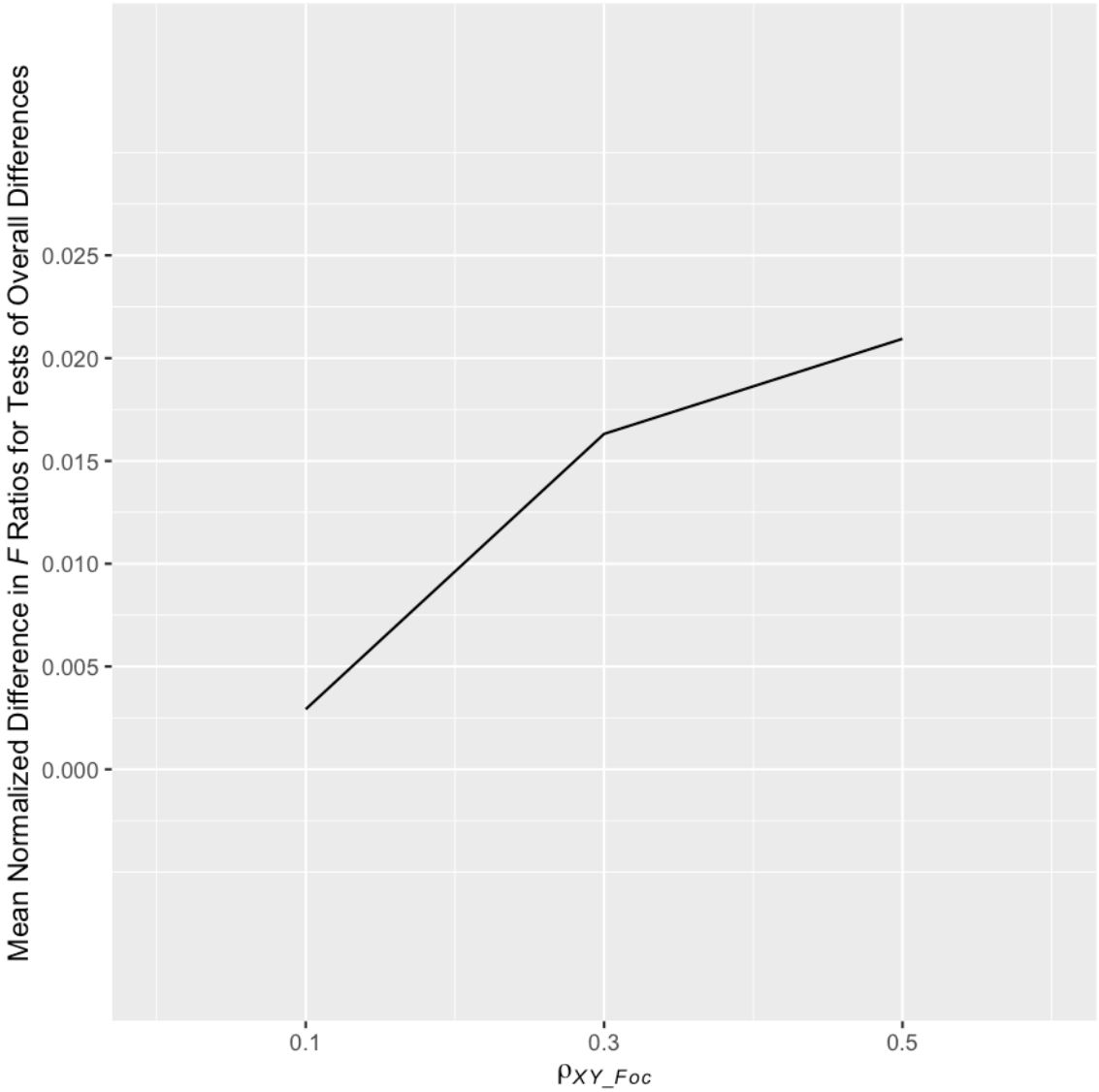


Figure 39

Main effect of ρ_{XY_Foc} on the power of tests of overall differences in prediction (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. ρ_{XY_Foc} = operational validity of *X* for predicting *Y* in the focal group's applicant population. Figure is based on data from 17,496 (88.9%) indirect range restriction conditions. Total η^2 of plotted effects = .01.

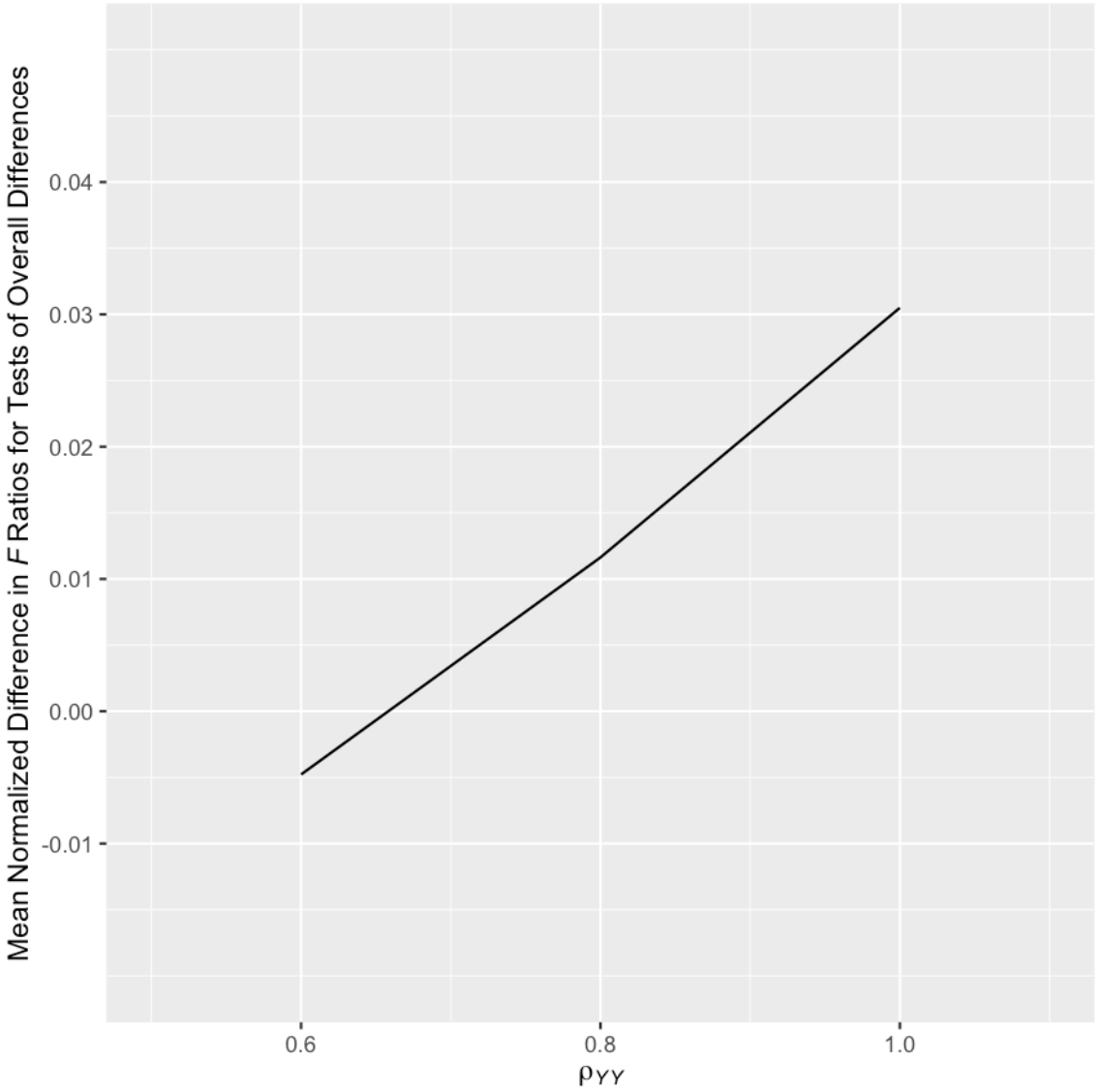


Figure 40

Main effect of ρ_{YY} on the power of tests of overall differences in prediction (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction. Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. ρ_{YY} = reliability of *Y*.

Figure is based on data from 17,496 (88.9%) indirect range restriction conditions.

Total η^2 of plotted effects = .04.

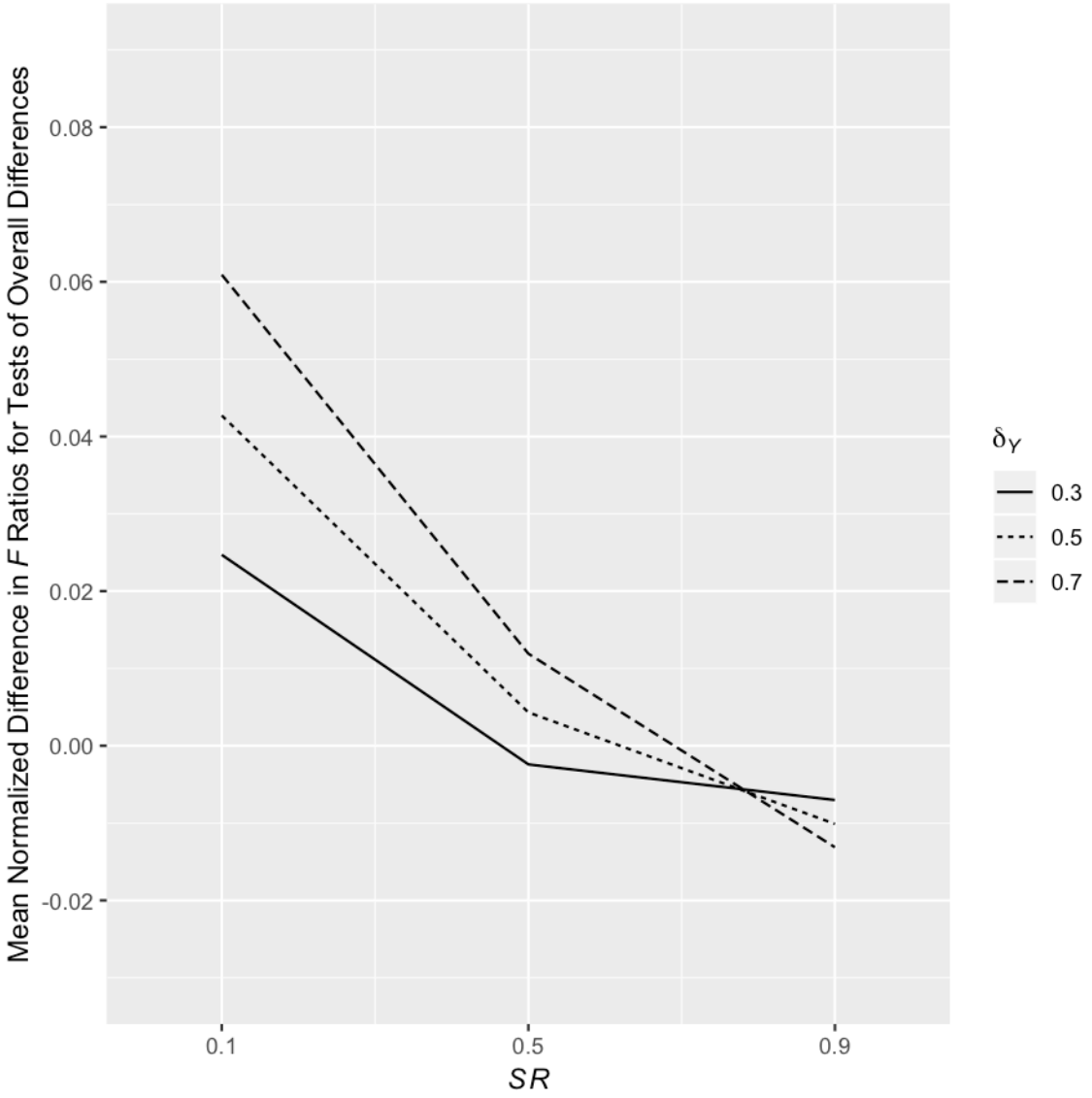


Figure 41

Effect of the two-way interaction between SR and δ_Y on the power of tests of overall differences in prediction (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; δ_Y = true-score standardized mean difference between the referent and focal groups on Y.

Figure is based on data from 17,496 (88.9%) indirect range restriction conditions.

Total η^2 of plotted effects = .11.

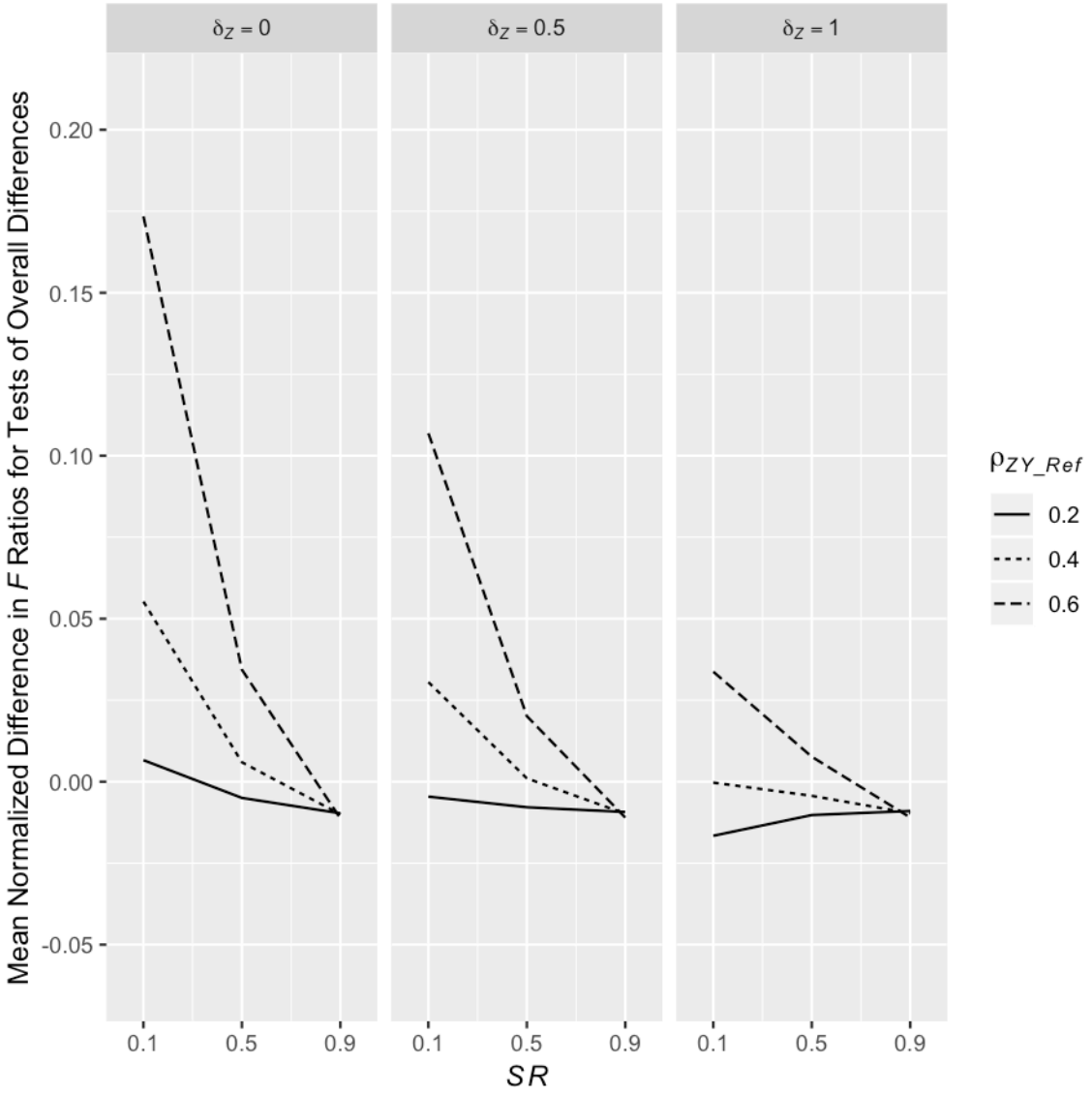


Figure 42

Effect of the three-way interaction among SR, ρ_{ZY_Ref}, and δ_Z on the power of tests of overall differences in prediction (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to *Z*; ρ_{ZY_Ref} = operational validity of *Z* for predicting *Y* in the referent group's applicant population; δ_Z = standardized mean difference between the referent and focal groups on *Z*.

