Received: 6 May 2023 | Accepted: 12 May 2023

DOI: 10.1111/peps.12593

PERSONNEL PSYCHOLOGY WILEY

ORIGINAL ARTICLE

Reducing subgroup differences in personnel selection through the application of machine learning

Nan Zhang[1] | Mo Wang[1] | Heng Xu[2] | Nick Koenig[3] | Louis Hickman[4,5] | Jason Kuruzovich[6] | Vincent Ng[7] | Kofi Arhin[6] | Danielle Wilson[7] | Q. Chelsea Song[8] | Chen Tang[2] | Leo Alexander III[9,10] | Yesuel Kim[11]

[1]Warrington College of Business, University of Florida, Gainesville, Florida, USA

[2]Kogod School of Business, American University, Washington, DC, USA

[3]Modern Hire, Cleveland, Ohio, USA

[4]Department of Psychology, Virginia Tech, Blacksburg, Virginia, USA

[5]The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[6]Lally School of Management, Rensselaer Polytechnic Institute, Troy, Michigan, USA

[7]Department of Psychology, University of Houston, Houston, Texas, USA

[8]Kelley School of Business, Indiana University, Bloomington, Indiana, USA

[9]School of Labor and Employment Relations, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

[10]Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

[11]Department of Psychological Sciences, Purdue University, West Lafayette, Indiana, USA

Abstract

Researchers have investigated whether machine learning (ML) may be able to resolve one of the most fundamental concerns in personnel selection, which is by helping reduce the subgroup differences (and resulting adverse impact) by race and gender in selection procedure scores. This article presents three such investigations. The findings show that the growing practice of making statistical adjustments to (nonlinear) ML algorithms to reduce subgroup differences must create predictive bias (differential prediction) as a mathematical certainty. This may reduce validity and inadvertently penalize high-scoring racial minorities. Similarly, one approach that adjusts the ML input data only slightly reduces the subgroup differences but at the cost of slightly reduced model accuracy. Other emerging tactics involve weighting predictors to balance or find a compromise between the competing goals of reducing subgroup differences while maintaining validity, but they have been limited to two outcomes. The third investigation extends this to three outcomes (e.g., validity, subgroup differences,

Correspondence
Louis Hickman
Email: louishickman@vt.edu

Funding information
National Science Foundation, Grant/Award Number: 2309853; Human Resource Research Center of the University of Florida

and cost) and presents an online tool. Collectively, the studies in this article illustrate that ML is unlikely to be able to resolve the issue of adverse impact, but it may assist in finding incremental improvements.

1 INTRODUCTION

The purpose of this article is to illustrate some of the current uses of machine learning (ML) to reduce the subgroup differences by race and gender in selection decisions. Subgroup differences are critically important in personnel selection because they can result in differences in passing/hiring rates by subgroup (called adverse impact), which is illegal if the selection procedures are not job related according to civil rights laws in the United States (Uniform Guidelines on Employee Selection Procedures, 1978). Even when selection procedures can be shown to be job related, the presence of adverse impact may be viewed as problematic by organizations attempting to increase the diversity of their workforces. As such, some researchers are attempting to use ML to help reduce subgroup differences. In order to present several relevant studies on the topic, only brief summaries are presented in this article. Interested readers should consult each study's Online Supplement for additional information on the study background, method, and supplemental analyses.

This article presents three complementary studies of this important problem. First, using mathematical proofs as well as simulated and real organizational data, Study 1 by Zhang and colleagues shows that (nonlinear) ML algorithms that make statistical adjustments to reduce subgroup differences must create predictive bias (also called differential prediction, which is the definition of unfairness in selection science), which may actually reduce validity and penalize high-scoring racial minorities. Study 2 by Hickman et al. illustrates one approach to reducing subgroup differences that involves adjusting the input data to be equivalent between races by oversampling higher-performing minorities during ML model training. The study shows that by statistically removing subgroup differences in the training data, one can only slightly reduce the differences in the resulting ML model but at the cost of slightly reduced accuracy. Third, attempting to increase the validity of statistical predictions and reduce subgroup differences at the same time is very difficult or impossible because the two outcomes are in conflict (i.e., increasing validity often increases subgroup differences of the predictor composites). Research in recent years has used Pareto-optimal analytic techniques to attempt to find the best compromise that maximizes both outcomes to the extent possible. The difficulty is that current techniques are limited to two outcomes, while there may be three (e.g., validity, subgroup differences, and cost). Study 3 by Song et al. presents a tool for achieving optimization for up to three objectives, which has many applications in selection.

2 STUDY 1: ARE FAIRNESS-AWARE ML ALGORITHMS REALLY FAIR? PREDICTIVE BIAS OF USING ML IN PERSONNEL SELECTION[1]

The past decade witnessed remarkable advances in the development of ML algorithms that automate the construction of prediction models (Sejnowski, 2018). These advances attracted interest from practitioners in applying ML to organizational decision-making processes, among which personnel selection is a prominent example (Hickman et al., 2022). Recognizing the importance of limiting adverse impact (Dastin, 2018), ML researchers devoted considerable attention to the development of *fairness-aware ML algorithms* (Barocas et al., 2019), which are designed to optimize for predictive accuracy while limiting the adverse impact of predictions. When used in personnel selection, these algorithms could offer mathematical guarantees in terms of an upper bound on the adverse impact of selection outcomes (Zafar et al., 2019). Yet what has received little attention is whether the predictions made by fairness-aware ML algorithms could suffer from *predictive bias* (also known as *test bias* or *differential predictions*; SIOP, 2018), that is, whether the relationship between the ML-predicted score and the criterion of interest (e.g., job performance) could be different for one demographic group than for another. Given the considerable attention afforded to predictive bias in personnel selection (Aguinis et al., 2010), this omission represents a significant issue in research and practice, and is thus the focus of this study.

The goal of this study is to assess the potential for predictive bias in predictions made by fairness-aware ML for personnel selection. We start with mathematical analysis showing that, unless a "plain" ML algorithm with no fairness constraint already satisfies the organizational requirement on adverse impact, predictions made by fairness-aware ML are almost always biased even when every predictor is free of predictive bias. Our mathematical findings also reveal a peculiar result. Contrary to the intuition that a racial minority candidate always stands to gain from the inclusion of fairness constraints, the opposite could be true for some racial minorities. Specifically, when an ML algorithm is designed to satisfy a fairness constraint, it could be inherently incentivized to "guess" whether an applicant belongs to a protected group. As a result, these ML predictions tend to unfairly penalize those racial minority candidates who "look like" racial majorities according to the predictor battery. When the mean criterion score of racial minorities is lower than the racial majorities, those racial minority candidates who "look like" the racial majorities could be those who are highly qualified for the job. In this case, the predictive bias of ML predictions could lead to the *exclusion* of these candidates who would have been selected had there been no fairness consideration in ML. In other words, the predictive bias of ML predictions could distort the selection outcomes so much that a fairness-aware ML algorithm introduces its own fairness issues in the process of reducing adverse impact.

After discussing the mathematical findings, we present Monte Carlo simulation results and a case study with real-world data that confirm our mathematical findings and demonstrate the prevalence of predictive bias in the predicted scores generated by a variety of fairness-aware ML algorithms. We conclude the study with a discussion of its practical implications.

2.1 | Preliminaries

2.1.1 | Fairness-aware ML algorithms

We note at the outset a distinction between the design of selection systems for personnel selection (De Corte et al., 2011) and that of fairness-aware ML algorithms. ML researchers, who are mostly computer scientists, rarely design an algorithm exclusively for one purpose such as personnel selection. Instead, they often cite "non-discriminatory hiring" (Friedler et al., 2019) as one of the most important goals of fairness-aware ML, while keeping open the possibility for the algorithm to be used for other purposes such as loan allocation (Feldman et al., 2015). Thus, while we review the existing fairness-aware ML algorithms in the context of personnel selection, it should not be interpreted as implying that they cannot be used in other relevant contexts.

In general, any algorithm can be characterized by its (1) input, (2) output, (3) requirement on the output, and (4) technical design for mapping the input to the output (Cormen et al., 2009, *p*. 5). In the passages that follow, we first review the input, output, and requirement on the output of fairness-aware ML, before briefly summarizing the existing algorithms for fairness-aware ML.

Input. A fairness-aware ML algorithm takes as input an incumbent dataset—known as *training dataset* in ML—collected from current or past employees of an organization. The composition of this dataset is similar to what is required for a local validation study in personnel selection. That is, for each incumbent employee, the dataset typically includes the predictor battery, a criterion score, and whether the employee is part of a protected group.

Output. The algorithm's output is a *prediction model* that relates the predictor battery of a candidate to a numeric predicted score, which we refer to as the *ML prediction*. The functional form of the model could vary widely, from a support-vector machine (Zafar et al., 2019) to a Gaussian process (Tan et al., 2020). Regardless of the functional form, the prevailing assumption in ML is that a prediction model serves as a drop-in replacement for the selection process. That is, once an organization applies the prediction model to a pool of applicants, it selects those applicants with the highest ML predictions.

Requirement on output. A key requirement on the output prediction model is to meet an organization's desired level of validity-diversity tradeoff (De Corte et al., 2011). That is, it needs to balance between (1) maximizing the expected criterion of selected candidates, and (2) minimizing the adverse impact of selection outcome. While adverse impact has been assessed with measures such as the adverse impact ratio (De Corte et al., 2011), the Fisher exact test (Siskin & Trippi, 2005), the Z_{IR} test (Morris & Lobsenz, 2000), and so on, once an applicant pool and selection rate are given, a threshold on one measure can be converted to another. Thus, we focus on the measure prescribed in the *Uniform Guidelines*, the Adverse Impact Ratio (AIR), which is the ratio between the selection rate of the racial minority group and the racial majority group (De Corte et al., 2011).

Algorithmic design. From an algorithmic perspective, fairness-aware ML falls under the general paradigm of learning with privileged information (Vapnik & Vashist, 2009). Whereas the algorithm does have access to the protected variable (e.g., race) of incumbents during training, it cannot include such a variable in the prediction model because, in the context of personnel selection, the use of protected variables in prediction is generally prohibited due to legal constraints in the United States. This makes the protected variable *privileged information* that is only available during training. The key technical challenge in algorithmic design then becomes how to leverage such privileged information in training the prediction model.

It is important to note the similarities in how protected variables are used in fairness-aware ML vis-à-vis traditional selection systems. Traditionally, human experts often evaluate the potential adverse impact of a selection-system design based on incumbent data (which includes protected variables for the assessment of adverse impact), and make the appropriate adjustments, such as revising the inclusion/exclusion of certain predictors or changing their weights. Yet, once a selection system is put into production, it has no access to any applicant's protected variables. If we draw an analogy between the design process for a selection system and the training of an ML model, then their use of protected variables is almost identical. That is, protected variables are used during training (manual training for traditional selection systems, algorithmic training for ML) but not when the selection system or ML model is deployed in practice.

To address this challenge, the general idea in fairness-aware ML is to revise a "plain" ML algorithm by assigning a penalty to a potential prediction model if it violates the fairness constraint (e.g., an upper bound on AIR). The more serious the violation is, the higher the penalty would be. Since this penalty can be assessed at training time using the privileged information, a fairness-aware ML algorithm would then be incentivized to adjust the output prediction model to avoid the penalty and satisfy the given fairness constraint. For example, Kamiran et al. (2010) revised a decision-tree algorithm, specifically the rules used by the algorithm to determine how to grow a branch, in order to minimize adverse impact. Similarly, researchers have integrated fairness constraints by revising algorithms for representation learning (Zemel et al., 2013), support-vector machines (Zafar et al., 2019), and natural language processing (Zhao et al., 2018). To the best of our knowledge, however, the predictive bias of predictions made by either plain ML or fairness-aware ML algorithm has not been systematically studied in the literature. Even though the selection of predictors has been examined, the prevailing view is to include all available predictors and leave feature selection to the fairness-aware ML algorithm (Kleinberg & Mullainathan, 2019).

2.1.2 | Predictive bias

Testing for predictive bias typically involves a moderated multiple regression framework known as Cleary's (1968) method (SIOP, 2018). Its precise description requires the introduction of a few mathematical notations. Let the criterion variable be Y, the (vector representation of) predictor battery be $\mathbf{X}$, the prediction generated by the fairness-aware ML algorithm be $\tilde{f}(\mathbf{X})$, and the group membership be G. For the sake of simplicity, we focus on two groups, with $G = 0$ being the racial majority and $G = 1$ being the racial minority (i.e., protected) group.

Consider the following linear models where a_0, b_0 and c_0 are the intercepts; a_1, b_1, b_2, b_3, c_1, and c_2 are unstandardized regression coefficients; and ε, ε', and ε'' are random error terms:

$$Y = a_0 + a_1\tilde{f}\,(\mathbf{X}) + \varepsilon,$$

$$Y = b_0 + b_1\tilde{f}\,(\mathbf{X}) + b_2 G + b_3\tilde{f}\,(\mathbf{X})\,G + \varepsilon',$$

$$Y = c_0 + c_1\,\tilde{f}(\mathbf{X}) + c_2 G + \varepsilon'',$$

Predictive bias exists if (1) $b_3 \neq 0$, indicating a slope difference between groups, and/or (2) $b_2 \neq 0$, indicating an intercept difference. The third equation further specifies whether a common regression line would, on average, over- ($c_2 < 0$) or under-predict ($c_2 > 0$) the criterion scores of racial minority candidates, with either indicating the existence of predictive bias. Statistical significance tests may be conducted directly over the regression coefficients (Sackett et al., 2003) or over the difference in R^2 between the first and second equations (Aguinis et al., 2010). In empirical literature, intercept differences are found to be more common than slope differences, with a common regression line typically overpredicting the criterion scores of racial minority candidates (SIOP, 2018). Note that, even though the Cleary's method tests linear models while ML may learn nonlinear functions, it remains an appropriate method for testing the predictive bias of ML predictions because, within each group, an applicant with a higher criterion score should be assigned a proportionally higher predicted score.

2.2 | Predictive bias of fairness-aware ML algorithms

2.2.1 | Key source of predictive bias: Prediction target

Designing a practical system with ML is a complex process (Barocas et al., 2019); and predictive bias could arise in many steps along the way, from making an improper selection of the ML algorithm to a lack of sufficient training samples (Buolamwini & Gebru, 2018). Since the purpose of this study is to investigate whether the introduction of fairness constraints could induce predictive bias in ML predictions, we need to ensure that our findings generalize to different implementations of fairness-aware ML regardless of their specific technical design. To this end, it is helpful to consider an idealized scenario in which fairness-aware ML produces the least possible amount of prediction error. If we could identify a source of predictive bias even in this idealized scenario, then the bias would likely generalize to all practical implementations of fairness-aware ML. We construct this idealized scenario with two assumptions as follows.

First, the ML algorithm being used should produce prediction models that are sufficiently complex to address the prediction task at hand (according to measures such as model capacity, Vapnik, 1998). For example, we would not consider the use of linear regression to fit a nonlinear predictor-to-criterion relationship, the problem of which was already noted in the literature (Bauer, 2005). With this assumption, any predictive bias we identify could not be easily fixed by switching to a more complex ML algorithm such as a non-parametric Gaussian process with unlimited model capacity (Rasmussen & Williams, 2006), which always satisfies this assumption.

