Received: 3 September 2021 | Revised: 18 May 2022 | Accepted: 18 May 2022

DOI: 10.1111/peps.12529

PERSONNEL PSYCHOLOGY WILEY

ORIGINAL ARTICLE

Investigating machine learning's capacity to enhance the prediction of career choices

Q. Chelsea Song[1] | Hyun Joo Shin[2] | Chen Tang[3] | Alexis Hanna[4] | Tara Behrend[1]

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

[2]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA

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

[4]College of Business, University of Nevada at Reno, Reno, Nevada, USA

Correspondence
Q. Chelsea Song, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN, 47907, USA.
Email: qcsong@purdue.edu

Abstract

Vocational interest measurement has long played a significant role in work contexts, particularly in helping individuals make career choices. A recent meta-analysis indicated that interest inventories have substantial validity for predicting career choices. However, traditional approaches to interest inventory scoring (e.g., profile matching) typically capture broad, or average relations between vocational interests and occupations in the population, yet may not be accurate in capturing the specific relations in a given sample. Machine learning (ML) approaches provide a potential way forward as they can effectively take into account complexities in the relation between interests and career choices. Thus, this study aims to enhance the accuracy of interest inventory-based career choice prediction through the application of ML. Using a large sample ($N = 81{,}267$) of employed and unemployed participants, we compared the prediction accuracy of a traditional interest profile method (profile matching) to a new machine-learning augmented method in predicting occupational membership (for employed participants) and vocational aspirations (for unemployed participants). Results suggest that, compared to the traditional profile method, the machine-learning augmented method

resulted in higher overall accuracy for predicting both types of career choices. The machine-learning augmented method was especially predictive of job categories with high base rates, yet underpredicted job categories with low base rates. These findings have practical implications for improving the utility of interest inventories for organizational practice, contributing to areas such as employee development, recruitment, job placement, and retention.

KEYWORDS
career choice, machine learning, profile matching, vocational interests

1 | INTRODUCTION

Career choices determine many subsequent work and nonwork outcomes and thus represent some of the most crucial decisions in a person's life. As a result, career choice is a critical outcome to predict and understand (Dierdorff, 2019; Hanna & Rounds, 2020; Hogan & Sherman, 2019). For organizations, career choice prediction has important implications for a host of human resource practices and personnel decisions, including employee development, recruitment, selection, promotion, and retention. As an example, organizations that accurately predict employees' career choices can use that knowledge to inform promotion decisions and support employees' career pathing within the organization, which are strategies that foster innovation (Malhotra et al., 2016) and employee retention (Benson & Rissing, 2020; Croteau & Wolk, 2010; Stroh, 1995). Similarly, organizations have an interest in understanding job seekers' and employees' vocational aspirations because people who work in careers that match their interests tend to have high job performance (Nye et al., 2012; Van Iddekinge et al., 2011), are more satisfied (Hoff et al., 2020), and are more likely to stay in the organization (Nye et al., 2012; Van Iddekinge et al., 2011). Critically, vocational interests are the strongest known predictors of career choices (Hanna & Rounds, 2020), and there have been calls for the use of interest inventories in selection batteries to predict occupational membership and job performance and to assist with job placement (Ingerick & Rumsey, 2014; Kirkendall et al., 2020; Nye et al., 2012; Oswald et al., 2019; Su et al., 2019). Because of the importance of career choices and their role in facilitating organizational practices, it is vital to predict them with the highest accuracy possible. Yet despite the strong relation between interests and career choices, traditional methods of interest inventory prediction leave much room for error.

Traditional methods of interest inventory prediction tend to capture *broad, or average*, relations between vocational interest scores and career choices; they follow the theoretical expectations of how interests should generally guide and motivate career choices in the population. However, these methods of prediction may not be accurate when predicting career choices for specific samples. Machine learning (ML) could suggest a way forward. Recent developments suggest that ML models can improve the prediction accuracy of personality tests for a wide range of outcomes by capturing complex and multidimensional relations (McCrae, 2015; Mõttus et al., 2017, 2019).

In this study, we propose an augmented method to predict career choices that incorporates the strengths of both theory-based interest measurement and the data-driven, ML approach, which together can optimize the overall accuracy of career choice predictions. Specifically, we make three primary contributions. First, we develop and describe an ML-augmented method to increase the accuracy of interest inventories in predicting career choices. Because interest inventories are used for career guidance and development for millions of people, even a small increase in prediction is important as it could influence career trajectories for substantial portions of the population. Thus, the proposed

approach has important implications in improving the utility of interest inventories for research, education, and workplace applications—including career guidance, employee development, and job placement. Second, we compare the predictive accuracy of traditional interest inventory scoring with ML-augmented scoring for predicting two different types of career choices: vocational aspirations and occupational membership. Finally, we provide Python code to implement the ML-augmented method in practice, easing the immediate incorporation of these methods into use for organizations, researchers, and career guidance professionals.

1.1 | Vocational interests and career choice

Vocational interests are trait-like preferences for particular work activities and environments (Rounds & Su, 2014). Interests have demonstrated importance for predicting outcomes such as job performance, income, turnover intentions, and career success (Nye et al., 2012; 2017; Rounds & Su, 2014; Stoll et al., 2017; Van Iddekinge et al., 2011). Yet, interest measures are most directly tied to career choices, which represent the culmination of career-related aspirations and decisions made across the lifespan (Ginzberg et al., 1951; Super, 1980). Vocational interest items are designed to assess a person's level of interest in engaging in different activities, such as building a birdhouse, doing an experiment in a lab, or teaching someone a new skill. Because vocational interests are always contextualized towards particular work-related activities (Rounds & Su, 2014), interest inventories play an integral role in vocational and career guidance to help individuals identify occupations that match their interests (Hansen, 1984; Strong, 1943; Zickar & Min, 2019).

1.1.1 | Theories of vocational interests and interest congruence

Several theories describe vocational interests and their fundamental link to careers. First, there have been a number of theoretical perspectives on the types and structure of vocational interests. Tracey and colleagues developed a spherical model of vocational interests emphasizing the importance of the ordered relationships among different interest types (Tracey, 2002; Tracey & Rounds, 1996). Tracey's model included eight interest markers based on the clustering of responses to interest items. More recently, Su et al. (2019) also developed an eight-category interest structure called the SETPOINT model, which can further be broken down into more specific, basic interest categories.

However, in terms of both interest structure and interest congruence, the most prominent and widely used vocational theory is Holland's (1997) Theory of Vocational Personalities and Work Environments. Holland's theory outlines six types of vocational interests that can be applied to both people and environments: *Realistic* interests involve manual labor and working with tools and objects; *Investigative* interests involve scientific activities, medicine, and technology; *Artistic* interests involve creative activities such as performing, visual arts, and music; *Social* interests involve helping, working with, and teaching others; *Enterprising* interests involve persuasion, leadership, and politics; finally, *Conventional* interests involve organization, attention to details, and data. These six interest types are often abbreviated as RIASEC. Similar to Tracey's (2002) model that specifies a structural ordering among interest categories, Holland's (1997) RIASEC model also contains a structural ordering of expected interrelations among the interest categories. However, whereas Tracey's (2002) model is a three-dimensional, spherical representation of interests, Holland specified a two-dimensional, hexagonal ordering in which interests that are closer to each other on the hexagon are more similar, and interests opposite from each other on the hexagon are least similar (see Figure 1).

According to Holland's theory, interests are motivational drivers that influence career choices. In other words, people choose careers that provide a good match to their own interests. This interest match, or *interest congruence*, then results in positive outcomes, such as job satisfaction, performance, and retention. Holland's (1997) theory has largely driven career guidance and vocational practice, including playing an important role in the career exploration of millions

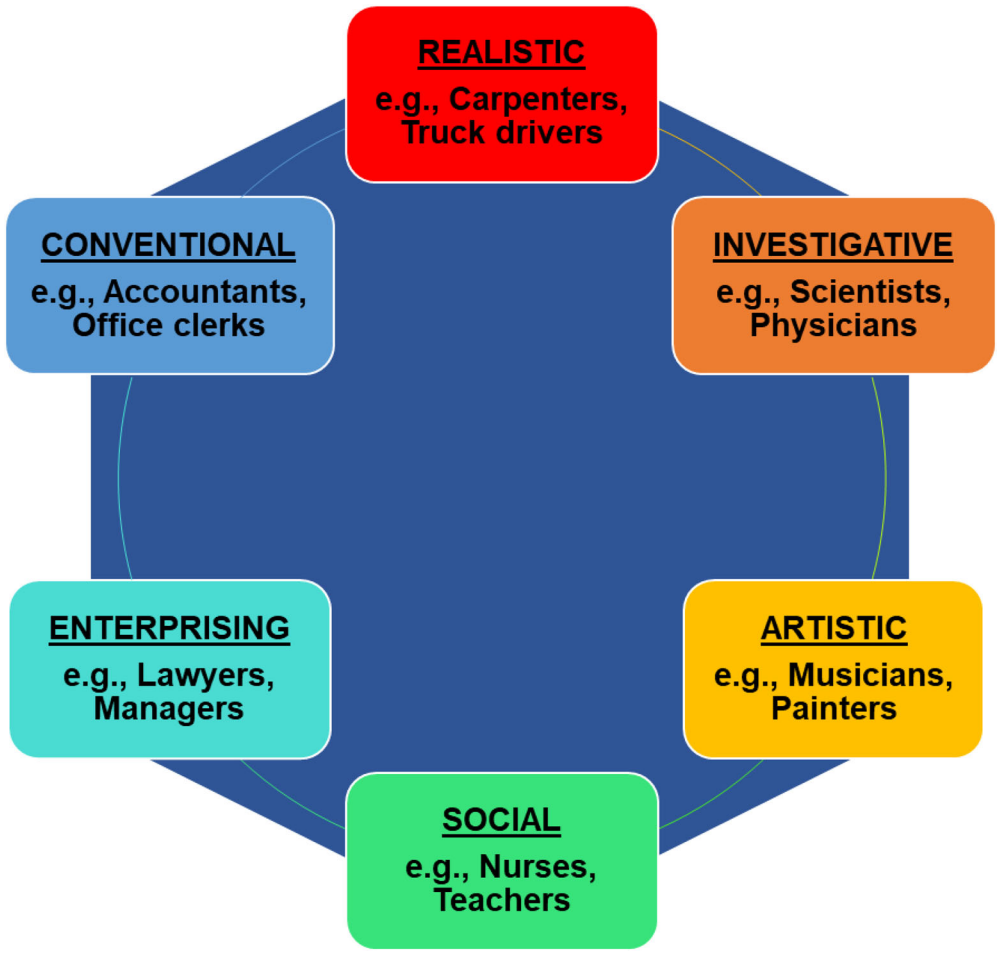

FIGURE 1 Holland's (1997) hexagonal ordering of the RIASEC interest types with example occupations

of people through the Department of Labor's Occupational Information Network (O*NET; U.S. Department of Labor, 2018).