Figure is based on data from 17,496 (88.9%) indirect range restriction conditions. Total η^2 of plotted effects = .31.

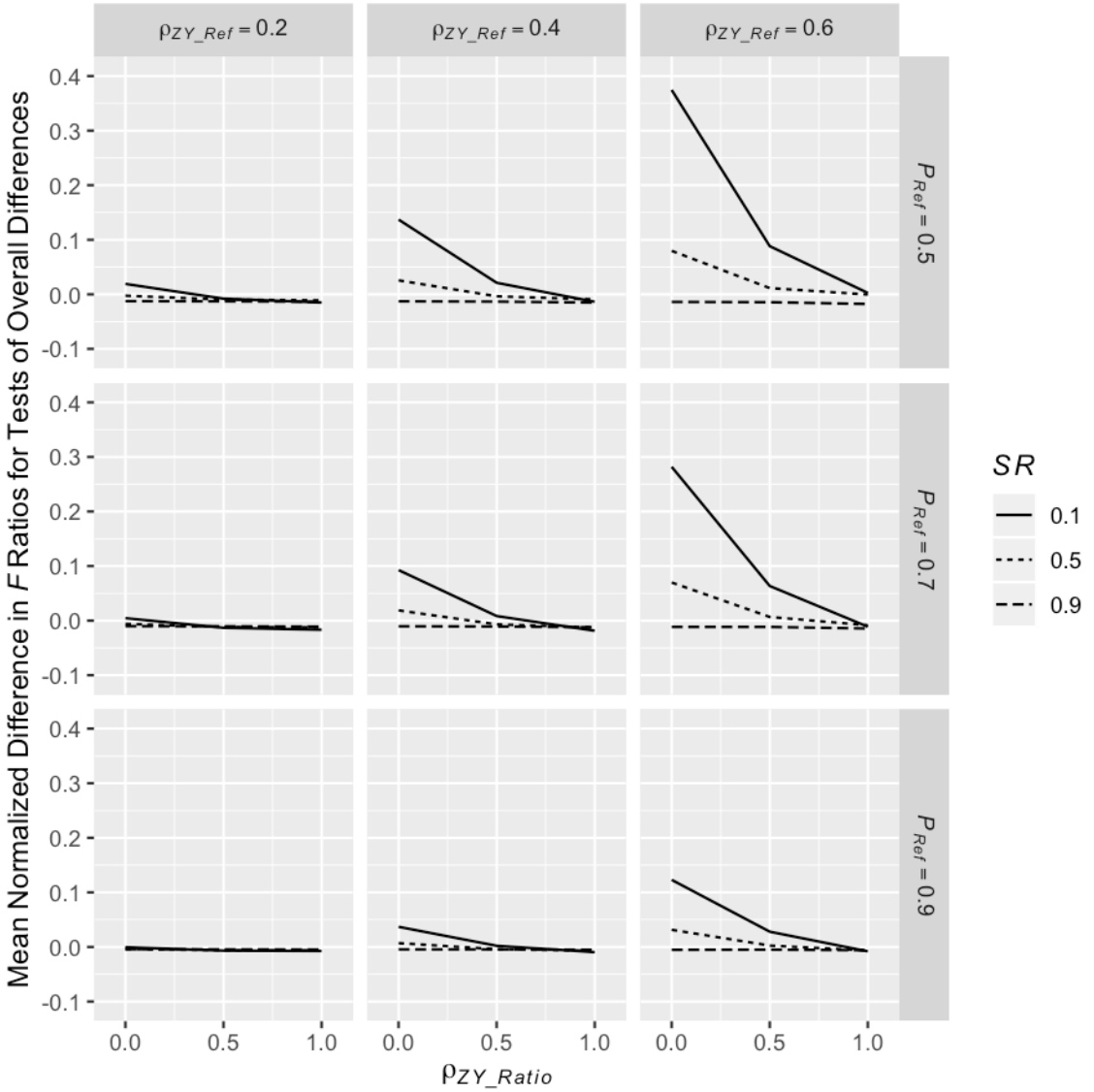


Figure 43

Effect of the four-way interaction among SR, P_{Ref}, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} on the power of tests of overall differences in prediction (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; P_{Ref} = proportion of referent-group members in the applicant population; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population. Figure is based on data from 17,496 (88.9%) indirect range restriction conditions. Total η^2 of plotted effects = .64.

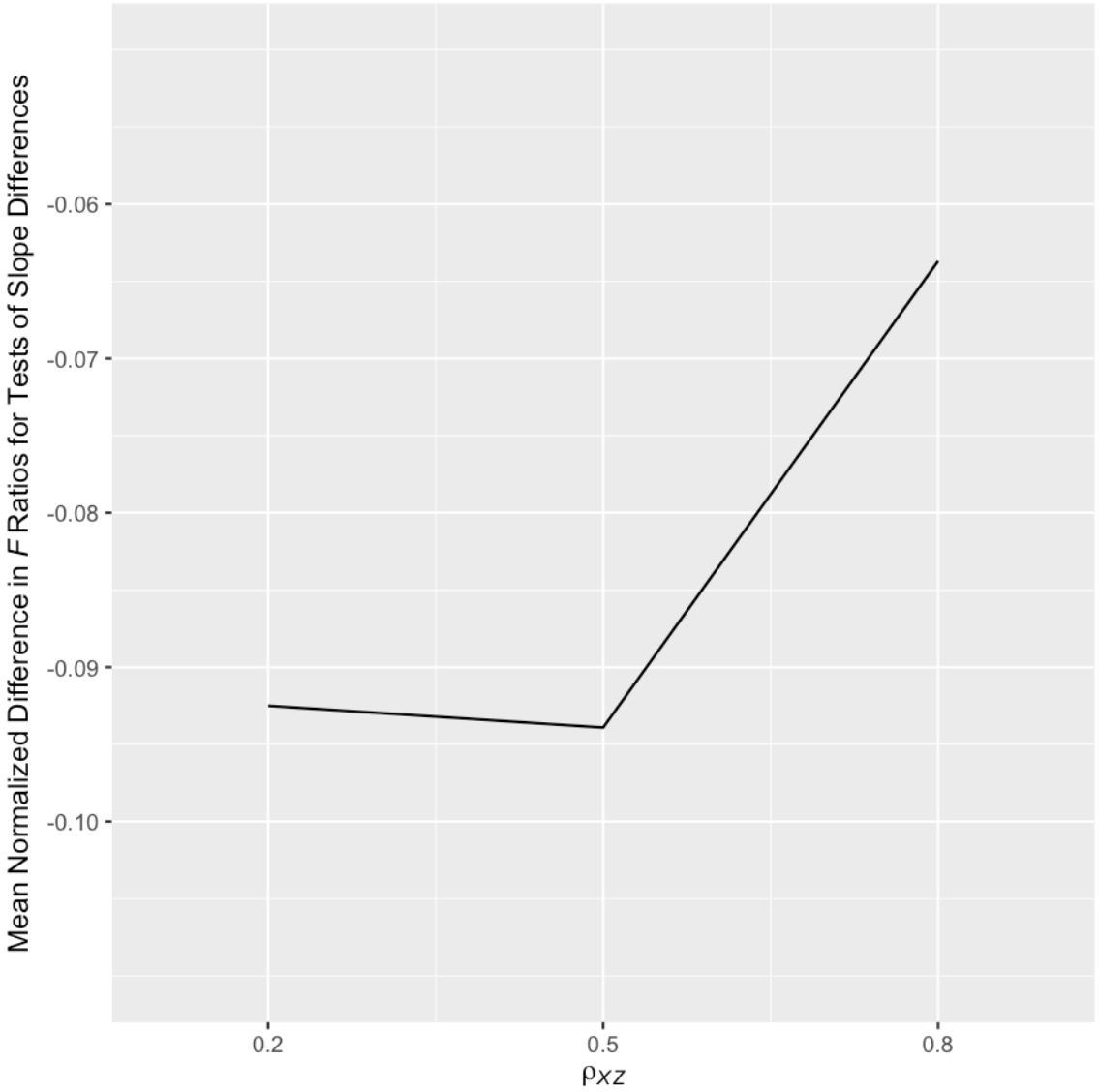


Figure 44

Main effect of ρ_{XZ} on the power of tests of slope differences (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. ρ_{XZ} = correlation between *X* and *Z* in the referent and focal groups' applicant populations.

Figure is based on data from 13,122 (66.7%) indirect range restriction conditions.

Total η^2 of plotted effects = .01.

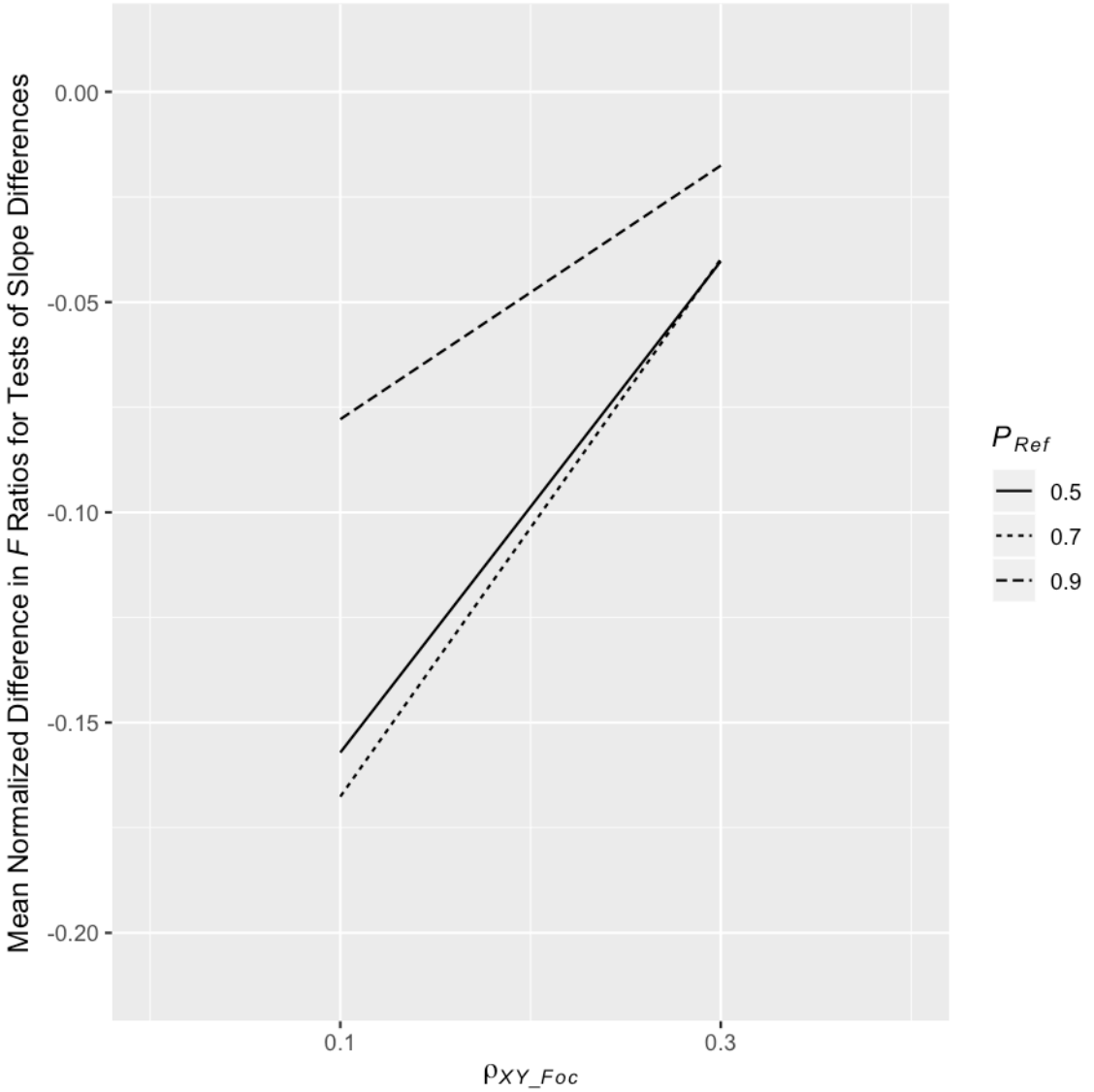


Figure 45

Effect of the two-way interaction between P_{Ref} and ρ_{XY_Foc} on the power of tests of slope differences (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. P_{Ref} = proportion of referent-group members in the applicant population; ρ_{XY_Foc} = operational validity of *X* for predicting *Y* in the focal group's applicant population.

Figure is based on data from 13,122 (66.7%) indirect range restriction conditions.

Total η^2 of plotted effects = .27.

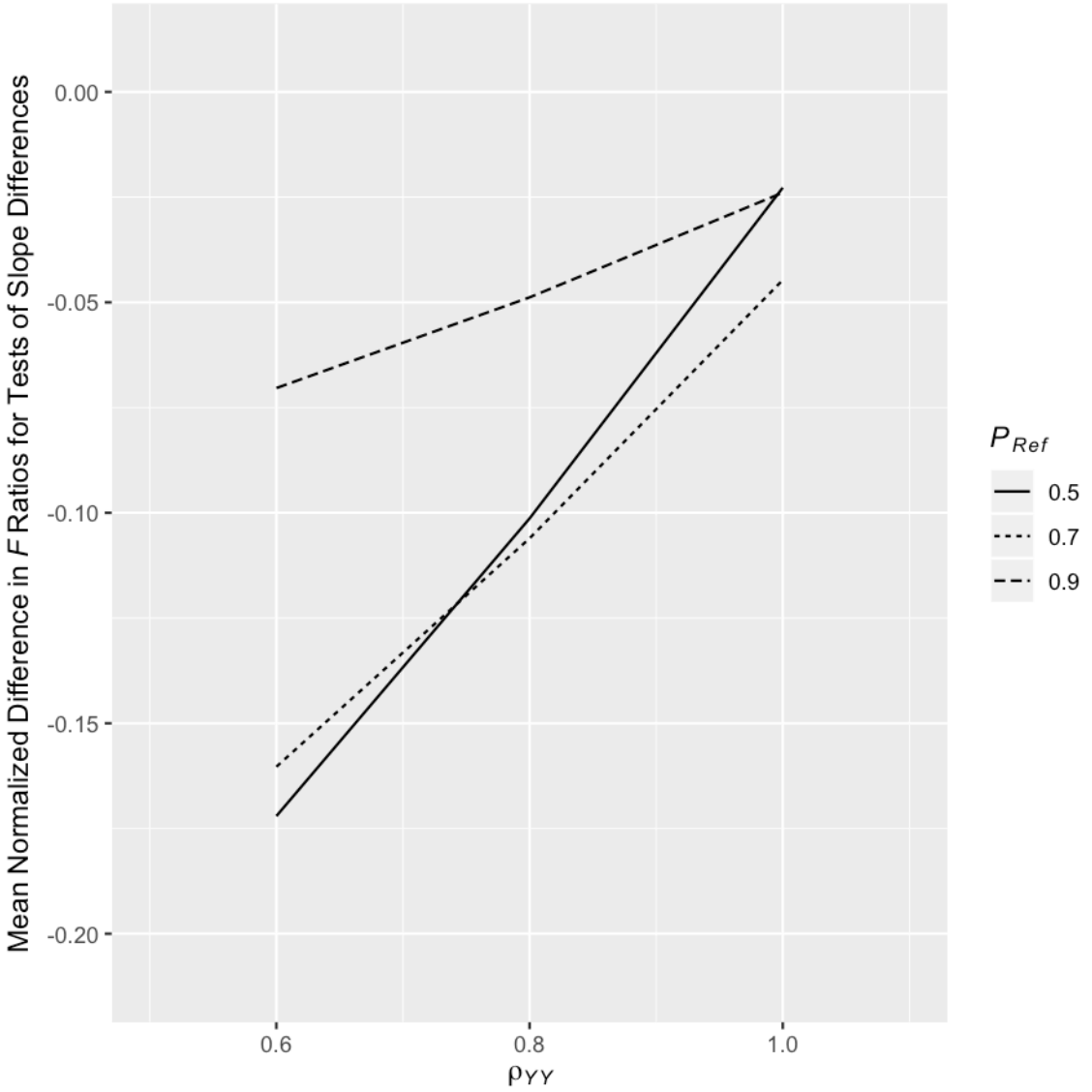


Figure 46

Effect of the two-way interaction between P_{Ref} and ρ_{YY} on the power of tests of slope differences (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. P_{Ref} = proportion of referent-group members in the applicant population; ρ_{YY} = reliability of *Y*.

