AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Applied Psychology

© 2023 American Psychological Association
ISSN: 0021-9010

2023, Vol. 108, No. 9, 1425–1444
https://doi.org/10.1037/apl0001084

New Strategies for Addressing the Diversity–Validity Dilemma With Big Data

Caleb Rottman, Cari Gardner, Joshua Liff, Nathan Mondragon, and Lindsey Zuloaga
HireVue, South Jordan, Utah, United States

The diversity–validity dilemma is one of the enduring challenges in personnel selection. Technological advances and new techniques for analyzing data within the fields of machine learning and industrial organizational psychology, however, are opening up innovative ways of addressing this dilemma. Given these rapid advances, we first present a framework unifying analytical methods commonly used in these two fields to reduce group differences. We then propose and demonstrate the effectiveness of two approaches for reducing group differences while maintaining validity, which are highly applicable to numerous big data scenarios: iterative predictor removal and multipenalty optimization. Iterative predictor removal is a technique where predictors are removed from the data set if they simultaneously contribute to higher group differences and lower predictive validity. Multipenalty optimization is a new analytical technique that models the diversity–validity trade-off by adding a group difference penalty to the model optimization. Both techniques were tested on a field sample of asynchronous video interviews. Although both techniques effectively decreased group differences while maintaining predictive validity, multipenalty optimization outperformed iterative predictor removal. Strengths and weaknesses of these two analytical techniques are also discussed along with future research directions.

Keywords: machine learning, diversity–validity dilemma, video interviews, selection

Supplemental materials: https://doi.org/10.1037/apl0001084.supp

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

The ability to both select a diverse workforce and have a valid assessment is one of the most difficult challenges in personnel selection. This challenge, often referred to as the diversity–validity dilemma, reflects the quandary where assessments with the highest predictive validity also tend to have the highest adverse impact (Ployhart & Holtz, 2008; Pyburn et al., 2008). This forces organizations to choose between hiring a diverse set of candidates or hiring the candidates most likely to succeed. An example of this trade-off can be found when cognitive ability (or general mental ability) tests are used to select applicants, since it simultaneously demonstrates high criterion-related validity (Schmidt & Hunter, 1998) and substantial group differences against Blacks and Hispanics (Roth et al., 2001). While numerous strategies have been proposed for addressing the diversity–validity dilemma (e.g., De Corte et al., 2011; Ployhart & Holtz, 2008), new techniques are emerging with the technological advancements in the fields of machine learning (ML; e.g., Zhang et al., 2018) and industrial and organizational psychology (IOP; e.g., Rupp et al., 2020; Tonidandel et al., 2018) that draw upon an increased ability to collect and process large amount of data and sophisticated algorithms to better analyze the resulting big data.

The exponential growth in the fields of ML, natural language processing, and deep learning has resulted in significant advances in the ability to create predictive algorithms. These techniques can now outperform humans in complex games such as chess, Jeopardy!, and Go (Greenemeier, 2017; Hern, 2017; Markoff, 2011). The field of IOP is in the initial stages of applying these new advances to reexamine assumptions and study organizational phenomena in new ways (Campion et al., 2016). Therefore, applying these new advances to reduce group differences while maintaining predictive validity is an area ripe for exploration. Although previous algorithmic methods for reducing group differences proposed in the selection literature have typically been limited to scenarios with a small number of variables per applicant, such as scores from personality and cognitive assessments (e.g., Hough et al., 2001; Potosky et al., 2005; Sackett & Ellingson, 1997; Sackett & Roth, 1996; Schmitt et al., 1997), new technologies such as social media analytics, resume analytics, web scraping, and video interviews are generating large data sets where each applicant can have hundreds to thousands of data points (Chamorro-Premuzic et al., 2016; Roth et al., 2016). In these scenarios, it is common for the number of predictors to exceed the sample size, resulting in unique analytical challenges.

Recent research has demonstrated the efficacy of using ML algorithms to score assessment tools (e.g., Campion et al., 2016; de Montjoye et al., 2013; Leutner et al., 2020); however, there is also an opportunity to explore diversity–validity trade-offs with ML. Accordingly, the purpose of the current research is to examine ways of reducing group differences while maintaining predictive validity in the context of big data and ML. The current research benefits the personnel selection literature in three primary ways. First, we present a unified framework of analytical methods for reducing group differences that incorporates techniques used in the fields of ML and IOP. Second, we demonstrate the effectiveness of two techniques for reducing group differences that are well-suited to big data. The first technique selectively removes predictors in a series of

This article was published Online First April 10, 2023.

Cari Gardner https://orcid.org/0000-0002-8092-4881

Correspondence concerning this article should be addressed to Cari Gardner, HireVue, 10876 S River Front Pkwy, South Jordan, UT 84095, United States. Email: cari.gardner@hirevue.com

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

iterative steps. While selectively removing predictors is a strategy that has historically been used in the field of IOP (e.g., Drasgow, 1987) and hiring companies commonly claim to use the strategy (Raghavan et al., 2020), we operationalize this strategy and demonstrate how it can be used in a big data scenario. The second technique is a new technique that incorporates a group difference penalty into the ML model optimization to explicitly optimize the diversity–validity trade-off. Third, we demonstrate the effectiveness of these two techniques using a field sample of applicants completing asynchronous video interviews. Altogether, the goal of this research is to introduce and develop novel and more effective ways to address the diversity–validity dilemmas that are emerging in the big data revolution.

New and Innovative Talent Identification Tools

The tools used to identify and select top talent are rapidly becoming more high-tech. The potential benefits in using these new tools include reduced administrative costs, less time required by candidates and hiring managers, improved applicant reactions, and improved ability to reach a more diverse slate of candidates (e.g., Brenner et al., 2016; Georgiou & Nikolaou, 2020; HireVue, 2017). An additional benefit of these new tools is their potential to collect new types of behavioral data. For example, while a traditional cognitive ability test may measure the number of correct answers within a specific time period, a gamified cognitive ability test can measure response time and accuracy for every interaction that the applicant has with the game, exponentially increasing the amount of data that can be used to generate a cognitive ability score. This results in data characteristics that are different from those found in traditional selection tests. In most traditional selection test scenarios, predictors are chosen based on theoretical and empirical relationships with the criterion, resulting in most predictors having a statistical relationship with the criterion. Big data scenarios, however, favor more predictors due to the expectation that unimportant predictors will be ignored by the ML model. Therefore, one of the main objectives of ML algorithms is to distinguish meaningful from spurious patterns, which are more likely to occur in big data scenarios. The ability to distinguish meaningful from spurious patterns becomes more complicated if group differences are also considered. If group differences exist in the criterion, the ML model attempts to predict these differences, and certain predictors may become important in the model merely because they provide information about group membership. For example, if males receive lower manager ratings of performance, then words or phrases such as “my husband” can become important predictors, not necessarily because they are related to true performance, but instead, because they help the model to identify the individual as most likely being female.

Perhaps the biggest opportunity that emerges from these new technologies and the corresponding rise in ML is the scoring of text-based data (e.g., resumes, open-ended application blanks, open-ended situational judgment tests, video interviews, written case study responses, scraped social media data). While an applicant can usually select only one answer with a multiple-choice question, with text-based data, an applicant has more autonomy both on what and how much they share. Natural language processing, the extraction of meaningful information from text-based data (e.g., Chowdhury, 2003), has drastically improved in recent years. Natural language processing methods include bag-of-words techniques where frequency of word usage is measured and the order of the words has no meaning (e.g., jumping jack is treated the same as jack jumping; El-Din, 2016), sentiment analysis where data dictionaries are used to determine the higher order meaning expressed by individuals (e.g., Serrano-Guerrero et al., 2015), and more cutting-edge techniques where word context is taken into consideration (Devlin et al., 2018). These new techniques provide improved and more nuanced interpretations of text-based data, which can then be used as inputs into analytical models.

The successful application of natural language processing and ML in a selection-based context is dependent upon the type of text-based data being used, the analytical techniques employed, and goals of the researcher. Despite the excitement around using ML and natural language processing to predict key outcomes in a selection context, given both privacy and legal concerns, it is recommended that researchers take a thoughtful and rigorous approach. For example, information gathered from an individual’s Facebook profile can successfully predict personality (Wald et al., 2012); however, not all applicants will have a Facebook profile, and using social media data may also result in privacy concerns and negative applicant reactions. Some examples of research that has used natural language processing and ML in a selection-based context include Sajjadiani et al. (2019) who used application form data to predict teacher performance and turnover, Hickman et al. (2022) who used transcribed text from video interviews to predict both self-reported and interviewer-rated personality traits, and Campion et al. (2016) who used text from candidate essays to predict human evaluator ratings. A challenge with using text-based data in a selection context is that different demographic groups use language differently; therefore, it is possible to identify demographic characteristics merely by knowing the words that an individual used (Gillick, 2010; Meier et al., 2020; Schler et al., 2006). Altogether, the nuances inherent in language and the relationship between language usage and demographic group membership can pose challenges in using text-based data in a selection context.

There are two primary benefits with using natural language processing and ML to evaluate text-based data. First, using human raters to systematically evaluate text-based data requires considerable time and effort. Natural language processing can reliably and accurately replicate human raters in a cost-effective manner (Campion et al., 2016). Second, human raters are prone to many explicit and implicit biases (Macan & Merritt, 2011). Traditional methods for decreasing these biases, such as training and standardization of evaluation methods, can be expensive and time-consuming. In contrast, ML techniques can be used to control for rater biases, ultimately reducing group differences. Thus, while ML can be applied to both text-based and nontext-based data, the application of ML to text-based data is especially promising given the large number of words that can be used and the potential for bias (e.g., Bohnet, 2016; Johnson et al., 2010) along with the cost inherent in using human evaluators.

Approaches to Reducing Group Differences

Within the field of personnel selection, numerous strategies have been proposed for reducing group differences while maintaining predictive validity. These strategies range from candidate-based

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

strategies, such as sourcing more diverse applicants and improving applicant reactions, to algorithmic strategies, such as combining scores from multiple assessments to minimize group differences (Ployhart & Holtz, 2008).

The field of ML has also been working on algorithmic solutions to the diversity–validity dilemma. ML algorithms are not inherently biased, but ML models mirror patterns found in the data. If group differences exist in the criterion, a model is incentivized to accurately mirror these group differences (Zhang et al., 2018). While there is overlap in strategies that are used in the two fields, the amount of data that is typically used in ML models presents unique challenges. For example, the field of personnel selection has suggested examining single items for group differences using differential item functioning (Hough et al., 2001). Although this may make sense when the number of items used in a selection test is in the range of several dozen to several hundred, this method quickly becomes untenable when there are tens of thousands of predictors. Therefore, the focus of those using ML has been on automating and optimizing the processes for systematically reducing group differences.

The approaches used to reduce group differences in the IOP and ML communities can be classified into three primary categories: premodeling techniques, weighting and modeling techniques, and postmodeling techniques (called preprocessing, in-processing, and postprocessing in the ML literature; Bellamy et al., 2019). The following section synthesizes IOP and ML analytical techniques into this framework, highlighting some of the challenges that may be present in commonly used ML strategies because hiring decisions are often subjected to equal employment opportunity laws (e.g., Title VII of the Civil Rights Act of 1964 in the United States).

The framework presented includes strategies that are commonly used both in the design of a single test (e.g., the selection of items) and in the development of an assessment test battery (e.g., combining different selection measures such as a work sample, a cognitive ability test, and a conscientiousness test). This is intentional, as the same strategy can be employed in both use cases. In addition, while the framework organizes the strategies into three categories, techniques from multiple categories can be used in tandem. For example, premodeling techniques (i.e., transforming data prior to modeling) can be used in conjunction with modeling techniques when developing an assessment.

Although the framework discusses techniques in a predictor context, some of the techniques can also be applied to criterion measures (Hattrup et al., 1997). For example, techniques can be applied to reduce group differences in a performance metric, which can then be used in a ML model, resulting in a prediction model with smaller group differences.