Other theories have also articulated the notion that interests are important aspects of people's motivation to pursue different career paths. For example, Gottfredson's (1981; 2002) theory of circumscription, compromise, and self-creation describes how interests are one of the earliest forces guiding the development of children's career aspirations, which then continue to narrow and refine during adolescence and young adulthood. Throughout lifespan career development, interests serve as a constant guide for navigating the realm of potential career options, taking into account other factors such as perceived sex roles, abilities, and general compatibility of different careers with the self.

Likewise, Schneider's (1987) Attraction–Selection–Attrition (ASA) model explicitly includes vocational interests as an important component of determining organizational behavior through initial attraction to a job, as well as potential attrition from that job if interest fit is lacking. Specifically, similar types of individuals are attracted to jobs, selected into organizations, and prefer to remain within the organization once selected. The similarity of these attributes influences organizational behavior, and interests play a role in initiating the ASA process by guiding people in the direction of particular jobs. In this way, the link between interests and career choices is not only important for understanding individual career development, but also for understanding broader organizational behavior and retention.

1.1.2 | Interest congruence in practice

Many interest inventories are based on Holland's six RIASEC interests, including the Self-Directed Search (SDS; Holland et al., 1994), ACT Interest Inventory (ACT, 1995, 2009), and the Interest Profiler (Rounds et al., 2021). These measures can be used to score *people's* interest in each RIASEC area. In addition, *occupational* interests represent the degree to which an occupation exemplifies each RIASEC type. The O*NET contains occupational interest profiles (OIPs; Rounds et al., 1999, 2013) composing scores on all six RIASEC interests for over 900 occupations. Using these OIPs, interest congruence is often assessed based on the degree to which a person's RIASEC interests correspond to the interests exemplified by their aspired or actual job. Specifically in practice, this correspondence may be determined in a variety of ways; for instance, congruence may be determined via assessment of high-point matches between a person's highest interests and those of their aspired or actual job, the degree of match between the full person and job interest profiles, or further exploration of job matches within RIASEC areas according to more specific, basic interests.

A recent meta-analysis found that the relation between vocational interests and career choices is one of the strongest empirical relations that exists between psychological attributes and behavioral outcomes (Hanna & Rounds, 2020). This meta-analysis empirically summarized career choice "hit rates." For a particular sample, the hit rate denotes the proportion of correct career choice predictions made by the interest inventory. There are many ways to examine career choice prediction depending on the interest inventory and type of scale. For example, a hit may be counted if a person scores highest on their matched occupational scale for their career choice, or a hit may be counted if a person scores above a designated cut score on their matched scale. The most common operationalization of a hit rate was based on a RIASEC high-point match ($k = 70$, Hanna & Rounds, 2020), such that a hit was counted each time a person's highest interest area on the inventory matched the highest interest area of their career choice. Based on these RIASEC high-point matches, interest inventories accurately predicted career choices about 45% of the time, which is an especially substantial link given the number of potential career choices. In addition, aspired and actual career choices are influenced by many factors (Super, 1980), so it is difficult for a single factor to account for a large share of the prediction. In light of these considerations, the meta-analytic results provide evidence for the validity of interest measures (Holland, 1997; Schneider, 1987).

Nonetheless, the credibility interval of the meta-analytic hit rate estimate was large ([17.4, 77.2]), suggesting that the hit rate varied across different samples (Whitener, 1990). In other words, although the overall hit rate in the population was substantial, the hit rate was not consistently high in all samples. Thus, there is further room to improve career choice prediction from interest measures. Increasing the prediction accuracy of specific samples would be particularly important for organizations wishing to use interest measures for recruitment, job placement, or career development with their job applicants and employees (Kirkendall et al., 2020; Nye et al., 2012; Oswald et al., 2019).

1.1.3 | Common methods of matching interests to careers

One way to match interests to career choices using Holland's (1997) RIASEC categories is based on whether the person's highest interest, or high-point code, matches the interest most exemplified by their job. In other words, interest congruence can be assessed via *high-point code matches*. For example, based on interest congruence theory, a person whose highest interest is Social would exhibit a high-point match in an occupation that exemplifies Social interests, such as nursing or teaching (Rounds et al., 2013). Within this conceptualization (based on a person's highest-ranked interest area), this match represents a form of interest congruence. On the other hand, that same person would not have a high-point code match with occupations such as accounting (Conventional) or plumbing (Realistic). Because of their simplicity and ease of use, high-point code matching has commonly been used to assess RIASEC interest congruence (Hanna & Rounds, 2020). However, this method has also been criticized for ignoring all other person- and job interest scores (e.g., Phan & Rounds, 2018; Tracey, 2012).

Another way to assess congruence is through *profile matching*. This method evaluates the degree of correspondence between the full profiles of RIASEC scores for a person and job, by correlating the sets of scores (i.e., profile correlation), calculating the Euclidean distance between the profiles, or using other indices of fit (for detailed discussions of fit indices, see Camp & Chartrand, 1992; Edwards, 1991). For instance, Euclidean distance assesses congruence by calculating the distance between a person's interest profile and their occupation's interest profile when the profiles are plotted on a plane overlaying Holland's hexagon (Tracey & Robbins, 2006; Xu & Li, 2020). In this way, less distance between the profiles indicates greater correspondence, and thus greater congruence. Compared to the use of a single score, a full profile of scores uses all available data to make the most accurate predictions from a person's interests. These profiles of scores also exhibit greater reliability than the use of high-point scores alone (Low et al., 2005; Tinsley, 2000).

Interest congruence, and specifically profile matching, has commonly been used to guide career choices in practice. For example, the O*NET Interest Profiler (Rounds et al., 2021) is a freely available interest inventory based on Holland's (1997) RIASEC theory. Individuals who take the Interest Profiler receive a score in each RIASEC domain based on their likes and dislikes, and each individual is asked to choose a job zone, or desired level of preparation required for their career. Using the full profile of participant scores, O*NET reports a list of "best fitting" career choices.

In sum, although there are several ways to use RIASEC interest inventories to predict career choices, it is most reliable and desirable to use the whole profile of scores. Despite the benefits of profile matching, however, there are documented methodological issues with relying on fit indices for career choice prediction. For instance, fit indices like profile correlations or difference scores collapse person and environment profiles into a single metric, which loses the unique contribution from each profile and does not take into account the magnitude of the scores when making predictions (for a full discussion, see Edwards, 1993). Thus, this work aims to incorporate all RIASEC interest scores in the prediction of career choices while simultaneously addressing limitations of traditional profile-matching methods (as we discuss below).

1.2 | Using machine learning to improve career choice prediction

Traditional approaches to interest inventory scoring (e.g., high-point codes, profile matching) follow Holland's (1997) theoretical assumptions and capture *broad*, or average relations between vocational interests and career choices in the population. However, due to their broad scope, these traditional approaches might not be accurate when predicting career choices for specific samples. A ML approach could contribute to improving the prediction accuracy by (1) incorporating both broad and specific levels of interest measurement and (2) potentially capturing the unique relation of interests and occupations in a specific sample.

First, compared to traditional approaches that are generally limited to incorporating interest scale scores, ML approaches can incorporate both broad and specific levels of measurement, allowing them to model complex and multidimensional relations. Recent studies on vocational interests suggest that specific interest scales, such as basic interest scales, tend to display higher hit rates than broad, RIASEC-based scales (Hanna & Rounds, 2020)[1]. Compared to RIASEC-based scales, specific interest scales capture more detailed nuances of individuals' interests, thus providing a better bandwidth-fidelity match to predicting specific career choices (Day & Rounds, 1997; Ralston et al., 2004; Su et al., 2019). For instance, Holland's Artistic interest area may contain basic interests such as Visual Arts, Applied Arts and Design, Performing Arts, Music, Writing, Media, and Culinary Art (Su et al., 2019). Nonetheless, there are advantages to using both broad and specific measures: broad interest measures provide an organizing, theoretical framework to guide research and validity evidence, whereas specific measures better capture the complex, multidimensional reality of interests, and provide stronger predictive power (Hanna & Rounds, 2020; Liao et al., 2008; Su et al., 2019; Van Iddekinge et al., 2011). Thus, rather than doing away with broader RIASEC scales in favor of more specific interest scales or items, we suggest combining both broad and specific interest measurements to improve career choice prediction.

Recent developments in ML provide a potential way forward: ML models can take into account complex relations and nuances at narrower levels of measurement; it is also capable of handling high-dimensional data in which the number of features is large relative to the sample size, enabling the integration of multiple interest items and scales (see Oswald et al., 2020; Putka et al., 2018; Song et al., 2020; Woo et al., 2020). Recent studies found that ML can effectively capture the unique and valid variances in personality items, contributing to improved prediction of job performance (McCrae, 2015; Putka et al., 2018). In other words, ML models are able to handle different types of data more easily than traditional prediction methods, allowing them to capitalize on the advantages of broad and specific interest measures.

Second, while the traditional approaches provide theoretical expectations on how interests should guide career choices, the ML approach improves prediction by taking into account relations specific to a given sample. For example, traditional methods based on Holland's (1997) theory would predict that individuals who are most interested in Artistic activities and environments should choose Artistic occupations. However, in reality, there are not nearly as many Artistic occupations as there are occupations in Enterprising and Conventional areas (DeCeanne et al., 2017), so it is possible that many people with high Artistic interests actually choose Enterprising careers. By using empirical samples to train the ML models, ML models can capture how interests are actually related to career choices in a given sample, beyond what theory would predict. In other words, compared to traditional scoring methods, ML models are more data-driven and do not rely on a priori rules to combine information (e.g., through certain difference indices or correlations), allowing for a more flexible examination of the underlying relation between interests and career choices in the data and enhanced prediction of career choices in the real world (e.g., Jiang et al., 2020).

In this study, we proposed an ML-augmented method that aims to combine the advantages of both traditional profile methods and ML to improve career choice prediction. Using a large sample, we then evaluated the prediction accuracy of the ML-augmented method compared to the traditional, profile method. By doing so, we aim to diagnostically explore the utility of ML in vocational interest-based career choice prediction.