Figure is based on data from 13,122 (66.7%) indirect range restriction conditions.

Total η^2 of plotted effects = .21.

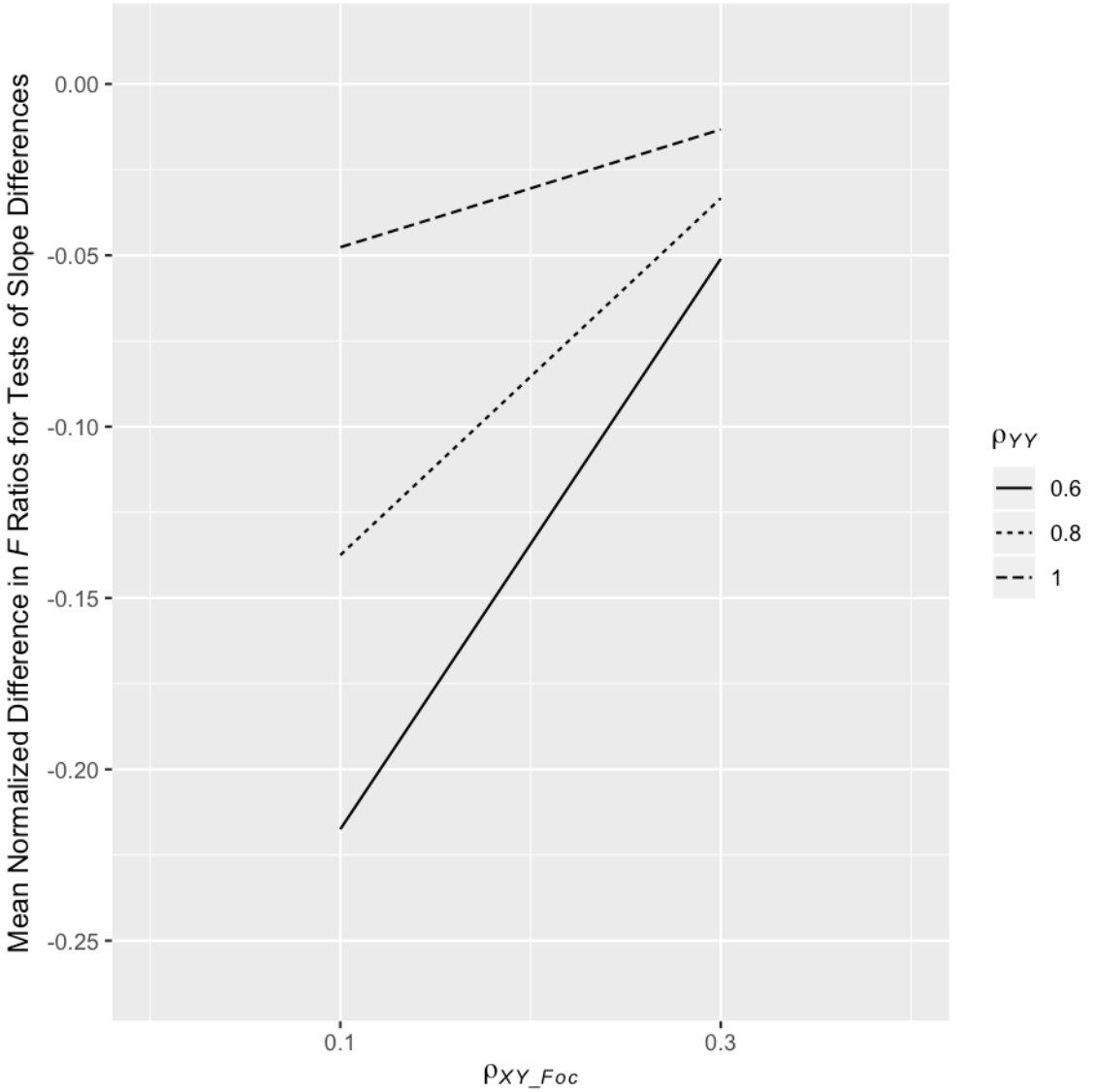


Figure 47

Effect of the two-way interaction between ρ_{XY_Foc} and ρ_{YY} on the power of tests of slope differences (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. ρ_{XY_Foc} = operational validity of *X* for predicting *Y* in the focal group's applicant population; ρ_{YY} = reliability of *Y*.

Figure is based on data from 13,122 (66.7%) indirect range restriction conditions.

Total η^2 of plotted effects = .39.

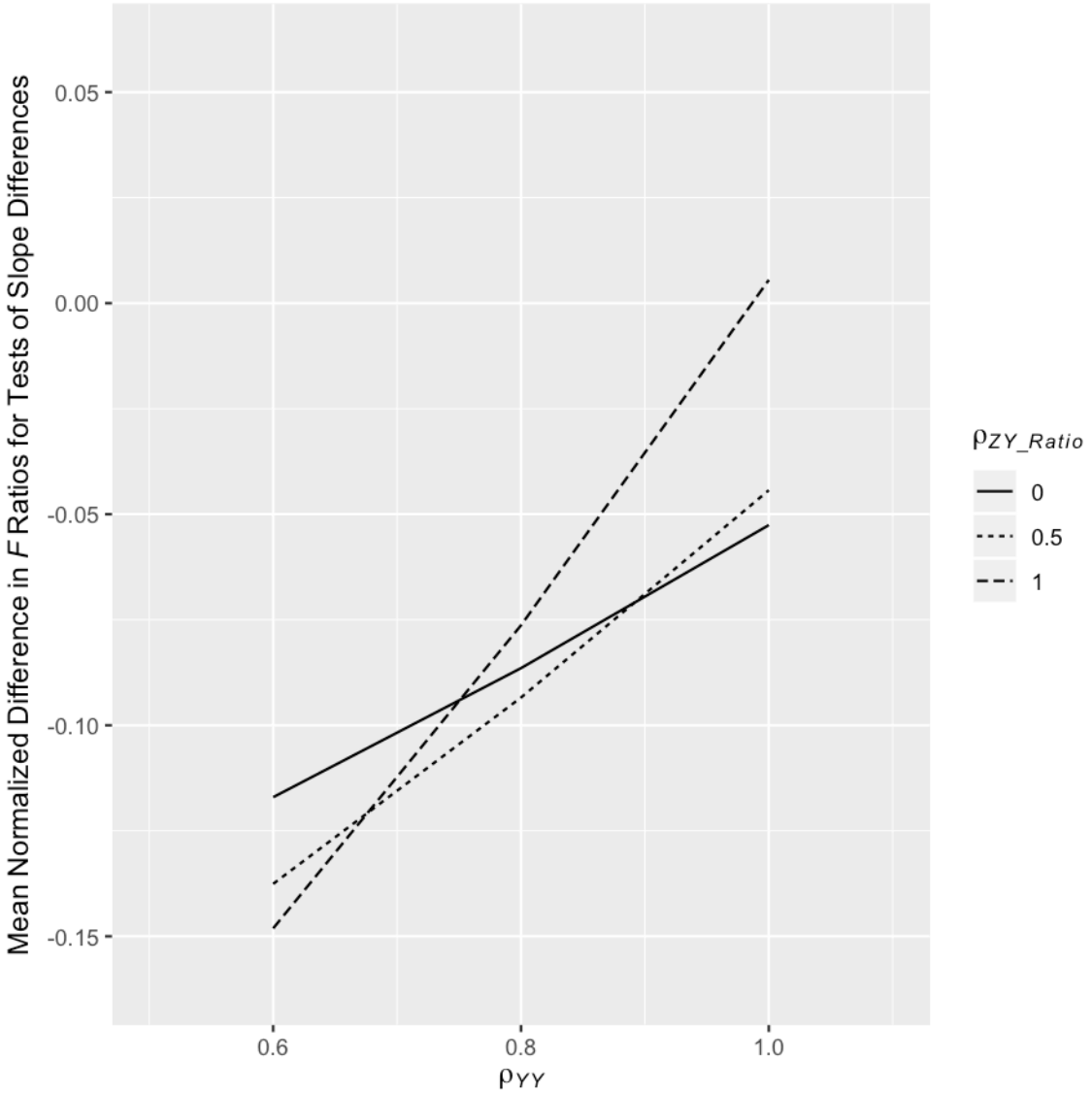


Figure 48

Effect of the two-way interaction between ρ_{ZY_Ratio} and ρ_{YY} on the power of tests of slope differences (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. ρ_{ZY_Ratio} = ratio of the validity of *Z* for predicting *Y* in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{YY} = reliability of *Y*.

Figure is based on data from 13,122 (66.7%) indirect range restriction conditions.

Total η^2 of plotted effects = .16.

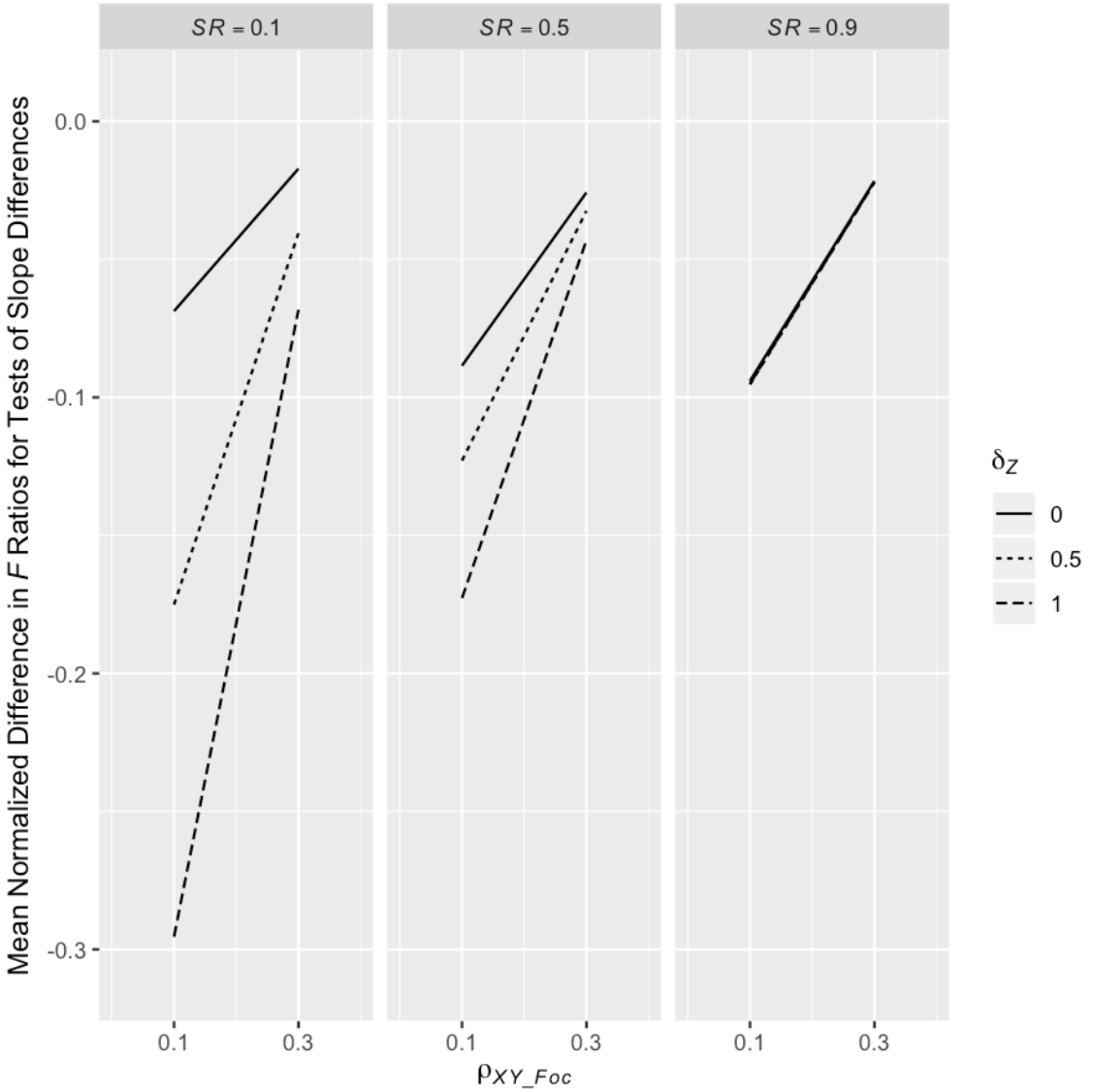


Figure 49

Effect of the three-way interaction among SR, ρ_{XY_Foc}, and δ_Z on the power of tests of slope differences (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; ρ_{XY_Foc} = operational validity of X for predicting Y in the focal group's applicant population; δ_Z = standardized mean difference between the referent and focal groups on Z.

Figure is based on data from 13,122 (66.7%) indirect range restriction conditions. Total η^2 of plotted effects = .38.