Premodeling Techniques

Premodeling techniques are analytical strategies that transform the predictor or criterion data. Premodeling techniques can be further broken down into group-agnostic and group-specific approaches. Group-agnostic premodeling algorithms apply an identical transformation to all predictors, regardless of group membership. In contrast, group-specific premodeling techniques apply different transformations for different groups. That is, the transformation of a candidate's predictors is dependent on that candidate's specific demographics.

Group-Agnostic. Predictor removal is an example of group-agnostic premodeling. For instance, if the goal is to reduce group differences between males and females, then removing predictors that are strongly related to gender may decrease the algorithm's ability to differentiate males and females, resulting in decreased group differences. In this example, the same gender-related predictors are removed for both males and females; hence, the premodeling step is group-agnostic. At the test battery level, a single test can be removed from the test battery and another test, which demonstrates lower levels of group differences, can be used instead (Ployhart & Holtz, 2008). At the single test level, test items can be removed, resulting in decreased levels of group differences.

While the actual removal of predictors is straightforward, the method of identifying which predictors to remove is more complicated. Group differences and differential prediction or predictive validity can be used to identify predictors to remove, with varying degrees of success (Berry, 2015; Drasgow, 1987; Hough et al., 2001). The selection of which predictors to drop is then based on the diversity–validity prioritization goals and judgment of the test developer. This is a challenging, time-consuming, and often subjective task, forcing the individual(s) creating the selection test battery or single test to evaluate the predictors one at a time. This process can become even more complicated when trying to decrease group differences across multiple demographic categories, such as gender and ethnicity (Ployhart & Holtz, 2008).

While IOP traditionally relies on expert judgment to select predictors, this approach rapidly becomes infeasible as the number of predictors increases. Given this, assessment technology companies have focused on automating and optimizing the process for systematically identifying and removing group differences, although the specific techniques that they use are often unknown (Raghavan et al., 2020).

Group-Specific. Group-specific data transformations systematically manipulate the predictor or criterion data so that the groups become more similar to one another. These transformations are often supplemented with additional mathematical constraints that try to minimize the magnitude of the transformation (e.g., Calmon et al., 2017; Feldman et al., 2015; Zemel et al., 2013). For example, Feldman et al. (2015) imposed a constraint to maintain the in-group ranking of applicants. A simple example of a group-specific data transformation is applying a z-score transformation for males and applying a separate z-score transformation for females for each predictor. These two transformations would result in the two groups having predictors with the same mean and standard deviation. Group-specific transformed data would then be used in a model that optimizes predictive validity. So far, most preprocessing techniques from the ML community are group-specific (e.g., Calmon et al., 2017; Feldman et al., 2015; Zemel et al., 2013).

While group-specific premodeling strategies can reduce group differences while maintaining validity, these strategies explicitly treat each group differently. Based solely on group membership, an applicant's score will be increased or decreased. Group-specific premodeling techniques, therefore, constitute disparate treatment in that the groups are not subjected to the same standard and should not be used to select applicants in the United States (Civil Rights Act of 1964). While group-specific premodeling techniques are not relevant in a personnel selection context, we believe that it is important to discuss these techniques as the field of ML may be unaware of their legal implications.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Weighting and Modeling Techniques

Traditionally, the field of IOP has also attempted to solve the diversity–validity dilemma by adding predictors that demonstrate lower levels of group differences and adjusting the weighting of predictors in a selection test battery (Bobko et al., 1999; Hough et al., 2001; Potosky et al., 2005; Sackett & Ellingson, 1997; Sackett & Roth, 1996; Schmitt et al., 1997). The weighting of the predictors can then be determined using statistical (e.g., regression-based techniques) or rational (e.g., weights derived from a job analysis) approaches. This strategy has been shown to be effective, although perhaps not as effective as commonly believed (Potosky et al., 2005; Sackett & Ellingson, 1997).

Using regression-based techniques to determine predictor weights maximizes the predictive validity of the model but does not model the specific trade-off that occurs when trying to optimize both diversity and validity. To model this trade-off, De Corte et al. (2007) presented pareto-optimization (see also De Corte et al., 2010, 2011; Rupp et al., 2020). Pareto-optimization explicitly models the trade-off between diversity and validity. The organization can then determine the diversity–validity trade-off they desire by examining the change in magnitude of criterion-related validity at different specified levels of adverse impact.

An alternative strategy for explicitly balancing diversity and validity was proposed by De Corte (1999) in the form of a constrained nonlinear program. In this approach, a constraint is applied where the model maximizes predictive validity without violating the four-fifths rule. As illustrated by De Corte (1999), using this technique can maximize the predictive validity of a model within this constraint.

Within the ML community, one technique is to modify the model objective (or cost) function. The model objective function is the metric that the model is trying to optimize. For example, in an ordinary least squares (OLS) regression, the model objective function is to minimize the sum of squared differences between the model prediction and the criterion. By changing the model objective function, one can better optimize "diversity"; however, much is dependent upon how the researcher operationalizes diversity (e.g., see Kamishima et al., 2012). Finally, recent advancements in adversarial deep learning, a specific type of modeling often used in the field of ML, have produced methods where a main neural network attempts to predict the criterion (e.g., job performance), while an adversarial neural network tries to predict group membership (Zhang et al., 2018). The overall objective is to maximize the main network's ability to predict the criterion while minimizing the adversarial network's ability to predict group membership, ultimately resulting in decreased group differences.

Postmodeling

Postmodeling involves the transformation of predicted scores to decrease group differences. One benefit of postmodeling techniques is that they do not require access to the predictors going into the model or the model itself. Only access to model predictions is needed. This is beneficial because an organization may only have access to final assessment scores. Like premodeling techniques, postmodeling techniques can be broken down into group-agnostic and group-specific techniques.

Group-Agnostic. Two commonly used group-agnostic postmodeling techniques include cut scores (Cascio et al., 1988; Hoffman & Thornton, 1997) and banding (Bobko et al., 2005; Cascio et al., 1995). Cut scores take the predicted scores and transform them into a binary variable (e.g., pass/fail) used for decision-making, while banding involves grouping test scores into ranges and treating scores within a particular range as equivalent when making personnel decisions. Both methods can be used to reduce group differences. While cut scores are commonly used in multiple-stage selection systems, banding is more controversial (see Campion et al., 2001; Schmidt, 1995; Schmidt & Hunter, 1995).

Group-Specific. Group-specific postmodeling techniques manipulate test scores based on group membership. Within-group percentile referral, where the percentile of each applicant within each group is calculated, is an example of group-specific postmodeling (Cascio et al., 1995) along with within-group norming (Sackett & Wilk, 1994). While beyond the scope of this research, group-specific postmodeling techniques also allow for the targeting of alternative fairness metrics not commonly used in IOP (e.g., Hardt et al., 2016; Pleiss et al., 2017). Like group-specific premodeling techniques, group-specific postmodeling also explicitly treats each group differently, constituting disparate treatment, and should, therefore, not be used in personnel selection.

New Machine-Learning Approaches to Addressing the Diversity–Validity Dilemma

There are two primary challenges in addressing the diversity–validity dilemma using current IOP techniques. The first challenge is in optimizing diversity for multiple protected demographic categories (e.g., gender and ethnicity). This is a crucial challenge to solve as organizations are required under equal opportunity employment laws to examine adverse impact across multiple protected demographic categories. The second challenge is the amount and type of data that can result from new assessment technologies. While the methods historically used in IOP may be effective with small data sets, different techniques are needed to effectively handle big data sets.

We demonstrate two techniques to optimize diversity for multiple demographic categories that can be used with big data sets. The first technique is a group-agnostic premodeling technique we are calling iterative predictor removal. Iterative predictor removal is a method that involves identifying and removing predictors that cause group differences while simultaneously accounting for their predictive validity. As previously mentioned, selectively removing predictors is a strategy that has been suggested in both the personnel selection and ML literature and is commonly cited as a strategy used by hiring technology companies (Raghavan et al., 2020). While the general concept of removing predictors is not novel, we are unaware of any previous research that has operationalized this technique in a big-data scenario where one is specifically looking to model the diversity–validity trade-off. As there may be tens of thousands of predictors in big data sets, it is necessary to develop techniques to remove predictors systematically. In addition, given the nature of big data, where numerous predictors are likely to demonstrate large group differences and low predictive validity, dropping these predictors may be especially effective. The second technique we developed is a new technique we are calling multipenalty optimization. This is a

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

modeling technique where the final model is determined by simultaneously optimizing both the validity and diversity of the model's predicted scores. The effectiveness of this technique was examined and compared to the iterative predictor removal strategy.

It is worth noting that both iterative predictor removal and multipenalty optimization require access to candidate demographics when building the model. Neither method, however, requires demographic information to be present when the model is used for decision-making. That is, group membership is not used as a predictor, which ensures that neither method constitutes disparate treatment.

Before either technique is used, it is first necessary for the researcher to examine the predictive validity of the initial model (i.e., the model before either technique is used). The initial model predictive validity sets the upper limit for the predictive validity once either technique is used. Therefore, if the initial model does not successfully predict the criterion, neither of these techniques will be viable.

Both iterative predictor removal and multipenalty optimization can be used with multiple types of ML models; however, their operationalization may change depending upon the specific type of ML model chosen. Therefore, we will first discuss the two techniques generally. In the Method and Results sections, we will then demonstrate how these techniques are operationalized using a ridge regression.

Iterative Predictor Removal

Iterative predictor removal is a technique that systematically removes predictors that cause the largest group differences relative to their overall predictive impact. This involves a series of iterations where a prespecified number of predictors are removed at each iteration. If the researcher is only interested in decreasing group differences for a single group pair (e.g., gender: male and female), then iterative predictor removal will focus only on removing features related to this group pair. If, however, the researcher is interested in multiple group pairs, it is important to consider that a predictor may demonstrate varying levels of group differences depending upon the specific group pair. For example, if White females have the highest predicted scores, the same predictor may yield higher predicted scores for Whites but lower predicted scores for females. Therefore, the utility of removing that predictor is mixed because removing it simultaneously decreases predicted score differences for one group (Whites) and increases predicted score differences for another group (females). It is, therefore, useful to measure the extent to which a single predictor impacts group differences across multiple group pairs when determining predictors to remove. Our strategy is to calculate group difference to predictive impact ratios for all demographic pairs and then combine them weighted by each demographic pair's predicted score Cohen's *d* (Cohen, 1988). The steps involved in iterative predictor removal are as follows:

- *Step 1: Train Model*. A model is trained on the original data, which includes all predictors. Predicted scores are then calculated.

- *Step 2: Identify Reference Groups and Calculate Reference-Comparison Group Pair Differences*. Group mean predicted scores are calculated. The highest scoring group in each demographic category is identified as the reference group (e.g., White for ethnicity, female for gender). Cohen's *d* values on predicted scores are calculated for every group pair that includes the reference group (e.g., Cohen's $d_{\text{Black vs. White}}$, Cohen's $d_{\text{Asian vs. White}}$, Cohen's $d_{\text{Latinx vs. White}}$, Cohen's $d_{\text{Male vs. Female}}$).

- *Step 3: Calculate Unique Contribution of Each Predictor to Group Differences*. The unique contribution of each predictor to predicted score differences for each group pair is calculated. The directionality of these contributions is preserved such that predictors that favor the lower scoring or comparison group (as identified in Step 2) are negative. For example, in an OLS regression model, the group means for every predictor are first calculated. Mean differences are then calculated for every reference-comparison group pair and these values are multiplied by the corresponding signed regression coefficient.

- *Step 4: Calculate Impact of Each Predictor to Model Performance*. Each predictor's predictive impact is obtained from the previously developed model (Step 1) and quantifies the unique contribution of each predictor to the current model's predictions. This value is nonnegative. Continuing the OLS regression example, we define this as the absolute value of the standardized regression coefficient.

- *Step 5: Calculate Group Difference-Predictive Impact Ratio*. The group difference-predictive impact ratio (f_{jk}) for each predictor and for each group pair is calculated by dividing each predictor's group differences contribution (the value obtained in Step 3) by its predictive impact (the value obtained in Step 4). This equation is:

$$f_{jk} = \frac{\text{Predictor Group Difference}}{\text{Predictor Predictive Impact}}, \tag{1}$$

  for the *j*th predictor and the *k*th group pair.

