Received: 2 June 2023 | Accepted: 3 June 2023

DOI: 10.1111/peps.12608

PERSONNEL PSYCHOLOGY WILEY

ORIGINAL ARTICLE

Improving measurement and prediction in personnel selection through the application of machine learning

Nick Koenig[1] | Scott Tonidandel[2] | Isaac Thompson[1] | Betsy Albritton[2] | Farshad Koohifar[1] | Georgi Yankov[3] | Andrew Speer[4] | Jay H. Hardy III[5] | Carter Gibson[1] | Chris Frost[1] | Mengqiao Liu[6] | Denver McNeney[6] | John Capman[6] | Shane Lowery[6] | Matthew Kitching[6] | Anjali Nimbkar[6] | Anthony Boyce[6] | Tianjun Sun[7] | Feng Guo[8] | Hanyi Min[9] | Bo Zhang[10] | Logan Lebanoff[11] | Henry Phillips[11] | Charles Newton[11]

[1] Modern Hire, Cleveland, Ohio, USA

[2] University of North Carolina at Charlotte, Charlotte, North Carolina, USA

[3] Development Dimensions International (DDI), Pittsburgh, Pennsylvania, USA

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

[5] Oregon State University, Corvallis, Oregon, USA

[6] Amazon.com, Inc, Seattle, Washington, USA

[7] Department of Psychological Sciences, Kansas State University, Manhattan, Kansas, USA

[8] Department of Psychology, University of Tennessee at Chattanooga, Chattanooga, Tennessee, USA

[9] Department of Psychology, Pennsylvania State University, State College, Pennsylvania, USA

[10] School of Labor and Employment Relations & Department of Psychology, University of Illinois Urbana-Champaign, Champaign, Illinois, USA

[11] Soar Technology, Inc, Orlando, Florida, USA

Correspondence
Scott Tonidandel, UNC Charlotte Ringgold standard institution, Charlotte 28223-0001, USA.

Abstract

Machine learning (ML) is being widely adopted by organizations to assist in selecting personnel, commonly by scoring

Email: scott.tonidandel@charlotte.edu

Papers in this article were submitted and evaluated individually in the traditional peer-review process by five independent reviewers, including regular editorial board members and a special board, judged to make a meaningful contribution after one or more rounds of revisions, accepted independently, and then combined into this thematic article. The co-editors of the special issue decided study ordering, which determined authorship ordering.

narrative information or by eliminating the inefficiencies of human scoring. This combined article presents six such efforts from operational selection systems in actual organizations. The findings show that ML can score narrative information collected from candidates either in writing or orally in response to assessment questions (called constructed response) as accurately and reliably as human judges, but much more efficiently, making such responses more feasible to include in personnel selection and often improving validity with little or no adverse impact. Moreover, algorithms can generalize across assessment questions, and algorithms can be created to predict multiple outcomes simultaneously (e.g., productivity and turnover). ML has even been demonstrated to make job analysis more efficient by determining knowledge and skill requirements based on job descriptions. Collectively, the studies in this article illustrate the likely major impact that ML will have on the practice and science of personnel selection from this point forward.

KEYWORDS
artificial intelligence/big data/machine learning, selection-methods, selection-validation

1 INTRODUCTION

The purpose of this combined article is to illustrate a range of current applications of machine learning (ML) to improve measurement and prediction in personnel selection. The goals of the applications are to solve historical major problems in personnel selection, commonly either to score narrative data or to make selection more efficient in terms of human rater effort or both. In order to present several relevant studies on the topic, only brief summaries are presented in this article. Interested readers should consult each study's Online Supplement for additional information on the study background, method, and supplemental analyses.

Study 1 used ML to address resource inefficiencies by examining an important potential obstacle when using ML to score narrative responses to open-ended prompts, which is the extent to which a model can be used across alternative prompts. If a unique algorithm had to be developed for each prompt, it would require more time and greater costs. Using ML, they show that algorithms can generalize to scoring responses from novel prompts, especially in certain conditions, which they call algorithmic construct generalizability. Study 2, Study 3, and Study 4 used ML to address resource constraints in assessing constructed responses (written or oral). Assessment center (ACs) exercises that collect unstructured narrative responses are highly desirable selection techniques due to typically good validity and low adverse impact, but they can be costly to score. Study 2 compared several algorithms to show that ML can score such responses with as much reliability and validity as humans or better.

A likely scenario for an organization considering ML is to add it to an existing assessment battery of traditional tests to score open-ended questions that would be too costly to score manually. Study 3 shows how ML can complement existing assessments by scoring open-ended questions as well as humans, but more efficiently while adding slightly to validity with little increase in adverse impact. Similarly, Study 4 developed an algorithm for scoring an audio constructed response simulation, which is becoming a much more common response format and costly to score by human raters. They found that ML can score audio constructed responses to a simulation assessment much more efficiently than humans with the same or greater reliability and even incremental validity. Study 5. addressed another practical issue that arises in personnel selection but has previously been difficult to solve, which is the situation in which the goal is to predict two different outcomes at the same time. Study 5 shows that ML algorithms can be created to predict multiple outcomes simultaneously (e.g., productivity and turnover). Finally, another common historical problem in personnel selection is the need to conduct job analyses to identify and justify the job requirements (e.g., knowledge and skills). Job analysis is a laborious process that could perhaps be assisted by ML to increase efficiency. Study 6 demonstrated that ML could be trained to extract knowledge and skill from job descriptions as well as humans, which may not fully replace traditional job analyses but could certainly make it more efficient as one component.

2 | STUDY 1: ALGORITHMIC CONSTRUCT GENERALIZABILITY: SCORING NOVEL OPEN-ENDED PROMPTS WITH DEEP LEARNING TRAINED ON ALTERNATIVE PROMPTS[1]

The age of automation is underway. One aspect of automation that is impactful to organizational psychologists is the automatic scoring of unstructured text. Historically, the analysis of textual data types was an extremely laborious task, requiring hours of subject matter experts' (SME) time to read and evaluate. Recent advances in natural language processing (NLP) and deep learning have provided models that can accurately replicate the ratings produced by SMEs and can do so quickly and efficiently. Extant literature on computer scoring of applicant data supports the reliability and validity of these models and argues for their value in research (M. C. Campion et al., 2016; Sajjadiani et al., 2019). Many questions remain, however, about how algorithms can be optimally built to satisfy existing psychometric standards of generalizability. Exploring the generalizability of how to optimally train algorithms to perform well in new scenarios and with new prompts would have substantial impact on our field. Currently, there are tremendous resource costs associated with repeatedly training and testing new models when hiring procedures change (e.g., new interview prompts to combat cheating or updated selection practices). During this process of updating algorithms for a new prompt, a large number of responses to the novel prompt must first be obtained and then qualified raters must label those responses. This creates a barrier to the continued use or application of well-performing algorithms to new problems or contexts. The purpose of this paper is to expand the current understanding of the generalizability of computer scoring of unstructured text via deep learning algorithms in selection contexts addressing the gaps in the literature. Specifically, this research focuses on the extent to which deep learning algorithms trained on a prompt can generalize and accurately score a new set of responses generated by a different prompt. Additionally, the impact of the assessment medium, sample size, prompt similarity, number of prompts used in the training set, and focal prompt seeding on generalizability are all investigated.

2.1 | Algorithmic construct generalizability

A key attribute of any selection test or system is generalizability. When we think about the generalizability of a selection system, we are referring to the generation of accurate scores in spite of specific changes, such as the specific prompts presented, the assessment medium, the rater, the directions, or the subjects (Crocker & Algina, 2008). In modern personnel selection, new generalizability concerns arise when a deep learning algorithm is trained to score open-ended responses on specific prompt(s), but new prompts are being introduced. Can an algorithm that has been trained on a set of responses generalize to accurately score candidate responses to a new prompt that is tapping into the same underlying construct? Will the introduction of a new or modified prompt require restarting the process from scratch and if not, what are the factors that impact the performance of existing algorithms when trying to score responses to novel selection prompts? If algorithmically scored assessments of unstructured candidate response are to be used in selection, answers to these questions are paramount.

