Received: 18 July 2023 | Revised: 11 September 2023 | Accepted: 12 September 2023

DOI: 10.1111/peps.12621

PERSONNEL PSYCHOLOGY WILEY

ORIGINAL ARTICLE

Machine learning applications to personnel selection: Current illustrations, lessons learned, and future research

Michael A. Campion[1] | Emily D. Campion[2]

[1] Purdue University, Daniels School of Business, West Lafayette, Indiana, USA

[2] University of Iowa, Tippie College of Business, Iowa City, Iowa, USA

Correspondence
Emily D. Campion, University of Iowa, 21 E. Market St., Iowa City, IA 52242, USA.
Email: emily-campion@uiowa.edu

[Correction added on September 29, 2023 after first Online publication: Location for author Michael A. Campion was updated from Denver, Colorado.]

Abstract

Machine learning (ML) may be the biggest innovative force in personnel selection since the invention of employment tests. As such, the purpose of this special issue was to draw out research from applied settings to supplement the work that appeared in academic journals. In this overview article, we aim to complement the special issue in five ways: (1) provide a brief tutorial on some ML concepts and illustrate the potential applications in selection, along with their strengths and weaknesses; (2) summarize findings of the four articles in the special issue and provide an independent appraisal of the strength of the evidence; (3) identify some of the less-obvious lessons learned and other insights that researchers new to ML might not clearly recognize from reading the special issue; (4) present best practices at this stage of the knowledge in selection; and (5) propose recommendations for future needed research based on the articles in the special issue and the current state of the science.

KEYWORDS
artificial intelligence, big data, machine learning, natural language processing, selection-methods, selection-validation

1 | INTRODUCTION

A major recent advancement in the study and practice of personnel selection is the use of artificial intelligence (AI) and specifically machine learning (ML) and related methods. Practitioners of personnel selection, being one of the largest areas of practice in Industrial and Organizational (I-O) psychology, have begun to adopt these techniques to improve assessments and other hiring procedures. ML has been less present in the academic research, but what has been published offers a glimpse into the variety of ways we can utilize these advanced methods to inform selection. Examples include scoring candidate essays (Campion et al., 2016), deriving selection content from work history (Sajjadiani et al., 2019), scoring biodata items (Putka et al., 2018), measuring organizational recruitment signals (Banks et al., 2019), and evaluating post-hire job performance (Speer, 2018). Non-selection examples can be found in the special issue edited by Woo et al. on AI, ML, and big data in *Personnel Psychology* on topics such as career choice (Song et al., 2022) and turnover (Min et al., 2022), coping with work-related stressors (Sajjadiani et al., 2022), and predicting occupational accident rates (Kumar & Burns, 2022).

Generating knowledge on ML applications to selection is especially important for two reasons. First, ML is currently being used in practice and we have not accumulated knowledge on the use and value-add of ML in our academic journals, which are intended to hold the most up-to-date and impartial knowledge of our science. Second, ML is used for staffing decision-making, and thus bears on crucial corresponding outcomes like organizational productivity and impact on workforce diversity.

As such, the purpose of this call was to draw out the research based on data from applied settings because we observed that organizations were conducting cutting edge work that was not yet reflected in our academic journals. We solicited research on new procedures, scoring methods, types of data, analytic strategies, and any other selection topics informed by ML. We also allowed submissions of brief study descriptions instead of full papers in order to be more inclusive of work that originated in practice. We received 44 submissions and, with the help of a special editorial board of 12 ML experts from practice and academia (listed in the Acknowledgments in the Appendix) as well as members from *Personnel Psychology*'s standing board, selected 11 projects that we present in four articles. These four articles include two traditional articles and two articles that combine the nine projects based on thematic issues to allow the greatest amount of research to be reported given limited journal space. While our hope is that this overview and the findings in this special issue are useful to organizational psychology and management scholars and practitioners of all skills levels, novice and intermediate users may find it particularly informative.

2 | OVERVIEW OF MACHINE LEARNING IN PERSONNEL SELECTION

AI's home discipline is computer science but is now often associated with data science (DS). DS is an interdisciplinary field that applies a wide range of techniques to data for myriad purposes across many domains. AI is a machine analogy to natural intelligence and refers broadly to the capacity of computers to exhibit or simulate intelligent behavior, such as sensing, learning, decision making, predicting, and so on. AI is commonly but not exclusively used in DS. Other disciplines also use AI, such as statistics and increasingly psychology, as well as many others. Machine Learning (ML) is a method of AI. ML refers to techniques that "learn" patterns in data to make predictions or summarize or score the data. It is not just the domain of DS. For example, ordinary least squares regression is a method of ML used for many years by psychologists when the weights from training are retained and applied to future data to create scores and make predictions. Natural Language Processing (NLP), another method of AI, refers to specialized techniques for analyzing text data, including both the words used and the relationships among words. NLP may be used by DS, but also many other disciplines such as linguistics and psychology. For example, sentiment analysis using word dictionaries is a simple method of NLP common in psychology. Finally, Deep Learning (DL) is a highly complex type of ML that can be used to analyze any type of data (numeric or text). Its use of neural networks and transformers that have many internal layers of analysis make the mathematical operations not possible to examine directly, thus leading to the "black box" characterization. However, one key lesson from this special issue is that while we may not be able to peer under the hood of the more advanced algorithms, by understanding our data at the bivariate level and interrogating the output, we can develop a fairly precise understanding of what the algorithm is doing. We generally use the term ML in this article because it captures the goal of selection by creating models that can learn to make future predictions, and DL and NLP are types of ML.

Broadly speaking, machine learning is either supervised or unsupervised.[1] Supervised machine learning refers to when there is a criterion (e.g., job performance, application scores, or selection decisions by the organization) used to create the model (i.e., select variables to include). In DS, such data are called "labeled data" because each case has a score or decision associated with it that we can use to train the model to predict and from which we can create weights to apply to score future data that are unlabeled. The criterion (label) can be either a continuous score or decision (e.g., passed, rejected, withdrew).