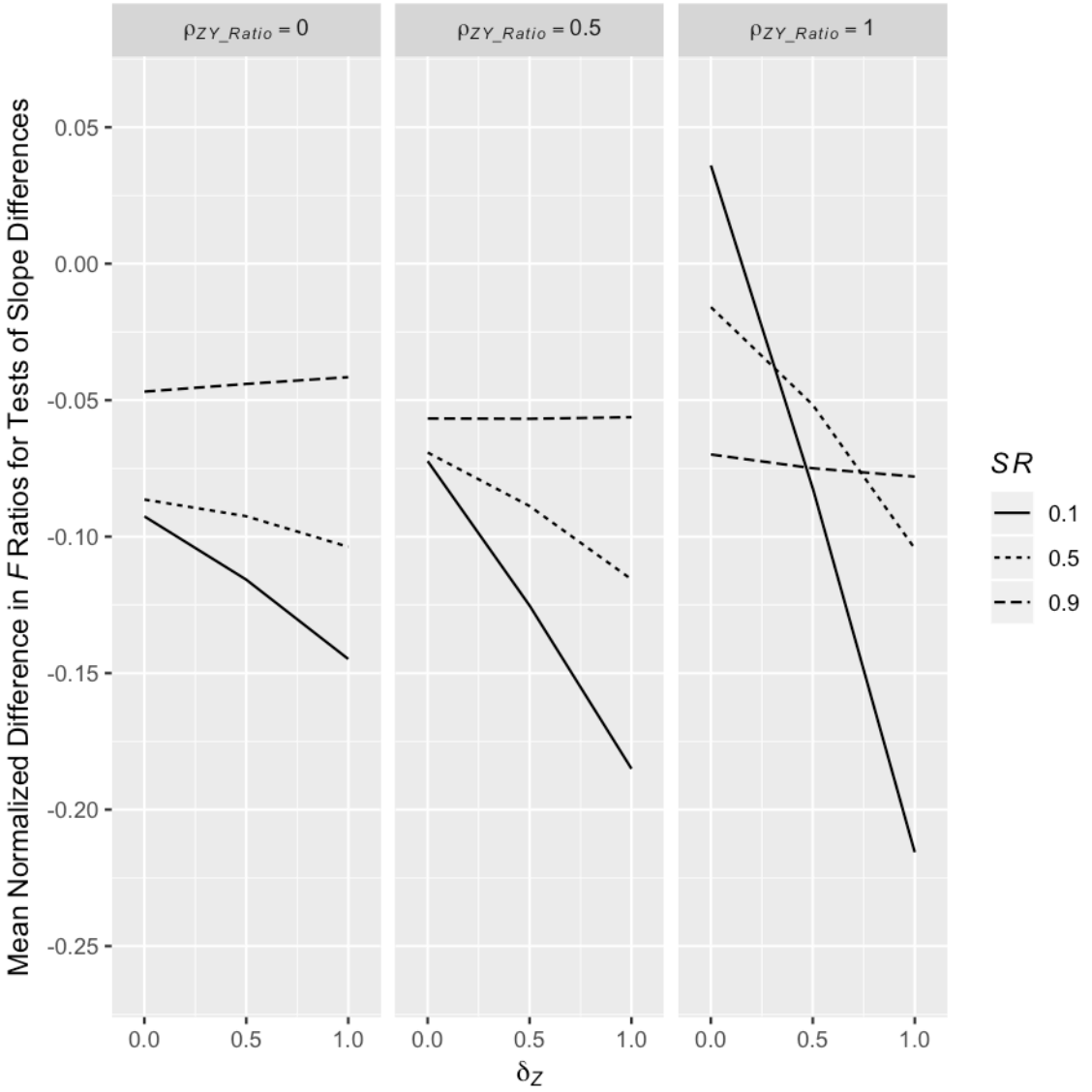


Figure 50

Effect of the three-way interaction among SR, ρ_{ZY_Ratio}, and δ_Z on the power of tests of slope differences (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; δ_Z = standardized mean difference between the referent and focal groups on Z.

Figure is based on data from 13,122 (66.7%) indirect range restriction conditions. Total η^2 of plotted effects = .18.

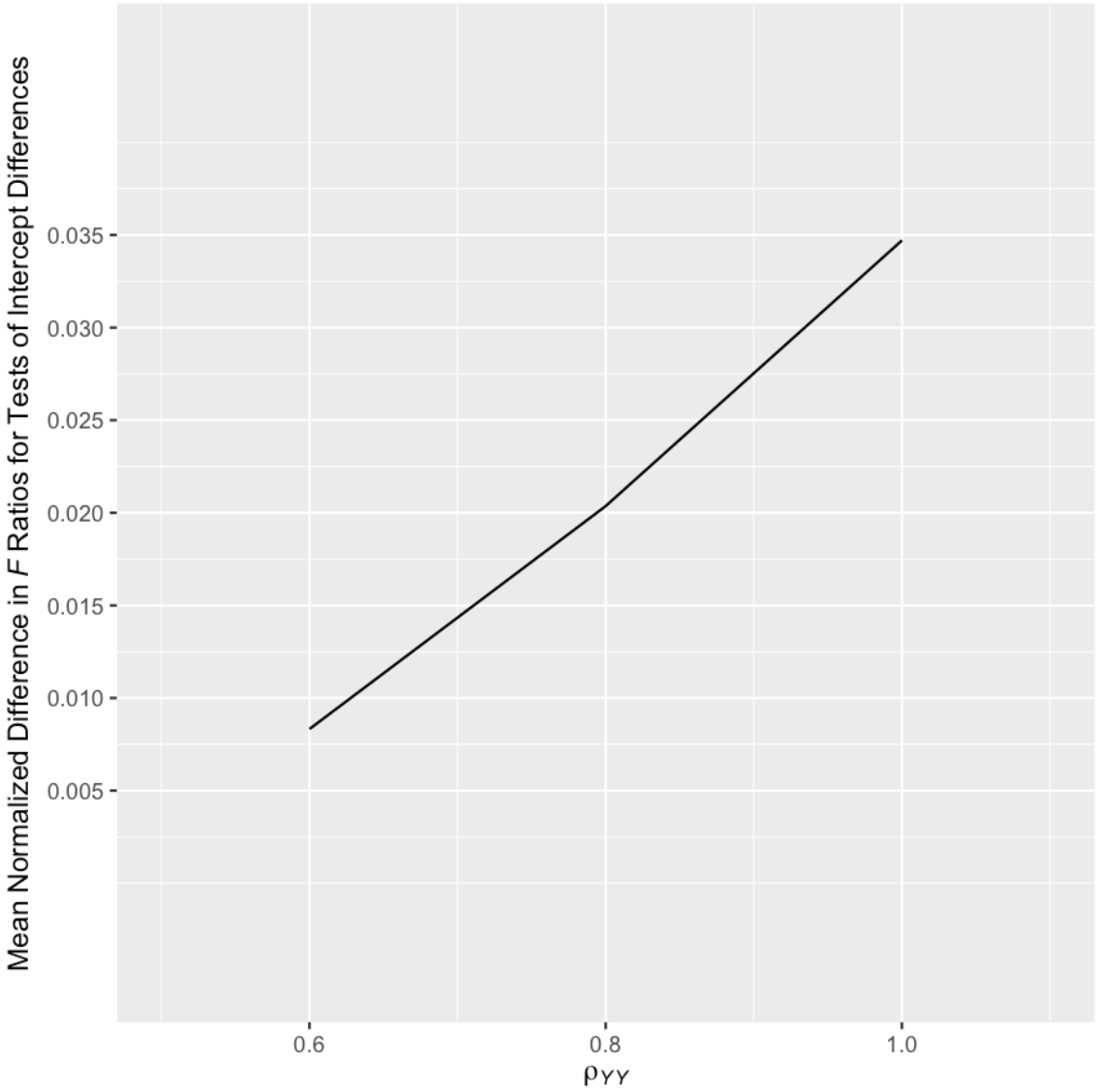


Figure 51
Main effect of ρ_{YY} on the power of tests of intercept differences (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction. Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. ρ_{YY} = reliability of *Y*.
Figure is based on data from 4,374 (22.2%) indirect range restriction conditions.
Total η^2 of plotted effects = .02.

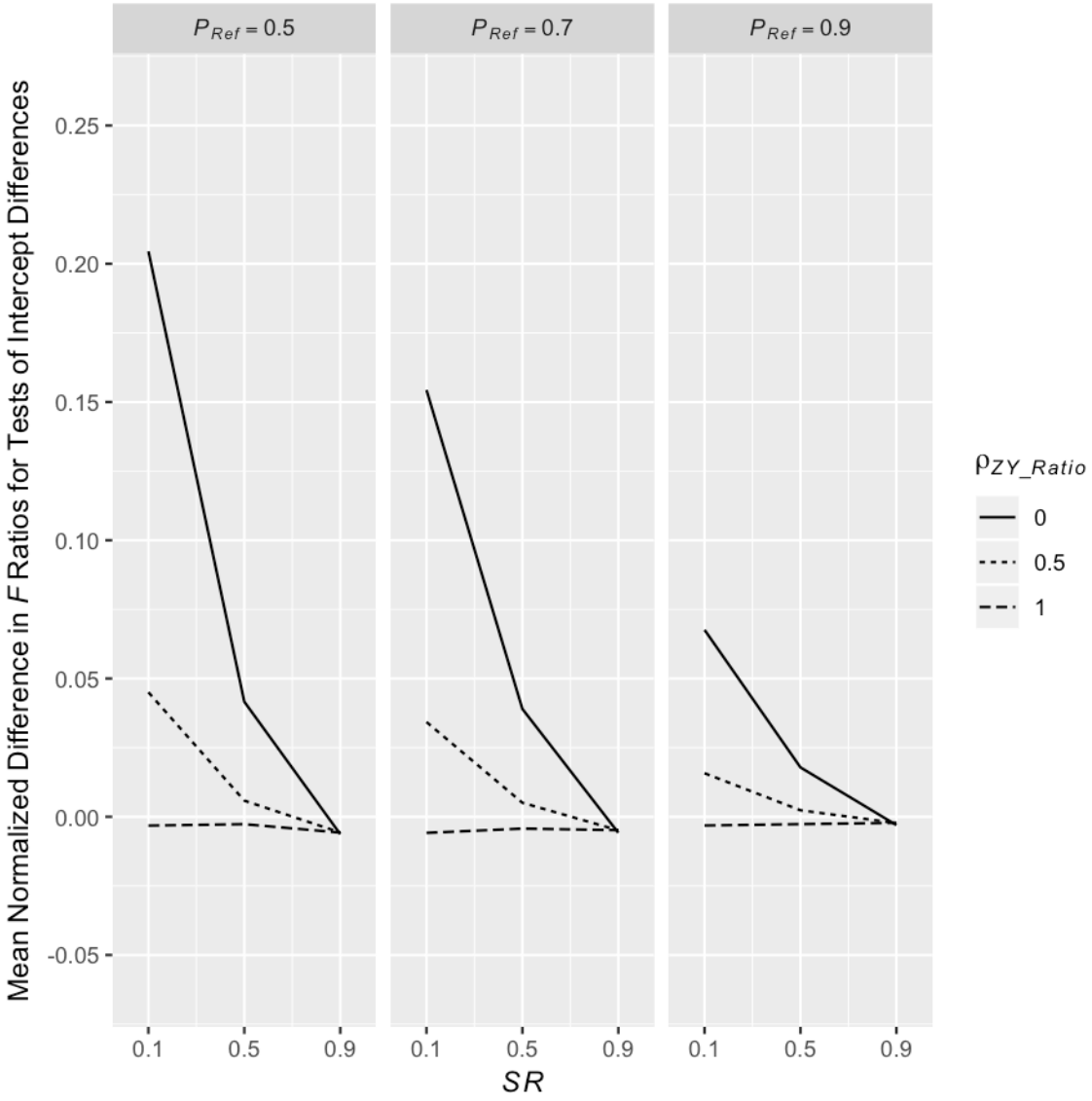


Figure 52

Effect of the three-way interaction among SR, P_{Ref}, and ρ_{ZY_Ratio} on the power of tests of intercept differences (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; P_{Ref} = proportion of referent-group members in the applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population.

Figure is based on data from 4,374 (22.2%) indirect range restriction conditions.

Total η^2 of plotted effects = .37.

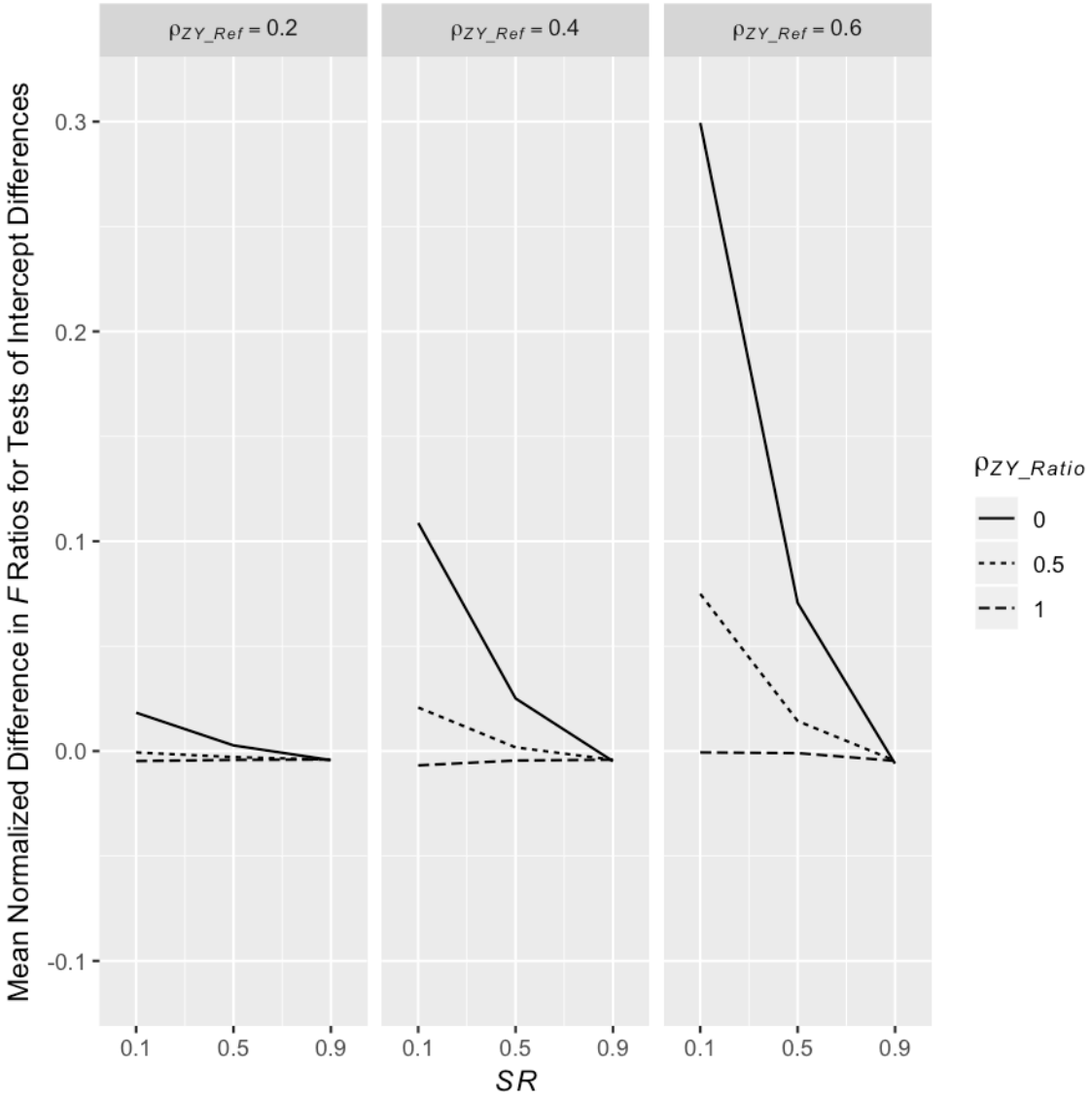


Figure 53

Effect of the three-way interaction among SR, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} on the power of tests of intercept differences (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population.

Figure is based on data from 4,374 (22.2%) indirect range restriction conditions.

Total η^2 of plotted effects = .58.