- *Step 6: Aggregate Difference-Predictive Impact Ratio Across Group Pairs*. For each predictor, we take a weighted sum of the group difference-predictive impact ratios, where each ratio is weighted by the absolute value of its corresponding Cohen's *d* (Step 2) which is defined as:

$$f_j = \sum_{k=1}^{D} |d_k| \times f_{jk}, \tag{2}$$

  for *D* total demographic pairs.

- *Step 7: Remove Predictors and Retrain Model on Reduced Data Set*. The predictors are then ranked based on their aggregate group difference-predictive impact ratio (f_j). Finally, the highest ranked predictors are removed from the data set. A new algorithm is then trained on the predictor-reduced data set (Step 1). This process (Steps 2–7) is repeated until the desired levels of diversity and validity are met. It is important to note that reference groups may change between iterations.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This process, as illustrated in Figure 1, can be used for any regressor or classifier model, provided suitable predictor group differences (Step 3) and predictive impact (Step 4) metrics can be obtained from the model. In the Methods section, we demonstrate this technique using a linear model (ridge regression). For a nonlinear model, we recommend using this technique with a global surrogate model (see Molnar, 2022, for more information).

Benefits. One benefit of iterative predictor removal is that, since it is a premodeling technique, it can be applied regardless of the ML algorithm used to build the final predictive model. An additional benefit is that this technique can use data that are missing either demographic or criterion data. While only cases with the criterion are required to train the model (Step 1), only cases with demographics (along with predictor data and model parameters) are needed to calculate the group difference-predictive impact ratio (Steps 2–7), such that cases with missing criterion data can still be used. This is especially beneficial in selection scenarios where applicant predictor and demographic data are often available on a broader sample prior to a selection decision.

Drawbacks. One drawback of iterative predictor removal is that it can delete predictors that are strongly related to the performance metric if they concurrently contribute to group differences. In addition, this technique offers no guarantee of removing the optimal predictors, as it is computationally infeasible to look at all possible combinations of predictors to drop. Another drawback is that there are two settings that govern this technique: the number of predictors

Figure 1

Example of the Iterative Predictor Removal Technique

	Predictor #	Predictor Group Differences	Predictive Impact	Group Difference-Predictive Impact Ratio	Predictor Removal
Predictors Removed: 0 **Predictive Validity (*r*):** 0.556 **Mean Group Difference (Cohen's *d*):** -0.343	64	1.277	-0.154	-0.121	Keep
	326	0.671	0.206	0.307	**Drop**
	525	5.162	1.911	0.370	**Drop**
	1199	70.231	16.745	0.238	Keep
	2881	47.967	11.225	0.234	Keep
	4033	36.978	5.322	0.144	Keep
	8449	0.355	0.020	0.057	Keep
	13099	0.217	-0.009	-0.040	Keep

	Predictor #	Predictor Group Differences	Predictive Impact	Group Difference-Predictive Impact Ratio	Predictor Removal
Predictors Removed: 100 **Predictive Validity (*r*):** 0.552 **Mean Group Difference (Cohen's *d*):** -0.313	64	0.446	-0.052	-0.118	Keep
	326	--------	--------	--------	--------
	525	--------	--------	--------	--------
	1199	77.648	16.858	0.217	**Drop**
	2881	54.597	11.272	0.206	**Drop**
	4033	39.900	5.387	0.135	Keep
	8449	0.379	0.020	0.053	Keep
	13099	0.229	-0.008	-0.036	Keep

	Predictor #	Predictor Group Differences	Predictive Impact	Group Difference-Predictive Impact Ratio	Predictor Removal
Predictors Removed: 500 **Predictive Validity (*r*):** 0.529 **Mean Group Difference (Cohen's *d*):** -0.237	64	1.779	0.182	0.102	**Drop**
	326	--------	--------	--------	--------
	525	--------	--------	--------	--------
	1199	--------	--------	--------	--------
	2881	--------	--------	--------	--------
	4033	50.069	5.334	0.107	**Drop**
	8449	0.463	0.019	0.041	Keep
	13099	0.269	-0.007	-0.025	Keep

Note. This figure provides selected predictor details to illustrate how the iterative predictor removal technique works. For predictor group difference calculations see Step 3; for predictive impact calculations see Step 4; for group difference-predictive impact ratio see Step 6. Predictor group differences and predictive impact are multiplied by 10,000 for clarity.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

dropped in each iteration as well as the total number of iterations. The optimal values for these settings depend on the data set, the amount of time it takes to train the models, and diversity–validity prioritization. Therefore, it is often necessary for the researcher to examine different settings to achieve their optimal solution. Removing 0.5% of total predictors at each iteration and conducting 100 iterations is usually a good starting point. With models that take longer to train, either due to the size of the data set or the type of model used, the researcher may consider increasing the number of predictors dropped at each iteration. If there is a very large decrease in group differences and predictive validity at each iteration, then the researcher may consider decreasing the percentage of predictors dropped at each iteration. The fine-tuning of the number of predictors to drop per iteration affects the performance of this technique, while the total number of predictors removed impacts of the diversity–validity trade-off. We recommend that researchers plot the group differences-predictive validity curve when exploring this technique. This will allow researchers to see the diversity–validity trade-off that is occurring when various numbers of predictors are removed. In addition, this plot allows the researcher to see the point of diminishing returns, where continuing to drop more predictors does not meaningfully change group differences while predictive validity continues to drop.

Other Considerations. The overall goal of iterative predictor removal is to create a final model that has maximally reduced group differences and minimally reduced predictive validity. In our technique, using the ratios of predictor group differences to predictor predictive validity mirrors this goal. We believe that this choice makes intuitive sense but it is also possible to use other heuristics to identify predictors to remove. For example, instead of the signed value, one could remove predictors using the absolute value of the group difference-predictive impact ratio. This would remove predictors regardless of if they increase or decrease the predicted scores of the lower scoring group. This heuristic may be appropriate in certain research scenarios; however, in our examination, we have found it to be less effective in reducing group differences.

As this technique involves dropping predictors that exhibit high group differences, this technique works best in big data sets where there are hundreds or thousands of predictors, justifying the predictor loss. This contrasts with more traditional assessments (e.g., cognitive ability test, personality tests, structured interview ratings), where the amount of data per individual is much smaller, resulting in a greater potential for decreased predictive validity when a single predictor is dropped.

Multipenalty Optimization

Data modeling techniques in both traditional IOP analyses and ML can be classified into two types: supervised and unsupervised modeling. Supervised refers to models that are trained to predict a specific criterion (e.g., OLS regression), while unsupervised refers to models that are used to analyze data without a specific criterion (e.g., cluster analysis). Traditionally, IOP has used supervised models, which have an objective of maximizing model accuracy (or equivalently minimizing prediction error). Mathematically, model accuracy is quantified using a data fit cost function and we denote it as C_{data}. For example, the data fit cost function of an OLS regression model is the sum of squared errors between the model's predicted scores and criterion scores. OLS regression models only optimize model accuracy which makes them prone to overfitting. To combat this, the field of ML has developed strategies to improve the generalizability of model predictions on new observations. Most ML models accomplish this by adding a second cost function which penalizes large magnitude model coefficients. This second cost function is referred to as a regularization cost function, and we denote it as C_{reg}.[1] Researchers can adjust the relative strength of this penalization by changing the regularization weight hyperparameter[2] (α) where a larger regularization weight will yield model coefficients closer to zero. Therefore, the total cost function trades-off minimizing model prediction error (C_{data}) with minimizing the degree of overfitting (C_{reg}).

The multipenalty optimization technique adds an additional cost function that penalizes group differences (e.g., male/female). To accomplish this, a group difference penalty (C_{diff}) is added to the original cost function, which includes data fit (C_{data}) and regularization (C_{reg}) terms. The total cost function (C_{total}) is written as:

$$C_{total} = C_{data} + \alpha C_{reg} + \beta C_{diff}. \tag{3}$$

The goal of model training is to solve for model parameters that minimize this cost function (C_{total}). Here, α is the original regularization weight and $\beta \geq 0$ is a tunable parameter representing the relative strength of the group difference penalty. Higher values of β place a larger penalty on models that exhibit group differences, whereas the special case of $\beta = 0$ is equivalent to the original, nonmultipenalty optimized model. Multiple groups can be included in the group difference penalty (e.g., male, female, Black, and White), simultaneously reducing group differences for multiple protected classes.

We define the individual group difference penalty as the squared difference between a single group (e.g., male) mean model prediction and the overall mean model prediction. The total group difference cost function (C_{diff}) is the sum of all individual penalties:

$$C_{diff} = \sum_{k=1}^{D} \left(\frac{1}{n_k} \sum_{i \in A_k} \hat{y}_i - \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i \right)^2, \tag{4}$$

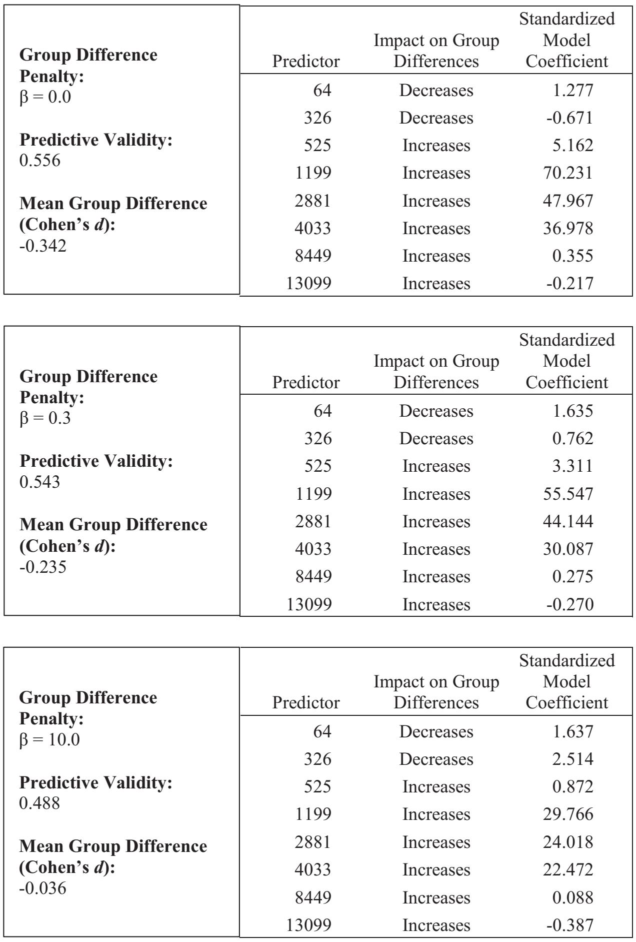
where D is the total number of groups across multiple demographic categories (e.g., male, female, White, Black, Latinx, and Asian), n_k is the number of members in the kth group, i is the ith case, A_k is the membership in the kth group, $\hat{y}$ is the predicted score, and n is the total sample size. As such, models that demonstrate a large difference in predicted mean group scores in relation to the overall mean predicted score are penalized (see Figure 2 for an illustration of this technique and the Appendix for a discussion on how to minimize the overall cost function).

[1] Examples of models that apply a regularization cost function include ridge regression (Hoerl & Kennard, 1970), least absolute shrinkage and selection operator regression (lasso; Tibshirani, 1996), elastic net (Zou & Hastie, 2005), neural networks (McCulloch & Pitts, 1943), support vector machines (Cortes & Vapnik, 1995), and l1- and l2-regularized logistic regression.