Models built to predict a decision are often called classification models. Researchers can use logit or probit regression to predict dichotomous or multiple categorical outcomes. It is also common in classification models to use tree models. Tree models can be conceptualized much like flow charts with a complex series of choices based on the relevance of each new variable considered that ultimately leads to a prediction. Researchers using classification will usually evaluate validity, which they might describe as accuracy, in terms of the agreement of the decisions between the computer model prediction and the actual decision ("label"). They most frequently use a framework for analyzing statistical accuracy based on Signal Detection Theory, which was developed to measure the ability to detect signals in a pattern of information from random noise or error (Green & Swets, 1966). This is visualized in Figure 1. The figure depicts the relationship between the recruiter decisions and the computer decisions. The horizonal axis is the computer score ranging from low to high, and the vertical axis is recruiter score ranging from low to high. The ellipse is the hypothetical plot of candidates on the two scores. As can be seen, those with higher computer scores tend to get higher recruiter scores, but the relationship is not perfect. The recruiter scores serve as a proxy for expected success on the job. Decisions above the horizonal line are considered correct detections (defined as successful candidates, if hired), while those below are considered incorrect detections (defined as less successful candidates, if hired). Those to the right of the vertical line would be hired based on the computer scores, while those to the left are not hired. This divides the figure into four quadrants that define four types of decisions: (a) correct acceptances (or true positives, TP) are candidates hired who turn out to be successful on the job; (b) incorrect acceptances (false positives, FP) are candidates hired who turn out to be unsuccessful on the job; (c) correct rejections (true negatives, TN) are candidates rejected who would have been unsuccessful on the job, if hired; and (d) incorrect rejections (false negatives, FN) are candidates rejected who would have been successful on the job, if hired.

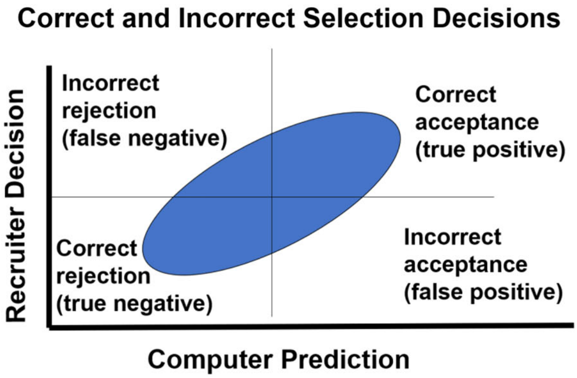
Correct and Incorrect Selection Decisions
Incorrect rejection (false negative)
Correct acceptance (true positive)
Correct rejection (true negative)
Incorrect acceptance (false positive)
Recruiter Decision
Computer Prediction

FIGURE 1 Selection decisions used by signal detection theory accuracy indices.

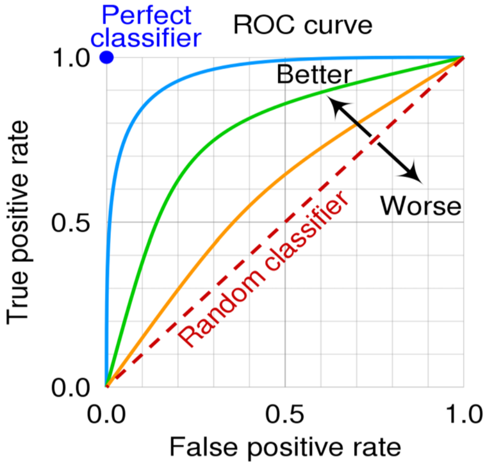

FIGURE 2 Receiver operating characteristic curve.

Researchers will then convert these consequences into metrics to evaluate accuracy based on the Receiver Operating Characteristic (ROC) curve (Fawcett, 2006). The ROC is a plot of the proportion of true positive decisions crossed with false positive decisions. The goal is to index the number of correct decisions compared to the number of incorrect decisions. The ROC Figure 2 ("Receiver operating characteristic", 2023) uses five metrics to analyze the accuracy of the decisions, which are defined below and also illustrated by comparing the equations to the areas in Figure 1:

a. Accuracy = (TP + TN)/total = proportion of correct predictions in total
b. Precision = TP/(TP + FP) = proportion of those hired that are TPs[2]
c. Recall (sensitivity) = TP/(TP + FN) = proportion of TPs of all those who would be successful on the job if hired
d. F1 = harmonic mean of precision and sensitivity
e. AUC (area under curve) = probability (ranging from 0 to 1, with random being .5) that the computer will score the TPs higher than the FPs based on the ROC curve (Figure 2).

[Correction added on September 29, 2023 after first Online publication: Placement of figure 2 was adjusted, a callout to figure 2 was changed to figure 1, and an additional callout to figure 2 was added.]

Some of the authors of the special issue use these metrics, while others use covariation-based metrics to evaluate validity that are more common in selection research (e.g., correlations and regressions). In the latter contexts, model efficiency is usually evaluated based on the correlation, multiple correlation, or multiple correlation squared.

Not all data are labeled (e.g., there is no criterion), however. In these instances, we can use unsupervised machine learning to uncover patterns within the data to summarize it (e.g., identify the topics) or to create measures (e.g., scores) that may be explored for their predictiveness of outcomes in the future. Unsupervised machine learning helps us solve problems through dimensionality reduction to cluster data in meaningful ways. The term "topic modeling" may be familiar and refers to using unsupervised machine learning on text data where the researcher determines an optimal number of topics based on how the model fits the unstructured text data using metrics such as coherence scores, as well as interpretability of the topics (Valtonen et al., 2022).

Table 1 provides an overview of the ML techniques that might be applicable to various selection situations to illustrate the potential value of ML. Many of the studies in the special issue illustrate these applications, but the table is not based solely on the special issue and the potential pros and cons are not meant to be a complete list or describe the specific articles in the special issue. The table leads to several key observations. First, there are many situations in

TABLE 1 Overview of illustrative machine learning applications in personnel selection.

Situation	Data types	Potentially applicable ML techniques to create model	Some potential pros	Some potential cons
1. Scoring resumes and employment applications	Numeric and text	Natural language processing (NLP) to analyze the text data and create numeric scores, and then combine with numeric data in applications using a wide range of supervised ML to predict selection scores and/or decisions.	Most common current application of ML in selection and can save time and cost in high-volume contexts.	Limited by quality of criterion data (e.g., past recruiter prescreening decisions) and special steps may be required to interpret algorithm (avoid "black box").
2. Scoring constructed responses to assessments (e.g., interviews, write-in test answers)	Primarily text	NLP to analyze the text to identify the content, sometimes using unsupervised ML, to create numeric scores to assess candidates, or to predict other outcomes (e.g., interviewer ratings) using supervised ML (e.g., Koenig et al., 2023, Studies 2–4).	Reliably scoring unstructured responses and giving equal consideration to all candidates (unlike typical human review).	Requires criterion data to train models to predict and may require special steps to interpret algorithm (if not topic modeling).
3. Combining scores to increase prediction	Numeric and text	Wide range of supervised ML to optimally predict outcomes (e.g., job performance) and may include many variables and curvilinear relationships (e.g., Koenig et al., 2023, Study 5; Landers et al., 2023).	May increase prediction compared to regression, especially when sample/ parameter (n/k) ratio is low.	Increase may be small in large samples given the added complexity.
4. Combining scores to reduce subgroup differences	Numeric and text	Pareto Optimal or similar ML, or make adjustments to analysis or data, to increase multiple outcomes simultaneously such as diversity and performance (e.g., Zhang et al., 2023, Studies 1–3).	May reduce adverse impact without much loss to validity.	May require smaller weights on cognitive predictors, capitalize on chance with small samples, make little improvement, reduce validity, and create prediction bias.