To compare the accuracy of predictions from the traditional profile method versus the ML-augmented method, we examined two kinds of career choices: vocational aspirations and occupational membership (Hanna & Rounds, 2020). Vocational aspirations represent ideal career choices, which reflect a motivational component in people's lifespan career development. On the other hand, occupational membership represents actual career choices, which reflect a behavioral component of choosing to work in a particular job or career. Together, vocational aspirations and occupational membership represent different manifestations within a full range of career choice operationalization (Hanna & Rounds, 2020). Thus, our hypotheses are as follows:

Hypothesis 1: The ML-augmented method using both scale- and item-level vocational interest scores will attain better prediction accuracy in *occupational membership* (among employed individuals) than the profile method using only scale-level vocational interest scores.

Hypothesis 2: The ML-augmented method using both scale- and item-level vocational interest scores will attain better prediction accuracy in *vocational aspirations* (among unemployed individuals) than the profile method using only scale-level vocational interest scores.

2 | METHOD

All analysis code and corresponding results for this project are available from the project page on the Open Science Framework (https://osf.io/27rbh/).

TABLE 1 Example of the occupational RIASEC scores and high-point code identification

Occupation	Realistic	Investigative	Artistic	Social	Enterprising	Conventional	High-point code
Example A	3.5	3.0	3.5	4.3	3.0	3.5	S
Example B	5.5	5.0	6.5	6.2	5.2	6.5	A, C
Example C	3.0	4.0	3.0	4.0	4.0	3.8	I, S, E

Note. High-point code was identified as the RIASEC domain with the highest score. For instance, in Example A, the Social interest score of 4.3 is highest among the six RIASEC domains, and thus the high-point code is S (Social). When multiple RIASEC domains were tied for the highest score (e.g., Examples B and C), there were more than one high-point codes. For instance, Example B has two high-point codes: A (Artistic) and C (Conventional); and Example C has three high-point codes: I (Investigative), S (Social), and E (Enterprising).

2.1 | Sample

Our initial sample included N = 84,349 participants who responded to a TIME Magazine online survey[2] in 2016 (Glosenberg et al., 2021). After data cleaning (e.g., removing observations that failed attention checks and observations that reported unrealistic age values [age value > 99]), the final sample size was N = 81,267. Among the participants, 75% were employed and 62% were female. The average age was M = 38 years (SD = 13.46).

2.2 | Measures

2.2.1 | Vocational interests

Vocational interests were measured using 16-items from the Personal Globe Inventory (PGI) Mini (see Tracey, 2019; a shortened measure based on the original PGI and the PGI-Short; Tracey, 2002, 2010). Participants were asked to respond on a 7-point Likert scale (1 = *Strongly dislike*; 4 = *Neutral*; 7 = *Strongly like*) to indicate their preference towards each activity. Example activities are "Seat patrons at a restaurant," "Install electrical wiring," and "Help children with learning problems."[3] Missing data were imputed using *k*-nearest neighbor imputation (Batista & Monard, 2003).

2.2.2 | Occupational high-point code

Employed participants reported their occupational membership (current occupation) by responding to the instruction, "Please enter your profession." Unemployed participants reported their vocational aspirations by responding to the question, "What is your dream job?" In both cases, the participants chose their occupational membership or vocational aspiration from an interactive drop-down menu of O*NET job titles. For each job, we obtained occupational RIASEC scores and high-point code using the O*NET occupational interest database[4] (see Table 1 for examples). The O*NET database contains scores for each occupation on all six RIASEC domains corresponding to how representative the occupation is for each interest (Rounds et al., 1999, 2013). The RIASEC domain with the highest score represents the occupation's high-point code. For instance, the high-point code of "electrician" is "Realistic," and thus "electrician" is classified as a Realistic occupation.

2.3 | Procedure and analytical strategy

Both the profile method and the ML-augmented method described below aim to predict the RIASEC scores of an occupation that best fits a participant's measured interests.

2.3.1 | Profile method

The profile method prediction is provided by O*NET and based on participants' six RIASEC scale scores (obtained from the interest inventory): O*NET uses profile correlations to find occupational choices that best match an individual's RIASEC interest profile (Rounds et al., 2021). Specifically, O*NET calculated the correlation between participants' RIASEC scores and the RIASEC interest scores of all occupations in the database. For each participant, O*NET recommended top five jobs that had the highest profile correlations; these occupations' RIASEC profiles yielded the closest match with the participant's interest profile. We treated the high-point code of the *first* O*NET-recommended job as the prediction made from the profile method because the first recommendation represents the occupation that provided the closest profile match to each participant's interest scores (based on the highest rank-order correspondence between participants' and occupations' interest profiles).

2.3.2 | ML-augmented method

The ML-augmented method of career choice prediction introduced in this paper uses the participants' interest item scores and scale scores to predict the interest scores of their career choice (occupational membership or vocational aspiration). First, the ML models "learn" the associations between the participants' interest scores and the interest scores of their career choices using a training sample. Then, the trained models are applied to a new sample to predict occupational interest scores. By training on a specific sample, the ML-augmented method should be able to better capture specific relations within the sample, and thus provide predictions with improved accuracy. The ML-augmented method is described in more detail below.

The ML-augmented method uses the participants' scores on the 16 PGI-Mini items and their six RIASEC scale scores to predict the six occupational RIASEC scores. Specifically, we implemented four different ML models, where each model was based on a different ML algorithm: neural network, *k*-nearest neighbors, elastic net, and random forest (see Appendix A for brief descriptions of the algorithms). These four models were separately trained with hyperparameter tuning (via random grid search and 10-fold cross-validation) to predict occupational RIASEC scores. Predictions from the four models were further ensembled with score averaging[5] to obtain the overall ML-augmented method prediction. Model training was conducted using Scikit-learn (Version 0.24.1, Pedregosa et al., 2011) and Python (Version 3.9.0). To mimic a typical organizational scenario, each ML model was trained and tested with a random sample of size 300 (see Bosco et al., 2015; Shen et al., 2011). We repeated this procedure 100 times and estimated the average prediction accuracy of the ML-augmented method (as well as the individual ML models).

2.3.3 | Evaluation: prediction accuracy

The prediction accuracy of the profile method and the ML-augmented method were evaluated in three ways: (a) high-point hit rates: using overall hit rates and differential hit rates, (b) Euclidean distance, and (c) profile correlations. The high-point hit rate focuses on the highest RIASEC code (i.e., high-point code); it is commonly used in career guidance (e.g., Hansen, 2019) and provides a benchmark against existing meta-analyses (e.g., Hanna & Rounds, 2020). Euclidean

TABLE 2 Example interest profile scenarios and corresponding evaluation metrics

Scenario	Example	Most relevant evaluation metric
1. A person is similarly interested in many RIASEC domains	Similarly high interest scores in Investigative, Artistic, Social, and Enterprising	– Euclidean distance
2. There is a tie for the highest RIASEC domains	Highest interest scores in Artistic and Social while scores of the other RIASEC domains are much lower	– Euclidean distance – High-point hit rate
3. A person has strongly differentiated RIASEC interest profile	High interest scores in Realistic, moderate interest scores in Investigative, Artistic, and Conventional, and low interest scores in Social and Enterprising	– Profile correlation – High-point hit rate
4. A person has strong dislikes	Very low interest scores in Realistic	– Profile correlation
5. A person has low interests across the board	Low interest scores in Realistic, Investigative, Artistic, Social, Enterprising, Conventional interests	– Profile correlation – High-point hit rate

Note. Scenario 1: A person is similarly interested in many RIASEC domains. In this scenario, predicting jobs that all have a high magnitude of scores (Euclidean distance) in those domains will be most relevant, as opposed to rank-order that would distinguish between the relevance of these areas unnecessarily. Scenario 2: There is a tie for the highest RIASEC domains. This scenario is akin to Scenario 1 where the person is similarly interested in multiple RIASEC domains. In this scenario, the magnitude of scores (Euclidean distance) in the top RIASEC domains will be relevant for considering jobs that have high occupational interests in each of the tied areas. In addition, high-point hit rates will also be helpful because a high-point match is attained for jobs that match one of the tied areas (e.g., jobs that have either a highest score in Artistic or highest score in Social would be considered a match). Scenario 3: A person has a strongly differentiated RIASEC interest profile. In this scenario, there is a clear rank-ordering among the RIASEC domains and a clear high-point interest, and thus profile correlation and high-point hit rates are most relevant. Scenario 4: A person has strong dislikes. In this scenario, it is important to identify occupations that de-emphasize their lowest interest areas, and thus profile correlations reflecting this rank-order are most relevant. Scenario 5: A person has low interests across the board. In this scenario, the person has "low interests" compared to others, yet there is still a relative (i.e., within-person) difference across RIASEC domains. Thus, when making a career choice, the within-person rank-order (profile correlation) and high-point (high-point hit rate) are more relevant than the absolute magnitude of scores (Euclidean distance). For example, even if the person's interest scores are low, if Conventional is still their highest rank-order, then a Conventional job may be the best choice.

distance and profile correlation utilize all six RIASEC scores to provide a profile-based comparison of predicted and reported career choices.

High-point hit rate describes *the match between the high-point code* of a person's career choice (occupational membership or vocational aspiration) and the predicted career choice (based on the profile method or the ML-augmented method). Euclidean distance describes *the absolute distance between the RIASEC profile* of a person's career choice and the predicted career choice. Profile correlation describes *the match between the rank-ordering of the RIASEC profile* of a person's career choice and the predicted career choice. These three metrics provide unique and complementary information that contributes to career choice prediction. Table 2 presents several scenarios that may occur in practice and highlights the most relevant evaluation metric to use in each case. Together, high-point hit rates, Euclidean distance, and profile correlations provide a comprehensive comparison of the accuracy of different prediction methods.

High-point hit rate. A "hit" signifies the match between an individual's career choice (occupational membership or vocational aspiration) and predicted occupational choice (based on the profile method or the ML-augmented method). In the profile method, we counted a hit when the O*NET-suggested occupation's high-point code was the same as the high-point code of the participant's occupational membership or vocational aspiration (see Table 3 for examples). In the ML-augmented method, each of the four ML models separately provided predictions of six occupational RIASEC scores. These scores were then aggregated across the four models using score averaging to obtain the final ML-augmented high-point code (see Table 4 for examples). When the ML-augmented occupational high-point code was the same as the high-point code of the participant's occupational membership or vocational aspiration, we counted it as a hit. If there was a tie in the occupational high-point code predictions (e.g., Example 3 in Table 4), it was counted as

TABLE 3 Example of the profile method hit operationalization

O*NET suggested occupation	High-point code		Hit
	O*NET suggested occupation	Self-reported occupation	
Electrician	R	E	No hit
Preschool teacher	S	S	Hit

a "hit" if at least one of the high-point code ties was the same as the high-point code of the occupational membership or vocational aspiration.