Second, we assume the training dataset to be sufficiently large and drawn from the same distribution as the applicant pool. Doing so allows us to sidestep a frequently arising issue in ML called *covariate shift*, which happens when a prediction model trained on one dataset is used for predicting over samples drawn from a different probability distribution. Covariate shift may potentially incur an increase of prediction error known as *generalization error* (Vapnik, 1998). While the reduction of generalization error is an important problem in ML and has been treated with methods such as importance sampling (Sugiyama & Storkey, 2006), it is tangential to our work because such errors are typically assumed to be independent and identically distributed Gaussian noise with no statistical difference between groups (Bishop, 2006, p. 29). In other words, they are unlikely to alter the predictive bias of ML predictions. Like the first assumption, this one ensures that any predictive bias we identify cannot be easily fixed by improving (e.g., increasing the size of) the input training dataset.

In this idealized scenario, an ML prediction model should be able to approximate any *prediction target* function that defines, according to the desired validity-diversity tradeoff, what the ML-predicted scores *should* look like for each given value combination of the predictor battery. Whether such a prediction target exhibits predictive bias thus becomes the key question for assessing the bias of fairness-aware ML. We address this question next.

2.2.2 | Existence of predictive bias

To assess predictive bias, we first need to derive a mathematical model for the prediction target of fairness-aware ML. To this end, it is helpful to start with a "plain" ML algorithm in which the sole objective for candidate selection is to maximize the mean criterion score

$$u(S) = \frac{1}{|S|} \sum_{\mathbf{x} \in S} \mathrm{E}\,(Y|\mathbf{X} = \mathbf{x})$$

of the selected candidates S for a given selection rate (i.e., a fixed $|S|$). For such an algorithm, Rambachan et al. (2020) proved that its prediction target function, denoted by $f_0(\mathbf{x})$, is simply the expected criterion score for the input predictor battery

$$f_0(\mathbf{x}) = \mathrm{E}\,(Y|\mathbf{X} = \mathbf{x})$$

because selecting those candidates with the maximum $f_0(\mathbf{x})$ is guaranteed to maximize $u(S)$. In other words, the overall task of candidate selection can be decomposed into the individual tasks of approximating the prediction target $f_0(\mathbf{x})$ for each candidate.

Compared with this "plain" algorithm, the prediction target for fairness-aware ML is more complex because it needs to balance the validity-diversity tradeoff. A common strategy is to pursue *Pareto-optimal* (De Corte et al., 2011) selection outcomes, that is, those that no other possible outcome can dominate on both expected criterion and AIR. To do so, $u(S)$ has to be maximized under a *fairness constraint* that S meets a given lower bound r on the AIR. Clearly, $f_0(\mathbf{x})$ is no longer a proper prediction target because selecting the top $|S|$ candidates with the maximum $f_0(\mathbf{x})$ might result in AIR $< r$. Since it may not be possible to assess whether AIR $\geq r$ without first assembling S, the introduction of the fairness constraint brings into question whether the task of candidate selection is still decomposable into approximating a prediction target for individual candidates. Fortunately, as proved in the following theorem, the Lagrange-multiplier method (Nocedal & Wright, 2006) provides an elegant solution that enables such a decomposition.

Theorem 1. For any $\lambda \geq 0$ and any selection rate, selecting the candidates with the maximum

$$f_\lambda(\mathbf{x}) = \mathrm{E}(Y\,|\mathbf{X} = \mathbf{x}) + \lambda \cdot \Pr\{G = 1|\,\mathbf{X} = \mathbf{x}\}$$

is Pareto-optimal on the validity-diversity tradeoff. Conversely, given any selection rate and any lower bound r on AIR, there must exist $\lambda \geq 0$, such that the selection outcome that maximizes $f_\lambda(\mathbf{x})$ also maximizes expected criterion score under the constraint of AIR $\geq r$. Further, there is

$$\mathrm{E}\left(\Pr\{G = 1\,|\mathbf{X} = \mathbf{x}\}\middle|\, G = 1\right) > \mathrm{E}\left(\Pr\{G = 1\,|\mathbf{X} = \mathbf{x}\}\middle|\, G = 0\right)$$

unless $\Pr\{G = 1|\mathbf{X} = \mathbf{x}\}$ *is constant for all* $\mathbf{x}$.

The mathematical proof is available in the Supplemental Materials. The theorem yields two insights. First, the overall task of seeking Pareto optimality is still decomposable to the individual tasks of approximating a prediction target

$f_\lambda(\mathbf{x})$ for each candidate, where λ, the *Lagrange multiplier*, is a function of the AIR requirement. In other words, Theorem 1 states that, to make a Pareto-optimal tradeoff, fairness-aware ML should produce predicted scores that resemble the prediction target $f_\lambda(\mathbf{x})$. Interestingly, the only difference between $f_\lambda(\mathbf{x})$ and the prediction target for plain ML is an additive term of λ times $\Pr\{G = 1|\mathbf{X} = \mathbf{x}\}$, that is, the likelihood for a candidate to be a racial minority given the observed predictors in $\mathbf{X}$.

Second, the theorem suggests that, even when every input predictor is unbiased, predictive bias could still emerge in the predicted scores of fairness-aware ML because the additive term in the prediction target function, $\lambda \cdot \Pr\{G = 1|\mathbf{X} = \mathbf{x}\}$, is systematically larger for the racial minority group in almost all cases. There are only two exceptions: 1) when $\lambda = 0$, or 2) when $\Pr\{G = 1|\mathbf{X} = \mathbf{x}\}$ is constant for all $\mathbf{x}$. Either exception would make $f_\lambda(\mathbf{x})$ equivalent with $f_0(\mathbf{x})$—because their difference would be either zero or the same across all candidates—implying that plain ML would achieve AIR $\geq r$ anyway.

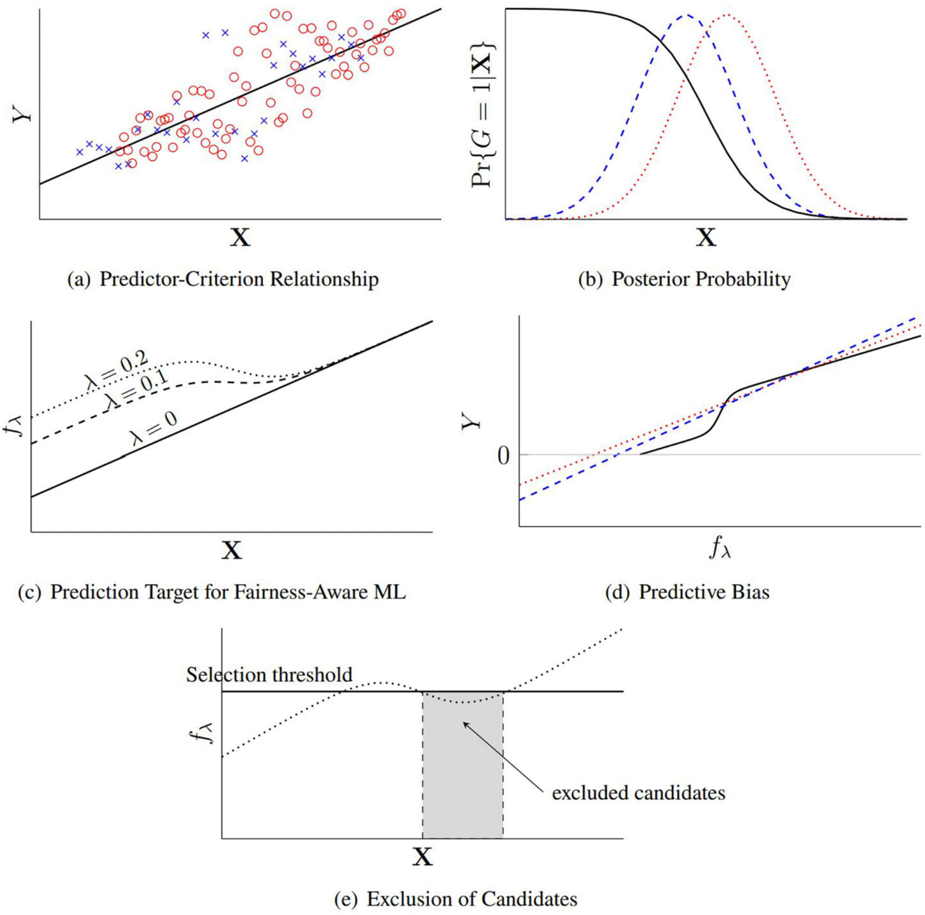
To understand the mechanism through which a between-group difference in the additive term manifests as predictive bias, consider a few idealized examples. The first is when λ is set to eliminate adverse impact by ensuring an equal mean of ML-predicted scores, say $\overline{f_\lambda}(\mathbf{x}) = F$, for both groups. Since a least-squares regression line linking $f_\lambda(\mathbf{x})$ and criterion Y always passes through the center-of-mass point $(\overline{f_\lambda}(\mathbf{x}), \overline{Y})$, the two groups' regression lines pass through $(F, \overline{Y}_0)$ and $(F, \overline{Y}_1)$, respectively, where $\overline{Y}_i$ are their mean criterion. According to the inequality in the theorem, there must be $\overline{Y}_0 > \overline{Y}_1$ when $\lambda > 0$, meaning that a common regression line has to overpredict the criterion score of racial minority candidates in this example.

Figure 1a provides a graphic illustration of another example where a predictor variable has no predictive bias but a slight mean difference between groups. Due to this mean difference, the additive term in $f_\lambda(\mathbf{x})$ (i.e., $\lambda \cdot \Pr\{G = 1|\mathbf{X} = \mathbf{x}\}$) becomes a reverse sigmoid function with $\mathbf{x}$, as shown in Figure 1b. Figure 1c shows how this reverse sigmoid function "bends" the prediction target f_λ to form a nonlinear relationship with the criterion. Since this bending is, by definition, more concentrated on racial minority candidates, the resulting nonlinearity is also more pronounced for them, resulting in the predictive bias shown in Figure 1d. In this specific example, a common regression line features a smaller slope and a larger intercept than the regression line for the racial minority group, leading to, on average, an overprediction of the criterion score for racial minority candidates.

The example also points to a negative consequence of predictive bias. As illustrated in Figure 1e, when $\lambda > 0$, fairness-aware ML *has to* exclude from selection *some* racial majority candidates who would have been selected if $\lambda = 0$. This exclusion is not a problem in and of itself, because it is necessary for achieving the given bound on AIR. What is problematic is that ML has no access to the group membership of a candidate. As such, it has no choice but to "guess," based on the observed predictors, whether a candidate is a racial majority who needs to be excluded. Recall from Figure 1b that the likelihood for a candidate to be in the racial majority increases with their criterion score. Thus, when fairness-aware ML needs to "guess" the candidates to exclude, it tends to pick some candidates with a higher criterion score. Unfortunately, such a guess is imperfect, meaning that some racial minorities could be inadvertently excluded too. The exclusion of these candidates *is* detrimental to fairness because it means that the adoption of fairness-aware ML leads to the exclusion of some qualified racial minority candidates who would have been selected had there been no fairness consideration in the first place. In other words, in attempting to reduce adverse impact in personnel selection, fairness-aware ML could inadvertently raise its own fairness issue through the introduction of predictive bias. Next, we present a simulation study and a case study to verify the findings based on Theorem 1.

2.3 | Simulation study

In the passages that follow, we describe the data-generating process for the simulation study, the fairness-aware ML algorithms tested, the simulation conditions, and the simulation results, respectively.



(a) Predictor-Criterion Relationship

(b) Posterior Probability

(c) Prediction Target for Fairness-Aware ML

(d) Predictive Bias

(e) Exclusion of Candidates

FIGURE 1 Examples of prediction target for fairness-aware ML. *Note.* Panel (a): **X** = predictor. *Y* = criterion score. ∘ = racial majority candidates. × = racial minority candidates. Solid line is the regression line connecting *Y* to **X**, which is identical for both groups. In other words, **X** has no predictive bias towards the criterion *Y*. Panel (b): Dashed and dotted curves are the probability density function (PDF) of **X** in racial minorities and majorities, respectively. The solid curve represents $\Pr\{G = 1|\mathbf{X} = \mathbf{x}\}$, which is a reverse sigmoid function that decreases with **X**. Panel (c) depicts the prediction target function f_λ, which is the sum of the solid line in Panel (a) and λ times the solid curve in (b). Note that, the larger λ is, the more f_λ becomes "bended" by the reverse sigmoid function, even to the point of losing monotonicity when $\lambda = .2$. Panel (d): Solid curve represents the relationship between f_λ and *Y*. Dashed line is the regression lines for the racial minority group, while dotted line represents the common regression line. The difference between these two lines indicates the presence of predictive bias. Observe from the panel that the regression line for racial minorities has a larger slope and a smaller intercept. On average, a common regression line would overpredict the criterion score for racial minority candidates. Panel (e) depicts the exclusion of candidates due to the fairness constraint, specifically. The dotted curve is the prediction target function f_λ when $\lambda = .2$. The solid line represents the selection threshold, that is, the minimum ML-predicted score for a candidate to be selected. The gray zone represents those candidates who would have been selected if $\lambda = 0$, but are excluded when $\lambda = .2$.

2.3.1 | Data-generating process

We generated two sets of data: (1) a varying-size training (i.e., incumbent) dataset, which was used for ML to learn its prediction model, and (2) a 1000-record testing dataset, which was used to assess the predictive bias of ML predictions. In terms of features, we followed Finch et al. (2009) to simulate (1) a battery of five predictors: biodata, cognitive ability, conscientiousness, integrity, and structured interview, (2) a criterion variable, which is job performance, and (3) a binary (i.e., racial majority or minority) group membership representing White and Black applicants, respectively. The detailed procedure and its potential limitations are discussed in the Supplemental Materials.

2.3.2 | Fairness-aware ML algorithms

As detailed in the Supplemental Materials, we tested four ML algorithms, Gaussian process (GP), support vector machine (SVM), regression-tree ensemble with least-squares boosting (BOOST), and a feed-forward, fully connected neural network (NN). While no qualitative difference emerged in results across algorithms, we found the first two (GP and SVM) to consistently outperform the latter (BOOST and NN) in terms of predictive accuracy. Thus, we focus on GP and SVM when reporting the simulation results. To set λ, we performed an iterative optimization like Google's TensorFlow Constrained Optimization (Cotter et al., 2019).

2.3.3 | Simulation conditions

We varied five parameters: algorithm, AIR bound, selection rate s, training dataset size N, and between-group difference ratio δ on predictors. The last parameter δ ($0 < \delta \leq 2$) served as a multiplicative factor for the standardized between-group mean difference for each predictor. We simulated two ML algorithms: SVM or GP; four levels for the selection rate: .1, .3, .5, .8; three for the lower bound on AIR: .5, .8, 1.0; three for the training dataset size: 1000, 2500, and 5000; and three for δ: .5, 1.0, 2.0. Overall, our simulation design consisted of 216 unique conditions or a 2 (algorithm) × 4 (s) × 3 (N) × 3 (δ) × 3 (AIR) factorial design. We repeated each condition 20 times, leading to a total of $216 \times 20 = 4320$ runs.

2.3.4 | Simulation results

Table A2 in the online supplement shows the marginal statistics for the prevalence of predictive bias in fairness-aware ML. In light of the different implementations of Cleary's method, we reported in the table both the regression-coefficient estimates for Race and the interaction between Race and ML prediction (Sackett et al., 2003) as well as ΔR^2, the increase in R^2 from adding Race and the interaction term as regressors (Aguinis et al., 2010). Both implementations identified predictive bias in an overwhelming majority of simulation runs. Specifically, ΔR^2 ($M = .026, SD = .017$) was statistically significant ($p < .05$) in 3871 (89.61%) out of all 4320 runs. Similarly, the coefficient estimates for Race ($M = -.350$, $SD = .183$) and the interaction term ($M = -.054$, $SD = .196$) were statistically significant in 3572 (82.69%) and 461 (10.67%) runs, respectively. Notably, the marginal means for both coefficients were consistently below zero across all conditions, echoing our earlier discussions that a common regression line constructed from the ML predictions would likely overestimate the criterion scores of racial minority candidates.