(Continues)

TABLE 1 (Continued)

Situation	Data types	Potentially applicable ML techniques to create model	Some potential pros	Some potential cons
5. Creating test questions	Text from existing questions to create model or use existing pre-trained language models	DL NLP (transformers using neural networks) to learn word patterns in past questions to create models that can produce similar questions or be used across similar questions (e.g., Hernandez & Nie, 2022; Koenig et al., 2023, Study 1) or to use available models pre-trained on large public language data sets.	Saves time and effort to write new questions; can be tuned to produce variety of items.	May require large training samples to create own model but can use pre-trained models for many purposes; untuned models may produce highly similar items; requires researcher judgment to select questions, and may be less effective for some types of items (e.g., situational).
6. Analyzing jobs to determine requirements	Text from job descriptions or job analysis data.	NLP to extract content to create scores used to predict job analysis ratings with supervised ML (e.g., Koenig et al., 2023, Study 6).	Saves time in job analysis, especially for low volume applications.	Dependent on having accurate job description and task text data as inputs, may only identify obvious requirements, and may be simpler methods if jobs can be linked to external frameworks (e.g., O*NET, ESCO).
7. Inferring skills and personality from narrative application information	Primarily text	Theoretically created closed-word dictionaries or NLP-created open-word dictionaries to score skill or personality in text data usually collected for other purposes to create scores (e.g., letters, statements of interest, responses to interviews or questions on applications, etc.).	Technically simple to use, many dictionaries available, longest history in I-O research compared to other ML, and can be used in many contexts.	May not be measures of constructs of interest available, evidence of construct validity required, and may predict less well than more sophisticated ML.

which ML may enhance personnel selection. These include prescreening in addition to primary selection procedures; scoring narrative as opposed to numeric data; scoring constructed responses (e.g., write-in comments) as opposed to structured responses (e.g., multiple choice); making tradeoffs and maximizing prediction; reducing subgroup differences; creating test questions; and deriving job requirements.

Second, there are many potential pros, but also some meaningful potential cons. For example, ML will increase efficiency in large-scale applications, but may not in small scale applications enough to justify the increased complexity, which is a tradeoff in all customized selection systems. Moreover, the prediction improvement of these procedures may be small, especially in large samples compared to traditional or more well-known procedures like regression. The value of ML may depend more on how it can help score data rather than increase prediction, such as text data and constructed responses. The reduction in subgroup differences may or may not be possible, and could actually create prediction bias, but more needs to be known on this front.

Third, there might also be other uses for ML, such as helping create assessment items or making other tasks easier like job analysis. Fourth, some applications like word dictionaries are actually simple ML that researchers are likely already familiar with, and dictionaries are easily available to novices. There is much yet to be learned and these are just illustrations to stimulate future research.

3 | FINDINGS OF THE SPECIAL ISSUE REGARDING THE STATE OF THE SCIENCE ON ML IN SELECTION

This is a brief description of the studies and some of their findings, which is necessary to identify the key conclusions of the special issue and helpful because two articles combined several studies. We also wanted to provide an independent appraisal of the evidence. Table 2 shows the authors of each study, the title, and the key findings. We follow this table with observations as to the findings' implications for the state of the science.

Taken together, the articles lead to two overarching conclusions. First, ML can have many potential contributions to our research and practice, such as saving time and effort, measuring new data types like text and other constructed responses, sometimes improving prediction, and balancing priorities among multiple objectives. Second, there are several important limitations, such as small expected improvement in prediction in many cases, imperfect or ineffective solutions to tradeoffs between validity and subgroup differences, added computational complexity, and decreased interpretability.

TABLE 2 Some key findings of the special issue articles.

Authors	Title	Key findings
Individual articles		
Hernandez and Nie (2022)	The AI-IP: Minimizing the Guesswork of Personality Scale Item Development Through Artificial Intelligence	ML models can be used to efficiently create personality item pools with reliability and construct validity similar to traditional methods.
Landers et al. (2023)	A Simulation of the Impacts of Machine Learning to Combine Psychometric Employee Selection System Predictors on Performance Prediction, Adverse Impact, and Number of Dropped Predictors	In a large-scale set of simulations, ML does not greatly predict beyond traditional methods like regression unless samples are small relative to parameters (i.e., an n-to-k ratio of less than 3 for scales or 14 for items), but there are many nuanced findings where ML may be better such as when item-level models are used

(Continues)

TABLE 2 (Continued)