Unfortunately, traditional psychometric theory falls short in answering these questions. Currently, there is no guidance for the aforementioned scenario which encapsulates a new form of generalizability called *algorithmic construct generalizability*. This type of generalizability refers to the ability of a model (i.e., set of weights) trained on other item responses to produce an accurate score on applicant responses generated from a novel item reflecting the same construct in the same population of participants. We distinguish this concept from the more general term algorithmic generalizability, the ability of a model to make accurate predictions on data not contained in the training dataset (Rivlin, 2019), which is similar to the traditional definition of cross-validation that refers to generalization of accuracy to a sample of new subjects (Hastie et al., 2017). Neither cross-validation nor algorithmic generalizability, however, account for the application of a model to a completely different measurement of the underlying construct. Algorithmic construct generalizability is also meaningfully different from parallel forms reliability. The primary concern in parallel forms reliability is the consistency of the rank order of participants across the two forms. Algorithmic construct generalizability, on the other hand, concerns a model's ability to accurately predict SME labels rather than the consistency of an individual's score across the different prompts that attempt to measure the same construct. For example, a participant could score differently across the two prompts, perhaps because the prompts sample different areas of the content domain, but evidence of algorithmic construct generalizability could still be high if the model can accurately reproduce the SME labeled score for the latent construct for the novel prompt. Thus, while parallel forms is a type of reliability, algorithmic construct generalizability assesses a unique type of algorithmic validity where the training data is generated from prompts that are different from the test prompt. Given these gaps in extant psychometric theory, the introduction and testing of algorithmic construct generalizability is necessary to both extend rigorous measurement standards to algorithmic contexts and provide guidelines for addressing a practical problem in personnel selection.

2.2 | Practical contribution of algorithmic construct generalizability in scoring unstructured text

Computer scoring of text has been shown to approach or even equal the accuracy of human raters in scoring text (M. C. Campion et al., 2016) enabling processing a larger number of applications with significant cost savings. M. C. Campion et al. (2016) estimated potential savings of $210,000 annually for one organization when computer scoring replaced human scoring. Algorithms can also be probed to understand their predictions despite common concerns that algorithms are black boxes with unforeseen, harmful biases. A growing body of research suggests that computer scoring can guard against human biases in personnel selection when we leverage existing tools that identify what is driving the decision or classification within the algorithm and those biased features are removed from the training data (Albritton & Tonidandel, 2020; Ribeiro et al., 2016; Tay et al., 2021). Given these advantages, practitioners continue to develop consulting firms and technology startups that rely on NLP, deep learning, and other ML approaches to improve and automate personnel selection. At the same time, companies continue to evolve their interview and AC prompts and often wish to apply new prompts or new scenarios. There is an ever-present need to consistently retire selection prompts to avoid cheating or applicant knowledge of the prompt prior to interviewing. In the process of creating new prompts, one must go through a series of steps to ensure an algorithm offers acceptable accuracy in scoring responses to the new prompt such as: collecting a sufficiently large sample of applicant responses, labeling those responses by human raters, training and testing a deep learning algorithm to accurately score the competency being assessed, and provide additional evidence of validity or test any potential unfairness in the tool. This is an iterative and time-consuming process that plagues practitioners and limits the adaptability of selection systems. The ability to apply existing algorithms to novel prompts would alleviate the heavy costs associated with training a new algorithm. Exploration of the generalizability of ML models in scoring applicant text data can further expand its use in selection contexts, optimize algorithms to score previously unseen selection prompts with greater accuracy, and allow companies to more easily incorporate new prompts without replacing their existing algorithm.

2.3 | Hypotheses

As the purpose of this paper is to investigate algorithmic construct generalizability in a selection context, a deep learning algorithm will be trained on text responses from prompts that are *different* from the ultimate prompt and responses one wishes to score, while all being conceptually aligned to the same competency.

One factor that may affect the generalizability of scoring algorithms is the assessment medium. Data were gathered from multiple assessment systems that implemented two distinct assessment mediums for obtaining responses: a virtual AC and a structured behaviorally based interview. In terms of algorithmic construct generalizability, predictions of novel prompts should be more accurate when an algorithm is trained on data collected using the same assessment medium as opposed to a different assessment medium. The impact of assessment medium, however, may not be the same across all items. In terms of algorithmic performance, Mitchell (1997) notes that "In general, learning is most reliable when the training examples follow a distribution similar to that of future test examples." (p. 6). This means that one would expect a greater degree of generalizability to the extent that the novel prompt produces responses that are more similar to the responses that are used to train the algorithm. Thus, we propose the following two related hypotheses:

Hypothesis 1: Predicted scores from deep learning algorithms on a novel prompt will correlate more highly with SME labels when trained on text obtained using a similar assessment medium than a different assessment medium (assessment center prompt vs. interview prompt).

Hypothesis 2: The influence of the assessment medium used for training will depend on item content similarity. Predicted scores from deep learning algorithms will demonstrate higher correlations with SME labels when the algorithms are trained on responses from prompts that produce responses that are more conceptually similar as opposed to prompts that produce responses that are more conceptually dissimilar.

The extant research on automated text analysis methods does not consider how the sample size of the dataset on which the algorithm was trained might influence computer scoring of *novel* unstructured text. Typically, if one is trying to score a prompt, the more quality labeled data available from that prompt, the better one will be able to score that prompt. What remains unknown is whether there is an improvement in scoring performance if additional labeled data from a *different prompt* is available. Given the general trend that more training data is better, we propose the following:

Hypothesis 3: Predicted scores from deep learning algorithms will correlate more highly with SME labels when the training data is larger (sample size) as opposed to smaller.

A unique boundary condition worthy of exploration is the impact of including a limited amount of focal prompt information in the training data. We refer to this as *seeding* the training data. Practically speaking, obtaining a small amount of labeled responses may be more feasible. In addition, training data that is more similar to the test data should lead to better algorithmic performance (Mitchell, 1997). Thus, it is anticipated that seeding the training data would increase the similarity between the training data and the test data, thereby improving algorithm performance.

Hypothesis 4: Predicted scores from deep learning algorithms will correlate more highly with SME labels when the training data is seeded with responses from the target prompt.

Another feature that might impact its generalizability is the diversity of prompts that are used for training. This combination of data sources mirrors what happens with algorithms using ensemble models (Dietterich, 2000). By leveraging a diversity of models, the performance of the overall model actually improves. Similar to ensembles, when training algorithms to generalize to novel prompts, a wider diversity of input would likely be beneficial.

Hypothesis 5: A wider variety of data (more prompts) will improve the ability of a deep learning algorithm to score novel prompts.

2.4 | Method

2.4.1 | Sample and context

This study involved 7017 responses from over 6000 unique participants who were job applicants from more than 50 different organizations that participated in different selection batteries. As part of the selection batteries, respondents participated in a Virtual Assessment Center experience (VAC) or structured behaviorally based interviews. As part of the VAC selection battery, applicants were asked to respond to four different hypothetical situations and describe the actions they would take given the situation. For the interview battery, behavioral interview questions that were part of an asynchronous video interview process were asked. For both the VAC and interview processes, different prompts were used by different organizations. The VAC system consisted of candidates for an entry level business consulting role and for a retail store management role. The interview candidates were from several dozen companies applying for roles from branch manager and technical underwriter to delivery driver and bookkeeper.

2.4.2 | Measures

Seven prompts across the two different selection systems were chosen for this study. A description of the four prompts utilized as part of the VAC process and the three prompts used as part of the interview process are available in the online supplemental materials (Online Supplement Table A2). Both the VAC prompts and the interview prompts were identified by a team of four Ph.D. Industrial-Organizational psychologists with multiple years of experience to be part of a single higher order competency – *Provides Exceptional Service*. Each response was independently evaluated by two trained SMEs on a 1–5 point behaviorally anchored rating scale. Consensus meetings were conducted on responses that had greater than one-point rating discrepancies between raters. Experts resolved such discrepancies by reviewing both competency definitions and the behaviorally anchored scale (BARs). Consensus would be reached once the experts came within one point of each other. The focal dependent variable is the aggregated SME ratings.

2.4.3 | Research design

The prompt used for testing is referred to as the *focal* prompt. To establish *baseline* performance, each open-ended prompt was trained and evaluated individually on all the available labeled responses for that prompt (i.e., training and testing occurred on the same focal prompt). Next, *within medium* generalizability was explored. For each focal VAC prompt, an algorithm was trained on all of the available labeled responses for the other three VAC prompts, not including the focal prompt, and then tested on the holdout focal prompt. This process was repeated, treating each VAC prompt as the focal prompt. A similar strategy was used for each interview prompt where an algorithm was trained on the other two interview prompts, holding out the focal prompt, and then tested on the focal prompt. *Between medium* generalizability was evaluated by training an algorithm on all of the prompts from the medium that differed from the focal prompt. Thus, if the focal prompt was a VAC prompt, the algorithm was trained on all the interview prompts, and if the focal prompt was an interview prompt, the algorithm was trained on all the VAC prompts. To further investigate generalizability independent of medium, an algorithm was trained on *all prompts*, excluding the focal prompt, and tested the performance of that algorithm on the focal prompt.