[2] Hyperparameters are model settings that govern how a model learns for a given problem and data set. They are distinct from a model's parameters (e.g., regression coefficients) that are solved for during model training. Hyperparameters are specified by the researcher building the model prior to model training. Multiple hyperparameter ranges are often examined to maximize out-of-sample accuracy but can include other goals (model speed, interpretability, etc.).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Figure 2
Example of the Multipenalty Optimization Technique



Group Difference Penalty:
β = 0.0

Predictive Validity:
0.556

Mean Group Difference (Cohen's *d*):
-0.342

Predictor	Impact on Group Differences	Standardized Model Coefficient
64	Decreases	1.277
326	Decreases	-0.671
525	Increases	5.162
1199	Increases	70.231
2881	Increases	47.967
4033	Increases	36.978
8449	Increases	0.355
13099	Increases	-0.217

Group Difference Penalty:
β = 0.3

Predictive Validity:
0.543

Mean Group Difference (Cohen's *d*):
-0.235

Predictor	Impact on Group Differences	Standardized Model Coefficient
64	Decreases	1.635
326	Decreases	0.762
525	Increases	3.311
1199	Increases	55.547
2881	Increases	44.144
4033	Increases	30.087
8449	Increases	0.275
13099	Increases	-0.270

Group Difference Penalty:
β = 10.0

Predictive Validity:
0.488

Mean Group Difference (Cohen's *d*):
-0.036

Predictor	Impact on Group Differences	Standardized Model Coefficient
64	Decreases	1.637
326	Decreases	2.514
525	Increases	0.872
1199	Increases	29.766
2881	Increases	24.018
4033	Increases	22.472
8449	Increases	0.088
13099	Increases	-0.387

Note. Impact on group differences indicates whether a higher model coefficient increases or decreases the group difference penalty (C_{diff}). Using predictor 326 as an example, as a larger group difference penalty is applied, the model coefficient increases due to a higher model coefficient value decreasing group differences. Using predictor 1,199 as another example, as a larger group difference penalty is applied, the model coefficient decreases due to a higher model coefficient value increasing group differences.

Benefits. The main benefit of this method is that it explicitly optimizes the diversity–validity trade-off. That is, the model objective is to maximize both diversity and validity where the trade-off between these two competing objectives is governed by the size of β. Similar to iterative predictor removal, multipenalty optimization effectively handles cases with missing criteria or demographic data. Cases with missing criterion are excluded from the data fit cost function and cases with missing demographic labels are excluded from the group difference penalty.

Drawbacks. The primary drawback of multipenalty optimization is that the β hyperparameter will require fine-tuning as the optimal solution is dependent upon the underlying data along with the prioritization of predictive validity and decreased group differences. An initial examination of β hyperparameters evenly sampled from the log-normal scale (0.1, 0.3, 1.0, 3.0, 10.0, 30.0, 100.0) is an effective starting point. Further examination can then be based on the initial results. Like iterative predictor removal, we recommend that researchers plot the group differences-predictive validity curve when exploring different values for β as this will allow researchers to see the diversity–validity trade-off that is occurring. Assuming a high enough level of β is reached, it will also allow researchers to see the point of diminishing returns, such that continuing to increase β does not meaningfully change group differences or predictive validity.

Examination of Iterative Predictor Removal and Multipenalty Optimization

The goal of the current research was to examine the effectiveness of iterative predictor removal and multipenalty optimization in decreasing group differences while maintaining predictive validity in a big data selection scenario. Therefore, we examined the effectiveness of these two techniques in a high-stakes field study where applicants completed asynchronous video interviews to apply for an educational instructor position.

Method

Participants and Procedure

Using a platform provided by a recruiting technology company, we collected the asynchronous video interviews from 32,989 applicants applying for primary and secondary educational instructor positions at a single organization in the United States. Of these 32,989 applicants, performance metrics were available for 13,518 individuals ultimately hired into the position. The video interviews were used in the hiring process and were evaluated by human raters; no algorithm-based scoring was used in the hiring process.

The recruiting technology company's proprietary demographic prediction algorithms were used to predict the applicant's gender (male and female), race/ethnicity (White, Black, Latinx, and Asian), and age (under 40 years and 40 years or older). To predict this, the algorithm used an applicant's name, which was obtained from their written application, and video thumbnails, which were obtained from their video interview. Importantly, the data used to predict demographic group membership are distinct from the interview data used in the modeling. The reported classification accuracy of the algorithm was 99% for gender, 87% for ethnicity, and 91% for age. Gender, ethnicity, or age could not be predicted for 6.9% ($n = 2{,}274$) of the applicants, often due to poor video thumbnail quality. Of all the applicants, 54.1% (n = 17,843) were White, 14.0% (n = 4,612) were Latinx, 13.3% (n = 4,375) were Black, 11.8% (n = 3,885) were Asian, 13.2% (n = 4,351) were age 40 years or older, and 52.7% (n = 17,391) were female (see Table 1).

Measures

All applicants completed an asynchronous video interview that consisted of four situational and behavioral standardized video interview questions which were developed by subject matter experts to measure teaching proficiency. The applicants' raw audio responses were transcribed to text using Rev AI's proprietary speech-to-text engine (Rev AI, 2021). Along with a binary bag-of-words approach where word usage was used as predictors (i.e., whether the applicants use a particular word at least once), a natural language processing procedure called Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach (RoBERTa; Liu et al.,

Table 1
Demographic Representation

Demographic groups	All applicants		Applicants with performance metrics		Applicants without performance metrics	
	n	%	*n*	%	*n*	%
Ethnicity						
White	17,843	54.1	6,878	50.9	10,965	56.3
Latinx	4,612	14.0	1,978	14.6	2,634	13.5
Black	4,375	13.3	2,093	15.5	2,282	11.7
Asian	3,885	11.8	1,582	11.7	2,303	11.8
Unknown	2,274	6.9	987	7.3	1,287	6.6
Age						
Under 40 years	26,365	79.9	11,038	81.7	15,327	78.7
40 years or older	4,351	13.2	1,494	11.1	2,857	14.7
Unknown	2,273	6.9	986	7.3	1,287	6.6
Gender						
Female	17,391	52.7	6,539	48.4	10,852	55.7
Male	13,325	40.4	5,993	44.3	7,332	37.7
Unknown	2,273	6.9	986	7.3	1,287	6.6
Overall	32,989	100.0	13,518	100.0	19,471	100.0

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

2019) was applied on the raw transcription data to instill meaning. RoBERTa considers the context of a word by using the preceding and anteceding words. This is important as the same word can have multiple meanings; for example, the phrase "the bat flew through the air" conveys a different message if the "bat" is a nocturnal mammal or a baseball bat. The RoBERTa procedure was also fine-tuned on other asynchronous video interview text to improve its predictive power. The average interview response consisted of 1,686 words (SD = 744). In total, 16,285 unique predictors (15,517 binary bag-of-words predictors and 768 RoBERTa predictors) were used in the modeling. The binary bag-of-words predictors were not standardized and the RoBERTa predictors were standardized (SD = 0.1). No video-based (e.g., facial expression) or audio feature-based (e.g., rate of speech or intonation) data were used.

Educational instructor performance was measured using a proprietary single-item measure of perceived instructor quality (M = 2.240, SD = 0.862). When a single instructor had multiple ratings, the average score was calculated by the organization. In the performance metric, small group differences were found for Latinx versus White (Cohen's $d = -0.148$), Asian versus White (Cohen's $d = -0.181$), 40 years or older versus under 40 years (Cohen's $d = -0.333$), and males versus females (Cohen's $d = -0.062$). Medium group differences were found for Black versus White (Cohen's $d = -0.444$; see Table 2).

Modeling

Model Building

The full data set (n = 32,989) was randomly split into training (70%, n = 23,091) and test (30%, n = 9,898) data sets. We examined the predictive validity of three regression-based models (ridge, lasso, and elastic net) using three predictor data sets: only bag-of-words predictors, only RoBERTa predictors, and both bag-of-words and RoBERTa predictors. The various data sets and models, along with multiple model hyperparameters, were evaluated for predictive validity using five-fold cross-validation[3] on the training set. A ridge regression with a regularization weight of α = 3,000 trained on both bag-of-words and RoBERTa predictors had the highest training data set out-of-sample predictive validity (see Tables 3 and 4). This model and data set were used as the baseline model for both the iterative predictor removal and multipenalty optimization techniques. While five-fold cross-validation was used for initial model and data set selection, all subsequent models were built using the full training set and results are shown applied to the test data set.

Iterative Predictor Removal

We followed the steps involved in iterative predictor removal as outlined in the Introduction. Decisions specifically pertaining to the current application of iterative predictor removal are as follows:

- *Step 1: Train Model*. As previously discussed, a ridge regression with a regularization weight of α = 3,000 was used on the complete training data set (before any predictors were removed).
- *Step 2: Identify Reference Groups and Calculate Reference-Comparison Group Pair Differences*. We identified reference groups and calculated group pair differences for ethnicity (White, Black, Latinx, and Asian), gender (male and female), and age (under 40 years and 40 years or older).
- *Step 3: Calculate Unique Contribution of Each Predictor to Group Differences*. The contribution of each predictor to the predicted score differences was defined as the difference in mean predictor values between the high scoring reference group ($\bar{x}_{jk}^{\text{high}}$) and low scoring comparison group ($\bar{x}_{jk}^{\text{low}}$) multiplied by the corresponding unstandardized model coefficient (w_j) for the jth predictor for the kth group pair.

[3] Cross-validation is a technique commonly used in the field of ML and involves splitting the sample into a training sample and a holdout sample. For five-fold cross-validation, the total sample is split into 80% training sample and 20% holdout sample five times. This technique was used as it demonstrates how well a model performs on data not used to train the model (see de Rooij & Weeda, 2020).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 2
Performance Means, Standard Deviations, and Demographic Differences

Demographic groups	*n*	*M*	*SD*	*df*	*t*	Cohen's *d*	Cohen's *d* 95% CI
Ethnicity							
White	6,878	2.351	0.826	—	—	—	—
Latinx	1,978	2.227	0.868	8,854	−5.819*	−0.148	[−0.199, −0.098]
Black	2,093	1.978	0.887	8,969	−17.771*	−0.444	[−0.493, −0.394]
Asian	1,582	2.200	0.866	8,458	−6.482*	−0.181	[−0.235, −0.126]
Age							
Under 40 years	11,037	2.284	0.848	—	—	—	—
40 years or older	1,494	2.000	0.900	12,530	−12.064*	−0.333	[−0.387, −0.278]
Gender							
Female	6,538	2.276	0.856	—	—	—	—
Male	5,993	2.222	0.862	12,530	−3.471*	−0.062	[−0.097, −0.027]

Note. CI = confidence interval. For ethnicity, Latinx, Black, and Asian were compared to White as they had the highest criterion scores.
* $p < .001$.

- *Step 4: Calculate Impact of Each Predictor to Model Performance*. The predictor unique contribution or predictive impact was defined as the absolute value of the predictor-standardized regression coefficient ($|w_j \times SD_{x_j}|$).

- *Step 5: Calculate Group Difference-Predictive Impact Ratio*. The group difference-predictive impact ratio for each predictor for each group pair was then calculated by dividing each predictor's group differences contribution in predicted score by its predictive impact:

$$f_{jk} = \frac{\text{Predictor Group Differences}}{\text{Predictor Predictive Impact}} = \frac{w_j(\bar{x}_{jk}^{\text{high}} - \bar{x}_{jk}^{\text{low}})}{|w_j \times SD_{x_j}|}$$
$$= sgn(w_j)\frac{\bar{x}_{jk}^{\text{high}} - \bar{x}_{jk}^{\text{low}}}{SD_{x_j}}. \tag{5}$$

- *Step 6: Aggregate Group Difference-Predictive Impact Ratio Across Group Pairs*. See Equation 2.

- *Step 7: Remove Predictors and Retrain Model on Reduced Data Set*. For each iteration, we removed the 100 highest ranked predictors and completed 40 total iterations. We stopped at 40 iterations as additional iterations no longer reduced group differences.