Authors	Title	Key findings
Composite Article 1: Improving measurement and prediction in personnel selection through the application of machine learning (Koenig et al., 2023)		
Study number (study authors within composite)	Study title	Study's key findings
Study 1 (Koenig et al., 2023)	Algorithmic Construct Generalizability: Scoring Novel Open-Ended Prompts with Deep Learning Trained on Alternative Prompts	ML algorithms can generalize to scoring responses from novel prompts, especially when the assessment is the same, when content is similar, and when training data are seeded.
Study 2 (Yankov and Speer)	Comparing Three Machine Learning Algorithms for Scoring Assessment Center Text Data	ML can score constructed responses to assessment center exercises with as much reliability and criterion-related validity as humans or better, and there are some differences by ML methods.
Study 3 (Hardy et al.)	Using Artificial Intelligence to Make Better Pre-Hire Assessments	ML can complement existing assessments by scoring open-ended questions as well as humans, but more efficiently, with slight criterion-related validity gains and also only slight adverse impact.
Study 4 (Liu et al.)	Developing and Validating Automated Scoring for an Audio Constructed Response Simulation	ML can score audio constructed responses to a simulation assessment with as much reliability and criterion-related validity as humans, and incremental validity beyond existing assessments.
Study 5 (Sun et al.)	Practical ML Algorithms for Selection Assessment Scoring: A Use Case Report on Multi-Outcome Prediction	ML can be used to predict multiple outcomes simultaneously (e.g., productivity and turnover), but the gains over traditional methods may only be marginal with highly structured data (e.g., multiple-choice).
Study 6 (Lebanoff et al.)	Naturalistic Extraction of Knowledge, Skills, Abilities and Other Characteristics using NLP with Human-Level Proficiency	ML algorithms can be trained to identify knowledge, skills, abilities, and other characteristics from tasks and other job descriptive text as well as humans.
Composite article 2: Reducing subgroup differences in personnel selection through the application of machine learning (Zhang et al., 2023)		
Study number (study authors within composite)	Study title	Study's key findings
Study 1 (Zhang et al., 2023)	Are Fairness-Aware ML Algorithms Really Fair? Predictive Bias of Using ML in Personnel Selection	Fairness-aware ML algorithms that statistically eliminate subgroup differences must create predictive bias mathematically, which may reduce validity and penalize high-scoring racial minorities.
Study 2 (Hickman et al.)	Oversampling Higher-Performing Minorities During Machine Learning Model Training Reduces Adverse Impact Slightly but Also Reduces Model Accuracy	Statistically removing subgroup differences in the training data only slightly reduces adverse impact ratios of the resulting ML model but also slightly reduces model accuracy (convergent validity in this study).
Study 3 (Song et al.)	Multi-Objective Optimization for Personnel Selection: A Guide, Tutorial, and User-Friendly Tool	Presents a tool for achieving optimization (Pareto optimal) for up to three objectives, which has many applications in selection.

4 | LESSONS LEARNED

At this early state of the science, it is important to identify some of the less-obvious observations and insights that researchers new to ML might not clearly identify from reading the special issue. In Table 3, we discuss some of the lessons learned.

TABLE 3 Some lessons learned.

1. Relevance of estimating the reliability of ML models. ML researchers do not always analyze reliabilities, but they should for all the same reasons we do otherwise. Alpha is not always relevant because the goal is prediction as opposed to creating a homogeneous measure of a construct. However, if the ML researcher wants to make statements about construct inferences from the algorithm, then alpha may be relevant, although this is perhaps more complex to estimate (as noted below). Test-retest reliability will likely be relevant in most cases to evaluate stability. Alternate/parallel forms reliability may also be useful if different versions are used (e.g., different prompts, which Koenig et al., 2023, Study 1, call algorithmic construct generalizability).

2. Influence of reliability on prediction. If models are trained against a criterion, then the correlation with that criterion is influenced by the reliability of the criterion. For example, achieving a correlation with human ratings as high as the interrater reliability is a common goal in ML. Although this is usually a limitation, several studies in the special issue found the ML algorithm to correlate with a human-rating criterion higher than the interrater reliability of the criterion (Koenig et al., 2023, Studies 2, 4, and 6). This is possible based on the classic psychometric formula for correcting observed correlations for reliability (e.g., Ghiselli et al., 1981).[3] This is also possible if the model is more reliable than the criterion on which it is trained, such as when the model more consistently measures the content (e.g., how past work experience is counted compared to human judgements) or if the model has higher validity by capturing more relevant content (e.g., more aspects of work experience). Moreover, comparisons may not be accurate due to differences in other factors such as the use of different samples, capitalization on chance, or other methodological reasons.

3. Criterion-related validity of ML algorithms in employment applications. Evidence currently suggests that we may be able to build ML models that demonstrate equivalent or better criterion-related validity. Criterion-related validity is essential to supporting the utility of an employment assessment. One common initial type of validity evidence in ML is to demonstrate the model's replicability of human scores. Several studies have shown that ML models can be developed to achieve criterion-related validity by replicating human ratings (Koenig et al., 2023, Studies 2, 3, 4, and 6). Another more direct measure of criterion-related validity is to demonstrate the model's predictive validity of performance and related organizational outcomes (e.g., turnover). Several studies in this special issue provide this kind of evidence (Koenig et al., 2023, Studies 2–4), as well as outside the special issue (e.g., Campion et al., in press).

4. Construct validity when scoring items. The potential relevance of internal consistency as construct information through alpha reliability is already noted above. Another concern is that building an ML model based on test items rather than total test scores or scales violates the inferences we can make about the construct validity, which are based on the total test score or scale. Landers et al. (2023) suggest that you cannot infer construct validity from a scale to its individual items but, if all the items are included in the model and scored separately, to what extent can we infer the model has the construct validity of the scale or total score that includes all the items? This issue may be even more complex if items differ not only in the weights they receive in a model but also if curvilinear relationships of items scored in the model. Potentially, some variance in total scores will be due to the construct and some due to item-level factors.

5. Improvement in prediction. When examining ML algorithms compared to traditional methods with large samples, the improvement in prediction is probably not going to be large in most instances. This is illustrated by Koenig et al. (2023) Studies 3–5, and especially by Landers et al. (2023). These studies were relevant for the special issue specifically because they *did not* report exceptional findings, which leads to overestimates in our literature and future meta-analyses. Instead, they illustrate realistic expectations that the improvements may be modest in some but not all large-scale circumstances. However, although the average improvement in prediction may not be large in these studies, it may be meaningful in a specific situation. As demonstrated in the Landers et al. (2023) simulation, the improvement may vary widely based on algorithm choice, number of variables, selection ratios, design choices, or analytic goals, as well as when *n*/*k* is small (see below). In DS, it is common practice to explore a range of algorithms and then select the best because they can vary meaningfully. Thus, these methods should be explored because they may make a large difference in individual contexts. A notable caveat to this point is when using ML to score new types of data. For example, using ML to score text data to combine with existing assessments may offer consequential gains in validity because the text data includes additional job-related constructs (Campion et al., in-press).

6. Role of *n*versus *k*. An important finding is that ML may help when the sample (*n*) is small compared to the number of parameters (*k*), which may occur if scoring a larger number of variables (e.g., a large amount of candidate application information, item-level scores, components of a simulation, individual text variables, etc.), as well as when samples are small, which is more common in selection contexts than in most DS contexts. This is demonstrated by Landers et al. (2023)'s simulation. They conclude that ML is better when the *n*/*k* ratio is 3 or smaller for scales, or 14 or smaller for items. In such instances, ML might be able to cross-validate better than traditional statistics.