Because the analyses above all included different numbers of prompts in the training set but used all of the available data, the size of the training set varied from condition to condition. Thus, the above analyses were repeated using a constant sample size of 800 for the training set. Finally, with the sample size constraint in place, the impact of *seeding* the training data with the focal prompt was investigated. For example, given the within or between medium generalizability condition, the focal prompt was included with the other VAC or interview prompts in the training set, or when all prompts were used, the focal prompt was included as part of the training set.

In summary, baseline performance along with within medium, between medium, and all prompt generalizability were all studied. All available data was used in one set of conditions and training sample size was constrained in another to provide more consistent comparisons between prompts. Finally, the impact of seeding focal prompt information into the training set when sample size is constant was explored (see Online Supplement Table A3).

2.4.4 | Content similarity

To identify any inherent characteristics of the various prompts that overlap and potentially contribute to algorithmic construct generalizability, a qualitative review of the VAC and interview prompts was conducted. A team of two SMEs independently reviewed the prompts as well as a sample of twenty text responses for each prompt and independently provided ratings of paired prompts on their content similarity on a 4-point Likert-type scale (1 = *Not similar at all*; 2 = *Not similar*; 3 = *Similar*; 4 = *Very similar*). Using these ratings and the aforementioned patterns within the ratings, the research team generated a priori hypotheses based upon content similarity for the generalizability of models trained using certain prompts compared to others (see H2a-c in the Online Supplement). These a priori hypotheses were preregistered through OSF prior to analyzing the data (osf.io/n2jm8).

2.4.5 | Deep learning architecture

A state-of-the-art approach to transformers, RoBERTa was utilized. Robustly Optimized BERT pretraining approach (RoBERTa; Y. Liu et al., 2019) is an expansion upon the basic transformer architecture Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2019), using the same architecture but further optimized. RoBERTa is pre-trained on a large body of texts from the internet (i.e., OPENWEBTEXT, Gokaslan & Cohen, 2019) and books (i.e., BOOKCORPUS, Zhu et al., 2015). RoBERTa architecture provides a structure to attend to information across the entirety of a document or response. Specifically, the embedding layer takes in the word embeddings or representations of the words produced by pre-trained RoBERTa along with a positional embedding, to locate the word/token within the response, and a token type, which represents things like the beginning of a response or the separation of a sentence within a response. This layer is passed to 12 hidden layers referred to as transformer blocks. These are identical and consist of a self-attention mechanism with a linear layer. Each of these layers has layer norming and a dropout weight, which for the purposes of consistency of comparisons for this research was consistently set to .10. Previous investigations by members of our research team have indicated that it tended to demonstrate the most consistent results. The output of this 12th transformer block is then fed into a pooling layer, which was then fed into a custom regression head to output a float between 1 and 5 to reflect the rating the SMEs provided for each competency. In summation, this architecture contains roughly 124 million parameters. RoBERTa was implemented in Python using the huggingface RoBERTa-base pretrained model (Wolf et al., 2019). It was then further trained on the downstream task of predicting each competency for each experimental condition on AWS spot instances that utilized Nvidia Tesla V100 GPUs. These spot instances had an installed Docker container with all packages and code necessary to train and evaluate the models. Since this transformer architecture was used, no form of text cleaning was done on the prompt responses. Additionally, our models were trained using the standard RoBERTa cutoff of 512 tokens per response. A minimal proportion of responses in our dataset exceeded a length of 512 tokens (Prompts 1–4: 0%, Prompt 5: 1%, Prompt 6: 1.7%, Prompt 7: 2.4%).

2.4.6 | Data configuration

For baseline conditions (i.e., models trained and tested on only responses from a single prompt), training and testing utilized an 8-fold cross validation strategy. The data for each focal prompt was split into four folds and three data sets: training, development, and test. These splits were developed using stratified random sampling on the label distribution to hold label distributions constant across all folds. The training data set, used to train the algorithm, consisted of three of the four folds. The final fold is split in half (half development and half testing). The development set was used for early stopping and for this reason was not considered as a fully independent holdout set. The test set was completely unseen by the model, even during model evaluation between epochs (to determine when optimization improvements had ceased to improve), and the only one used for evaluation. In an effort to ensure all data was part of the final unseen test set, the 50/50 split of the final fold was run twice, which involved flipping the development and test sets each time, thus creating an additional four-folds, and making it an eight-fold cross validation. The results of each of the eight folds were then averaged to create a holistic picture of the algorithms ability to predict unseen responses from the baseline prompt. For the generalizability conditions, because the data from these novel prompts only contained development/test sets, we only needed a two-fold cross validation where half of the holdout was first used for development and the other half was used for the test set and vice versa. As before, in all conditions where subsets of the data were selected, we used stratified random sampling to ensure we had an equal amount of labels from each prompt and that within the prompt the labels reflected the distribution of the population of labels for that prompt. This was performed on all of the available data as well as on a training sample that was capped at 800 for certain conditions. The results of each of the two folds were also averaged like in the baseline condition.

As part of this research, we were interested in understanding the impact of including a subset of responses from the focal prompt on algorithmic construct generalizability. We refer to this as the *seeding* conditions. For the seeding conditions, the training sample was capped at 800 responses. The training sample was composed of an equal proportion of the responses from prompts included in each model. For example, in the seeding condition where the model contained all four VAC prompts, there were 200 prompt responses representing the four VAC prompts in the training sample. Likewise, the models in the seeding condition for interview prompts contained three prompts, so they were each represented by 267 prompt responses. Like the prior conditions, responses were selected through a stratified random sampling technique and all responses not included in the training process were split in half and included first as the development set and then as the test set creating another two-fold cross validation methodology. Essentially, the model is trained and evaluated with the holdouts being swapped (i.e., validation/test, switches to become test/validation) and averaged, where each is evaluated as the true test set on separate model training instances.

2.4.7 | Analyses

The correlation between the computer-generated score and the SME label was the primary mode of evaluation of algorithmic construct generalizability under the various conditions described. An algorithm is considered accurate to the extent that it can reproduce the label generated by a knowledgeable expert coder (Thompson et al., 2023), with special interest to those algorithms that can meet a threshold of .60 and could thus be used in selection systems (M. C. Campion et al., 2016). This value of .60 is also nearly identical to the average correlation across prompts between two independent SME raters pre-consensus (*r* = .62). We first applied these standards to the correlation produced in the baseline condition when an algorithm is both trained and tested on the same focal prompt, as is current practice for training and testing models to score unstructured texts from selection prompts. Large deviations in algorithmic performance from that baseline would be indicative of low algorithmic construct generalizability, whereas being able to produce comparable levels of performance as the training set differs indicates high levels of generalizability. Because of the large sample sizes, we do not examine whether any two correlations are statistically significantly different. However, we do report the upper and lower bounds of the confidence interval to provide the reader some indication of sampling variability.

2.5 | Results

First, we wished to establish how well an existing deep learning algorithm could score text from novel prompts (compared to when an algorithm is trained specifically on that prompt) and what factors impact this performance. Results (see Table 1) indicate that across all prompts and all conditions, the correlation between SME and algorithmically predicted scores ranged from .75 to .04 (*M* = .42, *SD* = .20) when tested on a novel prompt. In terms of how closely these correlations matched those produced at baseline, these correlations were very similar, only 7% different, to widely divergent, up to 94% different. The deep learning algorithm does appear to accurately predict SME ratings of novel prompts in some situations, but in other situations the algorithmic predictions are severely inaccurate. Subsequent hypotheses explore potential reasons for these discrepancies.

Hypothesis 1 predicted scores from deep learning algorithms on a novel prompt will correlate more highly with SME labels when trained on text obtained using a similar assessment medium than a different assessment medium (AC prompt vs. interview prompt) and hypothesis 2 predicted that this effect would be impacted by similarity. Looking at the condition where all available data was used for training (i.e., top section of Table 1), this hypothesis appears to be supported for most, but not all, of the prompts. Predicted scores from prompts 1, 2, 4, 6, and 7 all correlate more highly with SME labels when the deep learning algorithm is trained on responses from novel prompts collected via a similar assessment medium compared to when the algorithm is trained on responses from novel prompts collected via a different assessment medium. On average for these five focal prompts, the correlation between predicted scores and SME labels was .29 higher for deep learning algorithms trained on novel prompts within medium compared to between medium.