To further examine how the simulated factors affect predictive bias, we conducted a five-way analysis of variance (ANOVA) with the dependent variable being ΔR^2 and the independent variables being the five simulation factors. Due to the space limit, we include the detailed results of ANOVA in the online supplement, and summarize the main findings here. Due to the large sample size (4320), we followed Steinley (2006) to only consider (main and interaction) effects with effect size $\eta^2 \geq .05$. In terms of main effects, ANOVA identified AIR (F (2, 4104) = 1819.98, $p < .01$), selection rate s (F (3, 4104) = 953.03, $p < .01$), and between-group difference ratio δ (F (2, 4104) = 557.49, $p < .01$). The ML algorithm and the amount of training data, on the other hand, do not have a pronounced effect on ΔR^2, consistent with our earlier finding that the predictive bias results from the intrinsic design of fairness-aware ML, specifically its prediction target, rather than the specific ML implementations.

In terms of the directions of the main effects, observe from Table A2 in the online supplement that ΔR^2 clearly increases with AIR and decreases with s. In other words, as the fairness constraint becomes more stringent with a larger AIR and/or in a "select in" scenario with a smaller s, fairness-aware ML has to "bend" the prediction target more to achieve a Pareto-optimal outcome, increasing its predictive bias. The relationship between ΔR^2 and δ is subtler and best qualified by a three-way interaction identified by ANOVA, $s \times \delta \times$ AIR ($F\ (12, 4104) = 97.34, p < .01$), as shown in Figure A1 in the online supplement.

The figure yields two observations. First, ΔR^2 was surprisingly small when $\delta = 2$, seemingly "capped" by an upper bound of around .025. This contradicts the intuition that, to reduce adverse impact when the between-group difference is large, fairness-aware ML has to increase its predictive bias. Interestingly, the contradiction speaks to a limit of using ΔR^2 to quantify predictive bias. When λ is so large that the prediction target is dominated by its second term, that is, the likelihood of a candidate being a racial minority, the prediction target becomes an approximate of the group membership rather than the criterion of a candidate. In this case, making the true group membership (i.e., race) a regressor alongside the prediction target adds little to the explanatory power (i.e., R^2) even when its coefficient is nonzero. Indicatively, Table A2 in the online supplement shows that, despite the low ΔR^2, all 1440 simulation runs with $\delta = 2$ returned statistically significant estimates for the coefficient of Race ($M = -.464$, $SD = .091$).

Second, an increase of δ from .5 to 1 actually reduced ΔR^2 when the fairness constraint was loose (e.g., a low AIR = .5 under a moderate selection rate $s = .3$ or .5). Upon further examination, we found a key reason to be how the Lagrange multiplier λ responded to an increase of δ: Under a loose fairness constraint, even though the higher δ reduced the selection rate of racial minorities, this reduced rate still met the AIR requirement. Thus, an increase of δ did not require an increase of the second term in the prediction target, that is, $\lambda \cdot \Pr\{G = 1|\mathbf{x}\}$. On the other hand, since a higher δ made it easier to distinguish between the two groups, the likelihood function $\Pr\{G = 1|\mathbf{x}\}$ became higher for the racial minorities. These two changes in combination drove down λ and thereby ΔR^2 in a loose-fairness regime.

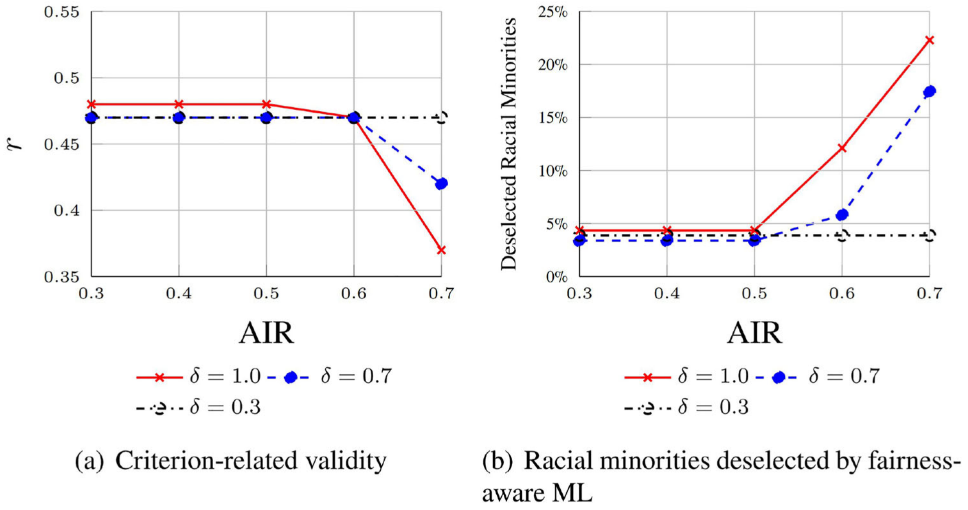
In sum, the simulation results showed the prevalence of predictive bias for multiple fairness-aware ML algorithms across diverse simulation settings. The results further suggested that ML predictive bias was the largest in a select-in scenario ($s = .1$) with a stringent AIR requirement (AIR = 1); and the smallest in a select-out scenario (e.g., $s = .8$) where the AIR requirement and the between-group difference in predictors are both small.

2.4 | Illustration of practical impact

To further illustrate the practical implications of using a fairness-aware ML algorithm in personnel selection, we examined three additional issues: (1) the existence of a validity-diversity tradeoff in the selection outcomes of fairness-aware ML; (2) the consequences of the predictive bias of fairness-aware ML, specifically the number of racial minorities who would have been selected by a plain ML algorithm but are excluded from selection by fairness-aware ML; and (3) the reliability of our findings over a real-world dataset.

Figure 2a demonstrates how the criterion-related validity of ML-predicted scores varies with the reduction of adverse impact when a fairness-aware ML algorithm is used over simulated datasets. As can be seen from the figure, fairness-aware ML lowered adverse impact at the expense of a potential decrease in criterion-related validity. Consistent with the understanding for traditional selection systems (Rupp et al., 2020), the magnitude of this validity-diversity tradeoff was more pronounced when the input data contained substantial between-group differences. For example, with a between-group difference ratio of $\delta = 1$, the criterion-related validity dropped from .48 to .37 when the AIR requirement increased from .3 to .7. While the existence of this tradeoff is technically obvious (e.g., given the Lagrangian objective function in Theorem 1), it indicates that fairness-aware ML is far from a silver bullet, but instead subject to the same validity-diversity "dilemma" (Pyburn Jr et al., 2008) and its associated practical challenges (Rupp et al., 2020) as traditional selection systems.

To illustrate how the use of fairness-aware ML could hurt rather than benefit certain racial minorities, Figure 2b depicts the number of "deselected" racial minorities, meaning those who would have been selected by a plain ML algorithm but were excluded by fairness-aware ML. As can be seen from the figure, the number of such candidates *increased* with AIR, reaching over 20% of all racial minorities being selected when AIR ≥ .7 ($\delta = 1$). This suggests that the deselected racial minorities were "sacrificed" by fairness-aware ML in pursuit of a lower adverse impact because,



(a) Criterion-related validity

(b) Racial minorities deselected by fairness-aware ML

FIGURE 2 Implications of fairness-aware ML. *Note.* r = (uncorrected) criterion-related validity of ML-predicted scores. AIR = adverse impact ratio. Deselected racial minorities = among racial minorities selected by plain ML (without fairness considerations), the percentage that are not selected by fairness-aware ML. The panels were generated with the SVM algorithm, selection rate $s = .3$, and training dataset size $N = 1000$. We also tested the GP algorithm and observed similar trends. All points on a validity-diversity tradeoff curve were generated over the same simulated dataset. Panel (a) shows that, consistent with earlier discussions, when the input data contains substantial between-group differences (e.g., $\delta \geq .7$), fairness-aware ML reduces adverse impact with a corresponding reduction in validity. Panel (b) shows that, contrary to the intuition that a racial minority candidate always stands to gain from the inclusion of fairness constraints (Zafar et al., 2019), the opposite could be true for some racial minorities, like those deselected ones depicted in the plot.

as discussed earlier, they “look like” racial majorities according to the predictor battery. Another observation from the figure is that the number of deselected racial minorities was higher when the input data contained larger between-group differences. This is consistent with our earlier discussions that, the more pronounced the validity-diversity tradeoff becomes (thanks to the larger between-group differences), the more likely it is for fairness-aware ML to incur predictive bias when reducing adverse impact.

Finally, to examine the reliability of our findings over a real-world dataset, we tested fairness-aware ML algorithms over a dataset containing the results of pre-employment tests used for entry-level positions in a Fortune 500 company. While we defer details of this case study to the online supplement, the main finding was that predictive bias remained prevalent over real-world data, even when the dataset features racial majority and minority candidates with very close criterion distributions (specifically, a standardized between-group mean difference of .11). While the smaller between-group differences led to a milder validity-diversity tradeoff and, in turn, a smaller magnitude of predictive bias (e.g., as measured by ΔR^2), the predictive bias identified over the real-world dataset featured a more pronounced slope difference than the simulation results, suggesting that fairness-aware ML likely had difficulty “guessing” the group affiliation of high-criterion candidates, leading to a flatter regression line for racial minorities.

2.5 | General discussion

Our findings are important because they speak to a substantive problem of using fairness- aware ML as a drop-in replacement for selection systems. Such a practice, while convenient, could lead to unintended consequences such as predictive bias. As summarized in Table 1, an important reason is the complexity of prediction models produced by ML. With a traditional selection system, prediction models (e.g., a weighted sum of predictors) take functional forms

TABLE 1 Comparison of prediction models for traditional selection systems and fairness-aware ML.

	Traditional Selection Systems	*Fairness-Aware ML*
Model complexity	Relatively simple. For example, the prediction model for a compensatory design features a linear combination of predictors.	Considerably more complex and generally nonlinear. For example, a deep learning model could contain millions of parameters learned from data.
Model interpretability	Relatively straightforward. For example, weights representing the relative importance of predictors.	Subject of ongoing research (Du et al., 2019), as many types of ML prediction models are not interpretable even by experts.
Predictor selection	Predictors are scrutinized (e.g., for predictive bias) in scientific literature and according to empirical evidence.	Prevailing view is to include all available predictors and leave selection to the algorithm (Kleinberg & Mullainathan, 2019).
Prediction target	The expected criterion score for a candidate based on the values of the input predictor battery.	Per Theorem 1, a linear combination of expected criterion score and the likelihood of the group membership.
Predictive accuracy	Linear models could incur large predictive errors when the predictor-to-criterion relationship is nonlinear.	Could achieve optimal accuracy given sufficient training samples, per universal approximation theorems (Goodfellow et al., 2016).
Source of predictive bias	A linear prediction model never incurs predictive bias when the input predictors are unbiased.	An ML prediction model, which is generally nonlinear, could incur predictive bias even when all input predictors are unbiased.
Prevention of predictive bias	Predictive bias is taken into account during the selection of predictors.	Generally impossible unless plain ML without any fairness constraint already satisfies the adverse-impact requirement.

with only a few parameters (e.g., weights). This allows researchers and practitioners to directly inspect a model and identify potential problems. For example, if regressing criterion scores over predictors returned a negative weight for a cognitive ability test, a natural response would be to scrutinize the test design rather than to just deploy the model in practice. With an ML algorithm, however, a prediction model may easily contain millions of parameters learned from data (Goodfellow et al., 2016). While such added complexity affords ML models with better accuracy, it also makes the model prohibitively expensive to scrutinize manually (Du et al., 2019). This is compounded with the prevailing view in ML that tasks like predictor selection should be done automatically by the algorithm rather than manually by experts (Kleinberg & Mullainathan, 2019). As a result, an ML prediction model could easily circumvent desirable, fairness-related, properties—for example, the absence of predictive bias—so long as these properties are not explicitly specified in the optimization goal. This is exactly what we observed in this study.

There are two ways to address this problem. One is to further the work of *Interpretable ML* (e.g., LIME; Ribeiro et al., 2016) to make ML algorithms and models open to manual scrutiny (Xu et al., 2020). Unfortunately, a wide gap exists between what is deemed “interpretable” in ML and what could be properly examined in personnel selection. For example, the prevailing view in ML is that decision trees are interpretable because they can be expressed as (long) sequences of if-then-else statements. Yet the adoption of decision tree in personnel selection means that the predictor battery used for different candidates may contain different predictors, be applied in different orders, and feature different cutoffs. All these variations raise new research questions in terms of their legal defensibility in the personnel selection context, for which our understanding is still nascent. Thus, to pursue this direction, we believe the participation of I-O psychologists is urgently needed in the future development of interpretable ML for personnel selection.

The second way to address the problem is by formulating fairness notions, like the absence of predictive bias, as mathematical constraints that can be formally entered into an ML algorithm. Recent work in ML has already started formulating fairness constraints beyond adverse impact, to include notions such as the statistical parity of predictive accuracy between groups (Feldman et al., 2015), the assurance that no protected group under one selection system would overwhelmingly prefer another system (i.e., “envy-freeness”; Zafar et al., 2019), and so on. While handling multiple such constraints is technically feasible (e.g., the method of Lagrange multipliers could assign a different λ to each constraint), the feasibility does not mean one could achieve all these constraints without increasing adverse impact or reducing the expected criterion score of selected candidates. Indeed, many of these constraints are conflictive with each other (Kleinberg et al., 2017) or other legal requirements (e.g., privacy; Xu & Zhang, 2022), even without considering criterion. In personnel selection, this conflict could lead to thorny questions. For example, is it better to condone predictive bias when removing it leads to an increase of adverse impact? Like the first solution, the future research for this one is also in urgent need of I-O psychologists’ participation, not only in using ML to further organizational research (e.g., Zhang et al., 2022), but also in developing fairness-related constraints and understanding the mathematical tradeoffs between them.

As both solutions require long-term research and development, we offer some suggestions that may apply in the short term. A popular belief among ML researchers is that it is always better for an ML algorithm to take in as many predictors as possible—regardless of whether a predictor exhibits predictive bias—and count on the algorithm to sort out the proper use of these predictors and to achieve the desired fairness properties. While this belief is mathematically correct when adverse impact is the *only* fairness concern (Kleinberg & Mullainathan, 2019), the literature of personnel selection has long noted the limitation of having a singular focus on reducing adverse impact when seeking a diverse work force (Kehoe, 2008). Our study further shows the danger of relying on fairness-aware ML for adverse-impact reduction. For example, our simulation results indicate that predictive bias tends to be greater when the AIR requirement is more stringent. This suggests that, while fairness-aware ML can indeed reduce the adverse impact of selection outcomes, in doing so it might also be forced to incur other types of fairness concerns, such as predictive bias. Our study also suggests that one way to address these concerns is to carefully attend to characteristics of the selection system, such as the selection rate and the composition of the predictor battery. For example, our simulation results indicate that predictive bias could be greater in highly selective scenarios, or when there are considerable between-group differences in predictors, especially under stringent AIR requirements. To this end, we submit that ML researchers and practitioners should not consider fairness-aware ML algorithms as silver bullets that work on any and all predictors, but carefully study the empirical evidence in the I-O psychology literature in determining which predictors to use in the context of personnel selection.