(Continues)

TABLE 3 (Continued)

7. Bias-variance tradeoff. The goal of ML is often the prediction of the total scores, not inferences about parameters of the model like the effect of given variables in a regression. DS is a science about mathematics and improving prediction and using human-generated data toward this end rather than making inferences about psychological constructs; meanwhile, psychology is a science focused on humans aimed at making inferences about psychological constructs and using ML as a tool to achieve this goal. In fact, many of the ML models do not have individual variables; the features are engineered by the process itself. With a focus on prediction, an important concept in ML is the "bias-variance tradeoff." It essentially refers to the tradeoff between reducing bias in parameters by tightly fitting (specifying) the model, which increases the variance in parameter estimates across samples. ML models may intentionally introduce some bias in estimates (e.g., like how ridge regression controls extreme weights) in order to achieve less variance (and more generalizability) in predictions. Underprediction can also result if models are under-fit because they do not fully utilize the data. Thus, both over-fitting and under-fitting can result in underprediction, and various ML models balance this tradeoff in different ways to find an optimal solution (Putka et al., 2018). Various techniques such as tuning hyperparameters may be used to control the ML process by constraining or scaling the complexity. Effectively managing the bias-variance tradeoff is a fundamental issue in ML. It may especially matter when sample sizes are small, thus explaining the benefit of ML in small samples, as noted above.

8. Scoring new types of data is the greatest opportunity. We believe the greatest opportunity afforded by ML is the ability to score data that have been relatively neglected—text data, as well as other unstructured responses—as opposed to improving prediction from traditional data such as the dominant multiple-choice and rating scale responses or other structured numeric responses in current assessments. Such data offers rich new information on candidate skills and other capabilities not always considered in hiring decisions, often because doing so requires a labor-intensive rating process. This information, if scored objectively and included in the selection decisions, might improve prediction, and perhaps reduce subgroup differences if the information shows smaller subgroup differences (Arthur et al., 2002; Campion et al., in press). This is illustrated by several studies in the issue (Koenig et al., 2023, Studies 3 – 5; and suggested by Landers et al. 2023).

9. Importance of word count. With regard to text responses, many people may consider response length as a confound factor that only reflects verbosity. However, it is often predictive of human scores. We contend that it may instead represent *the amount of content in the answers*. Candidates with more skills have more to describe. ML models should consider it explicitly. Its effects should not be hidden, for example, by just showing that text variables predict without knowing how much response length matters. It may not be a confound, but a part of total understanding (see also Koenig et al., 2023, Studies 3 and 4). If it is viewed as a confound, one potential solution is to control the allowed response length in the study design.

5 | EMERGING BEST PRACTICES

Because machine learning continues to rapidly develop, we provide a list of *emerging* best practices based on the special issue that may and likely will evolve over time. Table 4 presents best practices at this stage of the knowledge in selection.

TABLE 4 Emerging best practices at this stage of knowledge in personnel selection.

1. Try alternative ML methods. We recommend comparing to simple regression or other statistical procedures known to, and trusted by, selection researchers. Try alternative ML methods as well because some may work better than others in a given context as demonstrated by many of the articles in this issue, and each method comes with its own strengths and limitations. For example, if you are unsure about interactions and curvilinear relationships, then explore ML models that can detect such effects. As noted, trying many different algorithms is a routine practice in DS and should become commonplace in selection, as well. One distinction, however, is that DS selects the best-performing model. This practice is concerning in psychology because it may be viewed as cherry-picking, even though they cross-validate. Therefore, be purposeful and explain your rationale as opposed to simply trying any alternative method and retaining only one. Moreover, exploring different methods can improve trustworthiness of findings due to triangulation.

(Continues)

TABLE 4 (Continued)

2. Use the simplest necessary ML methods. Just like in traditional statistics, simpler ML methods may predict almost as well as more complex ML algorithms. As an added benefit, they may be easier for selection researchers to master and will be more explainable. For example, simple dictionaries can perform relatively well compared to deep learning models, although this may not be the case in the future with advancements in pre-trained language models.

3. Measurement and design are paramount. At times, ML has been viewed as having the ability to solve issues with data. ML will not fix inherent flaws in your data and is as vulnerable to the adage "garbage in, garbage out" as any other method (often described as "landfill in, landfill out" in DS due to the large amounts of data often analyzed). Pay as close attention to measurement and design as we have historically.

4. Be sure to focus on the prediction of job performance and not simply replicating human ratings. Demonstrating that the algorithm can score applicant data as well as humans by predicting human ratings is important when the goal is to increase efficiency and maybe replace a human rater. Nevertheless, the most important goal in personnel selection usually is criterion-related validity demonstrated by predicting job performance or other outcomes of relevance to the organization. This should be a priority in the development of ML models in selection.

5. Always look at subgroup differences. Evidence to date suggests two conclusions: (a) ML may not increase subgroup differences, and (b) ML is not a panacea for reducing subgroup differences while maintaining expected validity, but it may help a small amount. Be watchful of ML vendors who promise otherwise because it may not be correct or they might use procedures that are effectively illegal because they do not know the laws (such as score adjustments that are the same as within-group norming, which is prohibited by the 1991 Civil Rights Act in the U.S.).

6. Pay attention to interpretability. This is a major difference in selection applications versus other DS applications of ML. Candidates, hiring officials, and courts demand to know what is being measured when it comes to decisions about people. A small increment in validity by using more complicated methods might not be worth it if simpler methods can do almost as well and can be explained much easier. If there are legal actions and experts are called upon to explain the algorithms, the Daubert Federal Rule of Evidence 702 for expert witnesses requires a determination of the "reliability" of the evidence in the opinion of the court, which may depend heavily on interpretability to laypersons (Daubert v. Merrell, 1993).

7. Learn ML quality indices. As explained earlier, data scientists use Signal Detection Theory indices rather than correlations to evaluate accuracy in classification contexts (as opposed to continuous score contexts where correlational indices may still be used). Moreover, these indices may also complement correlations by being more explainable in terms of practical effects. For example, precision, as the proportion of those hired that are true positives to true positives and false positives, might be useful for predicting job performance, similar to expectancy tables. However, recall, or the proportion of true positives out of true positives and false negatives, might be useful for predicting low probability events like turnover or accidents.