The high-point hit rate is calculated as the sum of hits divided by the total number of individuals in the sample. We calculated separate hit rates for the outcomes of "occupational membership" (of the employed participants) and "vocational aspirations" (of the unemployed participants) and presented the average of overall hit rates across 100 test sets. In addition to the overall hit rate across the full sample, we also calculated hit rates for each RIASEC occupational category. The hit rate for each RIASEC category was calculated as the number of hits (e.g., correct predictions made by the profile method or the ML-augmented method) divided by the total number of participants whose reported job's high-point code is in that RIASEC category (see Hanna & Rounds, 2020). A higher high-point hit rate suggests a better prediction accuracy.

Differential hit rate. The differential hit rate is the difference between the high-point hit rate and the base rate in each RIASEC occupational category. In this context, base rates describe the distribution of career choices. The base rate of each RIASEC category is calculated as the number of participants whose occupational high-point code is within that category divided by the total number of participants in the sample (Dawis, 1996; Meehl & Rosen, 1955; Schmidt, 1974). For example, the base rate for Investigative occupational membership is the number of participants with Investigative jobs divided by the total number of participants in that sample. The differential hit rate is used to examine whether the predictions made by the profile method and the ML-augmented method are more accurate than random chance, or not using interest inventories at all (Bokhari & Hubert, 2015; Meehl & Rosen, 1955). To enable a direct comparison between the profile method and the ML-augmented method, we estimated the hit rates of the profile method and the ML-augmented method on the same samples.

Euclidean distance. In addition to hit rates based on high-point codes, prediction accuracy was also evaluated using Euclidean distance, which utilizes the full profiles of RIASEC interest scores (i.e., six RIASEC interest scores corresponding to the occupation). Both the profile method and the ML-augmented method predict the full RIASEC profile of individuals' career choices. Euclidean distance is calculated between each participant's *predicted* RIASEC occupational interest scores (from either the profile method or the ML-augmented method) and *reported* RIASEC occupational interest scores (using the scores from either their occupational membership or their vocational aspiration). A lower Euclidean distance suggests a better prediction accuracy.

We note that, in this study, we use Euclidean distance to *evaluate* the degree of correspondence between the interest profiles of the predicted and reported career choices. This is different from using Euclidean distance to predict career choices, which we described earlier. In other words, the present work uses Euclidean distance as a means of evaluation, rather than a means of prediction.

Profile correlation. Profile correlation evaluates prediction accuracy based on the *rank-order* of the predicted and reported interest profiles. When evaluating both the profile method and the ML-augmented method, for each participant, the six predicted RIASEC occupational scores were correlated with the six reported RIASEC occupational scores. A higher correlation indicates a better prediction accuracy. As described previously, profile correlations can be used as a metric of congruence for person and job interest profiles. Here, we use profile correlations to *evaluate* the accuracy of the profile method versus the ML-augmented method for predicting career choices.

TABLE 4 Example of the ML-augmented method hit operationalization

Example		Neural network	*k*-NN	Elastic net	Random forest	Average	ML-predicted high-point code	Self-reported occupation's high-point code	Hit
1	R	4.21	4.16	5.06	4.59	**4.50**	R	R	Hit
	I	2.89	3.01	1.88	2.28	2.51			
	A	3.76	3.15	3.06	3.36	3.33			
	S	1.95	2.73	3.28	4.39	3.09			
	E	5.05	4.46	4.53	3.73	4.44			
	C	3.41	3.83	3.93	4.26	3.86			
2	R	2.75	2.77	2.05	2.65	2.56	C	R	No hit
	I	3.46	3.38	3.00	3.25	3.27			
	A	2.34	2.82	3.27	3.38	2.95			
	S	3.79	3.79	4.53	4.12	4.06			
	E	4.86	4.51	4.23	4.04	4.41			
	C	4.44	4.49	4.69	4.32	**4.48**			
3	R	3.21	3.30	1.89	2.73	2.78	I and E	E	Hit
	I	4.23	4.36	4.78	4.33	**4.42**			
	A	3.68	3.32	2.98	3.37	3.34			
	S	2.16	2.59	3.87	3.40	3.00			
	E	4.75	4.57	3.84	4.51	**4.42**			
	C	3.42	3.40	4.62	4.24	3.92			
4	R	2.36	2.87	3.21	3.68	3.03	E and C	A	No hit
	I	3.44	3.62	3.97	4.14	3.79			
	A	2.89	2.97	2.30	2.40	2.64			
	S	3.59	3.30	3.19	3.45	3.38			
	E	4.41	4.30	4.95	4.46	**4.53**			
	C	4.84	4.40	4.53	4.34	**4.53**			

Note. This table provides four examples demonstrating how "hit" is operationalized for the ML-augmented predictions. In each example, the ML-augmented method first uses individual ML models to predict the job's RIASEC scores, and then average these predictions to obtain the ensembled ML-augmented method prediction. For instance, in Example 1, neural network, *k*-NN, elastic net, random forest each predicted the job's Realistic score to be 4.21, 4.16, 5.06, 4.59; the ensembled prediction is the average of the four scores, 4.50. The high-point code is the RIASEC domain with the highest ensembled prediction. For instance, in Example 1, Realistic interest's ensembled score is the highest among the six RIASEC domains, and thus Example 1's high-point code is Realistic. When multiple RIASEC domains are tied for the highest score, there will be multiple high-point codes. For instance, Example 3 has two highest scoring dimensions: Investigative and Enterprising, both with 4.42, and thus its high-point codes are Investigative and Enterprising. To determine "hit," the ML-predicted high-point code(s) is compared with the self-reported occupation's high-point code. When there are multiple high-point codes, a "hit" is determined if the self-reported occupation's high-point code matches at least one ML-predicted high-point code.

3 | RESULTS

Hypotheses 1 and 2 compare the prediction accuracy (high-point hit rate, Euclidean distance, profile correlation) between the ML-augmented method and the profile method in predicting occupational membership (Hypothesis 1) and vocational aspirations (Hypothesis 2). The results of the prediction accuracy (high-point hit rate, Euclidean dis-

TABLE 5 Overall hit rate, Euclidean distance, and profile correlation results

	Occupational membership			Vocational aspiration		
	Hit rate	Euclidean distance	Profile corr.	Hit rate	Euclidean distance	Profile corr.
Profile method	.28	5.18	.23	.34	4.97	.28
ML-augmented method (average across 4 algorithms)	**.34**	**4.07**	**.44**	**.36**	**4.25**	**.36**
Neural network	.33	4.13	.42	.36	4.28	.35
k-NN	.33	4.11	.43	.35	4.31	.33
Elastic net	.34	4.08	.44	.36	4.26	.36
Random forest	.33	4.08	.44	.36	4.26	.35

Note. Hit rate = high-point hit rate; Euclidean distance = Euclidean distance; Profile corr. = profile correlation. Higher high-point hit rate, lower Euclidean distance, and higher profile correlation suggest better prediction accuracy. In the first two rows ("Profile method" and "ML-augmented method"), the bolded values denote the better prediction accuracy results between the profile method and the ML-augmented method under each evaluation method (high-point hit rate, Euclidean distance, or profile correlation). "ML-augmented method" shows the results for the ensemble-averaged ML-augmented method predictions, and the individual ML-augmented model results are shown in "Neural network," "*k*-NN," "Elastic net," and "Random forest." For the ML-augmented method, in addition to simple ensemble averaging, we also evaluated the high-point hit rate of the ML-augmented method when majority vote was used to ensemble the individual ML models (Dietterich, 2000; Zhou, 2012). The high-point hit rate result of majority vote was highly similar to the result with simple ensemble average.

tance, and profile correlation) are shown in Tables 5 through 8. The results suggest that, in general (across high-point hit rate, Euclidean distance, profile correlation, and across the full sample), the ML-augmented method yielded better prediction accuracy than the traditional profile method (Table 5). This improvement in prediction accuracy for the ML-augmented method was especially substantial when predicting occupational membership than vocational aspiration.

However, when examining each RIASEC occupational category, the ML-augmented method tended to have better prediction accuracy in predicting RIASEC occupational categories with high base rate (e.g., Enterprising) while the profile method tended to have better prediction accuracy for RIASEC occupational categories with low base rate (e.g., Artistic). This pattern was observed when the prediction accuracy was assessed using high-point hit rate (Table 6) and profile correlation (Table 8) but not Euclidean distance (Table 7). The Euclidean distance results suggested that the ML-augmented method yielded prediction accuracy similar to, or higher than, the traditional profile method.

3.1 | High-point hit rates

As shown in Table 5, for both occupational membership and vocational aspirations, the ML-augmented method yielded higher high-point hit rates than the profile method. The overall hit rates for the ML-augmented method are .34 (vs. .28 for profile method) when predicting occupational membership, and .36 (vs. .34 for profile method) when predicting vocational aspirations. The overall hit rate results supported both Hypotheses 1 and 2.

High-point hit rate by RIASEC occupational category. Compared to the overall hit rates, the results for each RIASEC occupational interest category were more nuanced. Table 6 displays the comparison between base rates and hit rates in each RIASEC category. For each RIASEC occupational category, the profile method and the ML-augmented method differed in the accuracy of their career choice prediction. Although the ML-augmented method performed better than the profile method overall (see Table 5), it did not perform better than the profile method in all RIASEC occupational categories (see Table 6).

TABLE 6 High-point hit rate by RIASEC occupational category

	Occupational membership					Vocational aspiration				
		Hit rate		Differential hit rate			Hit rate		Differential hit rate	
	Base rate	Profile method	ML-augmented method	Profile method	ML-augmented method	Base rate	Profile method	ML-augmented method	Profile method	ML-augmented method
R	.07	.18	.00	**.11**	−.07	.08	.18	.04	**.10**	−.04
I	.17	.26	.04	**.09**	−.13	.21	.30	.24	**.09**	.03
A	.06	.45	.00	**.39**	−.06	.23	.42	.35	**.19**	.12
S	.16	.20	.05	**.04**	−.11	.12	.23	.04	**.11**	−.08
E	.35	.38	.79	.03	**.44**	.29	.42	.77	.13	**.48**
C	.19	.13	.23	−.06	**.04**	.07	.20	.03	**.13**	−.04

Note. Higher high-point hit rate suggests a better prediction accuracy. The differential hit rate is the difference between the hit rate and the base rate in each RIASEC category. For each RIASEC occupational category (i.e., each row), bolded values denote the higher differential hit rate between the profile method and the ML-augmented method.