3 | STUDY 2: OVERSAMPLING HIGHER-PERFORMING MINORITIES DURING MACHINE LEARNING MODEL TRAINING REDUCES ADVERSE IMPACT SLIGHTLY BUT ALSO REDUCES MODEL ACCURACY[2]

> “Is it going to have a disparate impact on different protected classes? That is the number one thing employers using artificial intelligence should be looking out for.” - EEOC Commissioner, Keith E. Sonderling (Strong, 2021)

Organizations are rapidly adopting tools that use artificial intelligence and ML for many purposes, including personnel assessment and selection (e.g., Campion et al., 2016; Hickman et al., 2022; Langer et al., 2020). However, significant concerns have been raised throughout society regarding the fairness and ethicality of ML assessments (Landers &

Behrend, 2023; Tippins et al., 2021). In the United States, a key legal concern for ML assessments is that personnel selection decisions that cause adverse (or *disparate*) impact—substantially different hiring rates between groups that disadvantage a legally protected group (Civil Rights Act, 1964)—constitute *prima facie* evidence of employment discrimination.

Several algorithmic solutions that adjust models to achieve equal group outcomes have been proposed to address group disparities in ML assessments (e.g., Calmon et al., 2017; Hardt et al., 2016; Kamishima et al., 2012; Kleinberg, Ludwig, Mullainathan, & Sunstein , 2018; Zemel et al., 2013), but many provide the final ML model with demographic information explicitly (e.g., by using demography as a predictor) or implicitly (e.g., by creating separate models for each group) during test administration. Both are likely illegal in the United States because they constitute disparate treatment (Civil Rights Act, 1964) and/or subgroup norming (Civil Rights Act, 1991) during test administration. Therefore, there is a pressing need to advance our understanding of the causes of and potential (legal) remedies to ML model adverse impact.

ML models tend to reflect subgroup differences in applicant attributes in the training data, which are then reflected in the ML model predictions. We investigate whether this tendency can be used to our advantage by examining whether removing (i.e., equal selection ratios) or reversing (i.e., selection ratios flipped to favor disadvantaged group members) subgroup differences in the training data reduces ML model adverse impact without sacrificing accuracy. To do so, we utilize a data preprocessing approach known as oversampling—techniques for resampling observations to address class imbalances (Chawla et al., 2002; Yan et al., 2020)—to manipulate adverse impact ratios in the training data. Then, we systematically examine how this affects the adverse impact and accuracy of ML models that use self-reports and interview transcripts to predict historical screening decisions.

The present study contributes to the literature on employment discrimination in several ways. First, we answer the special issue call to investigate adverse impact in artificial intelligence and ML personnel selection systems. Second, we answer calls to test the effects of oversampling minority groups to enhance diversity in training data (Hickman et al., 2022). We do so in a real-world, high-stakes dataset where adverse impact and group representation can be directly evaluated and altered. Oversampling to balance means and sample sizes has been shown to have small positive effects on ML model measurement bias (Yan et al., 2020) defined as equal accuracy across groups (Tay et al., 2022), but we are unaware of any studies of oversampling's effects on adverse impact. By doing so with both self-reports and interview transcripts, our study addresses the fairness of both traditional and modern selection systems. Further, we investigate the effects across a variety of text mining vectorization techniques and ML algorithms. This allows us to estimate the effect of oversampling on adverse impact across a variety of ML modeling approaches, reducing the chances that any observed effects are algorithm-bound. Third, we compare multiple oversampling strategies to inform future research and practice. Specifically, we compare the effects of (a) adjusting training data adverse impact versus adjusting training data adverse impact *and* equalizing sample sizes, as well as (b) oversampling real versus synthetic applicants. Doing so provides nuanced answers regarding how different oversampling methods affect the adverse impact and accuracy of ML model screening decisions.

3.1 | Indices of adverse impact

Adverse impact is often operationalized as an adverse impact (AI) ratio—or the ratio of the selection ratios of two subgroups. Selection ratios (SRs) are calculated as the number of applicants hired in a subgroup divided by the total number of applicants from that subgroup. The AI ratio is calculated by dividing one subgroup's SR by another subgroup's SR.

Adverse impact is concerned with equality of outcomes. The most common standard for identifying practically significant adverse impact and *prima facie* evidence of discrimination is the four-fifths rule, or that the SR of members of one legally protected subgroup should not be less than four-fifths the SR of members of another subgroup (Equal Employment Opportunity Commission, 1978).[3] Therefore, AI ratios should exceed .80. The AI ratio indicates the effect size of group differences in SRs and is commonly used, although significance testing is also relevant to discrimination claims (Morris, 2016). We chose to focus on the AI ratio because even minor subgroup differences in SRs become statistically significant when sample size is in the thousands, as in the present study.

3.2 | Origins of discrimination in ML models

ML models and their predictions reflect existing patterns in their training data. Therefore, to the extent that discrimination and/or adverse impact exist in the personnel data used to train ML models, the ML models may reflect those historical patterns (Barocas & Selbst, 2016). We now turn to summarize the standard ML model development and evaluation process, as illustrated in Figure B1 in the online supplement, and then explain the relevant sources of ML adverse impact that motivate our oversampling approach.

3.2.1 | Supervised ML in personnel assessment

Most ML assessments rely on supervised ML, which involves training an algorithm to predict some known individual-level outcome, such as historical screening or hiring decisions. To do so, individual behavior must be observed in an evaluative situation. Human observers then, either using the in situ behavior or a more holistic process involving additional information (e.g., resumés, cover letters), rate applicants and/or make selection decisions. A machine "perceiver" then observes and quantifies individual behavior, whether this behavior is performance in an evaluative situation (e.g., an interview), on a self-report scale, or on a test (e.g., of cognitive ability). For example, in automatically scored interviews, the unstructured, natural language of interviewee responses is transcribed, vectorized, and used in an algorithm to predict the outcome of interest (e.g., Hickman et al., 2022).

During ML model development, researchers often test multiple predictor-algorithm combinations. For example, in text mining, researchers may try out multiple vectorization techniques (i.e., methods for quantifying unstructured text data, such as closed and open vocabulary; Kern et al., 2016). To do so, the data are split into training and test datasets (e.g., Year 1 and Year 2, or *k*-fold cross-validation), the algorithm is fitted (or trained) on the training data, and the resulting ML model's accuracy is estimated on the test dataset. The predictor-algorithm combination with the highest cross-validated accuracy is often trained on all available data (i.e., both the training and test data). This final ML model is applied to future, unseen cases.

However, group differences in training data may affect the ML model, as reflected in the model parameters and its predictions. Two training data disparities affect ML models: (1) *group mean differences* on the outcome variable; and (2) *differential representation*, or underrepresentation of a subgroup (e.g., Barocas & Selbst, 2016; Kleinberg, Ludwig, Mullainathan, & Rambachan, 2018). Regarding group mean differences,[4] the concern is that if group mean differences in the training data are not representative of the population of applicants to which the model will be applied, this may alter the model weights in a way that favors one group of applicants over another (Barocas & Selbst, 2016). In our study, group means equal their SRs, and therefore, group mean differences represent adverse impact. We expect that training data AI ratios will affect ML model AI ratios, such that ML models trained with equal subgroup SRs (AI ratios = 1) or with subgroup SRs favoring the subgroup with the lower SR in the observed data (AI ratios > 1) will likely exhibit less adverse impact than models trained on data where AI ratios < 1.

Differential representation occurs when subgroups are unevenly represented in training data. In cases where the predictor-outcome relationships differ across groups,[5] unequal representation in the data will cause the algorithm to primarily reflect the most prevalent patterns in the training data—or those of majority group members (Barocas & Selbst, 2016). Therefore, we also investigate the effects of equally representing subgroups in training data, as it has been proposed elsewhere as a way of enhancing fairness and validity (Kleinberg, Ludwig, Mullainathan, & Rambachan, 2018).

3.2.2 | Diversity-validity tradeoff: ML edition

Making adjustments during ML model training to enhance fairness may negatively affect the model's convergent validity (in our study, accuracy; Barocas & Selbst, 2016). The so-called diversity-validity tradeoff is analogous to this concern (Ployhart & Holtz, 2008). The tradeoff occurs because some highly valid predictors of job performance (e.g., multiple choice cognitive ability tests) exhibit large subgroup differences, whereas selection procedures with smaller subgroup differences (e.g., personality traits) tend to less validly predict job performance. One common suggestion for addressing the so-called diversity-validity tradeoff is to find equally valid selection procedures with smaller subgroup differences (Ployhart & Holtz, 2008). Therefore, if oversampling to remove adverse impact in training data enhances ML model fairness without sacrificing convergence/accuracy, doing so may be preferable to training an ML model on raw historical data.

With these considerations in mind, our study addresses the following research questions:

Research Question 1a-b: How does adjusting the training data AI ratios affect ML model AI ratios when using (a) self-report scales, (b) text mined interview transcripts, or (c) both self-reports and interview transcripts to predict screening decisions?

Research Question 2: How does equally representing subgroups (i.e., equal *N*s) in training data affect ML model AI ratios?

Research Question 3: How does oversampling real versus synthetic observations affect ML model AI ratios?

Research Question 4: How does oversampling to remove adverse impact in training data affect ML model accuracy in the test data?

3.3 | Method

Figure 3 summarizes the present study's methods, and more detail is provided below and in the online supplement.

3.3.1 | Sample

Participants in our sample applied for US-based positions in a female-dominated service industry. The sample consists of 2501 applicants (71.9% female, 36.6% White, 28.3% Black or African American, 19.1% Hispanic, 6.3% two or more races, 4.1% Asian, and the remaining demographic groups each comprised <1% of the sample).

3.3.2 | Machine learning model predictor variables

Text mined interview transcripts. Participants recorded their answers to five interview questions using an online video platform, and computer software transcribed their responses. We applied six common vectorization techniques to convert the interviews to vectors for use as predictors in the ML models, as detailed in Figure 3 and the online supplement.

Self-report survey scores. The self-report survey included 16 proprietary multi-item, bipolar scales developed for the purposes of job selection that measure constructs including sociability, work ethic, and analytical mindset. These self-reports were scored in two ways: (1) as raw numerical scores and (2) as percentile scores based on norms derived by the survey vendor. The online supplement reports the scales' reliability and validity.

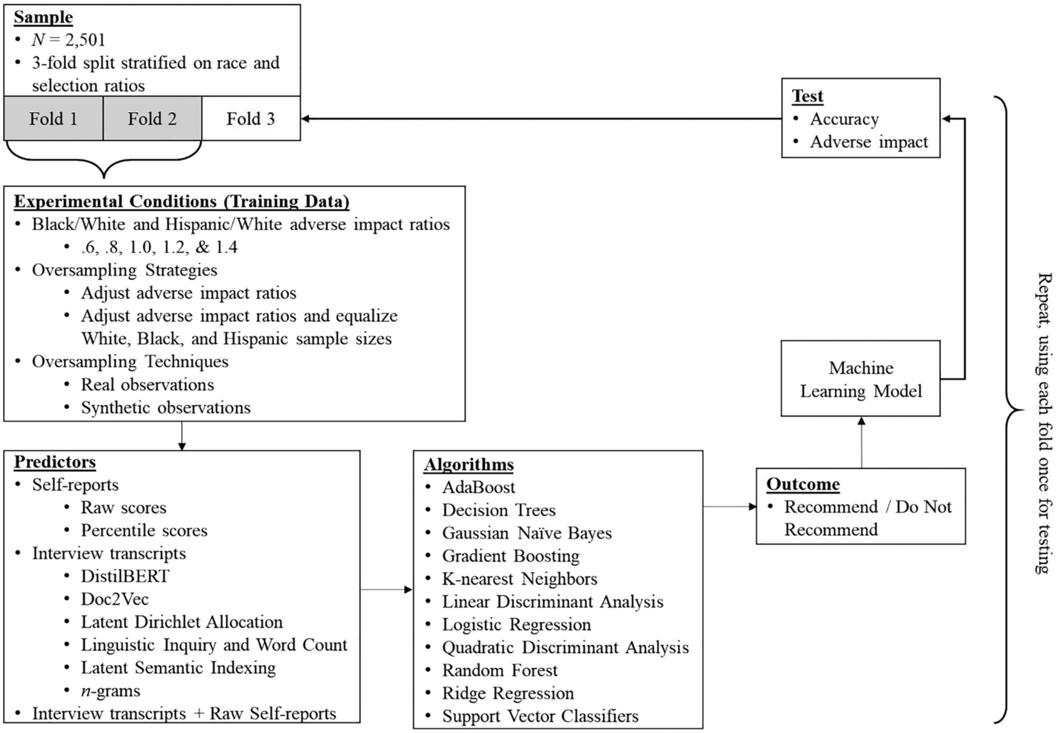

Sample
• N = 2,501
• 3-fold split stratified on race and selection ratios
Fold 1
Fold 2
Fold 3
Test
• Accuracy
• Adverse impact
Experimental Conditions (Training Data)
• Black/White and Hispanic/White adverse impact ratios
• .6, .8, 1.0, 1.2, & 1.4
• Oversampling Strategies
• Adjust adverse impact ratios
• Adjust adverse impact ratios and equalize White, Black, and Hispanic sample sizes
• Oversampling Techniques
• Real observations
• Synthetic observations
Machine Learning Model
Repeat, using each fold once for testing
Predictors
• Self-reports
• Raw scores
• Percentile scores
• Interview transcripts
• DistilBERT
• Doc2Vec
• Latent Dirichlet Allocation
• Linguistic Inquiry and Word Count
• Latent Semantic Indexing
• n-grams
• Interview transcripts + Raw Self-reports
Algorithms
• AdaBoost
• Decision Trees
• Gaussian Naïve Bayes
• Gradient Boosting
• K-nearest Neighbors
• Linear Discriminant Analysis
• Logistic Regression
• Quadratic Discriminant Analysis
• Random Forest
• Ridge Regression
• Support Vector Classifiers
Outcome
• Recommend / Do Not Recommend

FIGURE 3 Methods overview for the Study 2.

3.3.3 | Outcome variable (screening decisions)

We used the organization's screening decisions as the outcome variable. The decision is binary: applicants who passed proceeded to the next stage of the hiring process (screened in), and applicants who failed did not (screened out). The overall SR = .494, White applicant SR = .60, non-White applicant SR = .43, Black applicant SR = .46, and Hispanic applicant SR = .37. These SRs result in Non-White/White AI ratio = .72, Black/White AI ratio = .77, and Hispanic/White AI ratio = .62. A baseline model that always guesses "screened out" would have accuracy = .506, and this forms the 'baseline' accuracy against which ML model accuracy is judged.

3.3.4 | Train and test data splits for machine learning models

To compare multiple predictor-algorithm pairs, we created a stratified three-fold split of the raw data to conduct *k*-fold cross-validation. Specifically, we split the data into $k = 3$ folds such that White, Black, and Hispanic applicants had consistent *N*s and SRs in each fold, thereby maintaining the original data's properties. Across the three folds, White SR = .60, Non-White SR = .43, Black SR ranged from .45 to .47, and Hispanic SR ranged from .36 to .38. The three folds ranged in size from $N = 832$ to 835. For all experiments, we trained ML models on two folds then assessed them on the third fold and repeated the process three times, using each fold only once for testing. In total, we trained and tested: 5 (the adjusted training data AI ratios) * 2 (adjusting SRs or SRs and *N*s) * 2 (oversampling real or synthetic observations) * 154 (11 algorithms * 14 sets of predictors) + 154 (the 11 algorithms * 14 sets of predictors on the raw data) = 3234 models in each of the three folds, or 9702 models trained and tested.