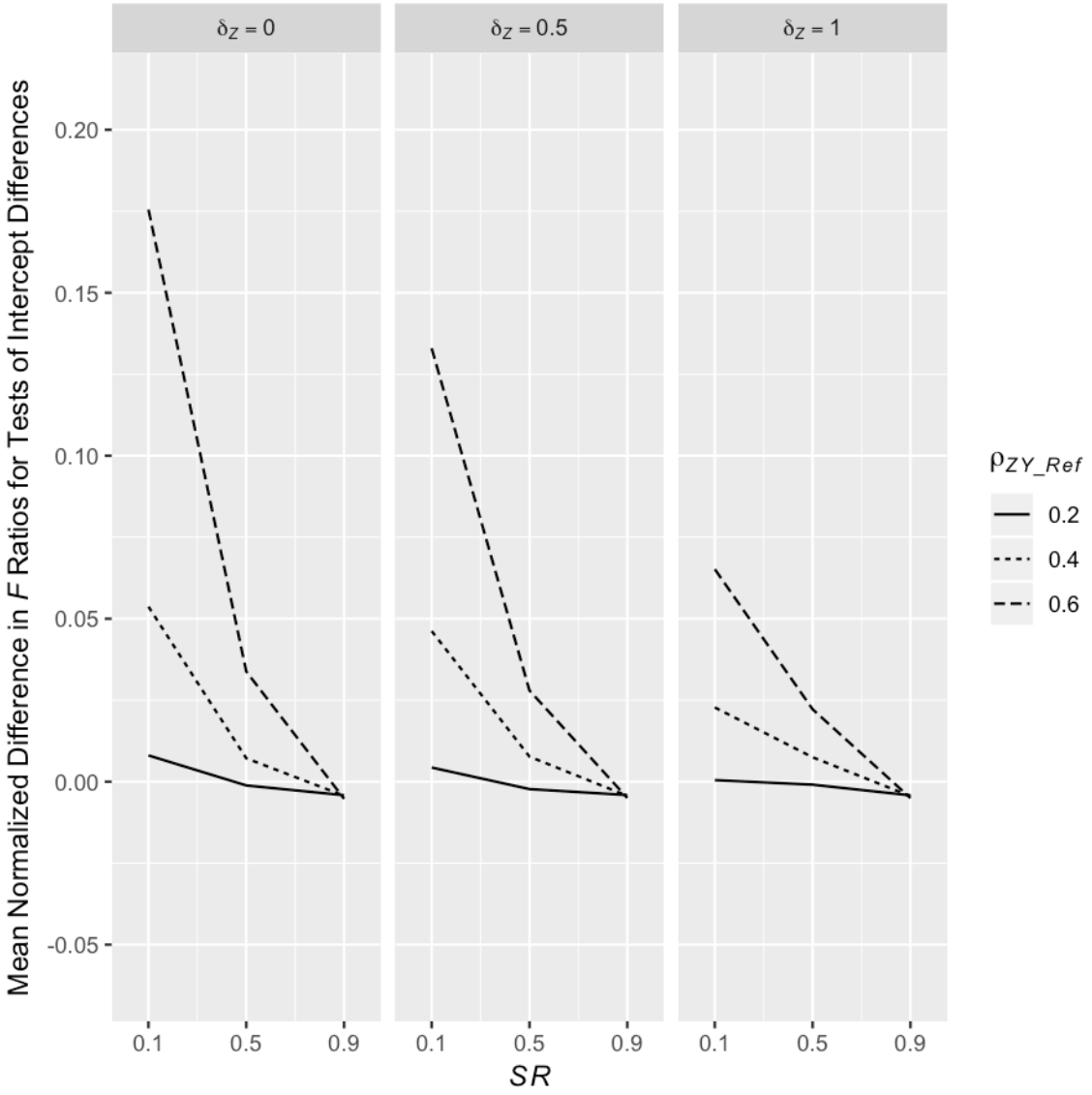


Figure 54

Effect of the three-way interaction among SR, ρ_{ZY_Ref}, and δ_Z on the power of tests of intercept differences (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; δ_Z = standardized mean difference between the referent and focal groups on Z.

Figure is based on data from 4,374 (22.2%) indirect range restriction conditions.

Total η^2 of plotted effects = .28.

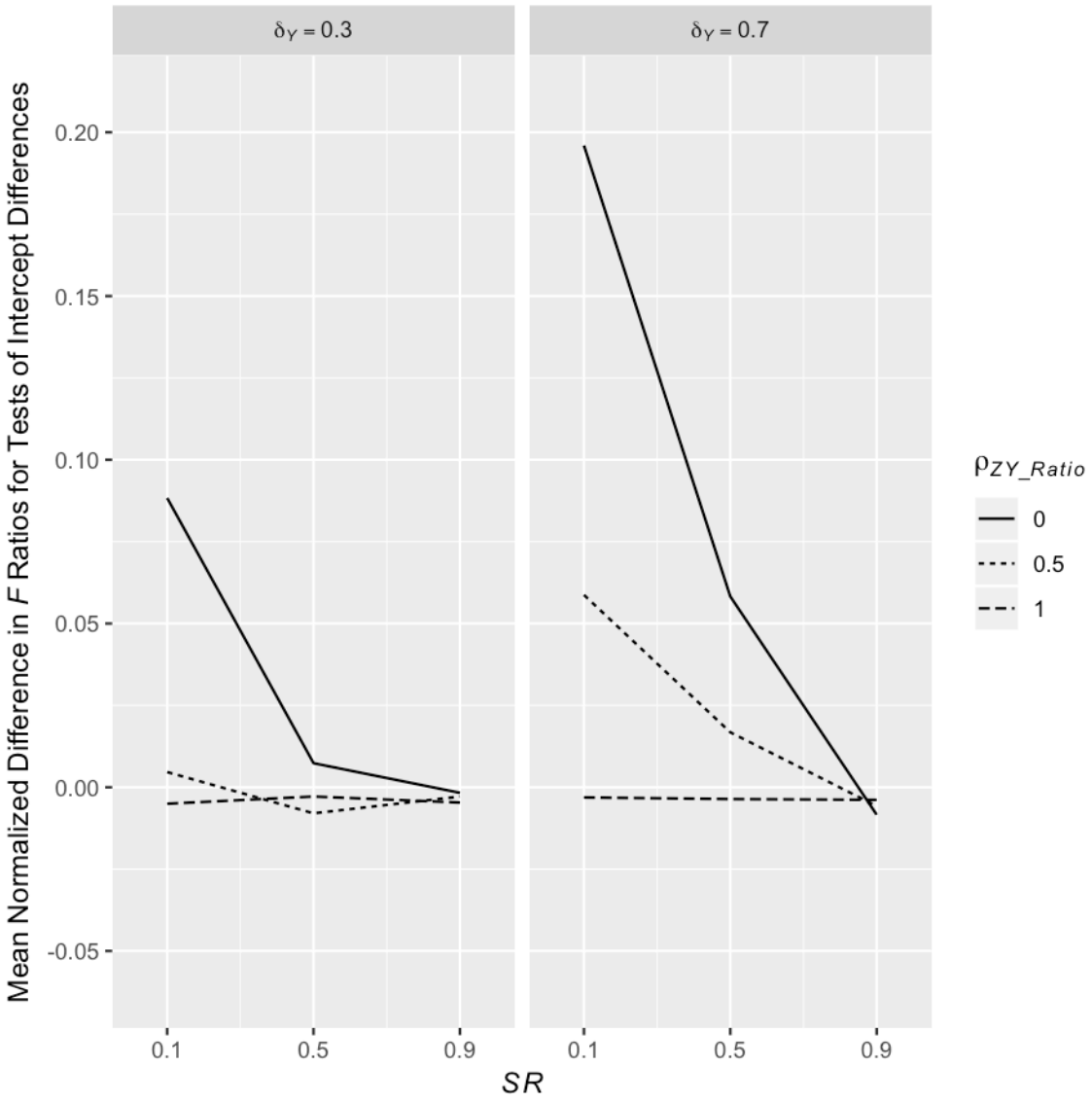


Figure 55

Effect of the three-way interaction among SR, ρ_{ZY_Ratio}, and δ_Y on the power of tests of intercept differences (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate greater power (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; δ_Y = true-score standardized mean difference between the referent and focal groups on Y.

Figure is based on data from 4,374 (22.2%) indirect range restriction conditions.

Total η^2 of plotted effects = .39.

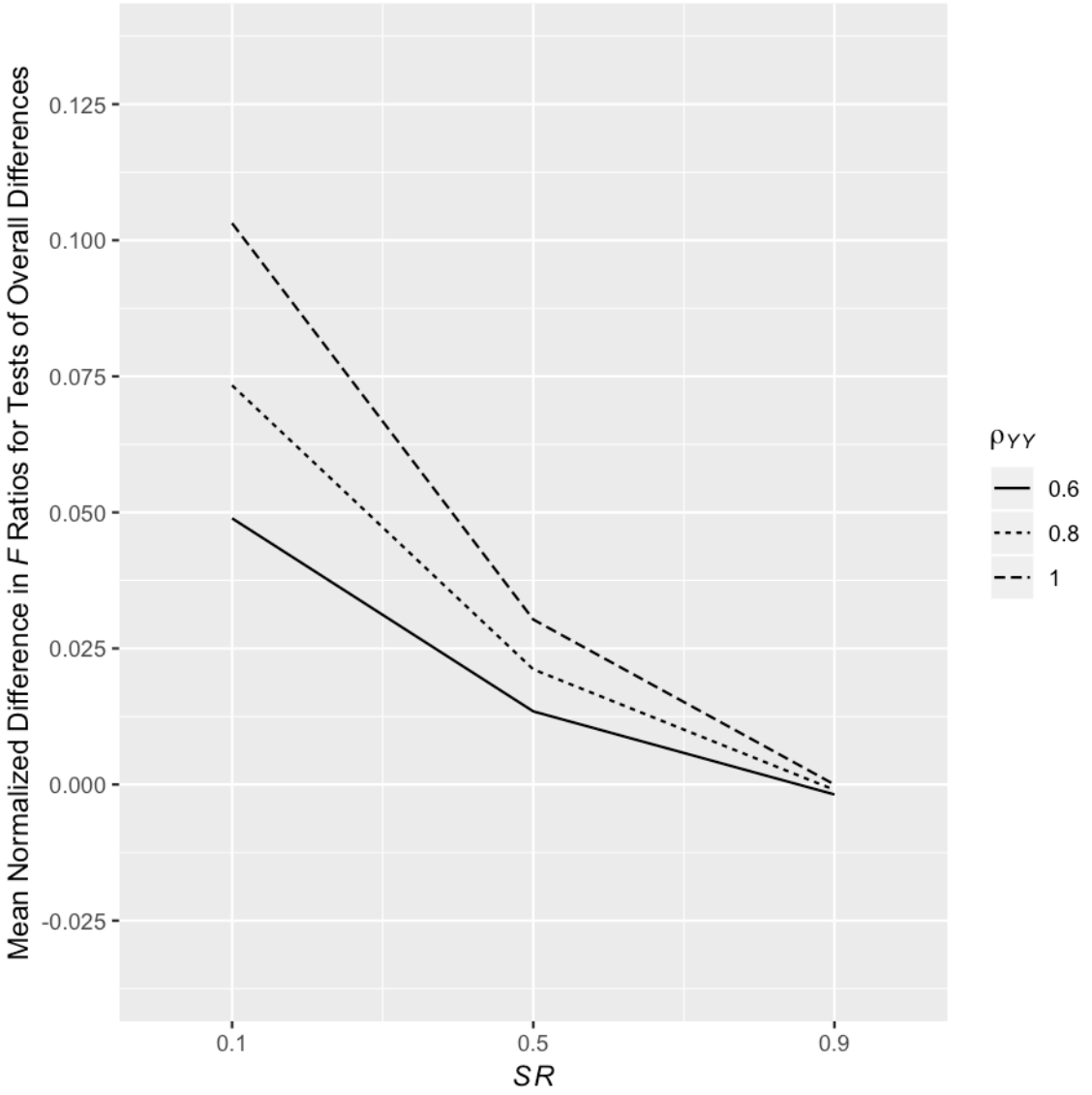


Figure 56

Effect of the two-way interaction between SR and ρ_{YY} on the Type I errors of tests of overall differences in prediction (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate elevated Type I error rates (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; ρ_{YY} = reliability of Y.

Figure is based on data from 2,187 (11.1%) indirect range restriction conditions.

Total η^2 of plotted effects = .15.

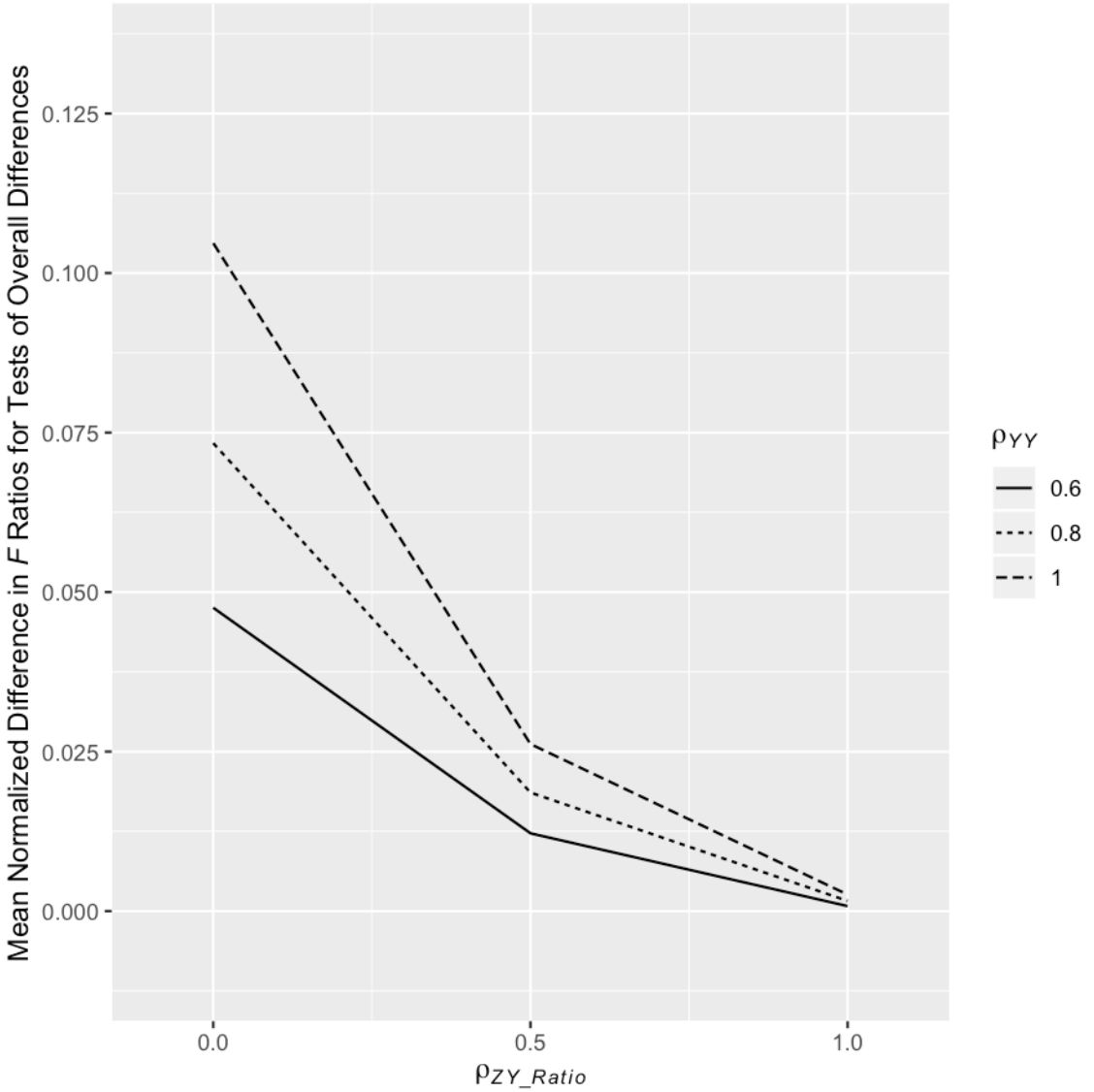


Figure 57

Effect of the two-way interaction between ρ_{ZY_Ratio} and ρ_{YY} on the Type I errors of tests of overall differences in prediction (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate elevated Type I error rates (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data.