TABLE 7 Euclidean distance by RIASEC occupational category

	Occupational membership		Vocational aspiration	
	Profile method	ML-augmented method	Profile method	ML-augmented method
R	5.79	**5.13**	5.80	**5.12**
I	5.07	**4.63**	5.00	**4.56**
A	**4.52**	4.96	4.76	**4.01**
S	5.29	**4.65**	5.06	**4.52**
E	5.00	**3.43**	4.81	**3.83**
C	5.49	**3.63**	5.08	**4.36**

Note. Lower Euclidean distance suggests a better prediction accuracy. For each RIASEC occupational category (i.e., each row), bolded values denote the lower Euclidean distance between the profile method and the ML-augmented method.

For the occupational membership, compared to the profile method, the ML-augmented method poorly predicted the RIASEC occupational categories with low base rate (Realistic, Investigative, Artistic, and Social), resulting in hit rates lower than the corresponding base rate (ML-augmented method Realistic differential hit rate = −.07; Investigative differential hit rate = −.13; Artistic differential hit rate = −.06; Social differential hit rate = −.11). On the other hand, the ML-augmented method outperformed the profile method in predicting the RIASEC occupational categories with high base rate (Enterprising and Conventional): the ML-augmented method's Enterprising hit rate was .79, about double the hit rate of the profile method (.38); the ML-augmented Conventional hit rate was .23, also about double the hit rate of the profile method (.13).

For vocational aspirations, similar to the occupational membership results, the ML-augmented method did not perform better than the profile method in all RIASEC categories, even though the ML-augmented method's overall hit rate was higher than that of the profile method. Specifically, the ML-augmented method only yielded higher hit rates for Enterprising vocational aspirations, yet the profile method hit rates for Realistic, Investigative, Artistic, Social, and Conventional vocational aspirations were higher than the ML-augmented method hit rate.

TABLE 8 Profile correlation by RIASEC occupational category

	Occupational membership		Vocational aspiration	
	Profile method	ML-augmented method	Profile method	ML-augmented method
R	**.04**	.02	**.02**	−.11
I	**.24**	.20	**.24**	.17
A	**.34**	−.14	.31	**.47**
S	**.16**	.11	**.23**	.12
E	.32	**.75**	.38	**.63**
C	.18	**.72**	.28	**.39**

Note. Higher profile correlation suggests a better prediction accuracy. For each RIASEC occupational category (i.e., each row), bolded values denote the higher profile correlation between the profile method and the ML-augmented method.

3.2 | Euclidean distance

As shown in Table 5, for both occupational membership and vocational aspirations, the ML-augmented method prediction yielded lower Euclidean distance (and thus better prediction accuracy) than the profile method. The overall Euclidean distance for the ML-augmented method was 4.07 (vs. 5.18 for profile method) when predicting occupational membership, and 4.25 (vs. 4.97 for profile method) when predicting vocational aspirations. The overall Euclidean distance results supported both Hypotheses 1 and 2.

Euclidean distance by RIASEC occupational category. Similar to the overall Euclidean distance, the ML-augmented results for each RIASEC occupational interest category were consistently better than the profile methods, except for Artistic occupational membership (see Table 7). For occupational membership, the ML-augmented method performed better than the profile method in most RIASEC categories, except for Artistic (profile method Euclidean distance = 4.52, ML-augmented method Euclidean distance = 4.96). For vocational aspiration, all of the Euclidean distances for the ML-augmented method were lower than that of the profile method, suggesting that the ML-augmented method was more accurate in predicting vocational aspirations than the profile method.

3.3 | Profile correlation

As shown in Table 5, for both occupational membership and vocational aspirations, the ML-augmented method prediction yielded higher profile correlation than the profile method. The overall profile correlation for the ML-augmented method was .44 (vs. .23 for profile method) when predicting occupational membership, and .36 (vs. .28 for profile method) when predicting vocational aspirations. The overall profile correlation results supported both Hypotheses 1 and 2.

Profile correlation by RIASEC occupational category. Compared to the overall profile correlation, the results for each RIASEC occupational interest category were more nuanced (see Table 8). For occupational membership, the ML-augmented method poorly predicted Realistic, Investigative, Artistic, and Social occupational memberships (ML-augmented method profile correlation for Realistic = .02; Investigative = .20; Artistic = −.14; Social = .11). On the other hand, the ML-augmented method outperformed the profile method in predicting Enterprising and Conventional occupational membership: the ML-augmented Enterprising profile correlation was .75, more than double the profile correlation of the profile method (.32); the ML-augmented Conventional profile correlation was .72, about four-times the profile correlation of the profile method (.18). For vocational aspiration, the ML-augmented method

yielded higher profile correlations for Artistic, Enterprising, and Conventional, but not for Realistic, Investigative, and Social occupational categories.

4 | DISCUSSION

Career choices are critical for individuals, organizations, and society as they represent the initial and ongoing decisions contributing to a host of work and life outcomes. For instance, since attraction to a job is the initial piece of the ASA process in an organization, the relation between interests and career choices is important for downstream outcomes including job performance and turnover (Schneider, 1987). This study advances research and practice on vocational interests and career choices by investigating ML's capacity to enhance the prediction of career choices.

Specifically, the current work makes three primary contributions. First, we proposed a new ML-augmented method of career choice prediction that can be implemented with any existing interest inventory. Second, we compared the accuracy of a more traditional interest profile-based method of career choice prediction utilized by O*NET to the new ML-augmented method that uses both broad and specific interest scores. Overall, the results showed that the ML-augmented method yielded higher high-point hit rates, lower Euclidean distances, and higher profile correlations than the traditional profile method in predicting both occupational membership and vocational aspirations. This ML-augmented method thus provides a significant contribution to a long history of research and practice linking interests and career choices (e.g., Hanna & Rounds, 2020; Hansen, 1984; Su, 2020; Zickar & Min, 2019). However, compared to the traditional profile method, the ML-augmented method was more sensitive to base rate, suggesting the need for future research to improve career choice prediction for occupations with low employment rates (which we will discuss below). Finally, we provide the Python code (see link at the beginning of the Method section) with instructions for implementation to allow immediate incorporation of this method into future research and practice.

The present findings revealed several notable patterns. To evaluate the accuracy and usefulness of the new ML-augmented method, it was important to not only examine its prediction accuracy on its own, but also to compare it to the most conventional methods that use interest inventories to predict career choices. Millions of people visit the O*NET for career exploration each year (U.S. Department of Labor, 2018), and O*NET uses profile matching to recommend well-fitting careers based on individual's vocational interests. The biggest overarching takeaway from our results comparing O*NET's predictions and predictions from the ML-augmented method is that overall, the ML-augmented method resulted in higher prediction accuracy for both occupational membership and vocational aspirations. In other words, this method makes career choice recommendations that match people's occupational membership and vocational aspirations more often than O*NET's recommendations do, implying that ML-augmented methods have the potential to provide better career guidance and more accurate prediction.

However, there were notable cases in which the ML-augmented method underperformed relative to O*NET's profile method. Beyond the overall prediction accuracy, we also examined the accuracies within each RIASEC occupational category to evaluate the profile method and ML-augmented method performed across different types of occupations. Here, results were less clearly aligned in favor of one method or the other. For some RIASEC categories, the ML-augmented method performed better; for others, the profile-based method from O*NET performed better.

In particular, an interesting pattern emerged: predictions from the ML-augmented method far exceeded the accuracy of the profile method in RIASEC categories with high base rates, whereas the ML-augmented method was less accurate for areas with low base rates (with the exception of the Euclidean distance evaluation metric). Based on the relevance of each prediction accuracy metric (high-point hit rate, Euclidean distance, profile correlation) in different career choice prediction scenarios (see Table 2), the pattern described above suggests the following recommendations in regard to using profile method and the ML-augmented method in practice. When a person is similarly interested in many RIASEC domains (Scenario 1 in Table 2), the ML-augmented method is more recommended than the profile method. This is because Euclidean distance is the most relevant evaluation metric in this scenario and the ML-augmented method resulted in low Euclidean distance (and thus better prediction accuracy) across most RIASEC

occupational categories. In all other scenarios, we recommend using profile correlation when the base rate of the predicted RIASEC category is low and ML-augmented method when the base rate is high.

The findings suggested that the ML-augmented method was fairly sensitive to base rates, whereas the traditional profile method was not. The reason for this sensitivity is that the ML models are trained to recognize the most and least common types of jobs. This training allowed the models to make more accurate predictions for the majority of the sample, which is why the overall prediction accuracy was higher than that of the untrained profile method. However, some accuracy was lost for members of the sample who work in, or aspire to, jobs in "less popular" RIASEC categories.

For example, the largest majority of the sample worked in Enterprising jobs (base rate = 35%). If we did not use an interest inventory and simply guessed that every member of the sample worked in an Enterprising job, we would be correct 35% of the time (i.e., we would attain a hit rate of 35% by chance alone). The profile method used by O*NET correctly predicted Enterprising career choices 38% of the time, which did not do much better than the base rate; on the other hand, the ML-augmented method correctly predicted Enterprising career choices 79% of the time, which far exceeded both the base rate and the profile method (see Table 6). On the other hand, the reverse was true for occupational categories with low base rates in this sample, such as Artistic and Realistic. For these occupational categories, the ML-augmented method performed poorly. As the most extreme example, the ML-augmented method hit rate for Artistic occupational membership was 0%, meaning that the ML-augmented method almost *never* accurately predicted when a person worked in an Artistic occupation. Essentially, because the base rate of Artistic employment was so low (base rate = 6%; i.e., a small proportion of people had Artistic jobs), the model rarely predicted that anyone worked in an Artistic occupation, which resulted in a very low hit rate in that occupational category.

Notably, these findings highlight the importance of using accurate training data that is representative of the target test sample for which predictions will be made. Based on the present findings, the underlying base rates (i.e., career choice distributions) in the training data appear to heavily influence model predictions by helping the models "learn" the most common and least common types of career choices. In this way, it is critical to ensure that the distribution of career choices in the training sample is highly similar to the distribution in the test sample. For instance, the types of career choices that are most common in urban areas likely differ from the most common types of career choices in rural areas, so the training data from these respective locations should ideally be locally collected.