Based upon SME similarity coding of items, it was predicted that prompt 1 would have the highest level of prompt generalizability when the algorithm is trained within assessment medium on the remaining VAC prompts. As evidenced in Table 1, prompt 1 did in fact have the largest correlation with SME labels (.70 vs. .62, .14, and .62). In addition, based upon SME ratings of content similarity, Prompt 3 showed the weakest content similarity to other VAC prompts, so it was predicted that prompt 3 would have the worst within assessment medium generalizability compared to the other VAC prompts. This exact pattern was found with Prompt 3 showing virtually no correlation (.14) with SME labels when trained on other VAC prompts whereas the other VAC prompts showed relatively good within assessment medium generalizability with correlations ranging from .62 to .70. Finally, it was predicted that prompt 5, given the higher levels of content similarity to the VAC prompts as rated by SME, would benefit more from cross assessment medium training than the other interview prompts. The pattern of results from Table 1 is consistent with partial support for this prediction. Prompt 5 displayed a large increase in the correlation with SME labels when moving from within assessment medium to between assessment medium (.04–.31) whereas the other two interview prompts showed a decrease (.57–.49 & .57–.34). Though this pattern is consistent with our predictions, what was not anticipated a priori was the extremely low level of correlation for prompt 5 within assessment medium (.04) and the fact that the between assessment medium correlations for the other interview prompts would still be higher than Prompt 5 (.49 & .34 vs.

TABLE 1 Algorithmic performance indexed via the correlation between predicted scores and SME labels.

		Correlation with SME label on holdout			
Condition	**Focal Prompt**	**Baseline**	**Within medium**	**Between medium**	**All prompts**
N_{train} = All Data; Unseeded	1	**0.79****	**0.70**** (0.67, 0.73)	0.23 (0.17, 0.29)	**0.70**** (0.67, 0.73)
N_{train} = All Data; Unseeded	2	**0.82****	**0.62*** (0.58, 0.66)	0.23 (0.17, 0.29)	**0.55** (0.51, 0.59)
N_{train} = All Data; Unseeded	3	**0.82****	0.14 (0.08, 0.20)	**0.50** (0.45, 0.55)	0.32 (0.26, 0.38)
N_{train} = All Data; Unseeded	4	**0.77****	**0.62*** (0.58, 0.66)	0.32 (0.26, 0.38)	**0.60*** (0.55, 0.64)
N_{train} = All Data; Unseeded	5	**0.68***	0.04 (–.01, 0.09)	0.31 (0.26, 0.36)	0.06 (0.01, 0.11)
N_{train} = All Data; Unseeded	6	**0.74****	**0.57** (0.53, 0.61)	0.49 (0.44, 0.54)	**0.50** (0.45, 0.54)
N_{train} = All Data; Unseeded	7	**0.63***	**0.57** (0.53, 0.61)	0.34 (0.28, 0.39)	**0.50** (0.45, 0.55)
	Mean =	0.75	0.47	0.35	0.46
		0.07	0.24	0.10	0.20
N_{train} = 800; Unseeded	1	**0.79****	**0.68*** (0.66, 0.73)	0.18 (0.64, 0.72)	**0.66*** (0.65, 0.72)
N_{train} = 800; Unseeded	2	**0.82****	**0.58** (0.65, 0.72)	0.42 (0.61, 0.69)	**0.64*** (0.70, 0.76)
N_{train} = 800; Unseeded	3	**0.82****	0.35 (0.68, 0.75)	0.27 (0.72, 0.78)	0.35 (0.49, 0.59)
N_{train} = 800; Unseeded	4	**0.77****	**0.63*** (0.55, 0.65)	0.32 (0.48, 0.59)	**0.54** (0.58, 0.67)
N_{train} = 800; Unseeded	5	**0.68***	0.16 (0.57, 0.65)	0.37 (0.58, 0.65)	0.38 (0.50, 0.58)
N_{train} = 800; Unseeded	6	**0.74****	**0.51** (0.56, 0.65)	0.27 (0.58, 0.66)	**0.56** (0.61, 0.69)
N_{train} = 800; Unseeded	7	**0.63***	**0.53** (0.54, 0.64)	0.29 (0.52, 0.61)	0.49 (0.48, 0.58)
	Mean =	0.75	0.49	0.30	0.52
		0.07	0.17	0.07	0.11
N_{train} = 800; Seeded	1	**0.79****	**0.70**** (0.65, 0.71)	**0.68*** (0.12, 0.24)	**0.69*** (0.62, 0.69)
N_{train} = 800; Seeded	2	**0.82****	**0.69*** (0.54, 0.62)	**0.65*** (0.37, 0.47)	**0.73**** (0.60, 0.68)
N_{train} = 800; Seeded	3	**0.82****	**0.72**** (0.29, 0.40)	**0.75**** (0.21, 0.33)	**0.54** (0.29, 0.40)
N_{train} = 800; Seeded	4	**0.77****	**0.60*** (0.59, 0.67)	**0.54** (0.26, 0.38)	**0.63*** (0.49, 0.59)
N_{train} = 800; Seeded	5	**0.68***	**0.61*** (0.11, 0.21)	**0.62*** (0.32, 0.42)	**0.54** (0.33, 0.43)
N_{train} = 800; Seeded	6	**0.74****	**0.61*** (0.46, 0.55)	**0.62*** (0.21, 0.33)	**0.65*** (0.52, 0.60)
N_{train} = 800; Seeded	7	**0.63***	**0.59** (0.48, 0.57)	**0.57** (0.23, 0.35)	**0.53** (0.44, 0.54)
	Mean =	0.75	0.65	0.63	0.62
		0.07	0.05	0.06	0.07

Note: The results for the Baseline conditon used all of the available data and are repeated in each section of the table for ease of compasion. 95% Confindence intervals are dislayed in parentheses. Bolded correlations exceed .50 indicative of nearing practical use standards; * indicates a correlation geater than .60, the target standard for practical use (Campion et al., 2016) ; ** indicates a correlation greater than .70, a lower-bound threshold for use in high-stakes testing (Campion et al., 2016).

.31). In general, predictions generated from items collected using the same assessment medium correlated more highly with SME ratings than items from different assessment mediums, supporting Hypothesis 1. Moreover, when variation between items existed, this variation was predictable a priori based upon item similarity ratings from SMEs, supporting Hypothesis 2.

Hypothesis 3 predicted that scores generated from deep learning algorithms on a novel prompt will correlate more highly with SME labels when the sample size of the training data is larger as opposed to smaller. To evaluate this hypothesis, we can compare the correlations at the top of Table 1 when the algorithm was trained on all of the data to the section right below that where the algorithm was trained on a restricted sample of 800 responses. On average, the reduction in sample size had only a minor impact on the correlation between SME labels and predicted scores. Surprisingly, the correlations to SME labels seemed to increase somewhat with the smaller sample sizes when the original correlations using all of the data were low and underperforming. Otherwise, reducing sample size generally led to a small decrease in the observed correlation. Given this small impact, Hypothesis 3 was not supported.

Hypothesis 4 predicted that scores from deep learning algorithms, when applied to a novel prompt, will correlate more highly with SME labels when the training data is seeded with information from the target prompt. To evaluate this hypothesis, one can compare the results from the middle section of Table 1 when the training set does not contain any data from the focal prompt to the lower section of Table 1 that seeds the training data with information from the focal prompt. Two consistent patterns regarding the impact of seeding emerge from this comparison. First, when the SME label for a focal prompt can be reliably predicted from a set of novel prompts (i.e., the correlation between SME labels and predicted scores is high), the seeding of the training data with responses from the focal prompt has a modest positive impact on the correlation with SME labels. However, seeding appears to be quite impactful at improving the performance of deep learning algorithms that are underperforming when just training on unseeded responses solely. For example, algorithms for Prompts 3 & 5, when trained on unseeded prompts from the same assessment medium, produced predicted scores that correlated only .35 and .16 with SME labels. In contrast, by seeding with some responses from the focal prompt, those correlations increased to .72 and .61. A similar pattern is exhibited across all of the prompts when the algorithm is trained across assessment mediums. The between assessment medium correlation between predicted scores and SME labels across all prompts averaged .30 (*SD* = .07). Those same correlations averaged .63 (*SD* = .06) when focal prompt responses were seeded into the training data. These results support hypothesis 4.

Our final hypothesis stated that a wider variety of data (more prompts) would impact the ability of a deep learning algorithm to score novel prompts. The column "All Prompts" in Table 1 contains results that speak to this hypothesis. In general, the addition of more prompts does not seem to improve algorithmic predictions. The average correlation observed when training an algorithm on all prompts tended to be about the same as the correlation when training an algorithm on just within assessment medium prompts. This was true across all three conditions of training data: training on all data, training on 800 responses and not seeding the training data, and training on 800 responses and seeding the training data. An improvement in the correlation was found when compared to the between assessment medium condition, but this improvement is likely due to assessment medium and not due to just having additional prompts. Based on these results, we conclude that adding more novel prompts to the training data will not aid in improved algorithmic predictions unless those prompts are more closely related to the focal prompt in some way (e.g., same assessment medium, more similarity).