Multipenalty Optimization

When the group differences penalty (Equation 4) was added to a ridge regression, the total cost function was:

$$C_{\text{total}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha\sum_{j=1}^{p} w_j^2$$
$$+ \beta\sum_{k=1}^{D}\left(\frac{1}{n_k}\sum_{i\in A_k}\hat{y}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i\right)^2, \tag{6}$$

where *y* is the criterion, *p* is the total number of predictors, and *D* is the total number of demographic groups. The first two terms

Table 3
Diversity–Validity Trade-Off for Machine Learning Models: Predictive Validity and Ethnicity Group Differences

			Latinx–White		Black–White		Asian–White	
ML model and data set	*r*	*r* 95% CI	Cohen's *d*	Cohen's *d* 95% CI	Cohen's *d*	Cohen's *d* 95% CI	Cohen's *d*	Cohen's *d* 95% CI
Ridge regression: Bag-of-words	0.547	[0.526, 0.569]	−0.215	[−0.275, −0.155]	−0.602	[−0.664, −0.541]	−0.291	[−0.353, −0.228]
Ridge regression: RoBERTa	0.513	[0.489, 0.535]	−0.262	[−0.322, −0.202]	−0.621	[−0.682, −0.559]	−0.286	[−0.349, −0.224]
Ridge regression: Bag-of-words + RoBERTa	0.556	[0.535, 0.577]	−0.226	[−0.287, −0.166]	−0.617	[−0.678, −0.555]	−0.289	[−0.352, −0.226]
Lasso: Bag-of-words	0.540	[0.518, 0.562]	−0.219	[−0.279, −0.159]	−0.590	[−0.651, −0.529]	−0.280	[−0.343, −0.218]
Lasso: RoBERTa	0.511	[0.488, 0.534]	−0.251	[−0.311, −0.191]	−0.612	[−0.673, −0.550]	−0.276	[−0.338, −0.213]
Lasso: Bag-of-words + RoBERTa	0.547	[0.526, 0.569]	−0.237	[−0.297, −0.176]	−0.604	[−0.665, −0.543]	−0.287	[−0.350, −0.224]
Elastic net: Bag-of-words	0.520	[0.497, 0.542]	−0.194	[−0.255, −0.134]	−0.551	[−0.612, −0.490]	−0.265	[−0.328, −0.202]
Elastic net: RoBERTa	0.499	[0.476, 0.052]	−0.242	[−0.303, −0.182]	−0.569	[−0.631, −0.508]	−0.235	[−0.298, −0.172]
Elastic net: Bag-of-words + RoBERTa	0.529	[0.506, 0.550]	−0.205	[−0.265, −0.145]	−0.563	[−0.625, −0.502]	−0.263	[−0.326, −0.201]

Note. ML = machine learning; CI = confidence interval; RoBERTa = Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach. Bag-of-words data set only includes bag-of-words predictors, RoBERTa data set only includes RoBERTa predictors, and bag-of-words + RoBERTa includes both bag-of-words and RoBERTa predictors.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 4

Diversity–Validity Trade-Off for Machine Learning Models: Age and Gender Group Differences

	40 years or older–under 40 years		Male–female	
ML model and data set	Cohen's *d*	Cohen's *d* 95% CI	Cohen's *d*	Cohen's *d* 95% CI
Ridge regression: Bag-of-words	−0.517	[−0.575, −0.458]	−0.057	[−0.098, −0.016]
Ridge regression: RoBERTa	−0.479	[−0.537, −0.420]	−0.100	[−0.141, −0.058]
Ridge regression: Bag-of-words + RoBERTa	−0.516	[−0.574, −0.457]	−0.065	[−0.107, −0.024]
Lasso: Bag-of-words	−0.481	[−0.540, −0.423]	−0.074	[−0.115, −0.033]
Lasso: RoBERTa	−0.484	[−0.542, −0.425]	−0.096	[−0.137, −0.055]
Lasso: Bag-of-words + RoBERTa	−0.479	[−0.537, −0.421]	−0.066	[−0.107, −0.025]
Elastic net: Bag-of-words	−0.527	[−0.585, −0.468]	0.001	[−0.040, 0.042]
Elastic net: RoBERTa	−0.493	[−0.552, −0.435]	−0.055	[−0.097, −0.014]
Elastic net: Bag-of-words + RoBERTa	−0.529	[−0.587, −0.470]	−0.008	[−0.049, 0.034]

Note. ML = machine learning; CI = confidence interval; RoBERTa = Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach. Bag-of-words data set only includes bag-of-words predictors, RoBERTa data set only includes RoBERTa predictors, and bag-of-words + RoBERTa includes both bag-of-words and RoBERTa predictors.

comprise the original ridge regression cost functions, while the third term is the group differences penalty (see Equation 3). We derive the closed-form minimizing solution to the total cost function (C_{total}) in the Appendix, but one key takeaway is that there is a unique solution (or single best fitting model). Practically, this makes the problem computationally straightforward.

As previously discussed, the regularization weight was $\alpha = 3{,}000$. We examined varying levels of β from 0 to 100. We penalized group differences for gender, ethnicity, and age simultaneously.

Model Performance

Model performance (r) and predicted score group differences (Cohen's d) were obtained for each iteration in the iterative predictor technique and at each level of β for the multipenalty optimization technique. In addition, both model performance and predicted score group differences were plotted (see Figure 3).

While the goal of iterative predictor removal and multipenalty optimization is to decrease group differences, from a measurement bias perspective, it is also critical to examine whether either of these methods leads to systematic differences in regression lines for each demographic group. Accordingly, we examined the impact of these two techniques on differential prediction (see Aguinis et al., 2010; Aguinis & Smith, 2007; Campbell, 1996; Cleary, 1968; Hough et al., 2001; Meade & Tonidandel, 2010). The steps outlined in Aguinis et al. (2010) were followed, where overall differential prediction, differences in model intercepts, and differences in model slopes were examined using moderated multiple regression. We also calculated standardized effect size as recommended by Dahlke and Sackett (2018; see also Nye & Sackett, 2017).

Transparency and Openness

We describe our data and analyses in the study and adhere to the *Journal of Applied Psychology* methodological checklist. Interview questions, interview transcripts, and the performance measure questions are not available due to their proprietary nature. Deidentified predictors, applicant demographics, the criterion, and code for the present research are available upon request. The study design and analyses were not preregistered. All research was conducted outside of academic institutions and used archival data; therefore, Institutional Review Board approval and informed consent were not obtained. All research complied with the American Psychological Association's Ethical Code of Conduct.

The majority of analyses were conducted using Python 3.8.10 (Van Rossum & Drake, 2009) and the following packages: Scikit-learn 0.21.1 (Pedregosa et al., 2011), pandas 1.2.3 (Reback et al., 2021), NumPy 1.19.5 (Harris et al., 2020), and sciPy 1.6.0 (Virtanen et al., 2020). Differential prediction standardized effect sizes were calculated in *R* 3.6.0 (R Core Team, 2022) using the psychmeta package (Dahlke & Wiernik, 2019). To facilitate usage of these two techniques, we provide Python code in the online Supplemental Material for the iterative predictor removal and multipenalty optimization techniques. This includes code for generating a simulated data set and code for conducting the iterative predictor removal and multipenalty optimization techniques on this simulated data set.

Results

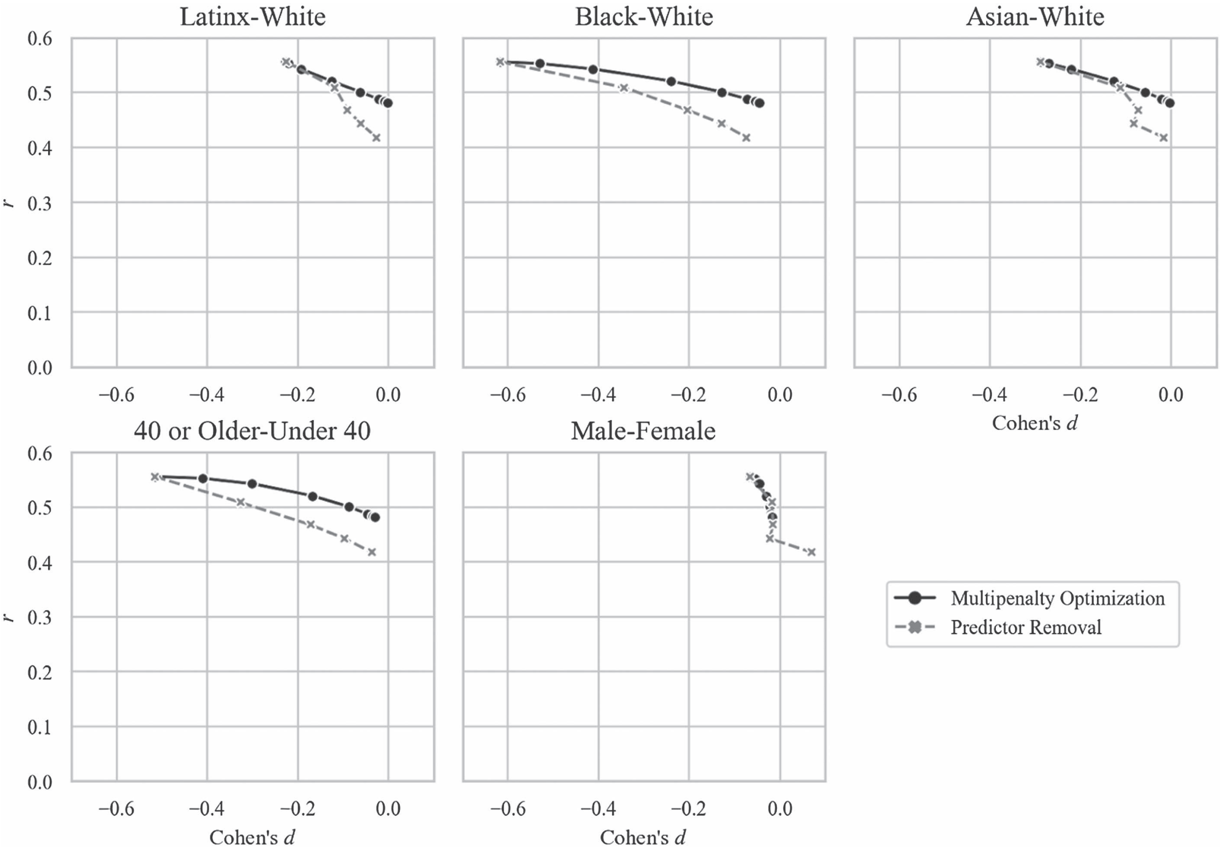
The effectiveness of iterative predictor removal and multipenalty optimization techniques on decreasing group differences while maintaining validity is illustrated in Figure 3 with requisite data presented in Tables 5–8.[4] When 4,000 predictors (25% of all predictors) were removed using iterative predictor removal, group differences decreased for all demographic comparisons. For example, Cohen's d for Black/White differences decreased from a starting value of −0.617 to −0.073. This drop in group differences came with a moderate cost to predictive validity ($r_0 = 0.556$, $r_{4,000} = 0.418$, $\Delta r = 0.138$, $z = 12.785$, $p < .001$). Multipenalty optimization also effectively minimized group differences but had less reduction in predictive validity. For example, with a β weight of 100, Cohen's d for Black/White differences decreased from a starting value of −0.617 to −0.046 ($r_0 = 0.556$, $r_{100} = 0.482$, $\Delta r = 0.074$, $z = 7.135$, $p < .001$). Multipenalty optimization was more effective at reducing group differences while maintaining predictive validity—iterative predictor removal yielded an 88% reduction of Black/White group differences with a corresponding r^2 reduction of 43%, whereas

[4] See online Supplemental Material, for predicted model score descriptive statistics.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Figure 3

Diversity–Validity Trade-Off for Both Iterative Predictor Removal and Multipenalty Optimization



Note. RoBERTa = Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach. Both iterative predictor removal and multipenalty optimization minimize group differences for ethnicity, age, and gender, so the curves in each subfigure correspond to the exact same set of models. Data set includes both bag-of-words and RoBERTa predictors.

multipenalty optimization yielded a 93% reduction of Black/White group differences with a corresponding r^2 reduction of 25%.