ρ_{ZY_Ratio} = ratio of the validity of *Z* for predicting *Y* in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{YY} = reliability of *Y*.

Figure is based on data from 2,187 (11.1%) indirect range restriction conditions.

Total η^2 of plotted effects = .15.

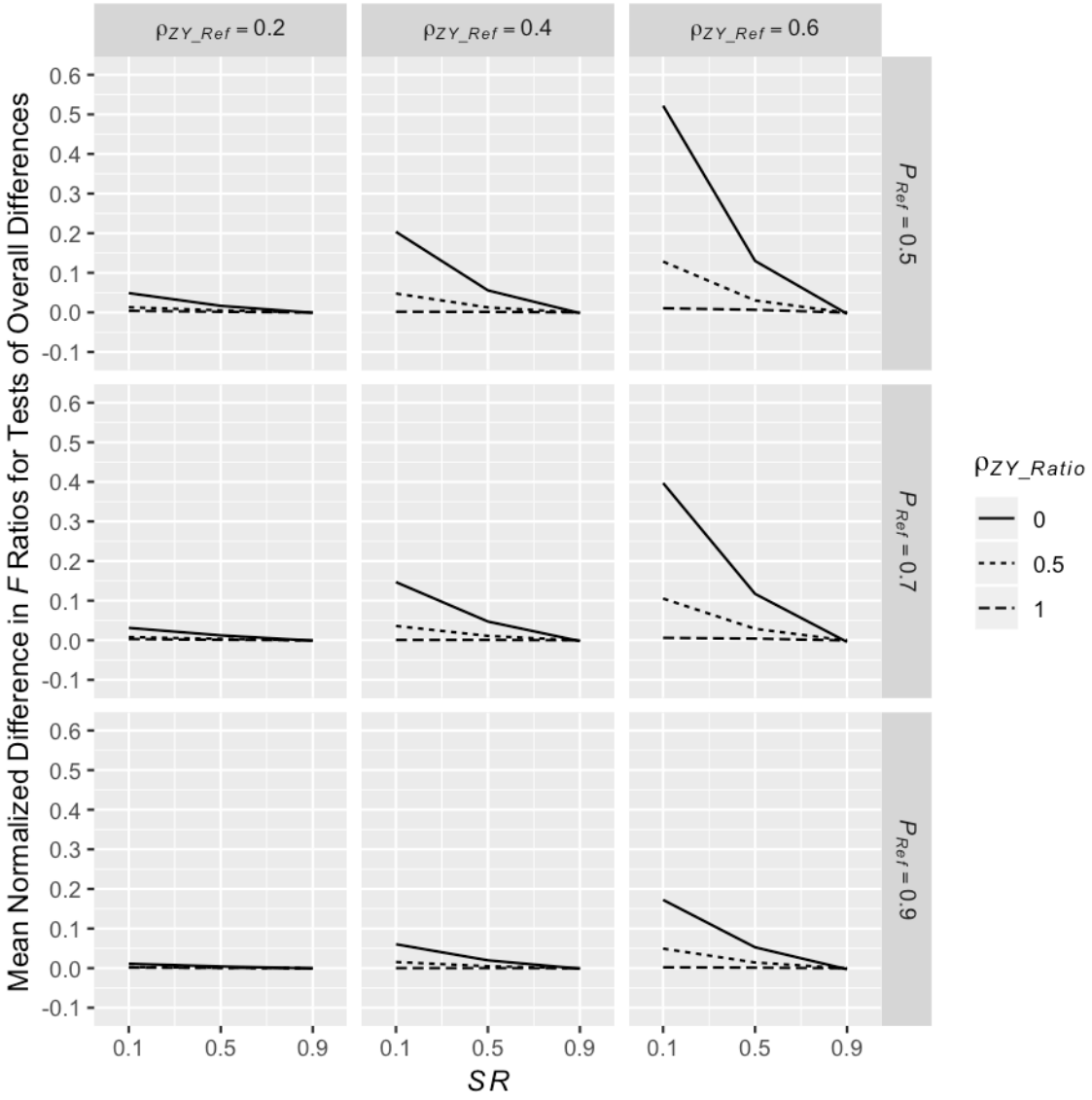


Figure 58

Effect of the four-way interaction among SR, P_{Ref}, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} on the Type I errors of tests of overall differences in prediction (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate elevated Type I error rates (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; P_{Ref} = proportion of referent-group members in the applicant population; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population.

Figure is based on data from 2,187 (11.1%) indirect range restriction conditions.

Total η^2 of plotted effects = .79.

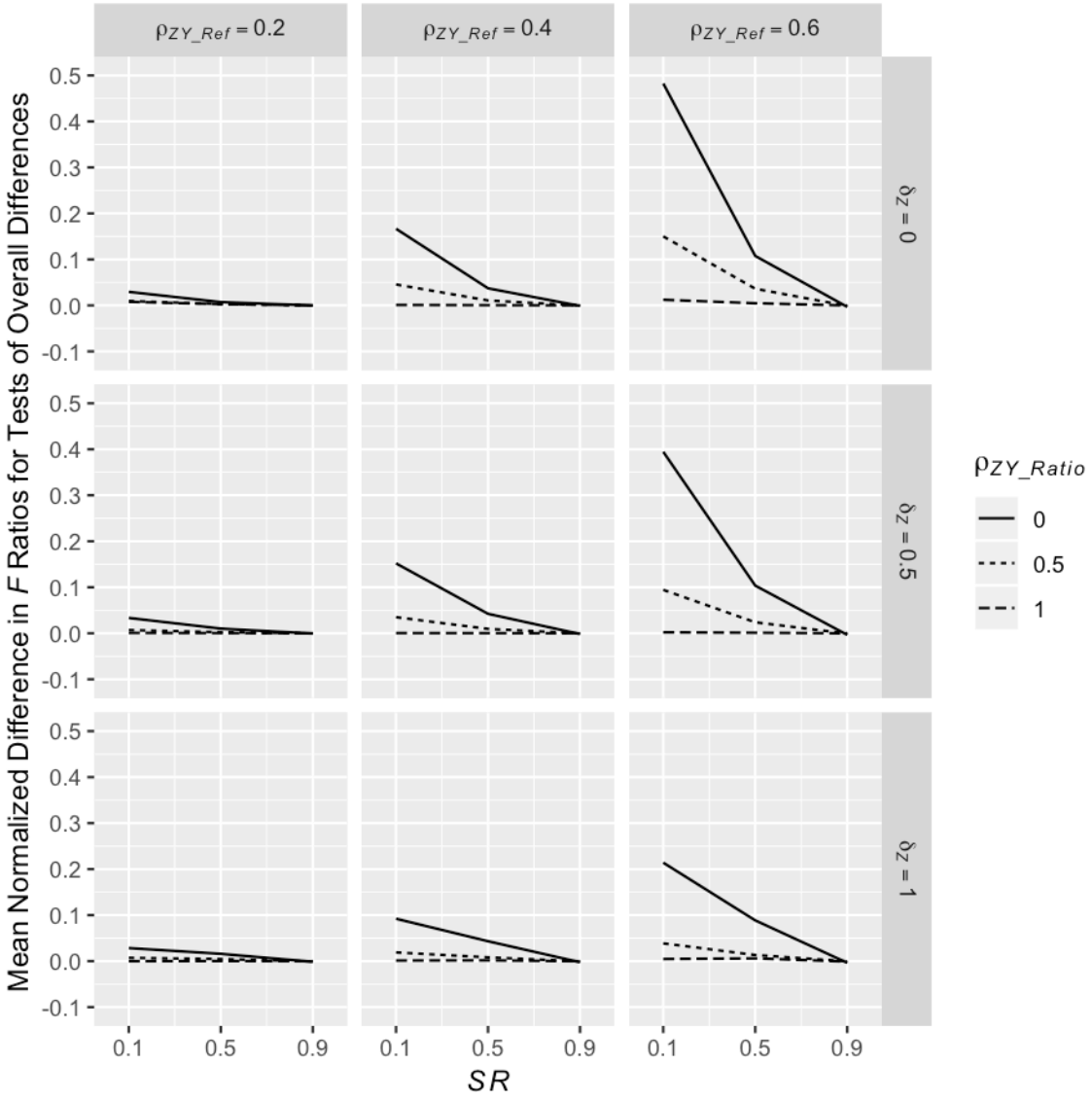


Figure 59

Effect of the four-way interaction among SR, ρ_{ZY_Ref}, ρ_{ZY_Ratio}, and δ_Z on the Type I errors of tests of overall differences in prediction (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate elevated Type I error rates (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; δ_Z = standardized mean difference between the referent and focal groups on Z.

Figure is based on data from 2,187 (11.1%) indirect range restriction conditions.

Total η^2 of plotted effects = .74.

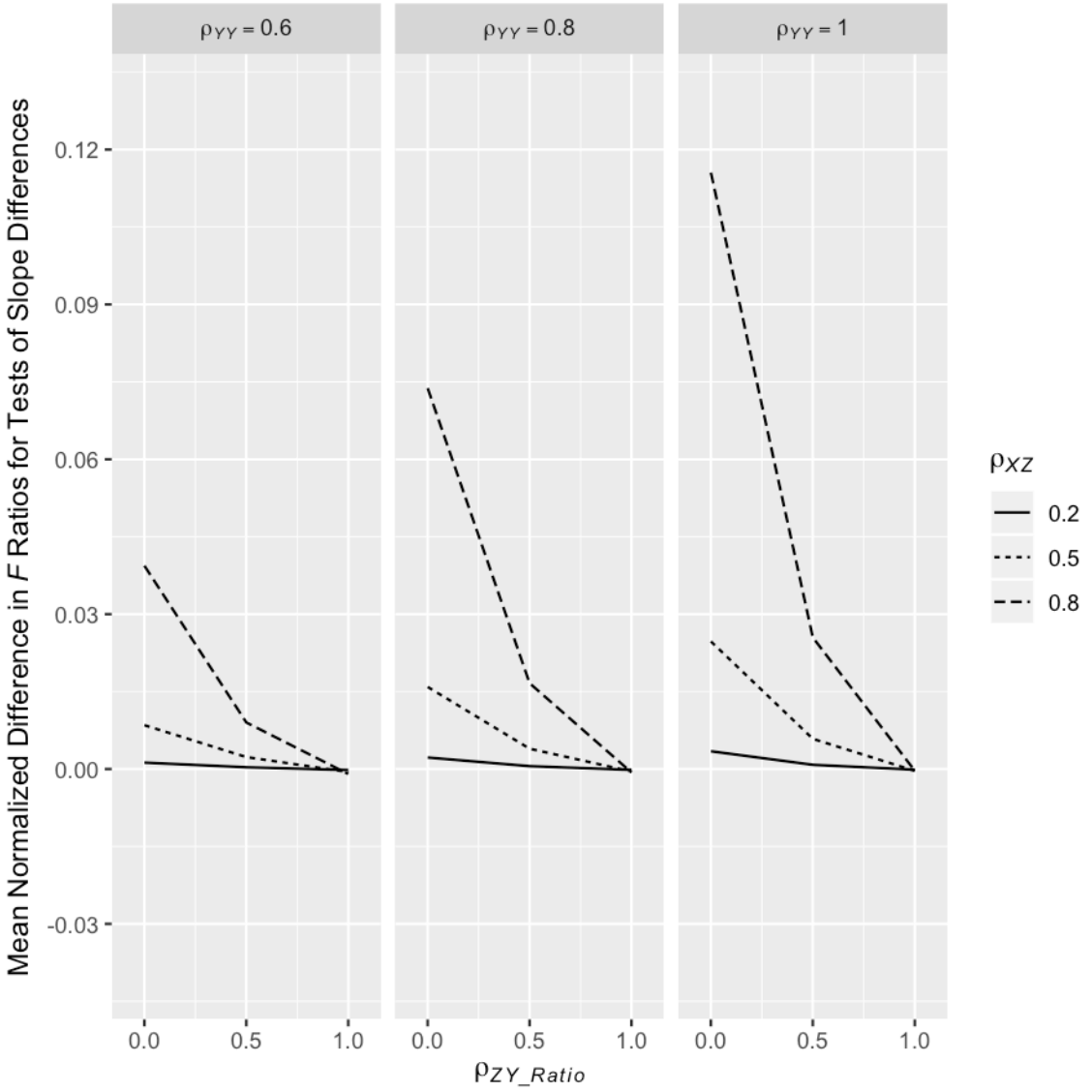


Figure 60

Effect of the three-way interaction among ρ_{ZY_Ratio}, ρ_{XZ}, and ρ_{YY} on the Type I errors of tests of slope differences (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate elevated Type I error rates (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. ρ_{ZY_Ratio} = ratio of the validity of *Z* for predicting *Y* in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between *X* and *Z* in the referent and focal groups' applicant populations; ρ_{YY} = reliability of *Y*.

Figure is based on data from 6,561 (33.3%) indirect range restriction conditions.

Total η^2 of plotted effects = .22.

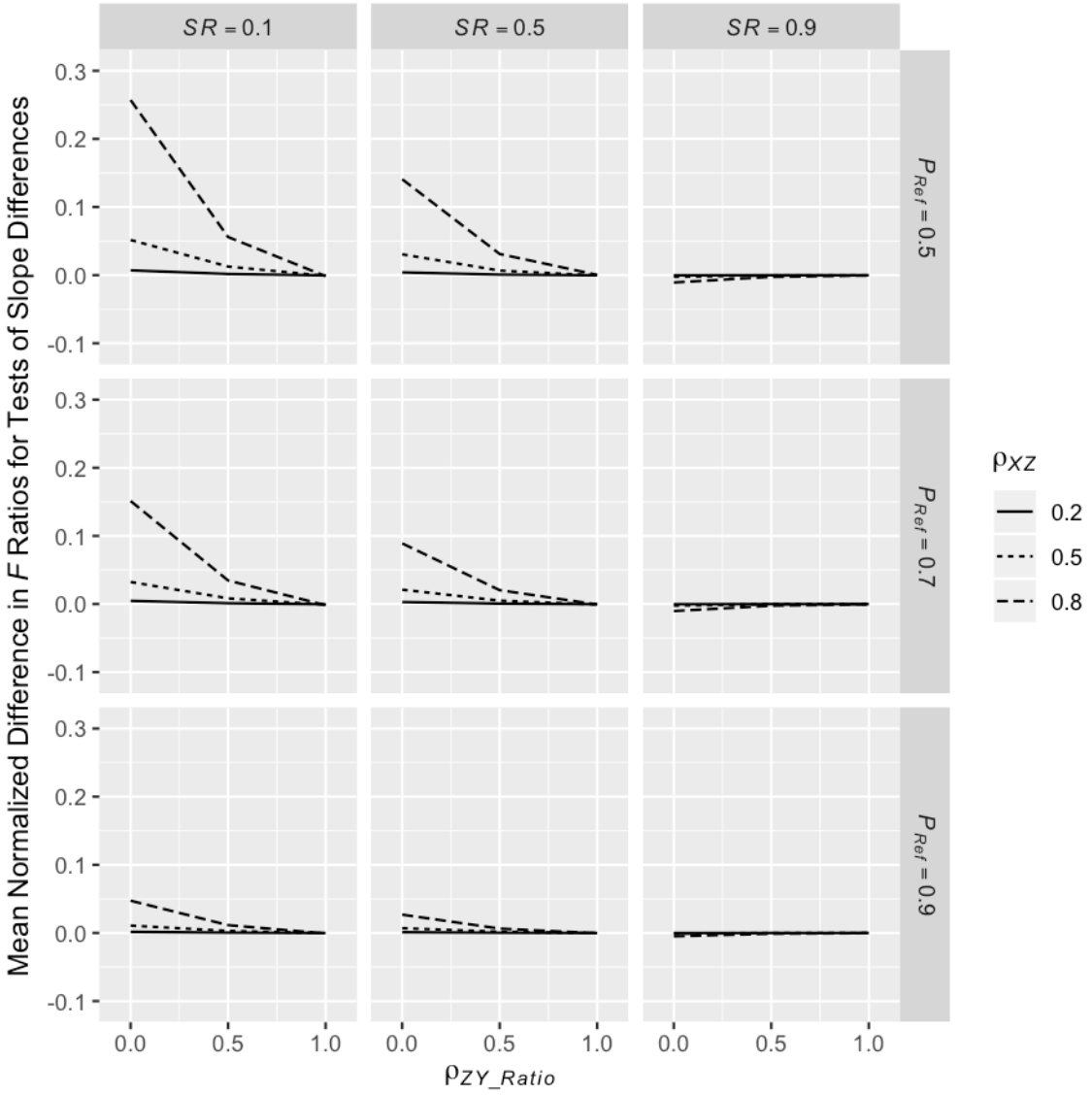


Figure 61

Effect of the four-way interaction among SR, P_{Ref}, ρ_{ZY_Ratio}, and ρ_{XZ} on the Type I errors of tests of slope differences (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate elevated Type I error rates (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; P_{Ref} = proportion of referent-group members in the applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between X and Z in the referent and focal groups' applicant populations.