8. Let us not standardize our approaches too soon. There is much to learn and standardizing too quickly might stifle discovery. For example, just because one previous study in our area does something one way, that should not become the expectation for all future studies. Authors should explicitly justify their choices and use online supplements and repositories (e.g., Open Science Framework [OSF]) to present technical details and alternative analyses. Experimentation with different methods is key at this stage of the science as we endeavor to better understand ML, it's place in our domain, and develop best practices.

6 | FUTURE RESEARCH

Table 5 presents some suggestions for future needed research based on the special issue and the current state of knowledge.

TABLE 5 Future research suggestions.

1. Scoring new types of data (e.g., text data):a) Can ML help uncover new constructs? In a review of the literature on the use of text analysis in the employment context, Campion and Campion (2020) found 28 studies relevant to selection. The constructs measured included individual skills, personality or orientations, and organizational or job characteristics. ML should be very useful for discovering new constructs, not only because of its data mining capabilities, but especially because it can analyze new types of data such as text.

(Continues)

TABLE 5 (Continued)

b) Will such data improve prediction of organizational selection decisions and especially job performance? There is some emerging evidence in the special issue (e.g., Koenig et al., 2023, Study 4) and elsewhere (Campion et al., in press). c) Can ML score other non-traditional types of data? The biggest opportunity is text data because of how much text data candidates create when they apply to jobs (e.g., applications, essays, interview responses) and its underutilization and costly human scoring. Moreover, as demonstrated in Koenig et al (2023).'s Study 4, audio-based (voice) data can be easily transcribed for analysis using automatic speech recognition products. However, scoring other types of data is also possible and a big potential opportunity. For example, aside from proprietary commercial products, there is some recent evidence in our literature that ML can be used to score paraverbal (e.g., tone and pitch) and nonverbal features (e.g., facial expressions and posture) of human communication to measure personality (e.g., Hickman et al., 2022). Conceivably, scoring other types of candidate information for personnel selection is also possible or easier with ML. Examples include, but are not limited to: tracing process steps or sequences in constructed responses like problem solving tasks or performing a procedure; scoring images in constructed responses like drawings, figures, and diagrams; scoring other visual information like candidate dossiers of creative work (e.g., pictures, designs, music) (e.g., Santos et al., 2021); scoring responses in other languages and foreign language skill; scoring presentations or other communication skills that simultaneously consider visual and audio data; scoring interactions in dyads like interviews or group exercises; or scoring the correctness of computer coding produced by software engineering candidates in response to hiring assessments like job samples, which is a growing need given the increasing hiring of such candidates and the difficulties of scoring by laypersons. ML may also more easily enable accommodations for candidates with disabilities like converting text to voice and vice versa and reading or displaying sign language.
2. Reductions in subgroup differences through ML:a) Will ML reduce subgroup differences without reducing validity? The Uniform Guidelines on Employee Selection Procedures stipulate that organizations must conduct a search for, and consider using, alternatives that reduce impact but not at the cost of reducing validity (Section 3B). b) If the answer is yes, what are the constructs, and will they improve understanding of how diverse candidates demonstrate job-related attributes? For example, do they demonstrate skills using different work and life experiences than other candidates? c) If the constructs are the same, might ML contribute to the reduction of subgroup differences because it assigns optimal weighting, or because it can more efficiently allow scoring of alternative response formats that are less cognitively loaded such as constructed responses (e.g., narrative descriptions)? d) On the other hand, could an ML model have greater subgroup differences than the data on which it is trained if it is more reliable or better measures the constructs?
3. Continued research on when improved prediction from ML might be worth the effort:a) What predictors might benefit most from ML? For example, perhaps measuring personality from text data would be less fakable than self-report personality inventories. Substantial text data are usually submitted as part of an application such as descriptions of work experiences, answers to application questions, letters of reference, and so on. Maybe ML-enabled text analysis can measure personality-related expressions more subtly than self-reports without requiring additional effort by candidates or hiring managers. b) What contexts might benefit from ML? Aside from contexts involving lower *n*/*k* ratios (Landers et al. 2023), curvilinear relationships, or when a decision tree ML model is better (e.g., classification contexts or categorical criteria), perhaps ML would be beneficial when research is truly exploratory (e.g., data mining situations). Putka et al. (2018) describe situations with structured data where various techniques may help.
4. Approaches to interpretation:a) What are the best approaches for interpretation with basic ML like dictionaries or algorithms that extract features? Should interpretation merely be based on examining the words or the features scored, or should research explore the use of subject matter experts to summarize the features or use visuals to more effectively communicate the content to users? The most rigorous approach might use all three, but these simpler methods might not require such significant effort to demonstrate or support interpretation. b) What are the best approaches for interpreting deep learning (neural networks)? Should interpretation be based on inputs and outcomes rather than trying to explain the math? For example, the reasonableness of the inputs to the model and what is specifically withheld (e.g., activities or accomplishments that are gender related) certainly bear on interpretation. Construct validity evidence based on correlations with measures of known constructs will be relevant, where available.

(Continues)

TABLE 5 (Continued)

5. How to do content validation with ML:a) How can ML models demonstrate content validity? Although ML models have historically depended on criterion-related validation, the Uniform Guidelines recognize content validation as a viable alternative (Section 5a), and it may be a useful approach if criterion-related validity is not possible to examine or weak. Moreover, validating only against the organization's past selection decisions may not be useful evidence if not done well (e.g., based on selections in the past by untrained recruiters) or when past decisions are under scrutiny, such as in many litigation contexts where validation evidence becomes critical. Feature importance output from many ML programs, as well as the input to the models, will be helpful for content validation. Also, the ability of ML to identify the relationships between individual components of selection procedures and criteria might inform researchers' understanding of how content validity relates to predictive validity, which is not as established as might be assumed (Murphy, 2009; Weekley et al., 2019).

6. Construct validation: Although not a focus of DS, this is very important to I-O Psychologists and others involved in selection. Typical construct validation by convergent and discriminant relationships is clearly applicable just like any other measurement context in psychology. Nevertheless, important research questions unique to ML remain.
a) As noted earlier, how does scoring individual items of assessments in many ML models influence construct validity inferences?
b) More broadly, are new construct validity issues becoming important, such as the construct validity of test items created by ML (Hernandez & Nie 2022), whether using different algorithms influences construct validity (Koenig et al., 2023, Studies 1 and 2), or the influence of Pareto optimization on the construct validity of an assessment battery (Zhang et al., 2023, Study 3)?