2.6 | Discussion

This paper examines the ability of a deep learning algorithm to score a novel selection prompt when the algorithm is trained on the responses from entirely different prompts, while still being conceptually linked via a competency framework. Using candidate responses from real selection systems, results indicate that algorithms trained on one set of prompts can generalize to new, never-before-seen selection prompts under certain conditions. Overall, algorithmic construct generalizability was higher when the responses to the novel prompts were collected using the same assessment medium as the responses used in the training data. These findings are similar to past research on ACs that consistently demonstrated additional variance in scores that was explained by exercise effects, rather than broad dimensions within exercises (Connelly et al., 2008; Hoffman et al., 2011). Additionally, algorithmic construct generalizability was higher for similar prompts (as rated by SMEs), even when those prompts were not rated as *highly* similar pairs. In fact, it was quite rare that prompt pairs in our study were rated as similar.

Seeding the training data with information from the novel prompt was an important feature for enhancing algorithmic construct generalizability, especially in instances when the algorithm was performing poorly at replicating the SME labels in the focal prompt. Interestingly, the two prompts that performed worst with respect to within assessment medium generalization without seeding, prompts 3 and 5, showed impressive performance gains under the same conditions when seeded with some data. However, these same two prompts once again showed the largest decrement in performance when training on all of the prompts even when seeding. The poorer performance of the algorithm in the seeded-all prompts condition was probably due to the decrease in the amount of available information being seeded into the training set from the focal prompt, further illustrating the importance of seeding.

Restricting the size of the training data to only 800 responses did not seem to impact algorithmic construct generalizability. The inconsequential addition of data to the training samples of models implies that algorithms might also reach a point of theoretical saturation (Tracy, 2013) where additional data is no longer improving the algorithmic construct generalizability or performance of a model. The use of limited data in language models has also been explored previously through few-shot learners (Brown et al., 2020). Few-shot learners learn using smaller training sets, also known as support sets, to identify similarities and differences in observations. Even with these smaller support sets, few-shot learners are shown to perform almost as well as standard, fine-tuned models (Brown et al., 2020). Our results are similar. We also believe that our use of a modern transfer learning architecture like RoBERTa is also at play here. In other work (Thompson et al., 2023), when training and testing language models on the same prompt, we found that the correlation between an algorithm's score and an SME's label on a training sample size of 800 was nearly .84. But reducing that training sample size to only 200 reduced the correlation to only .80. However, reducing sample size did impact the accuracy of less modern algorithmic approaches like LSTM. Although our approach here was different as it relied on training on a different prompt, we were using a similar RoBERTa architecture, which may explain why sample size had only a minimal effect. Similarly, adding additional prompts to the training data set had a minimal effect over and above the performance seen with a more limited set of prompts. A larger pool of more diverse prompts and responses would convey information to the deep learning algorithm allowing for enhanced predictions, but this was not supported by the data. This suggests that a limited number of prompts may be sufficient to achieve saturation.

2.6.1 | Practical implications

The main practical application of these results includes (a) facilitating the addition of new content to an AI scored systems and (b) reducing computational size of AI systems in production. The creation and addition of new content (prompts) to be scored automatically by AI systems is a key practical concern. Previously, each new prompt would require new candidate responses to be gathered, rated by SMEs, and then for a new model to be trained on this data. Results indicate that within a competency, if the assessment medium of a prompt (e.g., interview) is held constant, there is a good possibility that the model will successfully generalize to the new prompt. Practically speaking, additional assessment content (new prompts) can be added to preexisting systems without the need to gather additional data or start from scratch in training a new model. Our results indicate that the probability of such generalizability to new content will be increased by having additional similarities in prompts and can be insured with exposure to a small sample of "seeded data" or labeled response data from the new prompt and updating the model. Thus, when adding new AI scored opened-ended prompts to create parallel tests, fluctuate stems, add a new context, etc., practitioners are urged to consider the content similarity and if content diverges significantly, or the system is used for high-stakes employee selection, to "seed" their model with a few novel responses. This greatly reduces the effort and time required to update, improve, and evolve these AI systems.

Secondly, if each prompt requires a diverse model to be in production, the scalability of the AI scored systems is hindered. For example, a trained algorithm's size can be 1 GB, and when hosted in memory, engineering and financial considerations arise. By hosting an ever-growing number of new models that are created for new prompts, the engineering and financial burden of running computers necessary to host these models constantly in memory to score candidates 24-h a day quickly increases. The present research indicates that lumping together several prompts in the training of an algorithm did not hinder prediction on a singular prompt in that training. From this, there seems to be little downside in putting the training data together and thus reducing the total number of models needed to be hosted in production.

3 | STUDY 2: COMPARING THREE ML ALGORITHMS FOR SCORING AC TEXT DATA[2]

This study aimed to demonstrate the use of ML for scoring performance in online ACs (International Task Force, 2009), specifically applying NLP to a large database of operational AC responses. Online ACs contain a wealth of available data to review for each candidate, and it serves as a prime context to apply ML. There are considerable benefits to automating AC scoring, given the time and cost associated with assessor evaluations and report writing. However, there is little known published research where ML has been applied to automatically score operational ACs, and it is unclear whether automated ML methods can achieve adequate evidence of validity to support their use. In this study we test the validity of ML-derived AC scores, and in the process offer several contributions to the research literature.

First, this study serves as one of the first documented applications of ML and NLP to online ACs where participants' responses are text-based. Although some research has examined the validity of scoring simulations or interviews with ML, these are based on small samples or simulated, non-operational contexts (e.g., Hickman et al., 2021; Nguyen et al., 2014). Second, the choice of NLP algorithm is likely to impact the validity of derived model scores. In the organizational sciences, most research has used the "bag of words" (BOW) framework (e.g., M. C. Campion et al., 2016; Speer, 2018), which although simplistic and sufficient for certain NLP tasks, does not adequately capture the complexity of language. In this study we apply a multi-layered transformer neural network architecture (Devlin et al., 2019; M. Liu, 2019; Min et al., 2021; Vaswani et al., 2017; Zaheer et al., 2020), which better encapsulates the contextual sequence and meaning of words. We compare transformer-based scores to traditional BOW scoring, as well as a more rationally driven NLP approach that combines deductive, top-down winnowing of text along with empirical ML (Speer, 2020). Third, we examined the validity of ML-scored ACs using a wealth of validation data that included correlations with assessor ratings, general stability of scores, and criterion-related validity (i.e., correlation with supervisor-rated job performance).

In sum, our study is among the first to automatically score large, operational, and multidimensional AC data using NLP, doing so across multiple NLP methods and by examining the psychometric properties of scores in numerous ways.

3.1 | Applying ML to automate ACs

ACs have a long history of use, and a review of ACs can be found in the online supplementary materials. Here, we focus on the advantages of automating ACs. For one, automating candidate evaluation reduces total AC costs that occurs from using numerous human assessors. Two, ML scores are standardized and may capture nuances or complexity in the data that humans are less likely to consistently integrate into their judgments. There is high cognitive load when making AC judgments. Assessors have to constantly observe behavior, note the relevant behaviors for each dimension, and mentally appraise new behaviors being displayed, to eventually adjudicate participants' proficiency for each dimension (Jansen, 2012). This process is cognitively demanding and not easy on even the best of assessors. In contrast, ML scores will consistently consider all presented information, be trained to identify which information is relevant, and consistently integrate and combine information across AC respondents to arrive to judgments. Thus, there are potential benefits to automating AC scoring using ML.

3.2 | Different approaches to automatically score text using ML

The goal of the ML task in this study was to train an algorithm to use candidate text inputs from the AC to reproduce human assessor ratings of AC dimensions. We examined three different scoring strategies. These are discussed briefly here, but more detailed discussion can be found in the online supplementary materials (Online Supplement B). Each of the three strategies is a *supervised* ML algorithm, such that an ML algorithm is trained to recreate an existing target score (i.e., SME ratings of AC candidates).