Figure is based on data from 6,561 (33.3%) indirect range restriction conditions.

Total η^2 of plotted effects = .46.

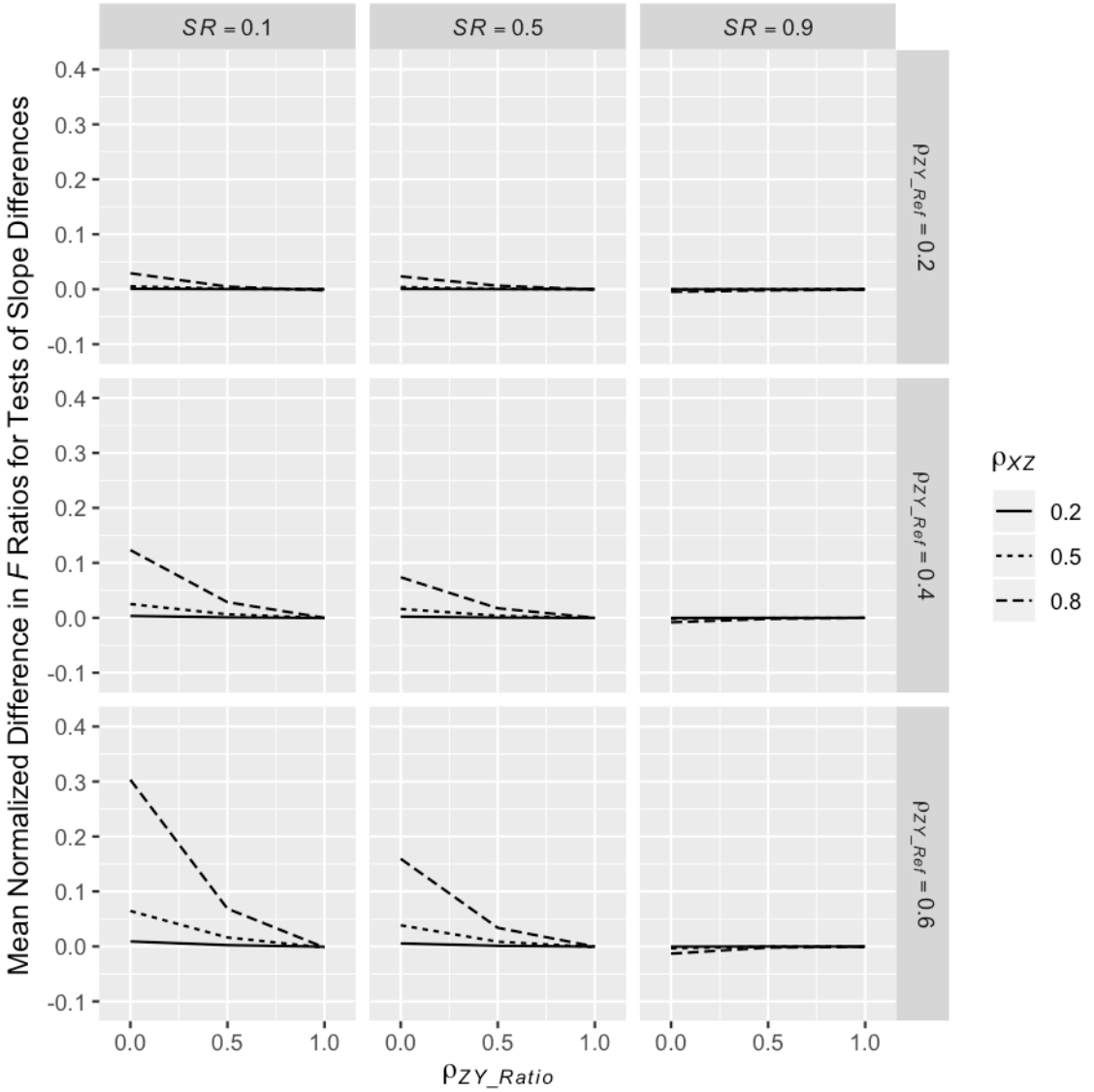


Figure 62

Effect of the four-way interaction among SR, ρ_{ZY_Ref}, ρ_{ZY_Ratio}, and ρ_{XZ} on the Type I errors of tests of slope differences (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate elevated Type I error rates (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between X and Z in the referent and focal groups' applicant populations.

Figure is based on data from 6,561 (33.3%) indirect range restriction conditions.

Total η^2 of plotted effects = .54.

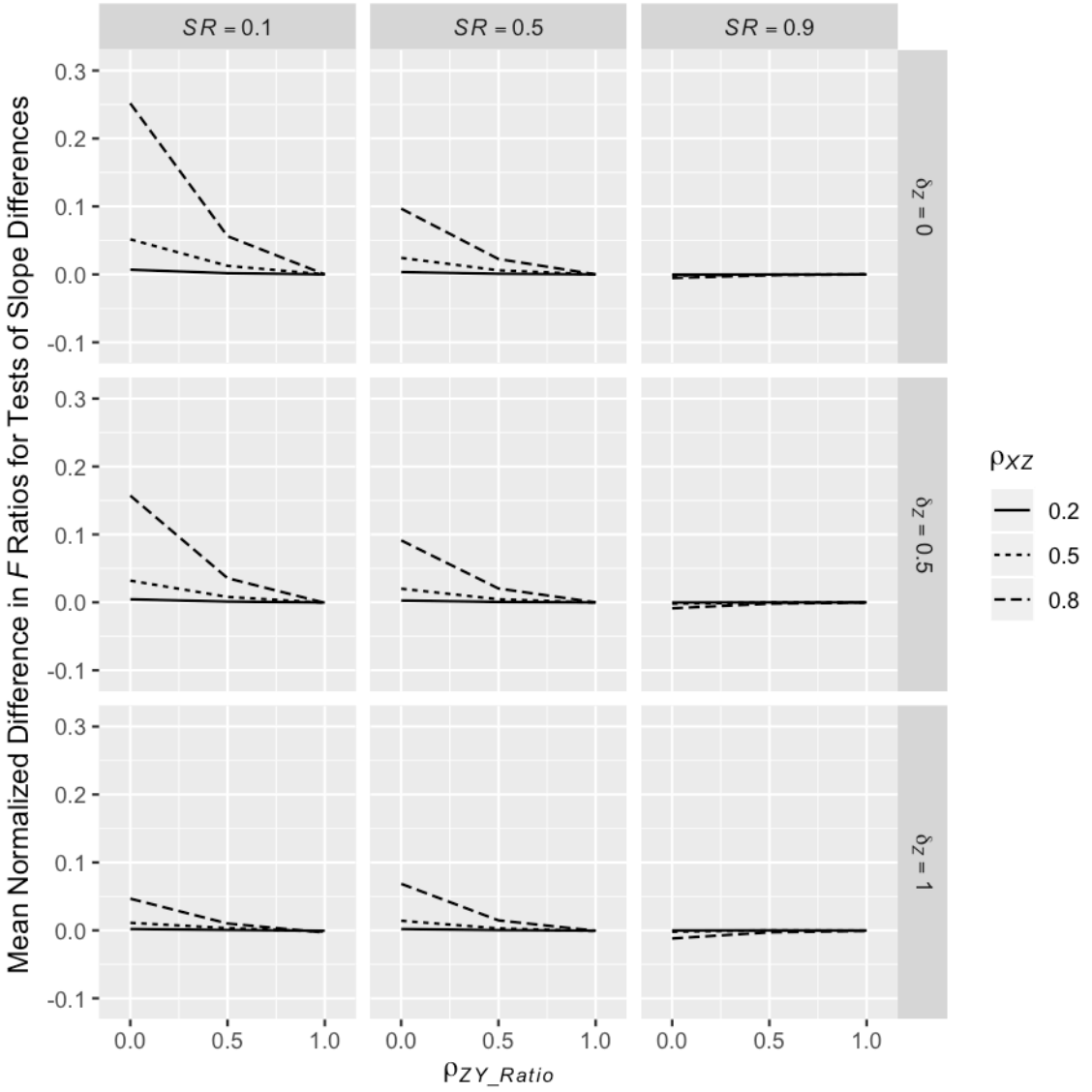


Figure 63

Effect of the four-way interaction among SR, ρ_{ZY_Ratio}, ρ_{XZ}, and δ_Z on the Type I errors of tests of slope differences (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate elevated Type I error rates (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{XZ} = correlation between X and Z in the referent and focal groups' applicant populations; δ_Z = standardized mean difference between the referent and focal groups on Z.

Figure is based on data from 6,561 (33.3%) indirect range restriction conditions.

Total η^2 of plotted effects = .43.

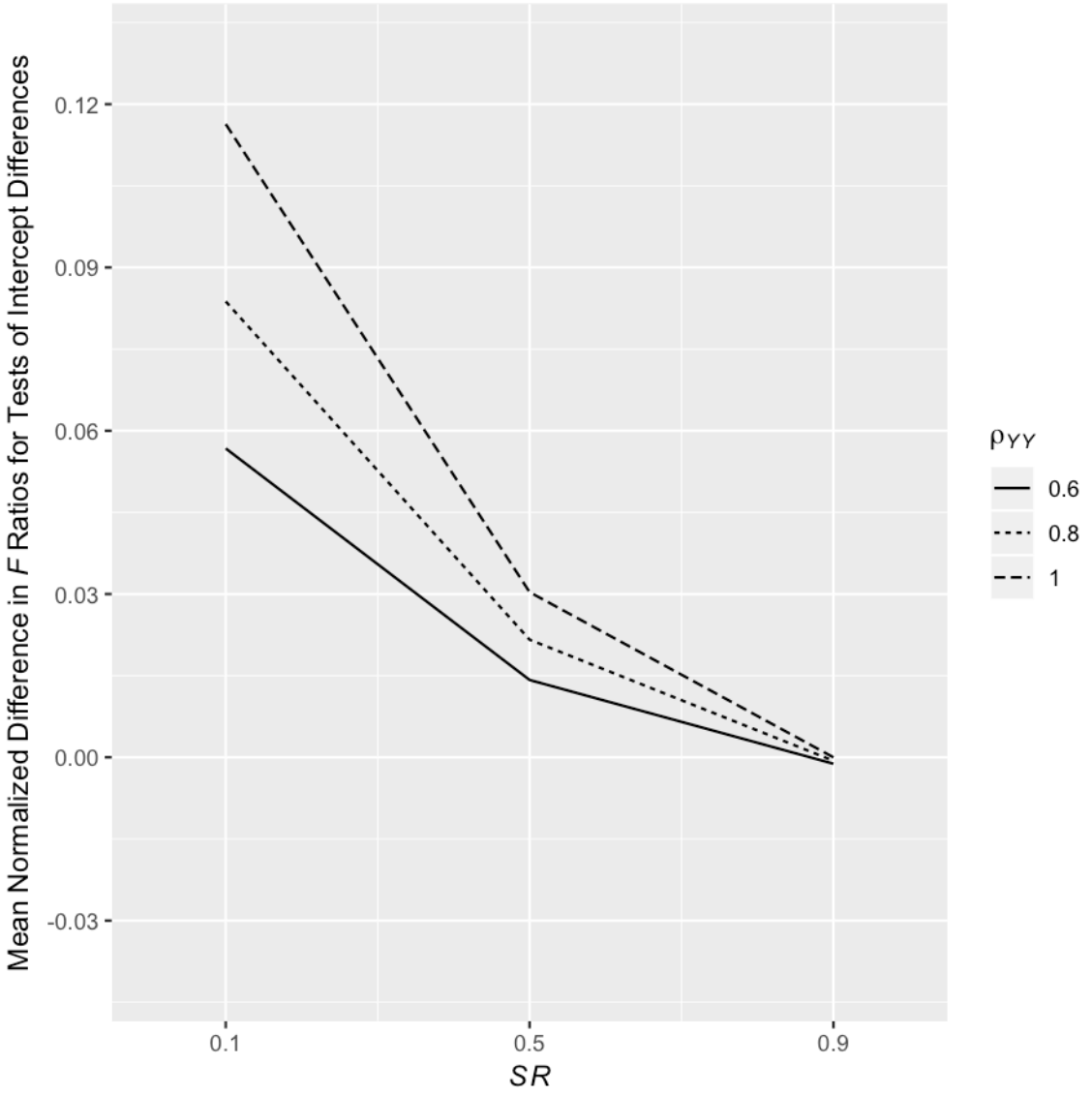


Figure 64

Effect of the two-way interaction between SR and ρ_{YY} on the Type I errors of tests of intercept differences (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate elevated Type I error rates (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; ρ_{YY} = reliability of Y.

Figure is based on data from 2,187 (11.1%) indirect range restriction conditions.

Total η^2 of plotted effects = .16.