7. Candidate acceptance:a) How can acceptance of ML be increased among candidates and the general public? This is not just how they react now because there is ample evidence they are interested and concerned, but instead how to improve acceptance and reduce concerns. For example, conducting experiments that manipulate instructions might show the effectiveness of communication to candidates such as explaining what ML measures (e.g., job-related skills) and how it does so more objectively and with less potential bias than human judgements, as it has done for other assessments (e.g., Truxillo et al., 2002).
b) Will candidates attempt to artificially increase their scores if they are aware ML is being used such as by inserting terms or phrases in their applications that they believe are scored by ML?

8. Technical details of ML in selection, such as what ML algorithms work best and when:a) Does deep learning offer sufficient improvement in prediction to justify the added complexity and lack of direct interpretability?
b) Do classification approaches (e.g., tree-based) provide a superior alternative to logit/probit in common selection contexts?
c) Does using ML to better model curvilinear relationships improve prediction enough and can we explain it?
d) Does modeling item-level predictors offer enough value beyond total scores, especially given construct validity and cross-validation issues?
e) How does the non-selection corpus a model is trained on influence its utility within selection? For example, BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2018) was pre-trained from BooksCorpus and English Wikipedia and while this model has been used in some selection scenarios, perhaps its development on a non-selection-related corpus bears on the results.

9. Leveraging, building, and managing generative text models.a) How can we utilize generative text models such as ChatGPT and others to develop assessment items or support other human resource functions (e.g., job analyses and descriptions)? In this special issue, Hernandez and Nie (2022) used GPT-2 to create a pool of one million potential personality items from which they built a model to identify a subset that they validated. We can envision such a process might be used for other types of assessments and human resource tasks.
b) How do we identify and manage potential misuse of generative text models and other sources that may influence how candidates prepare applications or respond to assessment questions? Applicants may use these models to help write essays, answer questions, practice and prepare for assessments, create other materials, and maybe other uses. We may need to build models to flag such submissions, generate guidance on what is and is not unethical use of generative models in employment applications, and develop processes to educate and direct applicants on these issues.

7 | CONCLUSION

ML may be the biggest innovative force in selection since the invention of employment tests and key insights such as the impact on population subgroups (and employment laws) and validity generalization. There could not be a more pivotal time in the current state of the science of selection. We hope this special issue will help promote rapid scientific development on this new frontier.

ACKNOWLEDGMENTS

The authors would like to acknowledge Daniel Putka and Richard Landers for their valuable feedback on this work.

CONFLICT OF INTEREST STATEMENT

We have no known conflict of interest to disclose.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Emily D. Campion https://orcid.org/0000-0003-1555-2089

ENDNOTES

[1] There are at least two other sub-types: (1) semi-supervised machine learning, which combines both supervised and unsupervised; and (2) reinforcement learning, which allows the model to learn right and wrong answers thus teaching (reinforcing) the model over time (e.g., IBM's Watson). For parsimony, we maintain the strict distinction between supervised and unsupervised.

[2] This is the same as the metric used in the well-known Taylor Russell Tables in personnel selection research.

[3] The true correlation is equal to the observed divided by the square root of the reliabilities. So if the human raters had a reliability of .6, the observed correlation could be up to .79 if the observed correlation was .6 (assuming computer scores had a reliability of 1). .79 = .6/sqrt(.6*1)

REFERENCES

Arthur, Jr, W., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology*, *55*, 985–1008. https://doi.org/10.1111/j.1744-6570.2002.tb00138.x

Banks, G. C., Woznyj, H. M., Wesslen, R. S., Frear, K. A., Berka, G., Heggestad, E. D., & Gordon, H. L. (2019). Strategic recruitment across borders: An investigation of multinational enterprises. *Journal of Management*, *45*(2), 476–509. https://doi.org/10.1177/0149206318764295

Campion, E. D., Campion, M. A., Johnson, J., Carretta, T. R., Romay, S., Dirr, B., Deregla, A., & Mouton, A. (in press). Using natural language processing to increase prediction and reduce subgroup differences in personnel selection decisions. *Journal of Applied Psychology*.

Campion, E. D., & Campion, M. A. (2020). Using computer-assisted text analysis (CATA) to inform employment decisions: Approaches, software, and findings. *Research in Personnel and Human Resources Management*, *38*, 287–327. https://doi.org/10.1108/S0742-730120200000038010

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*, 958–975. https://doi.org/10.1037/apl0000108

Daubert v. Merrell Dow Pharmaceuticals, In., 509 U.S. 579 (1993).

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

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. WH Freeman.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In D. E. Losada, & J. M. Fernández-Luna (Eds.), *Advances in information retrieval. ECIR 2005. lecture notes in computer science* (Vol. 3408, pp. 345–359). Springer. https://doi.org/10.1007/978-3-540-31865-1_25

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley.

Hernandez, I., & Nie, W. (2022). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. Advance online publication. *Personnel Psychology*, https://doi.org/10.1111/peps.12543

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

Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., Speer, A., Hardy iii, J. H., Gibson, C., Frost, C., Liu, M., McNeney, D., Capman, J., Lowery, S., Kitching, M., Nimbkar, A., Boyce, A., Sun, T., … Newton, C. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. *Personnel Psychology*. Advance online publication. https://doi.org/10.1111/peps.12608

Kumar, L. S., & Burns, G. N. (2022). Determinants of safety outcomes in organizations: Exploring O*NET data to predict occupational accident rates. *Personnel Psychology*. Advance online publication. https://doi.org/10.1111/peps.12560

Landers, R. N., Auer, E. M., Dunk, L., Langer, M., & Tran, K. N. (2023). A simulation of the impacts of machine learning to combine psychometric employee selection system predictors on performance prediction, adverse impact, and number of dropped predictors. *Personnel Psychology*. Advance online publication. https://doi.org/10.1111/peps.12587

Min, H., Yang, B., Allen, D. G., Grandey, A. A., & Liu, M. (2022). Wisdom from the crowd: Can recommender systems predict employee turnover and its destinations? *Personnel Psychology*. Advance online publication. https://doi.org/10.1111/peps.12551

Murphy, K. R. (2009). Content validation is useful for many things, but validity isn't one of them. *Industrial and Organizational Psychology*, *2*(4), 453–464. https://doi.org/10.1111/j.1754-9434.2009.01173.x