3.2.1 | Supervised scoring with bag of words

The most used NLP method within the organizational sciences is the BOW technique. BOW ignores the order of words in a document and instead splits text into vectors of words or word phrases, such that the analysis then takes place on whether and/or how often such word phrases occur. Once operationalized as a series of word vectors, these vectors can be used as predictor input features into an ML algorithm to predict target scores. For example, a BOW document term matrix can be used to predict AC assessor ratings using ML methods such as gradient boosted trees, which performs well in ML tasks (e.g., Chen & Guestrin, 2016). For this study, we label these as "Supervised BOW" scores, or SBOW.

3.2.2 | Supervised scoring using transformers

To better reflect the complexity of written language, contemporary NLP often uses neural network architectures that better encapsulate the contextual sequence of words. Since 2017, neural network transformer architectures (Vaswani et al., 2017) have emerged as the dominant NLP architecture in the computer sciences (Min et al., 2021). Transformer models handle sequential text input with contextual embeddings and specialized attention algorithms via deep neural networks. These models are composed of dense, multi-layered neural networks, with the layers capable of capturing meaning in language. These perform better than previously popular NLP architectures such as long-short-term-models and have achieved widespread use for numerous NLP tasks (Min et al., 2021). In the case of predicting human AC ratings, a top neural network layer can be stacked above the language layers, which was done in this study.

3.2.3 | Supervised scoring using theory-driven bag of words

A challenge with the previously described methods is that they are difficult to interpret. As such, we sought to compare those two empirical methods with a third method that is more deductively driven, called contextualized BOW (Speer, 2020). Instead of using all text within the ML process, contextualized BOW first filters each respondent's text to only those sentences that have a higher probability of being relevant to the focal construct. In this study we curated text by only keeping sentences which included a word that was theoretically linked to the targeted AC dimension(s). Once complete, a document term matrix is formed and used within a larger ML model to recreate target AC ratings. This provides a clearer understanding of the model inputs. A downside, beyond not accounting for the complexity of language like a transformer model does, is that with less text being used, contextualized BOW ultimately uses less information when estimating AC competencies. We label this method as CBOW.

3.3 | Current study and research questions

The purpose of this study was to compare the performance of the ML-derived AC scores in an operational context. In line with the unitarian view of validity (APA, 2018), we examined the psychometric properties of scores in numerous ways. Research questions are listed below.

Research Question 1. What is the correlation between ML scores (SBOW, CBOW, and transformer scores) and aligned assessor scores?

Research Question 2. What is the general stability for assessor scores and how does this compare to the stability for ML scores (SBOW, CBOW, and transformer scores)?

Research Question 3. How does the pattern of inter-correlations differ between assessor scores and ML scores (SBOW, CBOW, and transformer scores)?

Research Question 4. What is the correlation between AC scores and job performance ratings for ML-generated scores (SBOW, CBOW, and transformer scores) and for aligned assessor scores?

3.4 | Method

3.4.1 | Participants and data

We analyzed three archival data sets (primary, criterion, and test-retest) from a United States (US) assessment and development company. Each participant was part of only one data set. That is, the data sets were completely independent. The primary data set included 3152 applicant and incumbent frontline managers from different companies, all assessed in English. We split the data into a training set (for training and tuning the ML algorithms; $N = 2522$) and test set (for creating independent ML scores and comparing to the human assessor scores; $N = 630$).

The test-retest data set was composed of 164 participants: 50 applicants-applicants (i.e., applying for different companies at Time 1 and Time 2) and 114 applicants-incumbents. The AC was used for selection or development. Individuals who were tested at Time 1 for selection purposes took the AC before they joined the company. The average Time1-Time2 testing lag was 27.48 months ($SD = 20.02$; 5th – 95th percentile = 8–60 months). This is a lengthy time gap, likely resulting in true change in AC dimension scores and a lower expected correlation between administrations. For this reason, and also because the raters who assessed candidates differed from Time 1 to Time 2, the correlation between Time 1 and Time 2 scores is best labeled as *general stability* rather than a test-retest correlation.

The criterion data set was collected in 2010 and included 157 incumbents from four US companies—three in manufacturing and one in medical research. These companies were involved in various projects with the consulting company and had matched AC-performance data.

3.5 | Measures

3.5.1 | Assessment center

The AC simulates the work of a manager and was delivered in a virtual desktop environment. AC participants viewed emails, videos, and company materials and then responded via email responses. Thus, the only data analyzed in this project included written email responses. These responses were scored according to three AC dimensions: Coaching, Influencing, and Customer Focus, where each dimension was represented by exercises measuring the different content domains. For example, Coaching contained three exercises focused on coaching an irritated and continuously late employee, coaching a very bright and careless employee, and writing a career development plan. Notably, each exercise was designed to assess only one dimension. Note that we only received data for two dimensions (Coaching and Influence) for the general stability and criterion datasets.

Human assessors independently rated each exercise on multiple behavior indicators, which were then summed to create exercise scores and ranged from 0 to 16 points, where the higher the score the greater the skill demonstrated in the exercise. However, each exercise had different maximum scores because they contained different number of behavioral indicators. Different assessors rated each exercise, and assessors also varied across candidates.

The rating scale was: −1 (counterproductive behavior), 0 (behavior not observed), 1 (behavior observed), and 2 (excellent behavior)[3]. Assessors had examples for what constitutes a prototypic behavior for each of the scale points. There was no pre or post consensus meeting. To derive each exercise score, assessor ratings within each exercise were summed. Assessors regularly undergo calibration trainings, and company studies show that the average inter-rater reliability (ICC 1, 2) for the individual exercise scores is .67 (i.e., this is the average of the ICCs (1, 2) calculated for each exercise score), and that the single rater reliability averages .51 (ICC 1, 1). It should be noted that in practice scores are aggregated across exercises before reporting, and hence the composite reliability for dimensions (aggregated across exercises) is higher. For the individual ratings of behavioral indicators (where there are numerous within exercise), company studies have found that the average inter-rater agreement ranges from 75% to 85%.

3.5.2 | Job performance ratings

Supervisors of AC candidates rated candidate job performance using a multi-dimensional performance appraisal form completed for research purposes. The rating process occurred in the same range of four months as the AC testing. Scores were aggregated into an overall performance composite. See online supplementary materials for more details.

3.5.3 | NLP scores of AC responses

Algorithms were developed to recreate the human assessor scores. The algorithms were trained on the primary training sample. For all data sets (primary training, primary test, general stability, criterion), we used the same preprocessing and applied the same vectorization as developed in the training data set for that respective exercise.

3.5.4 | Developing supervised bag of words scores

SBOW scores were created by training a gradient boosted machine (XGBoost) to translate full document term matrices into assessor ratings[4]. We first transformed the raw text to a full document term matrix by removing generic stopwords, lemmatizing words with Python's spaCy library (https://spacy.io/api/lemmatizer), lowercasing all words, converting negative qualifying words to a common format (i.e., n't, never, and cannot became negative_word), stripping remaining punctuation, removing words appearing in less than 1% of texts, and removing white space. Some additional cleaning was performed to remove trailing non-ASCII characters. Then, *within* the training dataset we used a 10-fold randomized grid search to find the optimal XGBoost hyper-parameters[5]; the test set was not used for hyper-parameter tuning and kept completely independent.

3.5.5 | Contextualized bag of words scores

Contextualized BOW is similar to SBOW in that XGBoost was applied to a document term matrix. What differs is that for CBOW the ML algorithm is applied to a curated document term matrix based on text more likely to be relevant to the targeted AC dimension. To help identify the most relevant text to each of the three AC dimensions, we followed Speer's (2020) method to generate theme dictionaries for each AC dimension. However, we used a key innovation in generating an extended list of the theme vocabulary words. Specifically, we followed Li et al.'s (2021) method of using the word2vec NLP algorithm to generate an expanded theme dictionary by supplying a few seed words (i.e., words that are highly relevant) for each dimension, and then identifying theoretically similar words that were used in text. More information on this analytical process and data transformations can be found in the online supplemental materials.

After generating the theme dictionaries, for each AC dimension we then filtered the AC candidate text to only sentences that contained one of the theme words. These sentences have a higher probability of being relevant to the targeted dimension. This curated text was then transformed into a document term matrix based on the procedures outlined by Speer (2020), and then XGBoost was performed to train models to reproduce the exercise-dimension scores. This latter process was identical to that performed for the general SBOW scoring.