The output of our experiments used in all analyses in the manuscript and online supplement is available on OSF: https://osf.io/c46sp/?view_only=2ffb2172f8274968bf720429812deae4

3.3.5 | Algorithms

We trained a variety of common machine learning algorithms, as detailed in Figure 3 and the online supplement. No one algorithm is optimal for all tasks (i.e., no free lunch theorem; Wolpert & Macready, 1997), and these represent a sample of commonly used ML algorithms. We conducted hyperparameter tuning for each algorithm in each set of training folds of the original data, as detailed in the online supplement. We fully crossed these algorithms with the two methods for scoring the self-reports, the six text mining approaches applied to the interview transcripts, and the combined predictor set of the six text mining approaches plus the raw self-report scores. This allowed us to estimate the effect of oversampling on model outcomes across a variety of predictor-algorithm pairs, thereby ensuring that our results generalize across many ML models. Tables B8–B10 in the online supplement report the average accuracy obtained on the raw data when SR = .50 for all algorithms, predictors, and predictor-algorithm pairs, respectively.

3.3.6 | Oversampling ratios

We used under- and oversampling to investigate the relationship between training data AI ratios and ML model AI ratios in the test data. Prior to model training, we under- and oversampled minorities in the training data to achieve Black/White and Hispanic/White AI ratios ranging from .60 to 1.40, stepping by .20. In all cases, we kept Black and Hispanic SRs in the numerator. To achieve training data AI ratio = .60, we *under*sampled passing Black and Hispanic applicants to reduce their SRs. To achieve AI ratios = .80 to 1.40, we oversampled passing (i.e., screened in) Black and Hispanic applicants until the desired AI ratio was achieved.

3.3.7 | Oversampling strategies

We investigated two oversampling strategies: (1) oversampling to adjust SRs and (2) oversampling to adjust SRs and equalize sample sizes. The former case is described in the “Oversampling Ratios” subsection. In doing so, Black and Hispanic sample sizes increased by the number of cases added to achieve the manipulated AI ratio. To equalize sample sizes, we multiplied the White *N* by the desired SR (as determined by the desired AI ratio),[6] then we (a) oversampled passing Black and Hispanic applicants (respectively) to reach those values and (b) oversampled (or undersampled, if necessary) Black and Hispanic applicants who failed until White, Black, and Hispanic applicant *N*s were equal.

3.3.8 | Oversampling techniques

We, (1) oversampled real observations with replacement or (2) oversampled synthetically generated observations. We used the Synthetic Minority Over-Sampling TEchnique (SMOTE; Chawla et al., 2002) to generate synthetic (a) screened in Black applicants, (b) screened in Hispanic applicants, (c) screened out Black applicants, and (d) screened out Hispanic applicants. Figure B2 in the online supplement illustrates how SMOTE works. We used the first two categories to adjust AI ratios and all four categories when adjusting AI ratios *and* equalizing sample sizes.

3.3.9 | Test data selection ratio

As a robustness check, we analyzed our results at different overall SRs in the test data: .10 and .50. To do so, we had the ML models output class probabilities (i.e., continuous values ranging from 0 to 1) instead of binary predictions. Then, to achieve overall SRs = .10 and .50, we set the highest 10% and 50%, respectively, of the class probabilities to 1 (pass/screen in) and the remaining values to 0 (fail/screen out). AI ratios are more likely to violate the four-fifths rule as the overall SR decreases (Oswald et al., 2016). SR = .50 is very similar to the observed SR in our data, and SR = .10 represents a more competitive (e.g., later stage) selection procedure.

3.4 | Results

To investigate our research questions, we treated the 9702 sets of algorithmic predictions as observations for analysis and measured their accuracy and AI ratios at overall SR = .10 and .50. In the raw data, the models that used interview transcripts to predict screening decisions tended to be more accurate than models that used self-reports as predictors, and the models that used both interview transcripts *and* self-reports tended to be no more accurate than the interview models. Among models that used: interview transcripts as predictors, $\text{Accuracy}_{\text{Max}} = 69.6\%$ (averaged across the three folds); self-reports as predictors, $\text{Accuracy}_{\text{Max}} = 60.2\%$; and combined predictor sets, $\text{Accuracy}_{\text{Max}} = 70.6\%$.

Research Question 1 concerns the effect of training data AI ratios on ML model AI ratios when screening applicants in the test data. Table 2 reports the average ML model accuracy and AI ratios in the raw data and at each manipulated training data AI ratio for models that used self-reports (top), interview transcripts (middle), and both interview transcripts and self-reports (bottom) as predictors (Tables B5–B7 in the online supplement report the same information at overall SR = .10). On average, among models that used self-reports as predictors, changing training data AI ratios from .6 to 1.4 caused the ML model AI ratios to increase from a minimum of .11 (Hispanic/White AI Ratios) to a maximum of .16 (Black/White AI Ratios). Among models that used interview transcripts as predictors, the average increases were smaller, ranging from a minimum of .04 (Hispanic/White AI ratios) to a maximum of .07 (Black/White AI Ratios). On average, among models that used both sets of predictors, the AI ratios increased from a minimum of .06 (Black/White AI ratios) to a maximum of .08 (Non-White/White AI ratios). These findings that the effects were largest among models that used self-reports and smallest among models that used interview transcripts as predictors align with the magnitude of correlations between training data AI ratios and ML model AI ratios reported in Table B1 in the online supplement for each predictor set. Thus, for all three predictor sets, training data AI ratios affect ML model AI ratios.

Notably, however, the effects were sometimes small in magnitude. Table 3 and Figure 4 report the ML model AI ratios for the most accurate model from each predictor set because these are the models likely to have been selected for subsequent use. For example, the most accurate model included the Latent Semantic Indexing (LSI) operationalization of interview transcripts plus self-reports as predictors and used linear discriminant analysis for prediction. When it was trained on the raw data where Black/White AI ratio = .77, and Hispanic/White AI ratio = .62, it exhibited an average Black/White AI ratio = .715 and Hispanic/White AI ratio = .611, whereas when it was trained on data where AI ratios were adjusted via oversampling to equal 1, it exhibited an average Black/White AI ratio = .747 and Hispanic/White AI ratio = .649.

Research Question 2 regards whether equalizing subgroup *N*s further enhances ML model AI ratios. Tables B2-B4 in the online supplement report the average ML model accuracy and AI ratios, respectively, for ML models that used self-reports, interview transcripts, and the combined predictor set in each experimental condition at overall SR = .50 (Tables B5–B7 in the online supplement report the same at overall SR = .10). AI ratios tended to increase by about .01 (although did not always do so) from manipulating SRs *and* equalizing *N*s when compared to only manipulating SRs. These findings align with the correlations between a dummy variable for whether SRs or SRs and *N*s were manipulated

TABLE 2 Average ml model accuracy and adverse impact ratios (Overall SR = .50).

		Accuracy				AI Ratio		
Train AI ratio		Overall	White	Black	Hispanic	NW/W	B/W	H/W
Self-reports	Raw	.568	.584	.558	.569	.736	.692	.753
	.6	.564	.581	.552	.563	.724	.689	.737
	.8	.564	.577	.554	.562	.743	.711	.754
	1.0	.560	.570	.551	.560	.768	.745	.771
	1.2	.554	.560	.547	.552	.807	.789	.808
	1.4	.544	.545	.541	.540	.854	.844	.849
Interview transcripts	Raw	.625	.622	.623	.636	.844	.869	.795
	.6	.630	.628	.629	.643	.835	.858	.787
	.8	.630	.628	.630	.641	.845	.872	.793
	1.0	.627	.625	.628	.636	.853	.884	.795
	1.2	.620	.617	.620	.629	.867	.902	.809
	1.4	.611	.608	.612	.620	.891	.928	.831
Combined predictors	Raw	.633	.636	.628	.639	.782	.798	.739
	.6	.631	.635	.624	.642	.768	.785	.726
	.8	.631	.633	.627	.640	.781	.801	.735
	1.0	.628	.629	.624	.637	.799	.825	.747
	1.2	.623	.623	.621	.631	.818	.848	.762
	1.4	.613	.613	.612	.619	.850	.844	.793

For self-reports, *N* = 66 models on the raw data; *N* = 1,320 models when oversampling; for both interview transcripts and combined predictors, *N* = 198 models on the raw data, *N* = 3960 models when oversampling.

TABLE 3 Average accuracy and adverse impact ratios for most accurate models (Overall SR = .50).

			Test AI Ratios		
Model	Train AI ratio	Overall accuracy	NW/W	B/W	H/W
Self-reports (raw)	Raw	.602	.669	.627	.674
	1.0	.587	.725	.715	.720
LSI	Raw	.696	.695	.712	.625
	1.0	.697	.716	.742	.647
LSI + Self-reports	Raw	.706	.688	.715	.611
	1.0	.693	.719	.747	.649

Note: Results averaged across folds, oversampling methods and techniques (on the raw data, for each, *N* = 3; when train AI ratio = 1.0, *N* = 12 for each. NW/W = Non-White/White AI Ratio; B/W = Black/White AI Ratio; H/W = Hispanic/White AI Ratio. Logistic regression provided the highest accuracy for survey scores; ridge regression for LSI; and linear discriminant analysis for LSI + Self-reports.

and the ML model AI ratios, in that the correlations show a minimal positive effect of equalizing *N*s beyond manipulating SRs. Overall, equalizing sample sizes tended to exert minimal, positive effects beyond manipulating training data AI ratios.

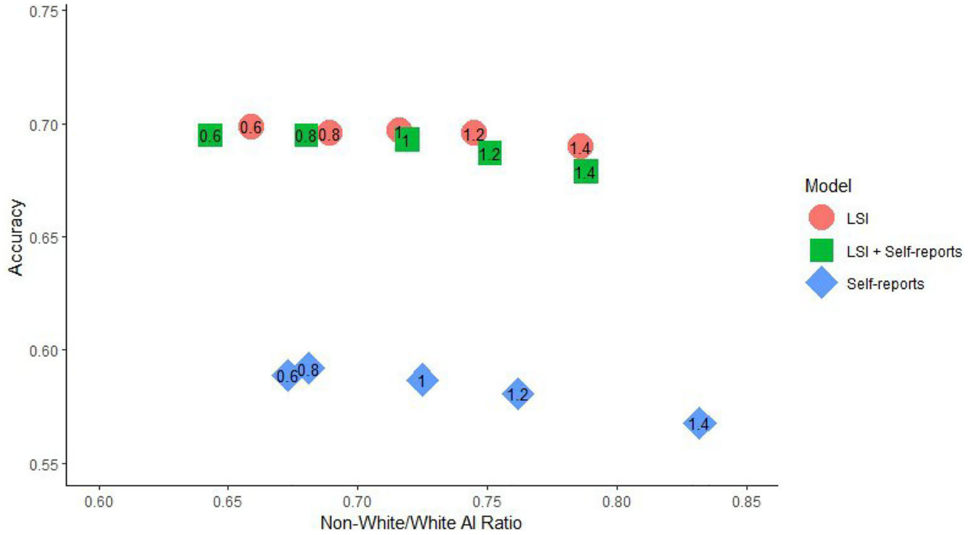


FIGURE 4 Average fairness-accuracy tradeoff for the most accurate models (selection ratio = .50). *Note*: Numbers in shapes indicate the training data Black/White and Hispanic/White AI ratios.

Research Question 3 regards whether different effects are observed from oversampling real versus synthetic observations generated by SMOTE. As reported in Tables B2–B4 in the online supplement, these effects tended to be even smaller in magnitude than the effect of manipulating SRs versus manipulating SRs and equalizing *N*s. Further, as reported in Table B1 in the online supplement, the effects were mixed across predictor sets, such that synthetic observations increased Non-White/White and Black/White AI ratios among models that used self-reports as predictors but decreased them and Hispanic/White AI ratios among models that used interview transcripts or both interview transcripts and self-reports as predictors. Therefore, oversampling real or synthetic observations provided similar effects.

Research Question 4 addresses the tradeoff between ML model accuracy and adverse impact. Table 3 reports the most accurate models' accuracy when trained on the raw data versus when training data AI ratios = 1, and Figure 4 illustrates the tradeoff between accuracy and Non-White/White AI ratios for these models when training data AI ratios = .6, .8, 1.0, 1.2, and 1.4 (on average across the other conditions). As Figure 4 shows, oversampling to adjust training data AI ratios tended to slightly decrease model accuracy. For example, for the models that used LSI and self-reports as predictors, AI ratios increased by .100 when training models on training data AI ratios = 1.4 compared to training on the raw data, and accuracy decreased by .027. Among models that used only LSI as predictors, AI ratios increased by .091 when training models on training data AI ratios = 1.4 compared to training on the raw data, and accuracy decreased by .006. This aligns with trends reported in Table 2 when all models' outputs were examined.

3.5 | Discussion

Adverse impact is a foundational concern for ML-powered selection tools, as they receive heightened scrutiny from applicants and policymakers. The present study investigated the effects of oversampling high-performing minorities, a technique being explored by data and computer scientists (Yan et al., 2020), on ML model AI ratios. Removing or reversing adverse impact in training data increased ML model AI ratios while reducing ML model accuracy, although the effect sizes were small.

3.5.1 | Theoretical and practical implications

Although adequate representation in training data is important for developing ML models that are equally accurate across demographic groups (Barocas & Selbst, 2016; Kleinberg, Ludwig, Mullainathan, & Rambachan, 2018), equal representation had very minor, positive effects on ML model AI ratios in our study. This may be because oversampling minority success already affected differential representation in our training data. Indeed, to create training data AI ratios = 1.0, 1.2, and 1.4, we necessarily oversampled many minority applicants, thereby increasing their sample size beyond the number of White applicants. Therefore, equal representation may have independent, positive effects, but our adjustments to training data AI ratios may have suppressed them.

Further, the observed effects were similar regardless of whether real or synthetically generated observations were oversampled. This is encouraging, because although binary classification problems tend to benefit more from oversampling synthetic than real observations (Chawla et al., 2002), there is something uncanny about using synthetic observations for personnel assessment. We encourage future work to continue to check if this holds true in other studies, but the current findings suggest that practitioners can reduce ML model adverse impact to a similar degree regardless of whether they oversample real or synthetic observations.

Although oversampling to adjust training data AI ratios slightly increased ML model AI ratios, doing so also tended to slightly reduce ML model accuracy. This issue is analogous to the so-called diversity-validity dilemma (Ployhart & Holtz, 2008) and a known limitation of methods of enhancing algorithmic fairness (Barocas & Selbst, 2016). Such decreases in accuracy limit the potential practical value of oversampling. Importantly, however, if ML models perfectly replicate historical human decisions, they could not reduce adverse impact. Future work is needed to determine how such adjustments affect validity and predictive bias, as these are more important than replicating historical decisions and our data did not include workplace outcomes.

3.5.2 | Limitations and future research

The generalizability of our findings to other personnel selection situations is limited by two primary properties of the dataset. First, our study focused on a subset of potentially useful predictors for personnel selection (i.e., self-reports and interview responses), yet many selection systems may have a broader array of predictors available, such as biodata and cognitive ability. Second, the sample size in our study is rather small for real-world ML applications, as they may be based, in practice, on tens of thousands of observations. The relatively low accuracy of our ML models may have enhanced the magnitude of oversampling's effects, whereas more accurate models may exhibit smaller effects from oversampling. Both a broader array of predictors and a larger sample size could potentially increase ML model accuracy and alter the effects of oversampling.

Due to the small effects of oversampling, future research should investigate additional approaches for addressing ML model adverse impact. For example, removing predictors that are predictive of group membership holds potential for reducing adverse impact (Booth et al., 2021), as does reducing the weight given to such predictors (Zhang et al., 2018). Future research is needed to determine the effects of these and other approaches in high-stakes settings.

The effects of oversampling high-performing minorities rely on an assumption that subgroup differences in predictors will be consistent in new data. If, however, subgroup differences on predictors were inconsistent between the training data and subsequent applicants, then training data AI ratios may have a weaker relationship with ML model AI ratios. This suggests another route to addressing adverse impact, regardless of whether ML models are used for assessment: enacting societal change to reduce subgroup differences in job-relevant qualifications (i.e., predictors). Mean racial subgroup differences in job-relevant qualifications begin at a young age (McDaniel et al., 2011), and without change, they may persist in society for another 90 years or more (Barrett et al., 2011). Addressing subgroup differences in job-relevant qualifications is the most direct route for reducing adverse impact across assessment methods.