As evident in Figure 3, as the number of predictors removed increased and as the β weight increased, there were diminishing decreases in group differences. At a certain point, removing more predictors or increasing the β weight would only minimally decrease group differences. Using Black/White differences as an example, removing the first 1,000 predictors (removing predictors 1–1,000) resulted in Cohen's *d* decreasing from −0.617 to −0.343 (ΔCohen's $d = 0.274$, $r_0 = 0.556$, $r_{1,000} = 0.509$, $\Delta r = 0.047$, $z = 4.617$, $p < .001$) while removing 1,000 predictors after 3,000 predictors were previously removed (removing predictors 3,000–4,000) only resulted in

Table 5

Diversity–Validity Trade-Off for Iterative Predictor Removal Models: Predictive Validity and Ethnicity Group Differences

			Latinx–White		Black–White		Asian–White	
Predictors removed	*r*	*r* 95% CI	Cohen's *d*	Cohen's *d* 95% CI	Cohen's *d*	Cohen's *d* 95% CI	Cohen's *d*	Cohen's *d* 95% CI
0	0.556	[0.535, 0.577]	−0.226	[−0.287, −0.166]	−0.617	[−0.678, −0.555]	−0.289	[−0.352, −0.226]
1,000	0.509	[0.486, 0.531]	−0.188	[−0.178, −0.058]	−0.343	[−0.404, −0.283]	−0.112	[−0.175, −0.050]
2,000	0.468	[0.443, 0.491]	−0.090	[−0.150, −0.030]	−0.203	[−0.263, −0.142]	−0.073	[−0.135, −0.010]
3,000	0.443	[0.418, 0.467]	−0.061	[−0.121, −0.001]	−0.127	[−0.188, −0.067]	−0.082	[−0.145, −0.020]
4,000	0.418	[0.392, 0.443]	−0.026	[−0.068, 0.034]	−0.073	[−0.133, −0.012]	−0.016	[−0.079, 0.046]

Note. RoBERTa = Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach; CI = confidence interval. Values in this table correspond to points on the dashed line of Figure 3. Data set includes both bag-of-words and RoBERTa predictors.

Table 6

Diversity–Validity Trade-Off for Iterative Predictor Removal Models: Age and Gender Group Differences

	40 years or older–under 40 years		Male–female	
Predictors removed	Cohen's *d*	Cohen's *d* 95% CI	Cohen's *d*	Cohen's *d* 95% CI
0	−0.516	[−0.574, −0.457]	−0.065	[−0.107, −0.024]
1,000	−0.327	[−0.386, −0.269]	−0.018	[−0.059, 0.023]
2,000	−0.172	[−0.229, −0.114]	−0.015	[−0.056, 0.026]
3,000	−0.097	[−0.155, −0.039]	−0.023	[−0.064, 0.018]
4,000	−0.037	[−0.095, 0.021]	0.070	[0.029, 0.111]

Note. RoBERTa = Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach; CI = confidence interval. Values in this table correspond to points on the dashed line of Figure 3. Data set includes both bag-of-words and RoBERTa predictors.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Cohen's *d* decreasing from −0.127 to −0.073 (ΔCohen's $d = 0.054$, $r_{3,000} = 0.443$, $r_{4,000} = 0.418$, $\Delta r = 0.025$, $z = 2.159$, $p = .031$). Similarly, for multipenalty optimization, increasing the β weight from 0 to 0.3 resulted in Cohen's *d* decreasing from −0.617 to −0.412 (ΔCohen's $d = 0.205$, $r_0 = 0.556$, $r_{0.3} = 0.543$, $\Delta r = 0.013$, $z = 1.310$, $p = .190$), while increasing the β weight from 30 to 100 only resulted in Cohen's *d* decreasing from −0.053 to −0.046 (ΔCohen's $d = 0.007$, $r_{30} = 0.484$, $r_{100} = 0.482$, $\Delta r = 0.002$, $z = 0.183$, $p = .854$). While more predictors could have been removed or higher β weights could have been examined, this would have resulted in minimal changes to group differences.

In the examination of differential prediction, the original model (before either iterative predictor removal or multipenalty optimization was applied) did not exhibit meaningful overall differential prediction for ethnicity, age, or gender ($d_{\text{Mod\_Signed}} < 0.14$; see Tables 9–14). As more predictors were removed or as a higher group differences penalty was applied, the amount of overall differential prediction increased. This increase was largely driven by an increase in intercept differences as opposed to slope differences, such that the models overpredicted for the groups with lower average performance scores and underpredicted for the groups with higher average performance scores. This increase in overall differential prediction was expected as both techniques decreased group's predicted score differences while criterion group differences remained constant, resulting in differential prediction driven by intercept differences (see Meade & Tonidandel, 2010, for further reading on the potential sources that contribute to differential prediction).

Discussion

Rapid advances in technology—the improved ability to collect and process large amount of complex data and ML algorithms—are opening up new ways of addressing the diversity–validity dilemma. While previous research has demonstrated the effectiveness of a wide variety of strategies (e.g., Ployhart & Holtz, 2008), the present research outlines two ways of addressing this dilemma using ML tools. While both iterative predictor removal and multipenalty optimization effectively removed group differences, multipenalty optimization more effectively controlled for group differences while maintaining predictive validity. Therefore, researchers with big data should consider using multipenalty optimization as opposed to commonly used predictor removal techniques.

Methodological Contributions

Both iterative predictor removal and multipenalty optimization techniques exhibit strengths that make their application in a selection

Table 7

Diversity–Validity Trade-Off for Multipenalty Optimization Models: Predictive Validity and Ethnicity Group Differences

			Latinx–White		Black–White		Asian–White	
β	*r*	*r* 95% CI	Cohen's *d*	Cohen's *d* 95% CI	Cohen's *d*	Cohen's *d* 95% CI	Cohen's *d*	Cohen's *d* 95% CI
0	0.556	[0.535, 0.577]	−0.227	[−0.287, −0.166]	−0.617	[−0.678, −0.555]	−0.289	[−0.352, −0.226]
0.1	0.553	[0.531, 0.574]	−0.221	[−0.281, −0.161]	−0.530	[−0.591, −0.469]	−0.271	[−0.334, −0.208]
0.3	0.543	[0.521, 0.565]	−0.193	[−0.254, −0.133]	−0.412	[−0.473, −0.351]	−0.221	[−0.284, −0.159]
1.0	0.521	[0.498, 0.543]	−0.126	[−0.186, −0.066]	−0.240	[−0.301, −0.180]	−0.127	[−0.190, −0.065]
3.0	0.501	[0.477, 0.523]	−0.062	[−0.122, −0.002]	−0.128	[−0.188, −0.067]	−0.057	[−0.120, 0.005]
10.0	0.488	[0.465, 0.511]	−0.022	[−0.082, 0.038]	−0.071	[−0.132, −0.011]	−0.021	[−0.083, 0.042]
30.0	0.484	[0.460, 0.507]	−0.008	[−0.068, 0.053]	−0.053	[−0.113, 0.008]	−0.008	[−0.071, 0.054]
100.0	0.482	[0.458, 0.505]	−0.002	[−0.062 0.058]	−0.046	[−0.106, 0.015]	−0.004	[−0.066, 0.059]

Note. RoBERTa = Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach; CI = confidence interval. Values in this table correspond to points on the solid line of Figure 3. Data set includes both bag-of-words and RoBERTa predictors.

Table 8

Diversity–Validity Trade-Off for Multipenalty Optimization Models: Age and Gender Group Differences

	40 years or older–under 40 years		Male–female	
β	Cohen's *d*	Cohen's *d* 95% CI	Cohen's *d*	Cohen's *d* 95% CI
0	−0.515	[−0.574, −0.457]	−0.065	[−0.106, −0.024]
0.1	−0.411	[−0.469, −0.353]	−0.056	[−0.097, −0.015]
0.3	−0.301	[−0.360, −0.243]	−0.045	[−0.086, −0.004]
1.0	−0.168	[−0.226, −0.110]	−0.031	[−0.072, 0.010]
3.0	−0.087	[−0.145, −0.029]	−0.023	[−0.065, 0.018]
10.0	−0.047	[−0.105, 0.010]	−0.020	[−0.061, 0.021]
30.0	−0.034	[−0.092, 0.024]	−0.019	[−0.060, 0.022]
100.0	−0.030	[−0.087, 0.028]	−0.018	[−0.060, 0.023]

Note. RoBERTa = Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach; CI = confidence interval. Values in this table correspond to points on the solid line of Figure 3. Data set includes both bag-of-words and RoBERTa predictors.

context beneficial, especially selection scenarios with large applicant counts and numerous data points per applicant. First, they allow the user to simultaneously reduce group differences for more than two groups (e.g., simultaneously reduce Black/White, Latinx/White, and male/female group differences). This provides the user flexibility when implementing these techniques in real-life selection contexts.

A second benefit of these strategies is their flexibility, both in terms of the data that can be used in the models and the models themselves. For instance, the ability to utilize applicant data with demographic labels but without performance criteria (or without demographic labels but with performance criteria) is a strength of both techniques as personnel selection scenarios are often constrained by small samples and missing criterion data resulting from applicants taking an assessment but not being hired. This ability to use data where the criterion is missing allows for improved reduction of group differences in the true applicant population.

While both strategies are effective at reducing group differences, the multipenalty optimization technique demonstrated less loss of predictive validity than the iterative predictor removal technique. There are several factors underlying why the multipenalty optimization technique outperformed the iterative predictor removal technique. First, iterative predictor removal is an "all or nothing" approach where a predictor is either included or not included in the model. In addition, iterative predictor removal does not guarantee that the optimal predictors are removed, as this would involve examining all potential combinations of removed predictors, which is computationally infeasible. In contrast, multipenalty optimization allows for the optimal weighting of predictors, and it is the combination of predictors that determines each predictor's respective model coefficient. Not only does multipenalty optimization provide a greater number of potential model solutions but it also guarantees the optimal solution given a specific group differences penalty.[5] Second, for iterative predictor removal, if group differences exist in the criterion, the ML model is incentivized to reproduce these differences as accounting for them results in higher model performance. Removing predictors that demonstrate high group differences merely makes it more difficult for the ML model to reproduce these differences. In contrast, multipenalty optimization removes the incentive to reproduce these differences by penalizing models that demonstrate high group differences. An additional benefit of multipenalty optimization is that it requires less predictors than the iterative predictor removal technique. That is, the effectiveness of the iterative predictor removal technique is highly dependent upon the number of predictors in the data set where the technique is less effective if there are fewer predictors. This results in iterative predictor removal being limited to data sets with a larger number of predictors, while multipenalty optimization does not share this constraint.

Substantive Contributions

There are numerous benefits in using an algorithmic strategy to reduce group differences, especially in scenarios that involve text-based data. Until recently, the only method for evaluating text-based data was using human evaluators who can be expensive and can introduce their own personal biases (Campion et al., 2016; Macan & Merritt, 2011). Being able to successfully replicate human evaluators and simultaneously minimize between-group differences using ML techniques, as demonstrated in this research, it can provide organizations viable ways of evaluating applicants. This can be especially useful in high-volume personnel selection contexts.

While the techniques proposed in the current research are two ways of reducing group differences, practitioners responsible for building selection systems should also consider using additional strategies. For example, using the techniques proposed here in tandem with other strategies, such as using interview questions less likely to demonstrate group differences, should reduce group

[5] Note that dropping predictors in iterative predictor removal is equivalent to setting the predictor coefficients to 0 in multipenalty optimization. As multipenalty optimization obtains the optimal predictor coefficients, coefficients of 0 can be used. This means that multipenalty optimization encompasses all potential solutions that could be obtained by any predictor removal technique.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 9
Ethnicity: Iterative Predictor Removal Differential Prediction

				$d_{\text{Mod\_Signed}}$		
Predictors removed	Overall ΔR^2	Intercept ΔR^2	Slope ΔR^2	Latinx versus White	Black versus White	Asian versus White
0	0.002**	0.001*	0.001	0.022	0.134	0.002
1,000	0.010***	0.009***	0.001*	0.108	0.287	0.107
2,000	0.016***	0.015***	0.002*	0.138	0.358	0.146
3,000	0.020***	0.019***	0.001*	0.160	0.399	0.149
4,000	0.022***	0.021***	0.001*	0.172	0.419	0.176

Note. RoBERTa = Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach. White was the reference group for all effect size calculations. $d_{\text{Mod\_Signed}}$ is the signed effect size for the over- and underprediction of the criterion based on the regression line. For further explanation on how to interpret the effect sizes, see Dahlke and Sackett (2018). Data set includes both bag-of-words and RoBERTa predictors.
* $p < .05$. ** $p < .01$. *** $p < .001$.

differences even further, as the effectiveness of these techniques is constrained by the data going into them.