3.5.6 | Transformer scores

Transformer models with transfer learning have become the dominant paradigm for most NLP tasks (for a review see Min et al., 2021), and they apply well to longer texts such as those used here. Our AC texts represent open-ended, loosely structured, long responses (average length of 100 to 200 tokens) to semi-specific exercise questions (e.g., "How are you going to maintain X's commitment to changing their behavior?", "How to gain Y's confidence in project Z."). They also have parallelly running dependencies of words/themes (discussing the importance of the behavioral problem, discussing the impact on colleagues). Due to problems of long-short term models (LSTM) and related models losing their attention over longer texts (cf. P. Liu et al., 2015; Xu et al., 2016), transformers are particularly well-suited for these data. Given the widespread adoption of deep neural network transformers, and given the nature of this AC's text, we used transformers for this research, and more specifically—the popular BERT (Devlin et al., 2019).

We applied a deep, pre-trained BERT language model and customized it by updating the upper neural network layers to account for language idiosyncrasies particular to this AC. The pre-trained BERT architecture includes an embedding layer and 12 hidden layers, and we stacked a custom top layer used to predict AC exercise scores. We allowed parameter updates to the upper layer parameters of the pretrained language model, in line with best practice transfer learning. This was performed using the Python programming language and the HuggingFace (Wolf et al., 2019, https://huggingface.co/)[6] pretrained BERT-base model with the PyTorch library. We found the best performance for models trained on batch sizes of 16, learning rate of .00005, and 3 epochs. Thus, we trained all BERT models with these hyperparameters. Tables with the results of the hyperparameter tuning are provided in the supplementary materials.

3.6 | Results

Results are presented here in the primary paper and also in the online supplementary materials. The latter contain descriptive statistics on the full (i.e., train and test) data set (Online Supplement Table B1), score intercorrelations within each ML algorithm (Online Supplement Table B2), intercorrelations between assessor and ML algorithm scores (Online Supplement Table B3), correlations between text lengths and all assessor and ML scores (Online Supplement B4), results from the BERT hyperparameter tuning experiments (Online Supplement Table B5), and age and gender group differences for all assessor and ML scores (Online Supplement Table B6). The supplement also contains information pertaining to important features and model tuning decisions. Here, we focus on the primary research questions.

3.6.1 | Convergence between ML scores and assessor ratings

Table 2 presents correlations between ML scores and assessor scores within the independent test set. Results are provided for individual exercises and also for each AC dimension, which aggregated exercise scores and are

TABLE 2 ML algorithms scoring results.

		Supervised BOW			Contextualized BOW			Transformer		
Dimension	Exercise	corr	M	*SD*	corr	M	*SD*	corr	M	*SD*
Coaching		**.67**	**12.24**	**2.70**	**.65**	**12.19**	**2.54**	**.71**	**12.93**	**3.22**
	Exercise 1	.54	5.58	1.35	.49	5.51	1.35	.62	5.93	1.56
	Exercise 2	.64	4.89	1.58	.61	4.90	1.46	.64	5.13	1.79
	Exercise 3	.51	1.77	.48	.43	1.78	.55	.55	1.88	.73
Influence		**.62**	**9.22**	**2.35**	**.62**	**9.20**	**2.29**	**.67**	**9.88**	**3.01**
	Exercise 4	.51	3.99	1.19	.54	3.98	1.11	.58	4.29	1.42
	Exercise 5	.67	2.04	1.01	.65	2.04	.99	.73	2.22	1.29
	Exercise 6	.48	3.19	1.03	.46	3.18	1.03	.57	3.37	1.31
Customer Focus		**.65**	**6.19**	**1.37**	**.58**	**6.25**	**1.41**	**.68**	**6.40**	**1.82**
	Exercise 7	.54	2.95	.87	.48	3.00	.83	.60	3.13	1.08
	Exercise 8	.57	3.24	.80	.52	3.25	.94	.64	3.26	1.12
Average Dimension Score		**.65**	**9.22**	**2.14**	**.62**	**9.21**	**2.08**	**.69**	**9.74**	**2.68**
Average Exercise Score		**.56**	**3.46**	**1.04**	**.52**	**3.46**	**1.03**	**.62**	**3.65**	**1.29**

Note. All correlations are $p < .01$. $N = 630$ selected at random from the calibration data set ($N = 3152$). The dimension scores were a sum of all exercises that measured that dimension.

therefore more reliable. As seen, scores from each NLP scoring method exhibited large correlations with aligned assessor dimension scores (SBOW = .65, CBOW = .62, transformer = .69). However, transformer scores exhibited more favorable convergence with assessor scores, with an average correlation of .69, and this value being significantly larger than SBOW (Steiger $z = 2.30$, $p < .05$) and CBOW (Steiger $z = 3.53$, $p < .01$). The average transformer correlation of .69 is in line with typical AC reliability coefficients when correlating scores from two raters (Lievens, 2009). This is also very similar to other applied NLP tasks in the organizational sciences. For example, M. C. Campion et al. (2016) found an average convergence of .64 for individual competencies when scoring achievement records. When compared to the reliability of a single AC assessor for a single exercise within this study (ICC 1,1 = .51), the observed correlation between transformer scores and aligned exercise scores ($r = .62$) provides further support for the psychometric properties of the transformer scores. This correlation is higher than the reliability for a single human rater, and if the reliability of human raters were higher in this AC, it is likely the trained transformer scores would have achieved even higher levels of convergence, given the attenuating effect of unreliability. Overall, the NLP algorithms exhibited acceptable correlations with assessor scores.

We also extracted the most important features driving the SBOW and CBOW scores. Appendix A in the online supplementary materials presents the 25 most important N-grams driving the predictions for the eight exercises. After a brief content review performed by the same raters who generated the seed words for the three AC dimensions, the consensus was that both algorithms were driven by similar language, although it seemed that the CBOW list included more contextualization to the exercise stimulus and more psychologically meaningful words.

3.6.2 | Inter-correlations

Tables B2 and B3 in the online supplementary materials provide the inter-correlations between the exercise scores produced by the three ML algorithms and the assessor scores. Assessor and NLP scores exhibited similar patterns of inter-correlations, with slightly higher correlations between exercise scores for the NLP algorithms. The average monotrait heteromethod (correlation between exercise scores within dimension, for example, exercises 1–3

TABLE 3 General stability of the ML scores.

	AC Dimension	
Scoring Method	Coaching	Influence
Assessor Scores	.34	.30
Supervised BOW Scores	.60	.46
Contextualized BOW Scores	.57	.47
Transformer Scores	.69	.54
Average ML Score	.62	.49

Note. All correlations are $p < .01$. $N = 164$. AC = assessment center.

for the Coaching dimension) and discriminant correlations (i.e., between exercise scores representing different AC dimensions) were .23 and .23 for assessor ratings, whereas they were .33 and .31 for NLP scores. Thus, the assessor scores and the NLP scores exhibited minimal differences between monotrait heteromethod and discriminant correlations. The NLP scores simply reproduced this assessor pattern, though with slightly higher intercorrelations overall.

3.6.3 | General stability of scores

Table 3 presents the correlations between Time 1 and Time 2 administrations for the assessor scores and ML dimension scores. As expected, given that respondent skills may naturally change and there was a long delay between testing administrations (~27 months), these correlations were low. The average correlation for assessor scores was just .32. Interestingly, we found that all ML algorithms outperformed the assessor scores in terms of stability correlations (Steiger $z = 4.13$, $p < .01$). On average, the NLP scores had a stability correlation of .56, with transformer scores once again displaying larger correlations ($r = .62$).

3.6.4 | Criterion-related validity

Table 4 presents the validity coefficients for the assessor scores and the three ML algorithmic scores in their prediction of job performance ratings. As seen, assessor scores had a moderate validity of .16 for Coaching and .19 for Influence. Validity coefficients for the ML scores were higher, averaging .26 for Coaching and .23 for Influence[7]. Validity coefficients were generally similar for the three NLP methods.

We ran separate multiple regressions to establish the incremental validity of ML-scores over the assessors scores in the criterion data set. In each regression model the assessor scores were included as the first predictor and the ML-generated scores (SBOW, CBOW, or transformer respectively) as the second predictor, and overall performance was the criterion variable. Results of the regressions are also presented in Table 3. As seen, the NLP scores exhibited stronger regression weights and explained unique variance in all comparisons but two. The best performing NLP model—transformer—explained unique variance in all cases.

3.6.5 | Response length

We found that the lengths of the written responses were strongly correlated with the AC exercise scores. The average correlation across all exercises was .50 with assessor scores, which reflects the need for candidates to thoroughly address all exercise requirements to perform adequately. This is consistent with past findings that essay length

TABLE 4 Criterion validity of the ML scores.