4 | STUDY 3: MULTI-OBJECTIVE OPTIMIZATION FOR PERSONNEL SELECTION: A GUIDE, TUTORIAL, AND USER-FRIENDLY TOOL[7]

Organizations often want to optimize multiple hiring objectives simultaneously, yet doing so can be difficult, especially for conflicting objectives (e.g., cost and effectiveness; job performance and diversity). As a result, there is a growing need to develop robust methods for making hiring decisions that consider multiple objectives. To date, however, most analytic approaches are limited to optimizing just one objective at a time (e.g., cost *or* effectiveness, instead of cost *and* effectiveness). Multi-objective optimization (MOO; a.k.a., Pareto-optimization) is a promising machine learning approach that can help organizations optimize multiple objectives. It generates predictor weights that optimize the value of one objective at given levels of the other objective(s); organizations then choose the set of predictor weights that best fulfills their needs and values. MOO has been used in personnel selection to address the diversity-validity dilemma by deriving hiring solutions that can as much as double the proportion of minority hires while maintaining the expected job performance of the new hires (Wee et al., 2014).

Nonetheless, existing MOO applications that are readily available for personnel selection are limited in two ways: (a) they have only been applied to optimize two objectives, and (b) those two objectives have always been task performance and diversity (in the form of adverse impact; e.g., De Corte et al., 2007; Song et al., 2017). Yet, organizations are often concerned with *multiple* hiring objectives *beyond task performance and diversity*, such as the likelihood of early turnover (Speer et al., 2019) and organizational citizenship behavior (OCB; Ployhart et al., 2017).

This study aims to provide a generalized MOO guide and tool that help users optimize multiple objectives in personnel selection. By doing so, we aim to make two contributions. First, we provide a guide to generalize the application of MOO to a wide range of objectives and enable the use of MOO in situations beyond the diversity-validity dilemma. The guide details how to use MOO for different personnel selection applications by explaining what types of problems are best addressed with MOO, how to define MOO problems, and how to implement, evaluate, and monitor MOO selection systems. Second, we introduce a user-friendly online application (Multi-Objective Selection Tool, MOST; https://orgtools.shinyapps.io/MOST/) to help a wide range of users explore MOO without requiring complex computer programming or mathematical knowledge of machine learning algorithms. We hope this study will foster the adoption of MOO and, thereby, improve hiring outcomes.

4.1 | Multi-objective optimization and current organizational applications

As the term “multi-objective optimization” suggests, the MOO approach consists of a variety of algorithms that simultaneously optimize multiple objectives. MOO is useful any time the objectives are in conflict—or whenever the objectives cannot be simultaneously optimized (De Corte et al., 2007). Buying a used car, for example, can be considered a MOO problem. Suppose our *objectives* for a car purchase are to (1) minimize price and (2) minimize mileage. Because price tends to decrease as mileage increases, there is a conflict between the two objectives, making it a typical MOO problem. Our purchase decisions are also often bounded by some conditions. For instance, we may only want to consider minivans; in MOO, this is called an *equality constraint* (i.e., body type = minivan). Also, we may only want to consider cars with a fuel efficiency of at least 16 miles per gallon; in MOO, this is called an *inequality constraint* (i.e., gas mileage ≥16 miles per gallon). Constraints reduce the range of feasible solutions (e.g., car options) to ones that are most likely to meet our needs. Thus, a typical MOO problem consists of a set of objectives (e.g., minimize price, minimize mileage) and, often, some equality constraints (e.g., body type = minivan) and inequality constraints (e.g., gas mileage ≥16 miles per gallon).

The goal of MOO is to identify Pareto-optimal solutions. A Pareto-optimal solution optimizes one objective, at a certain level of the other objective(s). In our car buying example, the Pareto-optimal solutions include minivans with the lowest price given a mileage, as well as minivans with the lowest mileage given a price. The choice of a final solution depends on the preference of the buyer. For instance, among all the Pareto-optimal minivan options, the buyer might decide to select the minivan that has the lowest mileage among all the medium-priced options.

The example above concerns discrete choices of used cars. In many applications (e.g., personnel selection), the final solution is a set of predictor weights. For example, MOO has been used to address the diversity-validity dilemma in personnel selection (e.g., De Corte et al., 2007). The diversity-validity dilemma concerns how common personnel selection predictors and procedures (e.g., cognitive ability tests) that validly predict job performance also tend to engender adverse impact. This creates a MOO problem where organizations have to optimize two conflicting objectives (in this example, diversity and validity). One way to address this problem is to assign different weights to predictors (e.g., structured interview, personality assessments) so that the resulting weighted predictor composite exhibits high validity in predicting job performance and low adverse impact (Outtz & Newman, 2010).

MOO uses a data-driven approach to generate multiple sets of predictor weights, each of which optimizes one hiring outcome (e.g., adverse impact ratio [AI ratio]) at a given level of the other outcome(s) (e.g., job performance). The final choice of predictor weights depends on organizational needs and values. For instance, organizations focused on complying with the four-fifths rule (Equal Employment Opportunity Commission, 1978) may select the solution that maximizes the expected job performance with expected AI ratio greater than or equal to .80; while organizations focused on enhancing social equality may select the solution that maximizes the expected job performance with the expected AI ratio equal to 1.00 (see Newman et al., 2022).

In summary, MOO is useful when the objectives are in conflict with each other, which occurs when two or more objectives “cannot be optimized by exactly the same weighting of the available selection predictors” (De Corte et al., 2007, p. 1382). By analyzing the relationships between the predictors and multiple objectives simultaneously, MOO provides a set of optimal solutions that organizations could choose from based on their specific needs and values.

4.2 | Growing demand in personnel selection for optimizing multiple objectives

There is a growing demand among organizations to simultaneously optimize multiple objectives in personnel selection. Recently, interest has increased in performance criteria beyond task performance, such as OCBs and CWBs (Ployhart et al., 2017), and non-performance criteria, such as turnover and employee well-being (Speer et al., 2019). OCBs contribute to, and CWBs detract from, positive organizational functioning (Van Iddekinge & Ployhart, 2008); high rates of voluntary employee turnover harm organizational outcomes (Park & Shaw, 2013); and employee well-being influences job satisfaction, job performance, and retention (Cleveland & Colella, 2010). However, it is often difficult to optimize these objectives simultaneously, as they are not perfectly related and sometimes are negatively or non-linearly related (e.g., curvilinear relationship between job performance and retention; e.g., Salamin & Hom, 2005). Thus, a strategy that optimizes one objective might not be optimal for, and could even hinder, another objective.

MOO can help organizations develop selection systems that optimize multiple hiring objectives. In the sections below, we provide a guide, a point-and-click R Shiny app, and an R package for obtaining MOO solutions for general personnel selection purposes.

4.3 | Guide for implementing multi-objective optimization for personnel selection

Myriad MOO algorithms exist for obtaining Pareto-optimal solutions, which are summarized in the online supplement. In this study, we focus on the normal boundary intersection (NBI) algorithm developed by Das and Dennis (1998), which has been most commonly used in personnel selection (e.g., De Corte et al., 2007, 2011; Newman et al., 2022; Song et al., 2017; see Rupp et al., 2020). Table 4 provides a checklist with the key steps for adopting MOO in personnel selection, and the online supplement provides a step-by-step example demonstration.

TABLE 4 A checklist of the key decisions for adopting multi-objective optimization for personnel selection.

Stage 1. Define the MOO Problem

a. Determine the hiring objectives
   - What are the hiring objectives (e.g., job performance, retention, diversity)? The choice depends on organizational needs and values.
   - How to operationalize each hiring objective?
   - Optimization goal: Decide the extent to which each objective will be optimized

b. Choose predictors and corresponding assessment methods
   - What predictors best predict the hiring objectives? The choice of predictors needs to be backed by job analysis.
   - How to measure/assess the predictors?

c. Set proper constraints
   - Consider practical and legal needs

Stage 2. Obtain MOO Solutions

a. Choose a MOO algorithm
   - Considerations for selecting MOO algorithm (See Online Supplement C, Table C1)
     - Are there relative weights associated with each hiring objective?
       - Yes: a priori algorithms
       - No: a posteriori algorithms

b. Prepare input statistics
   - Identify the calibration sample
     - What incumbent group to sample/archival data to use as the calibration sample? Calibration samples should closely match the target sample of interest (e.g., applicant pool).
     - What is the expected sample size? When possible, the calibration sample size should be large.
   - Collect data from the *calibration sample*
   - Compute input statistics for MOO
     - Predictor intercorrelations
     - Predictor-objective relationships

c. Obtain MOO predictor weights
   - Obtain the MOO predictor weights using the input statistics
   - Choose a solution that best satisfies the optimization goal(s)

d. Pilot trial (highly recommend)
   - Identify the pilot sample (e.g., applicant sample)
   - Collect predictor data from the pilot sample
   - Apply the MOO predictor weights to the pilot sample and estimate weighted predictor composites for each individual in the pilot sample (but do not use it to make actual hiring decisions)
   - Evaluate whether the MOO predictor weights result in hiring outcomes that satisfy optimization goals, if applicable
   - Identify other practical needs (e.g., communication, training)

Stage 3. Implementation

a. Use MOO weighted predictor composite to make hiring decisions
   - Assess predictor data from the target sample of interest (e.g., job applicants)
   - For each applicant, calculate weighted predictor composite scores using the MOO predictor weights
   - Make hiring decisions based on (or partially based on) the weighted predictor composite scores

Stage 4. Maintenance

- Evaluate the objectives in each round of implementation
- Monitor hiring outcomes across different samples and scenarios
- If change is required (e.g., due to updates from job analysis or substantial shrinkage), re-evaluate the MOO selection system starting from Stage 1

Note: All steps need to be documented for future validation and legal auditing.

4.3.1 | Stage 1. Define the MOO problem

The first stage is to define and operationalize the hiring objectives. Hiring objectives can be informed by organizational needs and values; for example, an organization may aim to improve employee diversity, retention, and performance. Stage 1 also typically involves setting the optimization goal (i.e., to what extent each objective should be optimized). For instance, some organizations may choose to select solutions that allow more favorable retention rates only to the extent that expected task performance is not substantially decreased; other organizations may choose to maximize new hire task performance and OCB only to the extent the adverse impact risks are low.

The hiring objectives then inform the choice of predictors. Each predictor should individually demonstrate evidence of job-relatedness, through job analysis and content or criterion-related validity (Equal Employment Opportunity Commission, 1978).[8] As each predictor can be assessed in multiple ways, the most suitable method of assessment can be determined with validation studies, (lack of) overlap with other assessments, and practical considerations. For example, should interpersonal skills be measured using situational judgment tests or structured interviews? Are there assessment tools that are already available, or do they need to be developed? One should also examine practical and legal considerations related to the selection predictors, such as constraining predictor weights to be non-negative to properly reflect job analysis results (see De Corte et al., 2007).

4.3.2 | Stage 2. Obtain MOO solutions

The second stage focuses on obtaining MOO (or Pareto-optimal) predictor weighting solutions. Multiple algorithms are available for implementing MOO, which could be broadly classified as a priori and a posteriori algorithms. If the optimization goal is clear, a priori algorithm should be used; if it is not clear, then a posteriori algorithm should be used (see online supplement for details).

MOO algorithms are supervised machine learning algorithms that train models (i.e., develop predictor weighting solutions) using calibration/training sample data. For example, when MOO is used for hiring, the calibration/training sample consists of employees with criterion data (e.g., job performance ratings), and the target/testing sample is the applicants.

MOO models generate multiple sets of predictor weights that optimize each objective at given values of the other objective(s). From them, the user chooses the solution (i.e., predictor weights) that best satisfies the optimization goals. In other words, MOO generates possible solutions; and from those solutions, the organization selects one based on their values, goals, and business necessity. Even when the optimization goal is loosely defined (e.g., to improve all three objectives to a reasonable degree), one can still narrow down the solution space based on the goal. They can, for instance, identify a subset of solutions with higher expected new hire job performance, retention rate, and AI ratio than the current practice, which can be presented to the organizational decision-makers for further consideration.

When possible, users should conduct a pilot trial to (a) evaluate the predictor weights in the target sample of interest and (b) identify any preparations needed to integrate the MOO system into selection practice. Specifically, collect predictor information from a pilot sample (e.g., applicant sample), use the MOO predictor weights to identify individuals who might be selected with the MOO selection system (but not yet use them to make actual hiring decisions), and evaluate the hypothetical hiring outcomes (e.g., whether the AI ratio satisfies the four-fifths rule). In addition, examine practical needs such as communication (e.g., how to explain to stakeholders) and training (e.g., how to train recruiters and hiring managers).

4.3.3 | Stage 3. Implementation

The MOO solution is then implemented to make hiring decisions. Specifically, collect predictor information from the job applicants and, for each applicant, use the MOO predictor weights to calculate weighted predictor composites.

The weighted predictor composite scores can be used to rank order job applicants to aid compensatory selection. They can also be used in conjunction with non-compensatory methods (e.g., minimum cut-score requirements), for instance, by first selecting out applicants that did not meet the minimum cut-score for certain predictors (e.g., education requirement, licenses, work experience requirements) and rank ordering the remaining applicants based on weighted predictor composite scores.

4.3.4 | Stage 4. Maintenance

The MOO selection system should be maintained through continuous validation. In each round of implementation, the hiring outcomes should be evaluated to determine whether they still satisfy the optimization goals and detect any changes in the hiring outcomes. For example, are the new hires' job performance and 6-month retention rates similar to previous implementations? Does the AI ratio still satisfy the four-fifths rule? Are there any changes across different locations or times of the year? When the validation suggests a need to revisit the selection system, the selection system must be re-evaluated, starting from Stage 1. This procedure supports the continued effectiveness of the selection system.

4.4 | Multi-objective selection tool (MOST): A user-friendly tool to implement MOO

The MOST online application (https://orgtools.shinyapps.io/MOST/) is a user-friendly and freely available R Shiny application that uses the NBI algorithm (Das & Dennis, 1998) to estimate predictor weights for optimizing three hiring objectives. We also provide a corresponding R package, "rMOST" that is available via the Comprehensive R Archive Network (CRAN; https://cran.r-project.org/) repository. Figure 5 provides an example MOST Shiny app interface and the online supplement provides a detailed manual for using the app.