Potential Limitations and Implications for Future Research

Researcher decisions that occur throughout the ML model-building process can impact both the initial levels of group differences found in the ML model along with the effectiveness of these two techniques. For example, previous research has found that ML techniques that attempt to embed word meaning can demonstrate bias (e.g., Brown et al., 2020; Caliskan & Lewis, 2020; Charlesworth & Banaji, 2021; Garg et al., 2018). In the current research, the initial group differences levels were comparable across ML models using only bag-of-words predictors, only RoBERTa predictors, and bag-of-words and RoBERTa predictors (see Tables 3 and 4) but different natural language processing decisions could result in initial data sets with varying levels of group differences.

While the results are not shown in the current research, we tested both techniques on only bag-of-word predictors and only RoBERTa predictors. These supplemental results were similar to the results presented in the current research. Both techniques successfully reduced group differences but multipenalty optimization outperformed iterative predictor removal. It is possible, however, that different data sets may be more or less amenable to these two techniques. For example, the structure of data associated with a cognitive ability test may allow for a unique application of these techniques and may present new breakthroughs in ways of maintaining the validity of cognitive ability tests while decreasing the group differences reported in the employee selection literature (Roth et al., 2001).

For iterative predictor removal, predictors were removed based on their contribution to group differences relative to their overall predictive impact where predictors high on this ratio were removed first. This is just one way to determine which predictors to iteratively remove. Future research should examine and compare other strategies for prioritizing predictor removal. For example, predictor frequency, reliability, and parameters obtained from each predictor's item characteristic curve could be used as inputs into the ranking of predictors.

The use of algorithmic-based demographics as opposed to self-reported demographics is also a limitation of the current research. While demographic categorization was based on data that were not used in the model (i.e., a video thumbnail and name from application) and is, therefore, unlikely to be systematically related to the data, the lack of self-reported demographics added error to the data. This error likely decreased both initial group differences and the ability of both iterative predictor removal and

Table 10
Age: Iterative Predictor Removal Differential Prediction

				$d_{\text{Mod\_Signed}}$
Predictors removed	Overall ΔR^2	Intercept ΔR^2	Slope ΔR^2	40 years or older versus under 40 years
0	0.000	−0.000	0.000	0.030
1,000	0.002**	0.001**	0.000	0.127
2,000	0.004***	0.004***	0.000	0.204
3,000	0.006***	0.006***	0.000	0.240
4,000	0.008***	0.007***	0.000	0.267

Note. RoBERTa = Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach. Under 40 years was the reference group for all effect size calculations. $d_{\text{Mod\_Signed}}$ is the signed effect size for the over- and underprediction of the criterion based on the regression line. For further explanation on how to interpret the effect sizes, see Dahlke and Sackett (2018). Data set includes both bag-of-words and RoBERTa predictors.
** $p < .01$. *** $p < .001$.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 11
Sex: Iterative Predictor Removal Differential Prediction

				$d_{\text{Mod\_Signed}}$
Predictors removed	Overall ΔR^2	Intercept ΔR^2	Slope ΔR^2	Male versus female
0	0.000	0.000	−0.000	0.046
1,000	0.001	0.001*	−0.000	0.063
2,000	0.001	0.001*	−0.000	0.066
3,000	0.001	0.001*	−0.000	0.062
4,000	0.002**	0.002***	−0.000	0.100

Note. Female was the reference group for all effect size calculations. $d_{\text{Mod\_Signed}}$ is the signed effect size for the over- and underprediction of the criterion based on the regression line. For further explanation on how to interpret the effect sizes, see Dahlke and Sackett (2018). Data set includes both bag-of-words and RoBERTa predictors.
* $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

multipenalty optimization to identify group differences. Ultimately, we would anticipate that using self-reported demographic data would improve results. This should be explored in future research.

One question remains around the increase in overall differential prediction driven by intercept differences that occur when either technique is used. If the criterion group differences are due to measurement bias, an increase in differential prediction resulting from intercept differences is desired as it reflects smaller group differences in predicted scores. If the criterion group differences are due to true performance differences, an increase in differential prediction may be less desired, as the model may not fully reflect these group differences. Overall, these two techniques decrease the likelihood of finding adverse impact and increase organizational diversity, with a cost of not fully accounting for criterion group differences. Therefore, the appropriate use of these two techniques depends on the source of criterion group differences along with organizational goals, objectives, and perceptions of fairness.

One of the critiques of artificial intelligence and ML is its “black box” nature (Gonzalez et al., 2019). It can be challenging to understand why a model arrives at a particular solution. This critique applies to both techniques, as it is difficult to fully interpret the meaning behind the items that are removed with iterative predictor removal and the changing weights of the items for multipenalty optimization. While some have argued that the efficacy of ML algorithms rests on their usefulness over time and explainability is not necessary (Norvig, 2017), we believe that future research should apply advances in the field of explainable artificial intelligence to shed light on what has changed from the initial model before any group-difference mitigation technique was applied to the final model after the technique was applied. Improved explainability would be beneficial as it could potentially provide justification of results along with improving the ability for researchers and practitioners to gain insights or help refine existing theories (Adadi & Berrada, 2018).

The present research demonstrated the effectiveness of using ML to improve the classic diversity–validity dilemma in personnel selection. Therefore, we strongly encourage the field of IOP to stay up-to-date on advances in ML and to apply these advances to further the field of IOP in new and innovative ways. While numerous advances are being made in ML to increase algorithmic fairness, we believe that a partnership between the fields of ML and IOP provides the optimal solution given the unique set of

Table 12
Ethnicity: Multipenalty Optimization Differential Prediction

				$d_{\text{Mod\_Signed}}$		
β	Overall ΔR^2	Intercept ΔR^2	Slope ΔR^2	Latinx versus White	Black versus White	Asian versus White
0.0	0.002**	0.001*	0.001	0.022	0.134	0.002
0.1	0.004***	0.003***	0.001	0.029	0.178	0.014
0.3	0.006***	0.006***	0.000	0.049	0.238	0.043
1.0	0.012***	0.012***	0.000	0.091	0.325	0.094
3.0	0.017***	0.017***	0.001	0.127	0.380	0.129
10.0	0.020***	0.019***	0.001	0.148	0.407	0.146
30.0	0.021***	0.020***	0.001	0.156	0.415	0.151
100.0	0.021***	0.021***	0.001	0.159	0.419	0.153

Note. RoBERTa = Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach. White was the reference group for all effect size calculations. $d_{\text{Mod\_Signed}}$ is the signed effect size for the over- and underprediction of the criterion based on the regression line. For further explanation on how to interpret the effect sizes, see Dahlke and Sackett (2018). Data set includes both bag-of-words and RoBERTa predictors.
* $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 13
Age: Multipenalty Optimization Differential Prediction

				$d_{\text{Mod\_Signed}}$
β	Overall ΔR^2	Intercept ΔR^2	Slope ΔR^2	40 years or older versus under 40 years
0.0	0.000	−0.000	0.000	0.030
0.1	0.001	0.000**	0.000	0.076
0.3	0.002**	0.001**	0.000	0.128
1.0	0.004***	0.004***	−0.000	0.193
3.0	0.005***	0.005***	−0.000	0.232
10.0	0.006***	0.006***	−0.000	0.252
30.0	0.007***	0.007***	−0.000	0.258
100.0	0.007***	0.007***	−0.000	0.260

Note. Under 40 years was the reference group for all effect size calculations. $d_{\text{Mod\_Signed}}$ is the signed effect size for the over- and underprediction of the criterion based on the regression line. For further explanation on how to interpret the effect sizes, see Dahlke and Sackett (2018). Data set includes both bag-of-words and RoBERTa predictors.
** $p < .01$. *** $p < .001$.

challenges associated with personnel selection. For example, we identified group-specific pre- and postmodeling analytic procedures in ML that could be used when attempting to decrease group differences. These procedures, while effective, would be in direct violation of the Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial Organizational Psychology, 2018), Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice (1978), and Title VII of the Civil Rights Act of 1964. Iterative predictor removal and multipenalty optimization techniques, however, were shown to help solve the diversity–validity dilemma by applying ML techniques that are appropriate to use in a personnel selection context. Thus, we believe that this article about newly identified methods to mitigate selection tool group differences would not have been possible without the convergence of both the ML and IOP fields.

Table 14
Sex: Multipenalty Optimization Differential Prediction

				$d_{\text{Mod\_Signed}}$
β	Overall ΔR^2	Intercept ΔR^2	Slope ΔR^2	Male versus female
0.0	0.000	0.000	−0.000	0.046
0.1	0.000	0.000	−0.000	0.047
0.3	0.000	0.000	−0.000	0.049
1.0	0.000	0.000	−0.000	0.052
3.0	0.000	0.000	−0.000	0.052
10.0	0.000	0.000	−0.000	0.053
30.0	0.000	0.000	−0.000	0.053
100.0	0.000	0.000	−0.000	0.053

Note. RoBERTa = Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach. Female was the reference group for all effect size calculations. $d_{\text{Mod\_Signed}}$ is the signed effect size for the over- and underprediction of the criterion based on the regression line. For further explanation on how to interpret the effect sizes, see Dahlke and Sackett (2018). Data set includes both bag-of-words and RoBERTa predictors.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access: Practical Innovations, Open Solutions*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

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

Aguinis, H., & Smith, M. A. (2007). Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. *Personnel Psychology*, *60*(1), 165–199. https://doi.org/10.1111/j.1744-6570.2007.00069.x

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Natesan Ramamurthy, K., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 435–463. https://doi.org/10.1146/annurev-orgpsych-032414-111256

Bobko, P., Roth, P. L., & Nicewander, A. (2005). Banding selection scores in human resource management decisions: Current inaccuracies and the effect of conditional standard errors. *Organizational Research Methods*, *8*(3), 259–273. https://doi.org/10.1177/1094428105277416

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology*, *52*(3), 561–589. https://doi.org/10.1111/j.1744-6570.1999.tb00172.x

Bohnet, I. (2016). *How to take the bias out of interviews*. Harvard Business Review. https://hbr.org/2016/04/how-to-take-the-bias-out-of-interviews

Brenner, F. S., Ortner, T. M., & Fay, D. (2016). Asynchronous video interviewing as a new technology in personnel selection: The applicant's point of view. *Frontiers in Psychology*, *7*, Article 863. https://doi.org/10.3389/fpsyg.2016.00863

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners*. ArXiv. http://arxiv.org/abs/2005.14165

Caliskan, A., & Lewis, M. (2020). *Social biases in word embeddings and their relation to human cognition*. PsyArXiv. https://doi.org/10.31234/osf.io/d84kg

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In U. von Luxburg & I. Guyon (Eds.), *Proceedings of the 31st international conference of neural information processing systems* (pp. 3995–4004). Curran Associates.

Campbell, J. P. (1996). Group differences and personnel decisions: Validity, fairness, and affirmative action. *Journal of Vocational Behavior*, *49*(2), 122–158. https://doi.org/10.1006/jvbe.1996.0038

Campion, M. A., Outtz, J. L., Zedeck, S., Schmidt, F. L., Kehoe, J. F., Murphy, K. R., & Guion, R. M. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology*, *54*(1), 149–185. https://doi.org/10.1111/j.1744-6570.2001.tb00090.x

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. https://doi.org/10.1037/apl0000108

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Cascio, W. F., Alexander, R. A., & Barrett, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional issues and guidelines. *Personnel Psychology*, *41*(1), 1–24. https://doi.org/10.1111/j.1744-6570.1988.tb00629.x

Cascio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1995). Statistical implications of six methods of test score use in personnel selection. *Human Performance*, *8*(3), 133–164. https://doi.org/10.1207/s15327043hup0803_2

Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2016). New talent signals: Shiny new objects or a brave new world? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *9*(3), 621–640. https://doi.org/10.1017/iop.2016.6

Charlesworth, T. E. S., & Banaji, M. R. (2021). Word embeddings reveal social group attitudes and stereotypes in large language corpora. In M. Dehghani & R. L. Boyd (Eds.), *Handbook of language analysis in psychology* (pp. 494–508). Guilford Press.