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/10.1177/1094428117697

Receiver operator characteristic. (2023). In Wikipedia. https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Sajjadiani, S., Daniels, M. A., & Huang, H.-C. (2022). The social process of coping with work-related stressors online: A machine learning and interpretive data science approach. *Personnel Psychology*. Advance online publication. https://doi.org/10.1111/peps.12538

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*, 1207–1225. https://doi.org/10.1037/apl0000405

Santos, I., Castro, L., Rodriguez-Fernandez, N., Torrente-Patino, A., & Carballal, A. (2021). Artificial neural networks and deep learning in the visual arts: A review. *Neural Computing and Applications*, *33*, 121–157. https://doi.org/10.1007/s00521-020-05565-4

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*. Advance online publication. https://doi.org/10.1111/peps.12529

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, *71*(3), 299–333. https://doi.org/10.1111/peps.12263

Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2002). Selection fairness information and applicant reactions: A longitudinal field study. *Journal of Applied Psychology*, *87*, 1020–1031. https://doi.org/10.1037/0021-9010.87.6.1020

Valtonen, L., Mäkinen, S. J., & Kirjavainen, J. (2022). Advancing reproducibility and accountability of unsupervised machine learning in text mining: Importance of transparency in reporting preprocessing and algorithm selection. *Organizational Research Methods*. Advance online publication. https://doi.org/10.1177/10944281221112494

Weekley, J., Labrador, J., & Campion, M. A. (2019). Job analysis ratings and criterion-related validity: Are they related and can validity be used as a measure of accuracy? *Journal of Occupational and Organizational Psychology*, *92*, 764–786. https://doi.org/10.1111/joop.12272

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

How to cite this article: Campion, M. A., & Campion, E. D. (2023). Machine learning applications to personnel selection: Current illustrations, lessons learned, and future research. *Personnel Psychology*, *76*, 993–1009. https://doi.org/10.1111/peps.12621

APPENDIX

Acknowledgments

We are incredibly thankful to our Special Issue Editorial Board. This team comprised the top machine learning experts in our field who graciously gave their time to provide substantive reviews to help authors improve their research in this special issue. Special Issue Editorial Board Members are listed below in alphabetical order.

Alexis Fink is Vice President, People Analytics and Workforce Strategy at Facebook. She has spent more than two decades leading Talent Analytics, Workforce Strategy, Talent Management and large-scale organizational change teams at global organizations, including Facebook, Microsoft and Intel. She has done extensive work applying advanced analytical methods to human capital problems.

Robert E. Gibby is IBM's Chief Talent Scientist responsible for data, technology, assessment, and learning solutions that help people choose the best candidates for opportunities across the employee lifecycle. He was one of three winners of IBM's 2016 Cognitive Build from over 8000 team submissions for ways to leverage IBM's Watson AI capability.

Genetha Grey manages the Data Science & Research group of the Employee Success Strategy & Analytics (ESSA) team at Salesforce. ESSA focuses on optimizing the employee experience through the use of data. Previously, she was an analytics research scientist at Intel Corporation where she worked on both engineering and people analytics applications.

Nick Koenig is the Director of Global Selection and Assessment at Walmart where he is responsible for the team that designs, validates, and implements the assessments and interviews that touch over 5 million candidates each year. Prior to working in this role, he was the Director of Data Science at Modern Hire. He also held previous roles as a Senior Research Scientist in Consumer Research and Data Science within Walmart Labs and as a Manager on Walmart's Global Selection and Assessment team.

Richard N. Landers is the John P. Campbell Distinguished Professor of Industrial-Organizational Psychology at the University of Minnesota and Principal Investigator of TNTLAB (Testing New Technologies in Learning, Assessment and Behavior). His research concerns the use of innovative technologies like games, gamification, machine learning, artificial intelligence, and virtual reality, to improve psychometric assessment, employee selection, adult learning, and research methods. He is president of Landers Workforce Science LLC (https://landers.tech), which focuses on auditing employee hiring systems incorporating artificial intelligence and machine learning.

Keith McNulty is Global Head of People Analytics for McKinsey & Company. He leads a large team focused on applying advanced analytics, machine learning, natural language processing and other artificial intelligence methods in the areas of hiring, performance, development, learning and engagement. He has built and implemented numerous AI processes in these fields including classification and reinforcement learning algorithms.

Nathan Mondragon is Chief I/O Psychologist at HireVue and has managed the creation and research of the first machine learning based interview assessment system used for selection, authored several articles under review, created additional novel approaches, and managed consultants to organizations on this topic.

Dan J. Putka is a Principal Scientist at the Human Resources Research Organization in Alexandria, Virginia. He focuses on advancing psychometric and analytic methods that are sensitive to the demands of applied research and practice, most recently in the area of machine learning and natural language processing. He has delivered numerous presentations on big data and data analytics at national conferences, and his work on machine learning prediction methods won the 2018 Organizational Research Methods (ORM) paper of the year award.

Daniel Schmerling is a Senior Research Scientist on the Talent Management team at Amazon AWS. In this role, Dr. Schmerling uses his dual background in I/O Psychology and Machine Learning to build models that predict employee engagement along with automatically identifying personalized career paths and development opportunities for the AWS employee population of around 100,000 "Builders." Before that, Dr. Schmerling served as the Head of HR Data Science at Prudential where he was responsible for leading all artificial intelligence initiatives and analyses in the HR space.

Kimberly Silva is a Senior Research Scientist at Talogy where she is the lead developer of natural language processing and machine learning systems for personnel selection. She also conducts extensive research on AI powered methods for optimizing selection and development assessments as well as the legal and ethical use of AI in Human Resource functions.

Andrew Speer is an organizational scientist and assistant professor at Wayne State University in Detroit, Michigan. In addition to consulting and conducting research on employee selection, personality, performance management, and employee turnover, Speer is also an expert in machine learning (ML) within the organizational sciences. He has numerous publications on the topics of machine learning and natural language processing and regularly consults with organizations on design, validation, and implementation of ML-based solutions. He also frequently serves as a reviewer for ML-related submissions to top journals.

Scott Tonidandel is Professor of Management in the Belk College of Business, a faculty affiliate of the School of Data Science, and Director of the Organizational Science Ph.D. program at the University of North Carolina—Charlotte. His recent work focuses on people analytics and the interface of big data and the organizational sciences. He co-edited the SIOP Frontiers series volume titled Big Data at Work: The Data Science Revolution and Organizational Psychology. Scott serves as an associate editor for the Journal of Business and Psychology, is a former associate editor for Organizational Research Methods, and is a fellow of SIOP, APA, and APS.