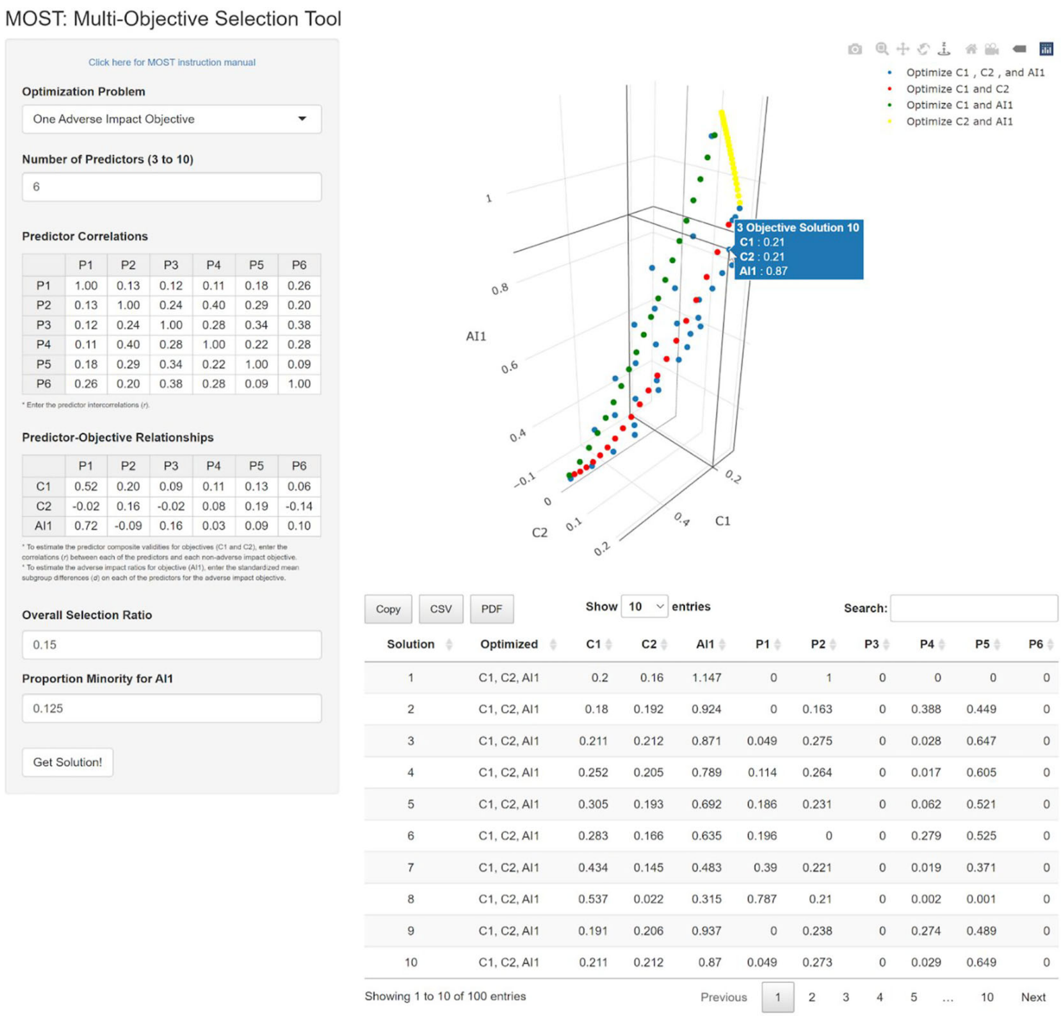
4.4.1 | Procedures to use the MOST Shiny app

Step 1. Define the MOO problem. The first step in implementing MOO in personnel selection is to define the hiring objectives (see Table 4). Common hiring objectives generally fall into two categories: adverse impact objectives and non-adverse impact objectives.[9] Adverse impact objectives relate to the proportion of selected applicants from legally protected groups (e.g., women, racial/ethnic minorities) relative to the proportion of selected majority applicants, and they are commonly operationalized with AI ratios (e.g., Oswald et al., 2016). Non-adverse impact objectives include all other objectives, such as dimensions of employee performance (e.g., task performance, OCB, CWB); they are commonly operationalized as supervisor ratings, peer ratings, and with objective employee records.

The MOST app can optimize predictor weights for: (1) three non-adverse impact objectives ("No Adverse Impact Objectives"), (2) two non-adverse impact objectives and one adverse impact objective ("One Adverse Impact Objective"), or (3) one non-adverse impact objective and two adverse impact objectives ("Two Adverse Impact Objectives"). Use the "Optimization Problem" drop-down menu to select one of the three options (see Figure 5). As the optimization function does not allow negative predictor weights (as suggested by De Corte et al., 2007), MOST requires that the predictors and non-adverse impact objectives be operationalized such that greater values indicate more desired outcomes (e.g., emotional stability instead of neuroticism; retention instead of turnover).

Step 2. Obtain MOO solutions.

Prepare input statistics. MOST (which implements NBI) takes the predictor intercorrelations and the predictor-objective relationships as input statistics, both of which can be estimated from the calibration sample. The predictor intercorrelations are represented by a correlation matrix of the predictors. The predictor-objective relationships are



MOST: Multi-Objective Selection Tool

Click here for MOST instruction manual

Optimization Problem

One Adverse Impact Objective

Number of Predictors (3 to 10)

6

Predictor Correlations

	P1	P2	P3	P4	P5	P6
P1	1.00	0.13	0.12	0.11	0.18	0.26
P2	0.13	1.00	0.24	0.40	0.29	0.20
P3	0.12	0.24	1.00	0.28	0.34	0.38
P4	0.11	0.40	0.28	1.00	0.22	0.28
P5	0.18	0.29	0.34	0.22	1.00	0.09
P6	0.26	0.20	0.38	0.28	0.09	1.00

* Enter the predictor intercorrelations (*r*).

Predictor-Objective Relationships

	P1	P2	P3	P4	P5	P6
C1	0.52	0.20	0.09	0.11	0.13	0.06
C2	-0.02	0.16	-0.02	0.08	0.19	-0.14
AI1	0.72	-0.09	0.16	0.03	0.09	0.10

* To estimate the predictor composite validities for objectives (C1 and C2), enter the correlations (*r*) between each of the predictors and each non-adverse impact objective.
* To estimate the adverse impact ratios for objective (AI1), enter the standardized mean subgroup differences (*d*) on each of the predictors for the adverse impact objective.

Overall Selection Ratio

0.15

Proportion Minority for AI1

0.125

Get Solution!

Copy CSV PDF Show 10 entries Search:

Solution	Optimized	C1	C2	AI1	P1	P2	P3	P4	P5	P6
1	C1, C2, AI1	0.2	0.16	1.147	0	1	0	0	0	0
2	C1, C2, AI1	0.18	0.192	0.924	0	0.163	0	0.388	0.449	0
3	C1, C2, AI1	0.211	0.212	0.871	0.049	0.275	0	0.028	0.647	0
4	C1, C2, AI1	0.252	0.205	0.789	0.114	0.264	0	0.017	0.605	0
5	C1, C2, AI1	0.305	0.193	0.692	0.186	0.231	0	0.062	0.521	0
6	C1, C2, AI1	0.283	0.166	0.635	0.196	0	0	0.279	0.525	0
7	C1, C2, AI1	0.434	0.145	0.483	0.39	0.221	0	0.019	0.371	0
8	C1, C2, AI1	0.537	0.022	0.315	0.787	0.21	0	0.002	0.001	0
9	C1, C2, AI1	0.191	0.206	0.937	0	0.238	0	0.274	0.489	0
10	C1, C2, AI1	0.211	0.212	0.87	0.049	0.273	0	0.029	0.649	0

Showing 1 to 10 of 100 entries Previous 1 2 3 4 5 … 10 Next

FIGURE 5 An example usage of the MOST Shiny App.

criterion-related validities (for non-adverse impact objectives; i.e., the correlation between each predictor and the measure of an objective) or subgroup difference (for adverse impact objectives; i.e., the standardized predictor mean score difference [Cohen's *d*] between minority and majority groups). For personnel selection applications, the intercorrelations and criterion-related validity need to be corrected for range restriction and criterion unreliability to reflect the relationships in the applicant sample (SIOP, 2018; see online supplement for recommended resources for range restriction correction).

To specify the inputs in MOST, begin by entering the number of predictors to be used in the field labeled "Number of Predictors." Next, enter the predictor intercorrelations in the table labeled "Predictor Correlations." The values entered into the table must be between −1 and 1. Simply click on a cell in the table and enter the relevant correlation; the predictors will be labeled "P" (for *predictor*) followed by a number representing their order in the table.[10]

Then, enter the relationships between the predictors and the objectives in the "Predictor-Objective Relationships" table. The non-adverse impact objective will be labeled "C" (for *criterion*) followed by a number representing their order among the non-adverse impact objectives in the table (e.g., "C1"); these predictor-objective relationships should be entered as correlations between −1 and 1. The adverse impact objectives will be labeled "AI" (for *adverse* *impact*) followed by a number representing their order among the adverse impact objectives in the table (e.g., "AI1"); these predictor-objective relationships should be entered as standardized mean subgroup differences of the predictor scores (subgroup *d*), with positive values favoring the reference group. If there are two adverse impact objectives, the subgroup *d*s entered should be the standardized mean difference between a minority group and the same reference group (e.g., both be minority groups compared to the White group, such as Black-White and Hispanic-White subgroup *d*'s).

If the MOO problem has "One Adverse Impact Objective" or "Two Adverse Impact Objectives," MOST will display additional fields requesting further input information. In both cases, the user needs to enter the expected overall selection ratio of the selection system (a value between 0 and 1) in the "Overall Selection Ratio" field. In addition, if the problem has "One Adverse Impact Objective," the user needs to enter the expected proportion of the minority subgroup in the applicant pool (a value between 0 and 1) in the "Proportion Minority for AI1" field. If the problem has "Two Adverse Impact Objectives," the user needs to enter the expected proportions of the two minority subgroups in the applicant pool in the "Proportion Minority for AI1" and "Proportion Minority for AI2" fields.

Obtain MOO predictor weights. After all the inputs are entered, click the "Get Solution!" button. MOST will visualize the MOO solutions in a 3-dimensional plot on the top right section of the screen and provide the predictor weights as well as the expected hiring outcomes (expected AI ratio and composite validities) associated with each solution, both in the plot and in a table on the bottom right section of the screen. The solutions in the table can be sorted in descending or ascending order based on each column, which facilitates solution comparison and selection. In addition to the results for the three-objective solutions, for each of the three pairs of objectives, MOST also provides the results of two-objective optimization solutions generated via NBI (e.g., C1 & C2, C2 & C3, C1 & C3).[11] The user can obtain more information on each solution by hovering the cursor over a point on the plot. For each point on the plot, MOST displays the solution number (which corresponds to the solution number in the table), predictor weights, and the expected hiring outcomes. The user can use both the plot and the table to select a solution that best fits their needs.

4.5 | Discussion

In this study, we highlighted organizations' growing demand to optimize multiple objectives in personnel selection and described how MOO can address this demand. We then introduced a guide for implementing MOO in personnel selection and provided a user-friendly online application and R package to help organizations implement MOO to optimize three hiring objectives.

4.5.1 | Utility of MOO over common personnel selection practices

To help readers consider the utility of MOO given their specific hiring scenarios and needs, we conducted a supplemental exploratory study to investigate the factors influencing the utility of MOO. Details of the study and recommendations are provided in the online supplement. Although the utility of MOO ultimately depends on the specific hiring scenario and the user's needs, the results of our exploratory study suggest that MOO tends to be more useful when predictors are differentially related to the objective. That is, MOO tends to be more useful over other predictor weighting methods when predictor-objective relationships vary widely across predictors; and MOO tends to be less useful over other methods when predictor-objective relationships do not vary much across predictors. When the predictor-objective relationships moderately vary across predictors, different users may find MOO useful to different degrees. Importantly, even in conditions where MOO does not seem to add much value, users can still choose to use MOO, as MOO generally performs (at the least) similarly to other predictor weighting methods. MOO solutions outperform unit weighting solutions; and MOO provides the same set of predictor weights as regression weighting at the endpoints (in addition to a range of optimal solutions that only MOO [but not regression] can provide).

4.5.2 | Shrinkage considerations with MOO and recommendations

When applying optimization methods such as MOO and regression weighting, we must consider shrinkage. Shrinkage refers to decreases in the weighted predictor composite's validity when predictor weights derived from one sample (calibration sample) are used in another sample (validation sample). In personnel selection, predictor weights are often obtained from an incumbent sample or archival data (calibration sample) and used to select applicants (validation sample).

Previous studies (De Corte et al., 2022; Song et al., 2017) suggested that shrinkage exists for MOO solutions. MOO shrinkage is influenced by calibration sample size and magnitude of the expected hiring outcome. First, shrinkage tends to be larger when the calibration sample size is small. Song et al. (2017), examining optimization of two objectives—job performance and diversity—found that shrinkage was sizable for a composite of common selection predictors (biodata, cognitive ability test, conscientiousness, structured interview, integrity test) when the calibration sample size was at or below 500; and shrinkage was sizable for a composite of cognitive subtest predictors when the calibration sample size was at or below 100. For a composite of common selection predictors, when calibration sample size was 500, validity shrinkage (difference in calibration and validation sample job performance validity) ranged between .00 and .01, and diversity shrinkage (difference in calibration and validation sample AI ratio) ranged between .00 and .08. In contrast, when calibration sample size was 100, validity shrinkage ranged between −.01 and .03, and diversity shrinkage ranged between .00 and .43 (see Song et al., 2017; Table 3).

Second, for a particular objective, across MOO solutions, shrinkage increases to the extent the objective is being maximized (Song et al., 2017). In other words, there tends to be the most shrinkage for objective C1 in the solution where C1 is maximized and the least (or no) shrinkage for C1 in the solution where C1 is least maximized. As an example, for a composite of common selection predictors, the diversity shrinkage (in terms of AI ratio) was as high as 1.24 (calibration sample size = 40, from AI ratio = 2.15 to .91; see Song et al., 2017; Table 3) for the solution where diversity was maximized but was approximately 0 (across all calibration sample size conditions) for the solution where diversity is least maximized (or where job performance was maximized; see Song et al., 2017; Table 3). Because the MOO solutions that maximize certain (single) objectives (i.e., the endpoints) are akin to regression solutions maximizing that same objective, compared to regression solutions, MOO solutions tend to be less or similarly susceptible to shrinkage [for the objective maximized by regression]. Specifically, compared to regression solutions that maximize a certain objective, MOO solutions at the endpoints are similarly susceptible to shrinkage [for that objective] while MOO solutions between the endpoints are less susceptible to shrinkage (see Song et al., 2017).

Based on these findings, we provide several recommendations regarding shrinkage when using MOO. First, when possible, use large calibration samples (e.g., with more than 100 or 500 individuals, depending on the predictors; see Song et al., 2017). Large calibration sample size can help reduce model overfit (or capitalizing on chance), thus reducing shrinkage (Song et al., 2021). Second, consider complementing MOO in conjunction with other approaches to improve hiring outcomes. For example, to ameliorate the diversity-validity dilemma in personnel selection and enhance organizational diversity, users could seek to develop and use predictor measures with smaller subgroup differences (e.g., Goldstein et al., 2010; Hough et al., 2001) and adopt recruitment methods that enhance diversity (e.g., Avery & McKay, 2006; Newman & Lyon, 2009).

Finally, consider using shrinkage formulas to approximate cross-validated hiring outcomes. MOO shrinkage formulas are recently developed to approximate cross-validated MOO (NBI) outcomes when optimizing two objectives (Song et al., 2023). The shrinkage formulas are developed on the basis that (1) classic shrinkage formulas (e.g., Browne, 1975; Claudy, 1978; Lord, 1950; Nicholson, 1960; Olkin & Pratt, 1958; Wherry, 1931) can be used to approximate cross-validated outcomes on the MOO solutions at the endpoints (e.g., solution where C1 is maximized); and (2) the solutions between the endpoints can be interpolated (based on the specific MOO algorithms). Although previous studies suggested that shrinkage formulas are effective for NBI-based MOO applications optimizing two objectives, future studies are needed to examine whether the MOO shrinkage formulas could be generalized to other scenarios (e.g., other MOO algorithms and optimizing more than two objectives). Such work will be instrumental for selection practices using MOO.