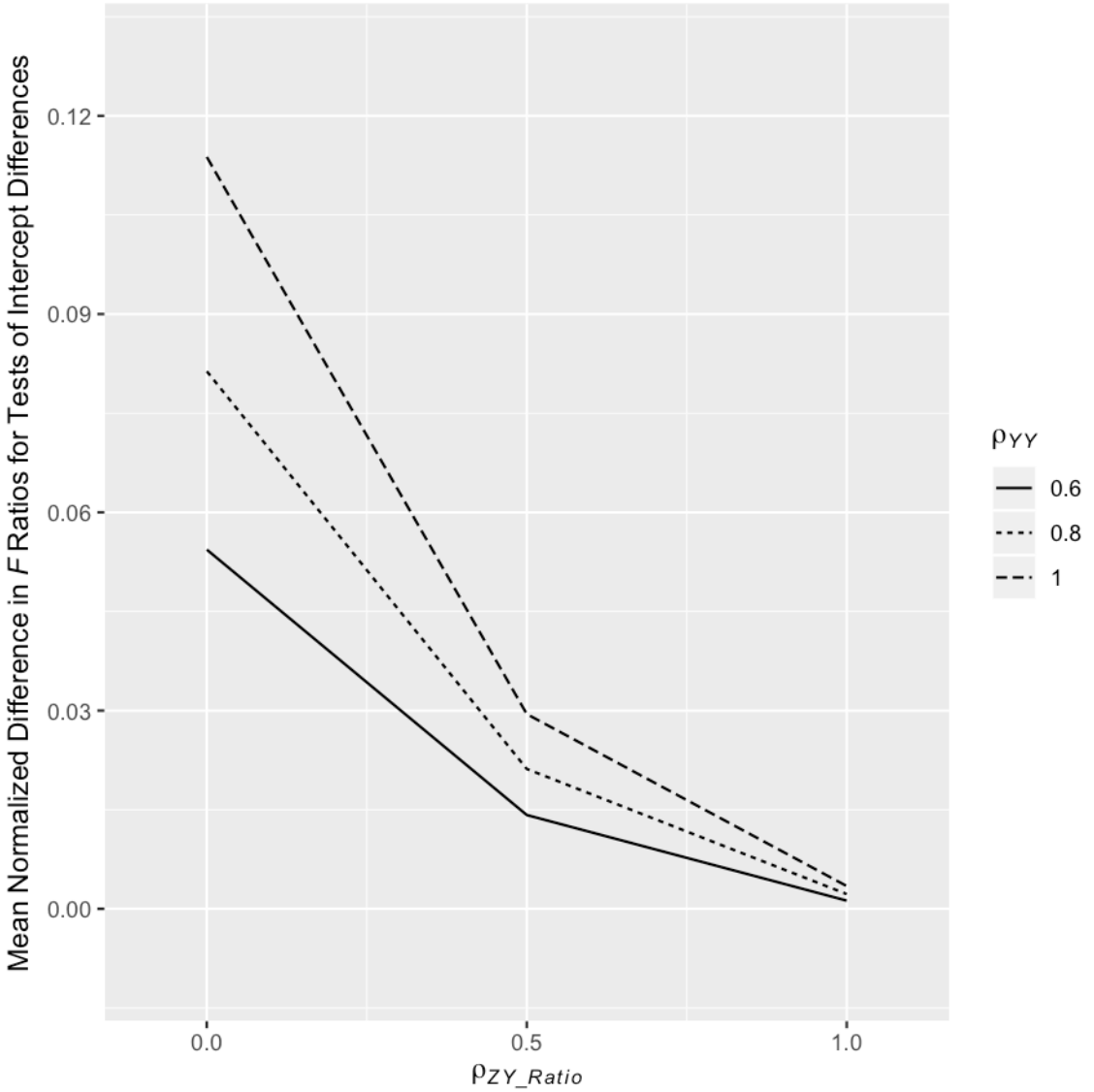


Figure 65

Effect of the two-way interaction between ρ_{ZY_Ratio} and ρ_{YY} on the Type I errors of tests of intercept differences (as indicated by normalized differences in *F* ratios) under conditions of indirect range restriction.

Positive values indicate elevated Type I error rates (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. ρ_{ZY_Ratio} = ratio of the validity of *Z* for predicting *Y* in the focal group's applicant population to the validity in the referent group's applicant population; ρ_{YY} = reliability of *Y*.

Figure is based on data from 2,187 (11.1%) indirect range restriction conditions.

Total η^2 of plotted effects = .14.

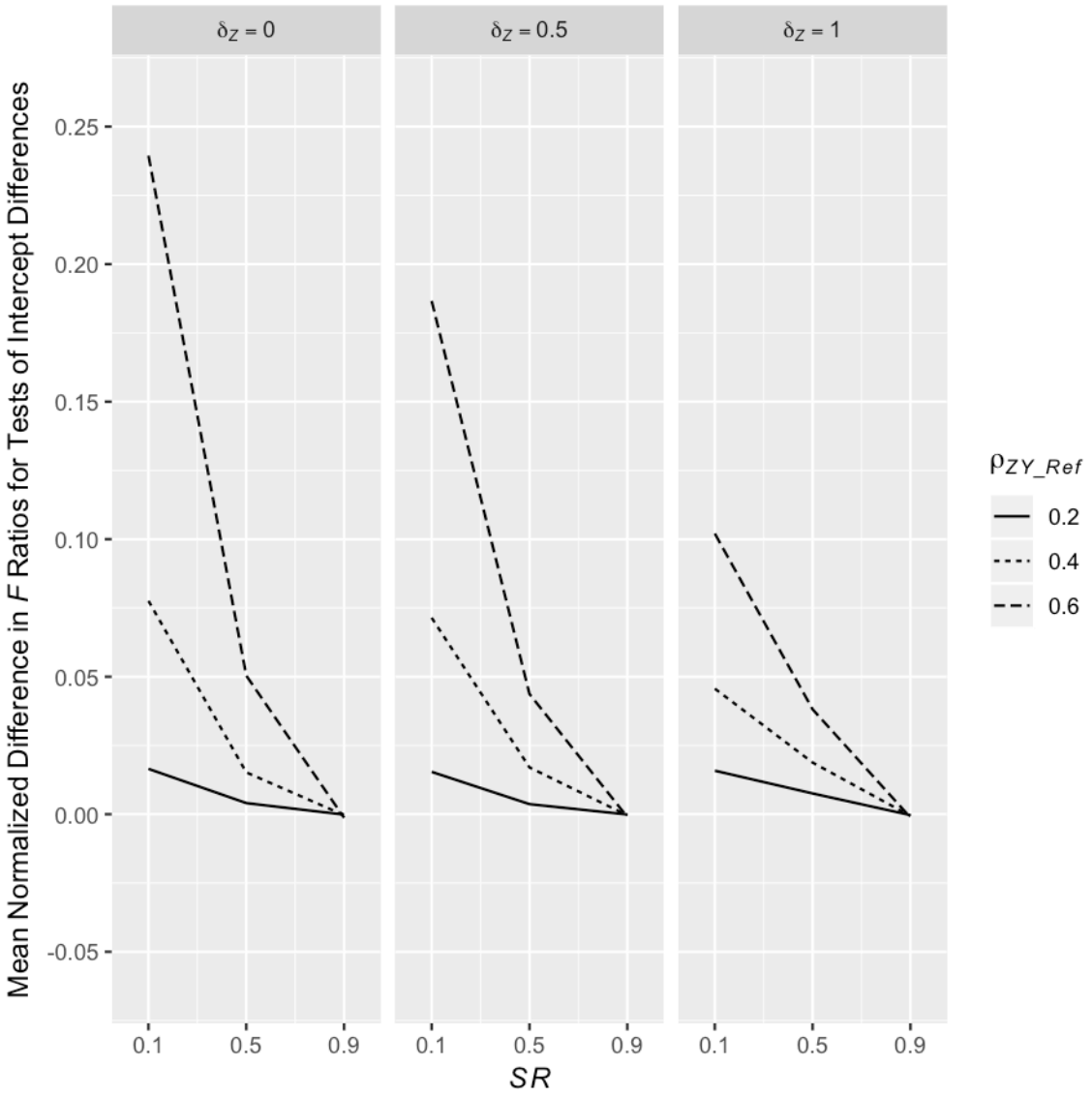


Figure 66

Effect of the three-way interaction among SR, ρ_{ZY_Ref}, and δ_Z on the Type I errors of tests of intercept differences (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate elevated Type I error rates (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; δ_Z = standardized mean difference between the referent and focal groups on Z.

Figure is based on data from 2,187 (11.1%) indirect range restriction conditions.

Total η^2 of plotted effects = .33.

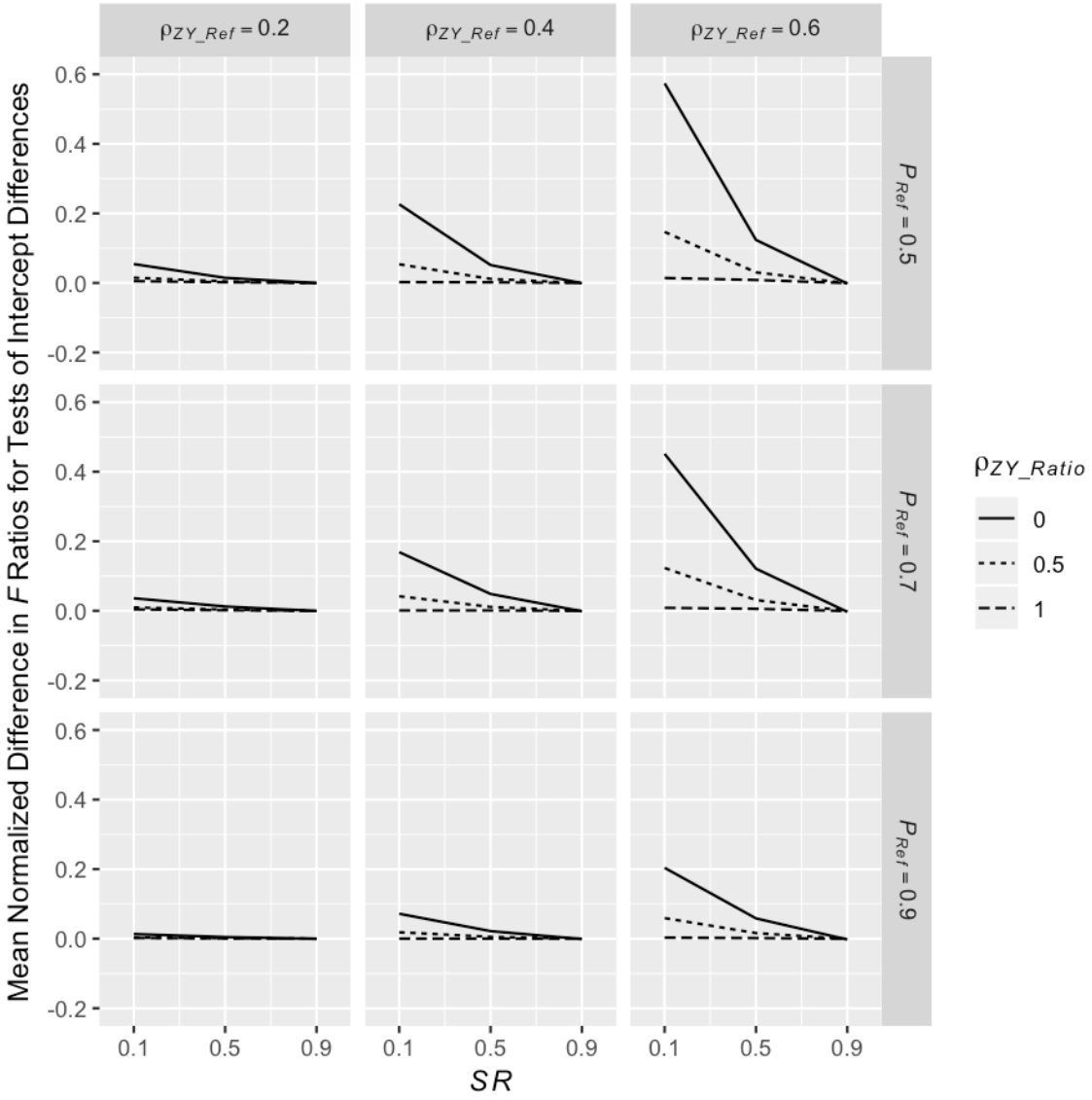


Figure 67

Effect of the four-way interaction among SR, P_{Ref}, ρ_{ZY_Ref}, and ρ_{ZY_Ratio} on the Type I errors of tests of intercept differences (as indicated by normalized differences in F ratios) under conditions of indirect range restriction.

Positive values indicate elevated Type I error rates (i.e., higher likelihood of significant results being estimated) for analyses based on observed data than operational data. SR = overall selection ratio applied to Z; P_{Ref} = proportion of referent-group members in the applicant population; ρ_{ZY_Ref} = operational validity of Z for predicting Y in the referent group's applicant population; ρ_{ZY_Ratio} = ratio of the validity of Z for predicting Y in the focal group's applicant population to the validity in the referent group's applicant population.

Figure is based on data from 2,187 (11.1%) indirect range restriction conditions.

Total η^2 of plotted effects = .80.

Appendix: Procedure for Combining Subgroup Mean Vectors and Covariance Matrices into an Interaction Matrix

To combine subgroup distributions into a mixture distribution that includes a dummy variable and an interaction term, it is first necessary to define how the dummy variable will be coded and which proportion of the total sample comes from each of the two groups. I represent the dummy code assigned to groups 1 and 2 as c_1 and c_2, respectively, and I represent the proportions of members from groups 1 and 2 as $p_1 = n_1/(n_1 + n_2)$ and $p_2 = n_2/n_1 + n_2$, respectively.

In the procedure described here, two continuous variables are contained within each subgroups multivariate distribution: A criterion denoted as *Y* and a predictor denoted as *X*. The variance of *Y* is indicated by s_Y^2, the variance of *X* is indicated by s_X^2, and the covariance between *X* and *Y* is indicated by s_{XY}; subgroup variances and covariances for groups 1 and 2 will be indicted by corresponding subscripts in the equations below.

To create a mixture distribution that includes a dummy variable and an interaction term, it is first necessary to organize subgroup's statistics into within-group covariance matrices as shown in Equation A1.

$$\mathbf{S}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & s_{Y_1}^2 & s_{XY_1} & c_1 \times s_{XY_1} \\ 0 & s_{XY_1} & s_{X_1}^2 & c_1 \times s_{X_1}^2 \\ 0 & c_1 \times s_{XY_1} & c_1 \times s_{X_1}^2 & c_1^2 \times s_{X_1}^2 \end{bmatrix} \quad \text{A1a}$$

$$\mathbf{S}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & s_{Y_2}^2 & s_{XY_2} & c_2 \times s_{XY_2} \\ 0 & s_{XY_2} & s_{X_2}^2 & c_2 \times s_{X_2}^2 \\ 0 & c_2 \times s_{XY_2} & c_2 \times s_{X_2}^2 & c_2^2 \times s_{X_2}^2 \end{bmatrix} \quad \text{A1b}$$

These subgroup matrices each have a row and column of zeroes that indicate where the dummy-variable information will go, they contain the variances and covariances of *X* and *Y*, and they contain a row and column in which the dummy codes c_1 and c_2 are used to define an interaction variable that represents the product of the dummy variable and the predictor. The subgroup matrices in Equation A1 need to be accompanied by the corresponding vectors of means shown in Equation A2.

$$\mathbf{m}_1 = [c_1 \quad \bar{Y}_1 \quad \bar{X}_1 \quad c_1 \times \bar{X}_1] \tag{A2a}$$

$$\mathbf{m}_2 = [c_2 \quad \bar{Y}_2 \quad \bar{X}_2 \quad c_2 \times \bar{X}_2] \tag{A2b}$$

The differences between the subgroups' means can be computed using Equation A3.

$$\mathbf{d} = \mathbf{m}_1 - \mathbf{m}_2 \tag{A3}$$

If the input matrices consist of unbiased estimates, the mixture matrix can be computed using Equation A4.

$$\mathbf{S}_{mix} = \left[p_1\mathbf{S}_1\left(\frac{n_1 - 1}{n_1}\right) + p_2\mathbf{S}_2\left(\frac{n_2 - 1}{n_2}\right) + p_1p_2\,\mathbf{d}'\mathbf{d}\left(\frac{|c_1 - c_2|}{|c_1| + |c_2|}\right)\right]\left(\frac{n_1 + n_2}{n_1 + n_2 - 1}\right) \tag{A4}$$

However, if the input matrices consist of maximum-likelihood estimates, the mixture matrix can be computed using Equation A5, where the adjustments for unbiased estimation are removed.

$$\mathbf{S}_{mix} = p_1\mathbf{S}_1 + p_2\mathbf{S}_2 + p_1p_2\,\mathbf{d}'\mathbf{d}\left(\frac{|c_1 - c_2|}{|c_1| + |c_2|}\right) \tag{A5}$$

Finally, the means of the can be computed using Equation A6.

$$\mathbf{m}_{mix} = p_1\mathbf{m}_1 + p_2\mathbf{m}_2 \tag{A6}$$

The $\mathbf{S}_{mix}$ matrix and the $\mathbf{m}_{mix}$ vector that result from this procedure can be used to compute a categorically moderated regression model, as required in the Cleary model of bias.