Because ML models are trained to underpredict groups or categories with low base rates, one potential way to deal with this issue is to consider predictions from both the ML-augmented and the profile methods. When the two methods predict different types of career choices, the ML-augmented method prediction can be used in cases where the base rates reflect high rates of aspirations or employment; on the other hand, the profile method prediction can be used in cases with low base rates. Future work may attempt to systematically test the viability of this decision rule in applied prediction contexts to evaluate the combined accuracy of using the ML-augmented predictions for high-base rate areas and profile method predictions for low-base rate areas.

4.1 | Theoretical and practical implications

Interest measurement has a rich history in organizational, vocational, and educational psychology (Su, 2020; Zickar & Min, 2019). The present findings suggest that ML can enhance the predictive validity of vocational interests, which provides a modern advancement in interest measurement. Despite meta-analytic evidence that interest measures can effectively predict career choices using traditional prediction methods (Hanna & Rounds, 2020), ML further improved the capacity of measured interests to predict both occupational membership and vocational aspirations. The proposed ML-augmented method combines the advantages of both traditional theory-driven methods that capture broad relations between interests and career choices and data-driven ML methods that enhance prediction for specific samples. The advanced predictive capacity of the ML-augmented method thus has implications for interest theory, research applications, and practice.

The ML-augmented method can effectively incorporate both broad and specific interests, using comprehensive interest information to enhance career choice prediction. In doing so, this method provides information about the modern-day validity and practical application of Holland's (1997) vocational theory, which has been widely used for career guidance and research purposes for decades. Although Holland's (1997) RIASEC categories holistically categorize hundreds of occupations and provide a comprehensive framework for evaluating the fit between people's interests and jobs, the traditional profile method based on RIASEC scale scores leaves room for improvement when predicting both occupational membership and vocational aspirations (Hanna & Rounds, 2020).

One possibility is that Holland's (1997) RIASEC categories may be too broad to meaningfully distinguish among nuanced career preferences (Day & Rounds, 1997; Ralston et al., 2004; Su et al., 2019). In line with this argument, more specific interests, such as basic interests, do tend to predict career choices better than broader interests (Hanna & Rounds, 2020). For these reasons, basic interests can provide fruitful avenues of future work in predicting career choices and providing career guidance. For example, more nuanced interest information can help job seekers narrow the range of possibilities within their highest RIASEC interest area, such as deciding among music, photography, and graphic design within the realm of Artistic career paths. Nonetheless, basic interest measures that contain more specific interest categories are typically much longer than RIASEC measures and more difficult to implement in person-environment fit research due to their lack of integration with common sources (e.g., O*NET's OIPs are based on RIASEC). The ML-augmented method provides a nice balance by capturing the ease of use and shorter survey length of RIASEC measures while also benefiting from the enhanced prediction of specific interests by incorporating interest items. Thus, this method both acknowledges the limitations of traditional, theory-driven methods in practice, while also capitalizes on the benefits of theoretically based measures and organizing frameworks.

The ML-augmented method also has a number of important implications for practice, particularly in organizations and career guidance contexts. First, this method can be used in career counseling to improve the accuracy of career choice recommendations. For example, career centers at universities around the United States typically have large stores of vocational interest data from previous students, as well as follow-up information denoting the types of jobs where those students ended up working after graduation. This data can be used to train ML-augmented models to better predict students' career choices and provide more useful guidance and recommendations to current and future students. In addition, by taking into account the imbalance in employment rates across different occupational categories and subsets of the population, the ML-augmented method can be used to increase diversity in different career orientations. For instance, models can be trained to identify more underrepresented individuals—who might not have been identified by traditional methods—to pursue certain career paths where increases in racial, gender, and other aspects of diversity would be beneficial. We will further explore this potential application in the "Limitations and Future Directions" section.

Relatedly, the ML-augmented method could inform targeted recruitment efforts by more accurately identifying potential qualified and interested job applicants. Because interest fit is related to important work outcomes such as job performance (Nye et al., 2012, 2017), it is desirable to recruit and hire individuals who are interested in the open positions. Organizations can measure the interests of current employees and train ML models to predict career choices within that particular setting. In doing so, organizations can have more localized and tailored information regarding the relation between interests and occupational membership among their particular population of employees. Using this information, they can reach out to broader candidate pools to improve expected performance and diversity through recruitment (Newman & Lyon, 2009). As an example, the ML-augmented method can be used in online recruitment and be incorporated with sourcing algorithms (e.g., Bogen & Rieke, 2018) to more effectively deliver job opening information to potential job applicants.

Organizations could also use the ML-augmented method to improve employee career development, job placement and retraining, selection, promotion, and retention. For instance, by improving the accuracy of career choice prediction, the ML-augmented method could facilitate lateral transfer to help keep high-performing employees in the organization when they would otherwise consider leaving, particularly when they view their current job as incongruent with their interests (Noe et al., 1988; Schneider, 1987). In addition, by capturing the ongoing changes in work tasks

and their relation with specific interests, the ML-augmented method helps adapt job placement and retraining efforts to meet the organizational needs.

Further, the ML-augmented method could enhance the use of vocational interests in personnel selection and promotion, which adds to the growing practice of using ML to advance selection practices (e.g., Campion et al., 2016; Hickman et al., 2021; Sajjadiani et al., 2019). Previous studies have found that certain interest dimensions (e.g., Enterprising, Social) are related to an individual's motivation to lead (Chan et al., 2000). The ML-augmented method could incorporate such information to facilitate managerial and leadership placement, particularly for internal promotion decisions. That is, in addition to facilitating job placements based on job tasks, the ML-augmented method can also improve placements based on leadership potential and motivation, as captured by vocational interests. Across all of these applications, the ML-augmented method can generally be used to further the career development of employees, which can help organizations increase employee retention (Croteau & Wolk, 2010; Malhotra et al., 2016).

Nonetheless, there are some caveats to implementing the ML-augmented method in practice. Aside from the issue of sensitivity to base rate that we mentioned above, small training samples might also hinder the accuracy of the ML-augmented method (e.g., Oswald et al., 2020; Song et al., 2021). Smaller organizations might have a more difficult time training accurate ML-augmented models based on a small local sample. In this case, the users could consider either using the traditional profile matching method or rely on ML-augmented models trained on a larger, nonlocal sample. There is value for future studies to explore solutions for improving career choice prediction when the base rate of the occupation is low, and when the sample size is low. Further, the complexity and data-reliant nature of the ML-augmented method could obscure construct-irrelevant idiosyncrasies or biases in career choice prediction (e.g., Oswald et al., 2020). In addition to prediction accuracy, future developments of ML-augmented methods should also consider other key criteria including bias and construct relevance.

4.2 | Limitations and future directions

Despite the strengths of this work, there are a number of limitations, which in turn suggest several future directions. First, despite the fact that the vocational aspirations from the unemployed participants were collected using the question, "What is your dream job?," it is possible that some participants responded with external considerations in mind (e.g., what occupations they could actually do, rather than what they ideally want to do). In this way, some of the self-reported career choices may have been influenced by contextual factors of the unemployed sample, rather than representing a pure assessment of ideal jobs. Further, although vocational aspirations represent motivational components of people's career choice trajectories (Hanna & Rounds, 2020; Holland & Gottfredson, 1975), it is unclear in this case whether these aspirations would guide the unemployed individual's future career choices in the same way as adolescent vocational aspirations, for example (Gottfredson, 1981, 2002). Because most previous studies of vocational aspirations focused on samples of youth and adolescents (Hanna & Rounds, 2020), future studies could consider studying vocational aspirations for unemployed participants, and follow-up with their vocational aspirations and occupational membership after employment.

Second, the match between an individual's interests and those of their occupation (i.e., interest congruence) is associated with positive work outcomes, including job satisfaction, performance, and retention (e.g., Hoff et al., 2020; Nye et al., 2012; Van Iddekinge et al., 2011). While the current work proposed an ML-augmented method to improve the match, we did not directly study how the proposed method could influence future work outcomes. Future studies are needed to explore the relation between the improved career choice prediction and work outcomes, particularly those assessed longitudinally. Such studies could enhance the theoretical understanding between interest congruence and work outcomes, as well as improve the practical utility of interest-based prediction for career development and work outcomes.

Third, occupations and aspirations are disproportionately distributed across the different RIASEC categories, so there are imbalances in career choices (see DeCeanne et al., 2017). This imbalance is also reflected in the current

sample. To further improve the accuracy of career choice prediction, future work could explore additional ML techniques to effectively address the imbalance in occupational distributions, as well as the possibility of considering the ML-augmented method together with profile methods that are less tied to base rates.

This effort to address imbalance in occupational distributions could also contribute to improved diversity in career orientations. For example, there is a major gender imbalance in STEM occupations, where STEM careers are more often pursued and attained by men than women (Morris, 2016; Su & Rounds, 2015). There are a number of reasons leading to this imbalance, among which career prediction and guidance are key to advance gender equality in STEM employment (for review, see Stewart-Williams & Halsey, 2021). The ML-augmented method proposed in this study could be extended to improve career guidance for STEM and non-STEM occupations. Specifically, future studies could explore ML-augmented methods to adjust for the imbalance in career choices by gender. These methods can be used to identify more women (who might not have been identified by traditional career choice prediction methods) to pursue STEM occupations, as well as identifying more men to pursue non-STEM occupations, such as nursing or teaching. Further, in practice, self-exploration also plays an important role in determining a person's career path. The ML-augmented method has the potential to help self-exploration by encouraging individuals to explore career directions they might not normally consider. In particular, practitioners and career counselors could use the ML-augmented method to provide such recommendations to clients, who can further explore these less-considered options to find a well-fitting career path. In sum, improved career choice prediction facilitates career development, recruitment, selection, retention, and contributes to diversity in both STEM and non-STEM occupations.

4.3 | Conclusion

This study aimed to enhance the accuracy of interest inventory-based career choice prediction through the application of ML. We proposed an ML-augmented method that combines the advantages of both interest theories and data-driven ML by incorporating information from both broader RIASEC interest scales and specific interest items. Results showed that, overall, the ML-augmented method yielded higher predictive accuracy than the traditional profile method in predicting both occupational membership and vocational aspirations. These findings suggest that ML can enhance the predictive validity of vocational interests, which provides an important contribution to both theory and practice. Nonetheless, the profile method outperformed the ML-augmented method for predicting career choices in occupational areas with low base rates. In this way, both the ML-augmented method and the profile method have important uses in practice, and base rates are critically important for understanding which method is preferable in future applications. Finally, by providing access to our analysis code and detailed instructions, our hope is that researchers, organizations, and counselors can easily implement the new ML-augmented method to predict career choices in both research and applied contexts.