4.5.3 | Future directions

MOO is a promising machine learning method to advance both organizational practice and research. In addition to the personnel selection applications exemplified in this study, MOO could be used in a number of other workplace applications. In practice, organizations often face a tradeoff between the cost and validity of a selection system. While some selection procedures (such as structured interviews and assessment centers) have high validity in predicting job performance, they could also be costly to develop and administer; and other inexpensive procedures (such as personality assessments) may have lower validity. MOO could be used to address this practical concern by reducing the cost of a hiring design while optimizing diversity and validity. Specifically, one can use MOO to optimize three objectives: cost, job performance, and diversity hiring outcome. The cost objective can be operationalized as the sum of the cost of the predictors with non-zero weights, and the validity and diversity objectives can be operationalized as described earlier (e.g., job performance validity, AI ratio). With appropriate algorithms and operationalizations, MOO can provide solutions with a select set of predictors that minimize cost, at a given level of job performance validity and diversity outcomes.

In addition to improving organizational practice, MOO could be used as a research method to advance the theoretical understanding of workplace phenomena. Examples include expanding the predictor space—to explore new predictors that contribute to optimizing multiple organizational objectives. Existing methodologies (e.g., regression) that can only analyze one outcome at a time have restricted the historical focus in personnel selection to predictors that have high correlation with task performance, such as cognitive ability, structured interviews, and biodata (Sackett et al., 2021). The vast majority of our understanding of personnel selection predictors are informed by meta-analyses, regression, and structural equation modeling studies that examine the criterion-related validity, composite validity, and/or incremental validity of predictor(s) in predicting a single workplace outcome—for example, job performance *or* retention. Our field as a whole has limited knowledge of how different predictors influence multiple workplace outcomes simultaneously— for example, job performance *and* retention. MOO, with its ability to systematically examine multiple objectives, holds promise to unveil more holistic predictor-criterion relationships. With MOO, we can identify novel predictors of important work outcomes and develop methods to enhance multiple hiring objectives simultaneously—a broader, diverse collection of predictors holds potential to improve overall hiring outcomes.

ACKNOWLEDGMENTS

Papers in this article were submitted and evaluated individually in the traditional peer-review process by five independent reviewers, including regular editorial board members and a special board, judged to make a meaningful contribution after one or more rounds of revisions, accepted independently, and then combined into this thematic article. The co-editors of the special issue determined study ordering, which determined authorship ordering. The composite article should be cited in the normal manner starting with Zhang and then listing all other authors in the order of the individual articles. Authors of the first study are listed first, authors of the second study are listed second, and authors of the third study are listed third. The authors of the second study are grateful to Angela Duckworth, Benjamin Manning, Q. Chelsea Song, Daniel Southwick, Louis Tay, and Kaitlyn Werner for their comments on an earlier version of the manuscript.

DATA AVAILABILITY STATEMENT

The data that was used for the demonstration of the MOO guide and example application shown in the online supplement is available in Table C5 in the online supplement. The data that support the findings of an exploratory investigation of selection scenarios is also detailed in the online supplement.

ENDNOTES

[1] Study 1 authored by Zhang, N., Wang, M., Xu, H., & Koenig, N.

[2] Study 2 authored by Hickman, L., Kuruzovich, J., Ng, V., Arhin, K., & Wilson, D.

[3] Even when selection procedures violate the four-fifths rule, employers can demonstrate the job relevance and business necessity of the selection procedure (Civil Rights Act, 1964). However, employers may also want to reduce adverse impact for ethical reasons and to reduce the likelihood of litigation (Oswald et al., 2016).

[4] In the present study, we do not consider differences in standard deviations/variances between groups because the standard deviation of binary variables (like screening decisions) is determined primarily by their means. Specifically, a binary variable's standard deviation = $(np(1\text{-}p))$^.5 where n = sample size and p = the observed mean. Further, in our study, group SRs are equivalent to group means.

[5] Differential prediction is rare in selection, and when it does occur, it tends to come in the form of overpredicting minority performance (Dahlke & Sackett, 2022).

[6] The "fail" Black applicants was $N = 4$ larger than the "fail" White applicants. To equalize sample sizes, we first oversampled four of the "fail" White applicants before oversampling Black and Hispanic applicants.

[7] Study 3 authored by Song, Q. C., Tang, C., Alexander III, L. Hickman, L., & Kim, Y.

[8] For benchmarks of criterion-related validity, we refer readers to Bosco et al. (2015), which provides an overview of effect size distributions for various bivariate relationships examined in applied psychology.

[9] The adverse impact objectives, which regard maintaining similar selection ratios across demographic groups, must be treated differently than other objectives because they are operationalized with group differences (or related statistics) rather than a validity coefficient.

[10] MOST will automatically update the corresponding cell on the other side of the diagonal with the same value so the correlation matrix remains symmetric. It will not allow the user to change the "1"s on the diagonal of the matrix. Note that in the cells that allow edits, negative values require a leading zero (e.g., "−0.20" instead of "−.20").

[11] Previous studies have suggested NBI's limitations in finding the entire Pareto front when optimizing more than two objectives (Burachik et al., 2017). Thus, the MOST app generates MOO solutions that optimize two objectives as well as the solutions that optimize three objectives, allowing the organization to choose from a range of solutions to satisfy their organizational needs.

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.

Avery, D. R., & McKay, P. F. (2006). Target practice: An organizational impression management approach to attracting minority and female job applicants. *Personnel Psychology*, *59*, 157–187.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. https://www.fairmlbook.org

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732.

Barrett, G. V., Miguel, R. F., & Doverspike, D. (2011). The Uniform Guidelines: Better the devil you know. *Industrial and Organizational Psychology*, *4*(4), 534–536.

Bauer, D. J. (2005). The role of nonlinear factor-to-indicator relationships in tests of measurement equivalence. *Psychological Methods*, *10*(3), 305.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021). Bias and fairness in multimodal machine learning: A case study of automated video interviews. *Proceedings of the 2021 International Conference on Multimodal Interaction* (ICMI '21).

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449.

Browne, M. W. (1975). A comparison of single sample and cross-validation methods for estimating the mean squared error of prediction in multiple linear regression. *British Journal of Mathematical and Statistical Psychology*, *28*, 112–120.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the Conference on Fairness, Accountability and Transparency*, 77–91.

Burachik, R. S., Kaya, C. Y., & Rizvi, M. M. (2017). A new scalarization technique and new algorithms to generate Pareto fronts. *SIAM Journal on Optimization*, *27*(2), 1010–1034.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017, December). Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3995–4004).

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology*, *101*(7), 958–975.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Civil Rights Act of 1964. (1964). Pub. L. No. 88–352, 78 Stat. 243.

Civil Rights Act of 1991. (1991). Pub. L. No. 102–166, 105 Stat. 1071.

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement*, *2*, 595–607.

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.

Cleveland, J. N., & Colella, A. (2010). Criterion validity and criterion deficiency: What we measure well and what we ignore. In J. L. Farr, & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 551–567). Routledge/Taylor & Francis Group.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms* (3rd ed.). MIT press.

Cotter, A., Jiang, H., Gupta, M. R., Wang, S., Narayan, T., You, S., & Sridharan, K. (2019). Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, *20*(172), 1–59.

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

Das, I., & Dennis, J. E. (1998). Normal-Boundary Intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems. *SIAM Journal on Optimization*, *8*(3), 631–657.

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/idUSKCN1MK08G

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.

De Corte, W., Lievens, F., & Sackett, P. R. (2022). A comprehensive examination of the cross-validity of pareto-optimal versus fixed-weight selection systems in the biobjective selection context. *Journal of Applied Psychology*, *107*(8), 1243–1260.

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.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68–77.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. http://uniformguidelines.com/uniguideprint.html

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Finch, D. M., Edwards, B. D., & Wallace, J. C. (2009). Multistage selection strategies: Simulating the effects on adverse impact and expected performance for various predictor combinations. *Journal of Applied Psychology*, *94*(2), 318–340.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 329–338.

Goldstein, H. W., Scherbaum, C. A., & Yusko, K. P. (2010). Revisiting g: Intelligence, adverse impact, and personnel selection. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 95–134): Routledge/Taylor & Francis Group.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hardt, M., Recht, B., & Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. *Proceedings of The 33rd International Conference on Machine Learning*, *48*, 1225–1234.

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

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

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. *Proceedings of the 2010 IEEE International Conference on Data Mining*, 869–874.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J., (2012, September). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35–50.

Kehoe, J. F. (2008). Commentary on pareto-optimality as a rationale for adverse impact reduction: What would organizations do? *International Journal of Selection and Assessment*, *16*(3), 195–200.

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods*, *21*(4), 507–525.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). (May). Algorithmic fairness. In *AEA Papers and Proceedings* (Vol., *108*, pp. 22–27).

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis*, *10*, 113–174.

Kleinberg, J., & Mullainathan, S. (2019). Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. *Proceedings of the 2019 ACM Conference on Economics and Computation*, 807–808.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science*, 43:1–43:23.

Landers, R. N., & Behrend, T. S. (2023). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*, *78*(1), 36–49. https://doi.org/10.1037/amp0000972

Langer, M., König, C. J., & Busch, V. (2020). Changing the means of managerial work: Effects of automated decision support systems on personnel selection tasks. *Journal of Business and Psychology*, *36*, 751–769.

Lord, F. M. (1950). *Efficiency of prediction when a regression equation from one sample is used in a new sample* (Research Bulletin No. 50-40). Princeton, NJ: Educational Testing Service.

McDaniel, M. A., Kepes, S., & Banks, G. C. (2011). The Uniform Guidelines are a detriment to the field of personnel selection. *Industrial and Organizational Psychology*, *4*(4), 494–514.

Morris, S. (2016). Statistical significance testing in adverse impact analysis. In S. M. Morris, & E. M. Dunleavy (Eds.), *Adverse impact analysis: Understanding data, statistics, and risk* (pp. 91–111). Routledge.

Morris, S. B., & Lobsenz, R. E. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology*, *53*(1), 89–111.

Newman, D. A., & Lyon, J. S. (2009). Recruitment efforts to reduce adverse impact: Targeted recruiting for personality, cognitive ability, and diversity. *Journal of Applied Psychology*, *94*(2), 298–317.

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

Nicholson, G. (1960). Prediction in future samples. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 322–330). Stanford University Press.

Nocedal, J., & Wright, S. (2006). *Numerical optimization*. Springer.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, *29*, 201–211.

Oswald, F. L., Dunleavy, E. M., & Shaw, A. (2016). Measuring practical significance in adverse impact analysis. In S. M. Morris, & E. M. Dunleavy (Eds.), *Adverse impact analysis: Understanding data, statistics, and risk* (pp. 112–132). Routledge.

Outtz, J., & Newman, D. (2010). A theory of adverse impact. In J. Outtz (Eds.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 80–121). Routledge.

Park, T.-Y., & Shaw, J. D. (2013). Turnover rates and organizational performance: A meta-analysis. *The Journal of Applied Psychology*, *98*(2), 268–309.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*(1), 153–172.

Ployhart, R. E., Schmitt, N., & Tippins, N. T. (2017). Solving the Supreme Problem: 100 years of selection and recruitment at the Journal of Applied Psychology. *Journal of Applied Psychology*, *102*(3), 291–304.

Pyburn, Jr, K. M., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity–validity dilemma: Overview and legal context. *Personnel Psychology*, *61*(1), 143–151.

Rambachan, A., Kleinberg, J., Mullainathan, S., & Ludwig, J. (2020). An economic approach to regulating algorithms (Tech. Rep.). National Bureau of Economic Research.

Rasmussen, C., & Williams, C. (2006). *Gaussian processes for machine learning*. MIT Press.

Ribeiro, M. T., Singh, S., & Guestrin, C., Why, S. (2016). uld I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rupp, D. E., Song, Q. C., & Strah, N. (2020). Addressing the so-called validity–diversity trade-off: Exploring the practicalities and legal defensibility of pareto-optimization for reducing adverse impact within personnel selection. *Industrial and Organizational Psychology*, *13*(2), 246–271.

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.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2021). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, Advance online publication.

Salamin, A., & Hom, P. W. (2005). In search of the elusive U-shaped performance-turnover relationship: Are high performing Swiss bankers more liable to quit? *Journal of Applied Psychology*, *90*(6), 1204–1216.

Sejnowski, T. J. (2018). *The deep learning revolution*. MIT Press.

Siskin, B. R., & Trippi, J. (2005). Statistical issues in litigation. *Employment Discrimination Litigation: Behavioral, Quantitative, and Legal Perspectives*, 132–166.

Society for Industrial and Organizational Psychology (SIOP). (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.). Society for Industrial; Organizational Psychology, Inc.

Song, Q. C., Tang, C., Newman, D. A., & Wee, S. (2023). Adverse impact reduction and job performance optimization via Pareto-optimal weighting: A shrinkage formula and regularization technique using machine learning. Journal of Applied Psychology. Advance online publication. https://doi.org/10.1037/apl0001085

Song, Q. C., Tang, C., & Wee, S. (2021). Making sense of model generalizability: A tutorial on cross-validation in R and Shiny. Advances in Methods and Practices in Psychological Science, *4*(1).

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.

Speer, A. B., Dutta, S., Chen, M., & Trussell, G. (2019). Here to stay or go? Connecting turnover research to applied attrition modeling. *Industrial and Organizational Psychology*, *12*(3), 277–301.

Steinley, D. (2006). Profiling local optima in *k*-means clustering: Developing a diagnostic technique. *Psychological Methods*, *11*(2), 178–192.

Strong, J. (Host) (2021, July 21). Playing the job market [Audio podcast episode]. *In Machines We Trust*. MIT Technology Review. https://podcasts.apple.com/us/podcast/in-machines-we-trust/id1523584878

Sugiyama, M., & Storkey, A. J. (2006). Mixture regression for covariate shift. *Advances in Neural Information Processing Systems*, *19*, 1337–1344.

Tan, Z., Yeom, S., Fredrikson, M., & Talwalkar, A. (2020). Learning fair representations for kernel models. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 155–166.

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in psychological assessment. Advances in Methods and Practices in Psychological Science, *5*(1).

Tippins, N. T., Oswald, F. L., & McPhail, S. M. (2021). Scientific, legal, and ethical concerns about AI–based personnel selection tools: a call to action. *Personnel Assessment and Decisions*, *7*(2), 1.

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology*, *61*(4), 871–925.

Vapnik, V. (1998). *Statistical learning theory* (1st ed.). Wiley-Interscience.

Vapnik, V., & Vashist, A. (2009). A new learning paradigm: Learning using privileged information. *Neural Networks*, *22*(5-6), 544–557.

Wee, S., Newman, D. A., & Joseph, D. L. (2014). More than *g*: Selection quality and adverse impact implications of considering second-stratum cognitive abilities. *Journal of Applied Psychology*, *99*(4), 547–563.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, *2*, 440–457.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82.

Xu, H., & Zhang, N. (2022). Implications of data anonymization on the statistical evidence of disparity. *Management Science*, *68*(4), 2600–2618.

Xu, H., Zhang, N., & Zhou, L. (2020). Validity concerns in research using organic data. *Journal of Management*, *46*(7), 1257–1274.

Yan, S., Huang, D., & Soleymani, M. (2020, October). Mitigating biases in multimodal personality assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, 361–369.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, *20*(75), 1–42.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *Proceedings of Machine Learning Research*, *28*(3), 325–333.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhang, N., Wang, M., & Xu, H. (2022). Disentangling effect size heterogeneity in meta-analysis: A latent mixture approach. *Psychological Methods*, *27*(3), 373–399.

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. C. (2018). Learning gender-neutral word embeddings. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhang, N., Wang, M., Xu, H., Koenig, N., Hickman, L., Kuruzovich, J., Ng, V., Arhin, K., & Wilson, D., Song, Q. C., Tang, C., Alexander, L., & Kim, Y. (2023). Reducing subgroup differences in personnel selection through the application of machine learning. *Personnel Psychology*, *76*, 1125–1159. https://doi.org/10.1111/peps.12593