Chowdhury, G. B. (2003). Natural language processing. *Annual Review of Information Science & Technology*, *37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Civil Rights Act of 1964, 42 U.S.C. § 7 (1964). https://www.govinfo.gov/content/pkg/STATUTE-78/pdf/STATUTE-78-Pg241.pdf#page=1

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, *5*(2), 115–124. https://doi.org/10.1111/j.1745-3984.1968.tb00613.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

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

Dahlke, J. A., & Wiernik, B. M. (2019). psychmeta: An R package for psychometric meta-analysis. *Applied Psychological Measurement*, *43*(5), 415–416. https://doi.org/10.1177/0146621618795933

De Corte, W. (1999). Weighing job performance predictors to both maximize the quality of the selected workforce and control the level of adverse impact. *Journal of Applied Psychology*, *84*(5), 695–702. https://doi.org/10.1037/0021-9010.84.5.695

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

De Corte, W., Sackett, P., & Lievens, F. (2010). Selecting predictor subsets: Considering validity and adverse impact. *International Journal of Selection and Assessment*, *18*(3), 260–270. https://doi.org/10.1111/j.1468-2389.2010.00509.x

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

de Montjoye, Y. A., Quoidbach, J., Robic, F., & Pentland, A. (2013). Predicting personality using novel mobile phone-based metrics. In A. Greenberg, W. Kennedy, & N. Bos (Eds.), *International conference on social computing, behavioral-cultural modeling, and prediction* (pp. 48–55). Springer. https://doi.org/10.1007/978-3-642-37210-0_6

de Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248–263. https://doi.org/10.1177/2515245919898466

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*(1), 19–29. https://doi.org/10.1037/0021-9010.72.1.19

El-Din, D. M. (2016). Enhancement bag-of-words model for solving the challenges of sentiment analysis. *International Journal of Advanced Computer Science and Applications*, *7*(1), 244–252. https://doi.org/10.14569/IJACSA.2016.070134

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register*, *43*, 38290–39315. https://www.ojp.gov/ncjrs/virtual-library/abstracts/employee-selection-procedures-adoption-four-agencies-uniform

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In L. Cao & C. Zhang (Eds.), *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259–268). Association for Computing Machinery. https://doi.org/10.1145/2783258.2783311

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Georgiou, K., & Nikolaou, I. (2020). Are applicants in favor of traditional or gamified assessment methods? Exploring applicant reactions towards a gamified selection method. *Computers in Human Behavior*, *109*, Article 106356. https://doi.org/10.1016/j.chb.2020.106356

Gillick, D. (2010). Can conversational word usage be used to predict speaker demographics? In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *Proceedings Interspeech* (pp. 1381–1384). ISCA. https://doi.org/10.21437/Interspeech.2010-421

Gonzalez, M. F., Capman, J. F., Oswald, F. L., Theys, E. R., & Tomczak, D. L. (2019). "Where's the IO?" Artificial intelligence and machine learning in talent management systems. *Personnel Assessment and Decisions*, *5*(3), 33–34. https://doi.org/10.25035/pad.2019.03.005

Greenemeier, L. (2017, June 2). *20 years after Deep Blue: How AI has advanced since conquering chess*. Scientific American. https://www.scientificamerican.com/article/20-years-after-deep-blue-how-ai-has-advanced-since-conquering-chess/

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 3315–3323). Neural Information Processing Systems Foundation.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Río, J. F. D., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020), Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hattrup, K., Rock, J., & Scalia, C. (1997). The effects of varying conceptualizations of job performance on adverse impact, minority hiring, and predicted performance. *Journal of Applied Psychology*, *82*(5), 656–664. https://doi.org/10.1037/0021-9010.82.5.656

Hern, A. (2017, May 25). Google's Go-playing AI still undefeated with victory over world number one. *The Guardian*. https://www.theguardian.com/technology/2017/may/25/alphago-google-ai-victory-world-go-number-one-china-ke-jie

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

HireVue. (2017). *Unilever finds top talent faster with HireVue assessments*. https://cdn.featuredcustomers.com/CustomerCaseStudy.document/hirevue_unilever_138410.pdf

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, *12*(1), 69–82. https://doi.org/10.1080/00401706.1970.10488635

Hoffman, C. C., & Thornton, G. C., III. (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology*, *50*(2), 455–470. https://doi.org/10.1111/j.1744-6570.1997.tb00916.x

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*(1&2), 152–194. https://doi.org/10.1111/1468-2389.00171

Johnson, S. K., Podratz, K. E., Dipboye, R. L., & Gibbons, E. (2010). Physical attractiveness biases in ratings of employment suitability: Tracking down the “beauty is beastly” effect. *The Journal of Social Psychology*, *150*(3), 301–318. https://doi.org/10.1080/00224540903365414

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In P. A. Flach, T. De Bie, & N. Cristianini (Eds.), *Joint European conference on machine learning and knowledge discovery in databases, part II* (pp. 35–50). Springer. https://doi.org/10.1007/978-3-642-33486-3_3

Leutner, F., Codreanu, S.-C., Liff, J., & Mondragon, N. (2020). The potential of game- and video-based assessments for social attributes: Examples from practice. *Journal of Managerial Psychology*, *36*(7), 533–547. https://doi.org/10.1108/JMP-01-2020-0023

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Settlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv. https://doi.org/10.48550/arXiv.1907.11692

Macan, T., & Merritt, S. (2011). Actions speak too: Uncovering possible implicit and explicit discrimination in the employment interview process. In G. P. Hodgkinson & J. K. Ford (Eds.), *International review of industrial and organizational psychology* (Vol. 26, pp. 293–337). Wiley-Blackwell. https://doi.org/10.1002/9781119992592.ch8

Markoff, J. (2011, February 16). Computer wins on ‘Jeopardy!’: Trivial, it’s not. *The New York Times*. https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Meade, A. W., & Tonidandel, S. (2010). Not seeing clearly with Cleary: What test bias analyses do and do not tell us. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*(2), 192–205. https://doi.org/10.1111/j.1754-9434.2010.01223.x

Meier, T., Boyd, R. L., Mehl, M. R., Milek, A., Pennebaker, J. W., Martin, M., Wolf, M., & Horn, A. B. (2020). Stereotyping in the digital age: Male language is “ingenious,” female language is “beautiful” - and popular. *PLOS ONE*, *15*(12), Article e0243637. https://doi.org/10.1371/journal.pone.0243637

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Nocedal, J., & Wright, S. J. (1999). *Numerical optimization*. Springer. https://doi.org/10.1007/b98874

Norvig, P. (2017). *Google’s approach to artificial intelligence and machine learning*. UNSW. https://www.youtube.com/watch?v=oD5Ug6uO0j8

Nye, C. D., & Sackett, P. R. (2017). New effect sizes for tests of categorical moderation and differential prediction. *Organizational Research Methods*, *20*(4), 639–664. https://doi.org/10.1177/1094428116644505

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in Neural Information Processing Systems 30*, 5684–5693. https://dl.acm.org/doi/10.5555/3295222.3295319

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. https://doi.org/10.1111/j.1744-6570.2008.00109.x

Potosky, D., Bobko, P., & Roth, P. L. (2005). Forming composites of cognitive ability and alternative measures to predict job performance and reduce adverse impact: Corrected estimates and realistic expectations. *International Journal of Selection and Assessment*, *13*(4), 304–315. https://doi.org/10.1111/j.1468-2389.2005.00327.x

Pyburn, K. M., Jr., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity–validity dilemma: Overview and legal context. *Personnel Psychology*, *61*(1), 143–151. https://doi.org/10.1111/j.1744-6570.2008.00108.x

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In M. Hildebrandt & C. Castillo (Eds.), *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 469–481). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372828

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 3.6.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reback, J., McKinney, W., jbrockmendel, Van den Bossche, J., Augspurger, T., Cloud, P., gfyoung, Hawkins, S., Sinhrks, Roeschke, M., Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Naveh, S., Garcia, M., patrick, Schendel, J., … h-vetinari. (2021). *pandas-dev/pandas: Pandas* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.4572994

Rev AI. (2021). *The world’s most accurate API for AI- and human-generated transcripts* (Version 1) [Computer software]. https://www.rev.ai/

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*(2), 297–330. https://doi.org/10.1111/j.1744-6570.2001.tb00094.x

Roth, P. L., Bobko, P., Van Iddekinge, C. H., & Thatcher, J. B. (2016). Social media in employee-selection-related decisions: A research agenda for uncharted territory. *Journal of Management*, *42*(1), 269–298. https://doi.org/10.1177/0149206313503018

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: Perspectives on Science and Practice*, *13*(2), 246–271. https://doi.org/10.1017/iop.2020.19

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*(3), 707–721. https://doi.org/10.1111/j.1744-6570.1997.tb00711.x

Sackett, P. R., & Roth, L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology*, *49*(3), 549–572. https://doi.org/10.1111/j.1744-6570.1996.tb01584.x

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, *49*(11), 929–954. https://doi.org/10.1037/0003-066X.49.11.929

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, *104*(10), 1207–1225. https://doi.org/10.1037/apl0000405

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. *American association for artificial intelligence spring symposium on computational approaches for analyzing weblogs* (Vol. 6, pp. 199–205). AAAI Press.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Schmidt, F. L. (1995). Why all banding procedures in personnel selection are logically flawed. *Human Performance*, *8*(3), 165–177. https://doi.org/10.1207/s15327043hup0803_3

Schmidt, F. L., & Hunter, J. E. (1995). The fatal internal contradiction in banding: Its statistical rationale is logically inconsistent with its operational procedures. *Human Performance*, *8*(3), 203–214. https://doi.org/10.1207/s15327043hup0803_6

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

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology*, *82*(5), 719–730. https://doi.org/10.1037/0021-9010.82.5.719

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, *311*, 18–38. https://doi.org/10.1016/j.ins.2015.03.040

Society for Industrial Organizational Psychology. (2018). Principles for the validation and use of personnel selection procedures. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *11*(S1, Supl. 1), 1–97. https://doi.org/10.1017/iop.2018.195

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Methodological*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods*, *21*(3), 525–547. https://doi.org/10.1177/1094428116677299

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … the SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wald, R., Khoshgoftaar, T., & Sumner, C. (2012). *Machine prediction of personality from Facebook profiles* [Conference session]. 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI), Las Vegas, NV, USA, 109–115. https://doi.org/10.1109/IRI.2012.6302998

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *Proceedings of the 30th International Conference on Machine Learning*, *28*(3), 325–333. https://dl.acm.org/doi/10.5555/3042817.3042973

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In J. Furman, G. Marchant, & H. Price (Eds.), *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society* (pp. 335–340). Association for Computing Machinery. https://doi.org/10.1145/3278721.3278779

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Appendix

Minimizing Solution for Multipenalty Optimization

To solve for the optimal model coefficients in multipenalty optimization, we rewrite Equation 4 in vector form. Using the fact that $\hat{y} = Xw$, where X is the predictor matrix and w is the vector of model coefficients, the overall cost function is:

$$C_{total} = ||y - Xw||_2^2 + \alpha||w||_2^2 + \beta\sum_{k=1}^{D}\left(\frac{1}{n_k}a_k^T Xw - \frac{1}{n}1^T Xw\right)^2, \tag{A1}$$

where a_k is a dummy vector that indicates membership in the kth group. For notational simplicity, we define:

$$z_k = X^T\left(\frac{1}{n_k}a_k - \frac{1}{n}1\right), \tag{A2}$$

which simplifies the group differences penalty to:

$$C_{diff} = \sum_{k=1}^{D}\left(z_k^T w\right)^2. \tag{A3}$$

The optimal w that minimizes this cost function is unique and can be written in closed form. By taking the first variation of the total cost function with respect to w and setting it equal to zero, the minimizing solution is:

$$w = \left(X^T X + \alpha I + \beta\sum_{k=1}^{D}(z_k z_k^T)\right)^{-1} X^T y. \tag{A4}$$

There are several ways to compute w but due to performance and memory constraints, we implemented this using a Line Search Newton–Conjugate Gradient solver (Nocedal & Wright, 1999).

Received April 5, 2021
Revision received January 23, 2023
Accepted January 29, 2023 ■