ORCID

Q. Chelsea Song https://orcid.org/0000-0003-0910-1281
Hyun Joo Shin https://orcid.org/0000-0002-6791-5681
Chen Tang https://orcid.org/0000-0003-3502-9906
Alexis Hanna https://orcid.org/0000-0001-8869-0437
Tara Behrend https://orcid.org/0000-0002-7943-5298

ENDNOTES

[1] This is parallel to support for using Big Five facets in personnel selection (e.g., Hogan & Roberts, 1996; Paunonen et al., 1999).
[2] https://time.com/4343767/job-personality-work/
[3] The PGI-Mini is based on a circumplex model, which was empirically supported by previous studies (e.g., Glosenberg et al., 2019). Specifically, Glosenberg et al. (2019) evaluated the circumplex model using a randomization test of hypothesized order

relations (that compares the fit of the data with the hypothesized circumplex model against a randomized, permuted data). They used a correspondence index (CI) to examine the fit, where CIs of 0.5 are considered good fit (meaning 75% of the theoretical predictions were supported; see Rounds et al., 1992). Results suggested a statistically significant fit between PGI-Mini and the RIASEC circumplex structure (CI = .89, $p = .017$). Realistic was calculated as: (item 5 + item 15)/2. Investigative was calculated as: (item 6 + item 16)/2. Artistic was calculated as: (item 7 + item 17)/2. Social was calculated as: [2 × (item 8 + item 18) + (item 1 + item 11)]/3. Enterprising was calculated as: [2 × (item 2 + item 12) + (item 1 + item 11)]/3. Conventional was calculated as: [2 × (item 4 + item 14) + (item 3 + item 13)]/3. See Tracey (2019) for details.

[4] The O*NET database with high-point codes and RIASEC scores for each occupation is available at https://www.onetcenter.org/dictionary/25.3/excel/interests.html

[5] Averaging is a common ensemble learning technique to ensemble the prediction of multiple ML models. The technique could increase the prediction accuracy and model generalizability (Dietterich, 2000; see Zhou, 2012, pp. 68–69 for details).

[6] The term "bias" here refers to a numerical value that is part of the input signal processing/calculation. It is analogous to the intercept in multiple regression.

CONFLICT OF INTEREST

Authors have no known conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

The data and analysis codes that support the findings of this study are openly available in the Open Science Framework at https://osf.io/27rbh/. The data were derived from the following resources available in the public domain: https://doi.org/10.17605/OSF.IO/BE5JA (Glosenberg et al., 2021) and https://www.onetcenter.org/dictionary/25.3/excel/interests.html (O*NET).

REFERENCES

ACT, Inc. (1995). *Technical manual: Revised unisex edition of the ACT Interest Inventory (UNIACT)*. Iowa City, IA: Author.

ACT, Inc. (2009). *Technical manual: ACT Interest Inventory*. Iowa City, IA: Author.

Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, *17*(5–6), 519–533. https://doi.org/10.1080/713827181

Benson, A., & Rissing, B. A. (2020). Strength from within: Internal mobility and the retention of high performers. *Organization Science*, *31*(6), 1475–1496.

Bogen, M., & Rieke, A. (2018). *Help wanted: An examination of hiring algorithms, equity, and bias*. Upturn https://www.upturn.org/reports/2018/hiring-algorithms/

Bokhari, E., & Hubert, L. (2015). A new condition for assessing the clinical efficiency of a diagnostic test. *Psychological Assessment*, *27*, 745–754. https://doi.org/10.1037/pas0000093

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. https://doi.org/10.1037/a0038047

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Camp, C. C., & Chartrand, J. M. (1992). A comparison and evaluation of interest congruence indices. *Journal of Vocational Behavior*, *41*(2), 162–182.

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.

Chan, K. -. Y., Rounds, J., & Drasgow, F. (2000). The relation between vocational interests and the motivation to lead. *Journal of Vocational Behavior*, *57*(2), 226–245.

Croteau, J. D., & Wolk, H. G. (2010). Defining advancement career paths and succession plans: Critical human capital retention strategies for high-performing advancement divisions. *International Journal of Educational Advancement*, *10*(2), 59–70.

Dawis, R. V. (1996). Vocational psychology, vocational adjustment, and the workforce: Some familiar and unanticipated consequences. *Psychology, Public Policy, and Law*, *2*(2), 229–248. https://doi.org/10.1037/1076-8971.2.2.229

Day, S. X., & Rounds, J. (1997). A little more than kin, and less than kind": Basic interests in vocational research and career counseling. *The Career Development Quarterly*, *45*(3), 207–220.

DeCeanne, A., Lewis, P., & Rounds, J. (2017). A RIASEC snapshot of the modern U.S. workforce. In *Proceedings of the Poster Presentation at the 32nd Annual Conference of the Society for Industrial and Organizational Psychology*, Orlando, FL.

Dierdorff, E. C. (2019). Toward reviving an occupation with occupations. *Annual Review of Organizational Psychology and Organizational Behavior*, *6*, 397–419. https://doi.org/10.1146/annurev-orgpsych-012218-015019

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Proceedings of the* International workshop on multiple classifier systems (pp. 1–15). Berlin, Heidelberg: Springer.

Edwards, J. R. (1991). Person-job fit: A conceptual integration, literature review, and methodological critique. In C. I. Cooper & I. T. Robertsons (Eds.), *International review of industrial and organizational psychology* (Vol. 6, pp. 283–357). J. Wiley & Sons.

Edwards, J. R. (1993). Problems with the use of profile similarity indices in the study of congruence in organizational research. *Personnel Psychology*, *46*(3), 641–665.

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, *57*(3), 238–247. https://doi.org/10.2307/1403797

Ginzberg, E., Ginsburg, S. W., Axelrad, S., & Herma, J. L. (1951). *Occupational choice*. Columbia University.

Glosenberg, A., Behrend, T. S., Tracey, T. J. G., Blustein, D. L., & Foster, L. (2021). *Vocational interest data from Time magazine*. https://doi.org/10.17605/OSF.IO/BE5JA

Glosenberg, A., Tracey, T. J. G., Behrend, T. S., Blustein, D. L., & Foster, L. L. (2019). Person-vocation fit across the world of work: Evaluating the generalizability of the circular model of vocational interests and social cognitive career theory across 74 countries. *Journal of Vocational Behavior*, *112*, 92–108.

Gottfredson, L. S. (1981). Circumscription and compromise: A developmental theory of occupational aspirations [Monograph]. *Journal of Counseling Psychology*, *28*, 545–579. https://doi.org/10.1037/0022-0167.28.6.545

Gottfredson, L. S. (2002). Gottfredson's theory of circumscription, compromise, and self-creation. In D. Brown (Ed.), *Career choice and development* (pp. 85–148). Jossey-Bass.

Hanna, A., & Rounds, J. (2020). How accurate are interest inventories? A quantitative review of career choice hit rates. *Psychological Bulletin*, *146*(9), 765.

Hansen, J. C. (1984). The measurement of vocational interests: Issues and future directions. In S. D. Brown & R. L. Lent (Eds.), *Handbook of counseling psychology* (pp. 99–136). Wiley.

Hansen, J.-I. C. (2019). Interest inventories. In G. Goldstein, D. N. Allen, & J. Deluca (Eds.), *Handbook of psychological assessment* (pp. 169–190). Academic Press. https://doi.org/10.1016/B978-0-12-802203-0.00006-7

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2021). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, Advanced online publication. https://doi.org/10.1037/apl0000695

Hoff, K. A., Song, Q. C., Wee, C. J. M., Phan, W. M. J., & Rounds, J. (2020). Interest fit and job satisfaction: A systematic review and meta-analysis. *Journal of Vocational Behavior*, *123*, 103503.

Hogan, J., & Roberts, B. W. (1996). Issues and non-issues in the fidelity-bandwidth trade-off. *Journal of Organizational Behavior*, *17*, 627–637.

Hogan, R. T., & Sherman, R. A. (2019). New(ish) directions for vocational interests research. In C. D. Nye & J. Rounds (Eds.), *Vocational interests: Rethinking their role in understanding workplace behavior and practice. SIOP organizational frontiers series* (pp. 189–204). Routledge. https://doi.org/10.4324/9781315678924-10

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments*. Psychological Assessment Resources.

Holland, J. L., & Gottfredson, G. D. (1975). Predictive value and psychological meaning of vocational aspirations. *Journal of Vocational Behavior*, *6*(3), 349–363. https://doi.org/10.1016/0001-8791(75)90007-X

Holland, J. L., Powell, A. B., & Fritzsche, B. A. (1994). *The self-directed search (SDS)*. Psychological Assessment Resources.

Ingerick, M., & Rumsey, M. G. (2014). Taking the measure of work interests: Past, present, and future. *Military Psychology*, *26*(3), 165–181. https://doi.org/10.1037/mil0000045

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: A brief primer. *Behavior Therapy*, *51*(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Kirkendall, C. D., Nye, C. D., Rounds, J., Drasgow, F., Chernyshenko, O. S., & Stark, S. (2020). Adaptive vocational interest diagnostic: Informing and improving the job assignment process. *Military Psychology*, *32*(1), 91–100.

Liao, H.-Y., Armstrong, P. I., & Rounds, J. (2008). Development and initial validation of public domain Basic Interest Markers. *Journal of Vocational Behavior*, *73*, 159–183.

Low, K. S. D., Yoon, M., Roberts, B. W., & Rounds, J. (2005). The stability of vocational interests from early adolescence to middle adulthood: A quantitative review of longitudinal studies. *Psychological Bulletin*, *131*(5), 713–737.

Malhotra, N., Smets, M., & Morris, T. (2016). Career pathing and innovation in professional service firms. *Academy of Management Perspectives*, *30*(4), 369–383.

Mccrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review*, *19*(2), 97–112.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194–216. https://doi.org/10.1037/h0048070

Morris, M. L. (2016). Vocational interests in the United States: Sex, age, ethnicity, and year effects. *Journal of Counseling Psychology*, *63*, 604–615. https://doi.org/10.1037/cou0000164

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & Mccrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *112*(3), 474–490. https://doi.org/10.1037/pspp0000100

Mõttus, R., Sinick, J., Terracciano, A., Hřebíčková, M., Kandler, C., Ando, J., Mortensen, E. L., Colodro-Conde, L., & Jang, K. L. (2019). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *117*(4), e35–e50. https://doi.org/10.1037/pspp0000202

Newman, D. A., & Lyon, J. S. (2009). Recruitment efforts to reduce adverse impact: Targeted recruiting for personality, cognitive ability, and diversity. *Journal of Applied Psychology*, *94*(2), 298–317. https://doi.org/10.1037/a0013472

Noe, R. A., Steffy, B. D., & Barber, A. E. (1988). An investigation of the factors influencing employees' willingness to accept mobility opportunities. *Personnel Psychology*, *41*(3), 559–580.

Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2012). Vocational interests and performance: A quantitative summary of over 60 years of research. *Perspectives on Psychological Science*, *7*(4), 384–403. https://doi.org/10.1177/1745691612449021

Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2017). Interest congruence and performance: Revisiting recent meta-analytic findings. *Journal of Vocational Behavior*, *98*, 138–151.

Oswald, F. L., Hough, L. M., & Zuo, C. (2019). Personnel selection and vocational interests: Recent research and future directions. In C. D. Nye & J. Rounds (Eds.), *Vocational interests: Rethinking their role in understanding workplace behavior and practice*, *SIOP organizational frontiers series* (pp. 129–141). Routledge. https://doi.org/10.4324/9781315678924-7

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big Data in industrial-organizational psychology and human resource management: Forward progress for organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*(1), 505–533. https://doi.org/10.1146/annurev-orgpsych-032117-104553

Paunonen, S. V., Rothstein, M. G., & Jackson, D. N. (1999). Narrow reasoning about the use of broad personality measures for personnel selection. *Journal of Organizational Behavior*, *20*, 389–405.

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

Phan, W. M. J., & Rounds, J. (2018). Examining the duality of Holland's RIASEC types: Implications for measurement and congruence. *Journal of Vocational Behavior*, *106*, 22–36.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, *21*(3), 689–732. https://doi.org/10.1177/1094428117697041

Ralston, C. A., Borgen, F. H., Rottinghaus, P. J., & Donnay, D. A. C. (2004). Specificity in interest measurement: Basic Interest Scales and major field of study. *Journal of Vocational Behavior*, *65*(2), 203–216.

Rounds, J., & Su, R. (2014). The nature and power of interests. *Current Directions in Psychological Science*, *23*(2), 98–103. https://doi.org/10.1177/0963721414522812

Rounds, J., Hoff, K., & Lewis, P. (2021). *O*NET interest profiler manual*. National Center for O*NET Development.

Rounds, J., Smith, T., Hubert, L., Lewis, P., & Rivkin, D. (1999). *Development of occupational interest profiles for O*NET*. National Center for O*NET Development.

Rounds, J., Su, R., Lewis, P., & Rivkin, D. (2013). *Occupational interest profiles for new and emerging occupations in the O*NET system: Summary*. National Center for O*NET Development.

Rounds, J., Tracey, T. J., & Hubert, L. (1992). Methods for evaluating vocational interest structural hypotheses. *Journal of Vocational Behavior*, *40*, 239–259.

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.

Schneider, B. (1987). The people make the place. *Personnel Psychology*, *40*(3), 437–453. https://doi.org/10.1111/j.1744-6570.1987.tb00609.x

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schmidt, F. L. (1974). Probability and utility assumptions underlying use of the Strong Vocational Interest Blank. *Journal of Applied Psychology*, *59*(4), 456–464. https://doi.org/10.1037/h0037324

Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (2011). Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology*, *96*(5), 1055–1064. https://doi.org/10.1037/a0023322

Song, Q. C., Liu, M. Q., Tang, C., & Long, L. (2020). Applying principles of big data to the workplace and talent analytics. In *Big Data in Psychological Research* (pp. 319–344). APA Books.

Song, Q. C., Tang, C., & Wee, S. (2021). Making sense of model generalizability: A tutorial on cross-validation in R and Shiny. *Advances in Methods and Practices in Psychological Science*, *4*(1), 1–17. 2515245920947067.

Stewart-Williams, S., & Halsey, L. G. (2021). Men, women and STEM: Why the differences and what should be done?. *European Journal of Personality*, *35*(1), 3–39.

Stoll, G., Rieger, S., Lüdtke, O., Nagengast, B., Trautwein, U., & Roberts, B. W. (2017). Vocational interests assessed at the end of high school predict life outcomes assessed 10 years later over and above IQ and Big Five personality traits. *Journal of Personality and Social Psychology*, *113*(1), 167–184. https://doi.org/10.1037/pspp0000117

Stroh, L. K. (1995). Predicting turnover among repatriates: Can organizations affect retention rates?. *International Journal of Human Resource Management*, *6*(2), 443–456.

Strong Jr, E. K. (1943). *Vocational interests of men and women*. Stanford University Press.

Su, R. (2020). The three faces of interests: An integrative review of interest research in vocational, organizational, and educational psychology. *Journal of Vocational Behavior*, *116*, 103240.

Su, R., & Rounds, J. (2015). All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields. *Frontiers in Psychology*, *6*, 189. https://doi.org/10.3389/fpsyg.2015.00189

Su, R., Tay, L., Liao, H.-Y., Zhang, Q., & Rounds, J. (2019). Toward a dimensional model of vocational interests. *Journal of Applied Psychology*, *104*(5), 690–714. https://doi.org/10.1037/apl0000373

Super, D. E. (1980). A life-span, life-space approach to career development. *Journal of Vocational Behavior*, *16*(3), 282–298. https://doi.org/10.1016/0001-8791(80)90056-1

Tinsley, H. E. A. (2000). The congruence myth: An analysis of the efficacy of the person–environment fit model. *Journal of Vocational Behavior*, *56*(2), 147–179.

Tracey, T. J. G. (2002). Personal Globe Inventory: Measurement of the spherical model of interests and competence beliefs. *Journal of Vocational Behavior*, *60*(1), 113–172.

Tracey, T. J. G. (2010). Development of an abbreviated Personal Globe Inventory using item response theory: The PGI-Short. *Journal of Vocational Behavior*, *76*(1), 1–15.

Tracey, T. J. G. (2012). Problems with single interest scales: Implications of the general factor. *Journal of Vocational Behavior*, *81*(3), 378–384.

Tracey, T. J. G. (2019). *Personal globe inventory: PGI, PGI-Short, and PGI-Mini. (Manual Version 1.4)* [Unpublished Manual]. https://PGI.asu.edu/technical

Tracey, T. J. G., & Robbins, S. B. (2006). The interest–major congruence and college success relation: A longitudinal study. *Journal of Vocational Behavior*, *69*, 64–89.

Tracey, T. J. G., & Rounds, J. (1996). The spherical representation of vocational interests. *Journal of Vocational Behavior*, *48*(1), 3–41.

U.S. Department of Labor, Employment and Training Administration. (2018). O*NET Data Collection Program, Office of Management and Budget clearance package supporting statement. Part A: Justification.

Van Iddekinge, C. H., Roth, P. L., Putka, D. J., & Lanivich, S. E. (2011). Are you interested? A meta-analysis of relations between vocational interests and employee performance and turnover. *Journal of Applied Psychology*, *96*(6), 1167–1194.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology*, *75*(3), 315–321.

Woo, S. E., Tay, L., Jebb, A. T., Ford, M. T., & Kern, M. L. (2020). Big data for enhancing measurement quality. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychological research* (pp. 59–85). American Psychological Association. https://doi.org/10.1037/0000193-004

Xu, H., & Li, H. (2020). Operationalize interest congruence: A comparative examination of four approaches. *Journal of Career Assessment*, *28*(4), 571–588.

Zickar, M. J., & Min, H. (2019). A history of vocational interest measurement. In C. D. Nye & J. Rounds (Eds.), *Vocational interests: Rethinking their role in understanding workplace behavior and practice, SIOP organizational frontiers series* (pp. 59–79). Routledge. https://doi.org/10.4324/9781315678924-4

Zhou, Z. H. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press.

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

SUPPORTING INFORMATION

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

How to cite this article: Song, Q. C., Shin, H. J., Tang, C., Hanna, A., & Behrend, T. (2022). Investigating machine learning's capacity to enhance the prediction of career choices. *Personnel Psychology*, 1–25. https://doi.org/10.1111/peps.12529

APPENDIX A

BRIEF DESCRIPTION OF MACHINE LEARNING ALGORITHMS USED IN THE STUDY

Neural network

Neural networks mimic natural neurons in the human brain, where each natural neuron takes input signals, activates based on the strength of the signals, and yields a possible answer (Schmidhuber, 2015). The neural network model used in the current paper is a multiple-layer perceptron, a kind of feedforward neural network. Artificial neurons in this model are connected and grouped into multiple layers. At the beginning of model training, the input data (predictors) are processed through the layers using randomly initiated parameters (e.g., weights on each connection between neurons, bias[6] of each neuron). Given the prediction error between the predicted and the observed outcomes, the neural network iteratively updates the parameters until prediction error is minimized. Details about neural networks can be found in Chapter 11 of Hastie et al. (2009).

k-nearest neighbors (*k*-NN)

The *k*-nearest neighbors algorithm uses the unknown data points' *k* nearest neighbors to determine which group the data belongs to, assuming similar data points are close to each other (Fix & Hodges, 1989). For example, if $k = 5$, to predict the outcome value of the unknown data point, 5-NN algorithm calculates the distance (e.g., Euclidean distance) between the predictor scores of the unknown data point and the known data point in the training dataset. It then selects five observations in the training dataset with the lowest distance. The algorithm makes the prediction by taking the average of the outcomes corresponding to the selected observations.

Elastic net

Elastic net is a type of regularized regression algorithm that aims to prevent model overfitting by limiting (penalizing) the magnitude of regression coefficients (Zou & Hastie, 2005). Elastic net regularization combines the advantages of ridge (which is capable of dealing with high multicollinearity) and LASSO (which is capable of performing variable selection; Putka et al., 2018) regularizations by including the ridge and LASSO penalty terms in the algorithm.

Random forest

The random forest algorithm fits and combines multiple decision tree models on bootstrapped samples. It uses a random subset of the predictors to train the nodes in each decision tree (Breiman, 2001), which increases the diversity of each decision tree and helps improve the overall prediction accuracy of the random forest (Putka et al., 2018).