	Coaching						Influence					
Assessment Center Scoring Method	*R* Performance	ΔR	*R*	*R2*	*β* AS	*Β* NLP Scores	*R* Performance	ΔR	*R*	*R2*	*β* AS	*Β* NLP Scores
Assessor Scores	.16*						.19*					
Supervised BOW Scores	.25**	.09*	.25**	.063	.05	.22*	.22**	.05	.24*	.058	.10	.18
Contextualized BOW Scores	.26**	.10**	.26**	.068	.05	.23**	.22**	.05	.24*	.058	.11	.16
Transformer Scores	.27**	.11**	.27**	.073	.04	.26**	.24*	.06*	.25**	.063	.07	.20*
Average ML Score	.26						.23					

Note. $^{*}p < .05$; $^{**}p < .01$. $N = 156$. The first step in regression was to enter human assessor scores, and then the natural language processing (NLP) scores in step 2. β = standardize regression weight. Responses were based on exercises 1 and 5, which were the exercises those with job performance data had responses to. AS = assessor scores.

strongly correlates (in the .60s—.70s) with essay scores from various university entrance exams (Kobrin et al., 2011). It should be acknowledged that response length was more strongly correlated with the NLP scores than with assessor scores. For example, the BERT scores correlated on average .74 with response length. Thus, NLP-based scoring seems to favor texts that are longer, and which more thoroughly address all the demands of the exercises. Overall given the complexity of solving the interpersonal and business challenges posed to participants in the AC stimulus, we believe that response length reflects not just verbosity but dimension-relevant content and skill by candidates.

3.7 | Discussion

This research explored the viability of automatically scoring AC data using ML. Working with a large, operational, and text-rich AC dataset, we compared the psychometric properties of the derived ML scores to human assessor ratings and found several noteworthy results.

First, transformer scores exhibited superior psychometric evidence than the other NLP methods. Although the BOW-based methods exhibited generally favorable psychometric evidence, even small improvements in validity are meaningful in high-stakes setting. If researchers and practitioners are considering applying NLP to automatically score ACs, we therefore suggest they consider using transformer models, given the superiority of transformer scores in this study, their widespread use in other NLP tasks (e.g., Min et al., 2021), and the ability to apply transformer models on smaller datasets via transfer learning (Wolf et al., 2019).

Likewise, it is worth considering why SBOW outperformed CBOW. We wouldn't expect CBOW to outperform SBOW in terms of convergent correlations because SBOW has larger document term matrices. By not filtering text, SBOW can use all of an AC candidates' text in reproducing human judgments, and modern ML algorithms like XGBoost will naturally identify the word phrases that should receive the most weight. In comparison, CBOW may have been harmed by excluding text that potentially could have been useful. This could have occurred if there was any imprecision when identifying dictionary words that filtered text, and it is also possible that some of the surrounding text that was removed via CBOW was in fact relevant to the construct at hand. Either way, BERT more effectively handles the goal of CBOW, which is to provide a more sophisticated representation of text when trying to predict ground truth scores.

Second, the transformer scores exhibited strong convergence with human assessor scores (.69), higher general stability than assessor scores, and explained unique variance in job performance ratings above and beyond assessor scores. The strong convergence is particularly favorable and is similar or higher than other efforts in the organizational sciences to automatically score text-based assessments using NLP (e.g., M. C. Campion et al., 2016; Hickman et al., 2021). It is also in line with inter-rater reliability of human-to-human AC ratings (Lievens, 2009). The value of .69 suggests that NLP and assessor scores share overlap, and given this, companies may benefit from leveraging NLP scoring into their AC assessment process. This might be especially useful when multiple human assessors rate each AC candidate, and the NLP scorer might then replace one of the humans as a method to save costs.

However, beyond the simple cost savings to using NLP to automatically score AC responses, the results suggest that the derived NLP scores captured consistent and meaningful variance. There are likely several reasons for this. Most notably, humans are prone to inconsistency when evaluating targets; with a great deal of information presented during the AC, it is cognitively challenging for humans to thoroughly consider all information and weight it consistently for every AC candidate (cf. Zedeck, 1986). On the other hand, the NLP algorithm evaluates and then scores all text for all candidates in the exact same way. This likely contributed to NLP scores that correlated higher with human ratings than human ratings correlated with each other. NLP scores will not differ based on assessor mood or inattention. Although automated algorithms will likely have deficiencies in judgment and may not outperform human assessors in all cases, the resulting standardization of ML algorithms is a major benefit.

3.7.1 | Limitations

There are several limitations to this research. First, validity is a unitarian concept (Binning & Barrett, 1989), and the ML scores exhibited evidence of validity according to multiple sources (convergence with assessor scores, reliability, correlations with job performance ratings). Nonetheless, some of the results do raise questions about the variance being captured by the ML algorithm. It was notable that the ML scores exhibited higher reliability estimates and higher criterion-related validity than the assessor scores. It is unclear exactly what language style the NLP algorithms detected that accounted for consistency (i.e., correlations of scores over time) and substantive prediction of future job performance. It is unclear whether the gains in validity were due to capturing some unique, substantive language factor or whether the gains were primarily because of the standardization of ML-based scoring. This is a foundational question for the use of ML for scoring ACs and one that further research should address.

The sample sizes for the general stability sample and the criterion-related validity sample were also both small. Although this is one of the first studies to test the operational validity of NLP-scored ACs, larger samples would lend more confidence in findings. Additionally, there were issues with the reliability estimates, in that the time between AC administrations was very large and different assessors rated the same candidates between testing times. Future research would ideally examine the test-retest reliability of AC scores using a shorter interval between test administrations and holding the assessors constant across the AC candidates.

Finally, it may be worthwhile to investigate the necessary sample sizes required to train ML scores from AC responses. Our training dataset was rather large, and many companies may not possess large enough databases to train reliable ML models. Future research should investigate just how large the training data needs to be to obtain acceptable psychometric properties for the derived ML scores.

3.7.2 | Additional implications

In conclusion, we discuss some implications about the interpretability of the ML scores (and particularly the Transformer scores) and how ML scoring can be practically used in ACs.

BERT takes into consideration the context of text and how words/topics/themes relate to one another. From a high level, transformer AC scores capture AC candidate mastery of the exercise as a whole and the quality of candidate reasoning when responding to it. Although transformers are likely to produce scores with superior psychometrics (in comparison to BOW-based methods), one of the downsides to transformer models is a lack of transparency in what is ultimately causing scores to be high or low. With BOW, researchers can at least examine how specific word phrases relate to the target scores and apply methods such as partial dependency plots to examine how changes in word phrase usage influence model predictions. There is no easy way to do this with BERT solutions. That said, if hiring managers or other HR professionals wished to better understand what drives scoring, providing them with examples of texts scored very low and very high by the model may be insightful in gaining an understanding. There probably isn't a one-size-fits all approach in this regard, but inspection of several narratives and their subsequent NLP scores could inform end-users of what's driving scoring. Nonetheless, the diminished interpretability of transformer scoring has its drawbacks, and organizations, may be willing to sacrifice the modest gains in validity of using a more complex ML algorithm for the easier interpretation afforded by other methods. Additionally, we find it important to reiterate that the inputs to the transformer models are themselves work-related. With ML solutions, it is important to control what variables are ultimately included as predictor features, therefore ensuring that ML scoring is not allowing work-irrelevant variables to impact derived scores.

Second, we would like to provide some guidance on how this automated scoring approach might differ across different AC designs. Within this study, each exercise was designed to assess only one dimension. When ACs exercises measure more than one dimension though, the transformer scoring approach used here could still be applied in a similar manner. Assuming the SME ratings are reliable and accurately reflect the targeted constructs, numerous BERT models could be formed within each exercise, or one for each dimension in that exercise. Likewise, it would be possible to train a shared BERT language model using all AC exercise text and apply multiple dimension-specific output heads to predict each of the exercise dimension scores. That said, it is worth noting that the quality of the derived scores would depend upon the quality of SME ratings. There is a long and contentious history surrounding the construct validity of AC ratings (e.g., Lievens, 2009). In brief and in relation to this point, it is common for different dimensions to exhibit large correlations with one another within exercise. Thus, any ML-based score would likely recreate this trend, and it would therefore be important to ensure that the quality of SME judgments is acceptable before embarking upon any ML-based scoring.

4 | STUDY 3: USING ARTIFICIAL INTELLIGENCE TO MAKE BETTER PRE-HIRE ASSESSMENTS[8]

The modern hiring process is being transformed by the rapid integration of deep learning and artificial intelligence (AI). The impetus behind this paradigm shift lies in the potential of deep learning's analytical prowess to enhance the efficiency and efficacy of hiring decisions. The promise of leveraging open-ended applicant text data collected during pre-hire assessments is particularly enticing to hiring managers as a tool for further optimizing hiring outcomes. Although interest in this type of data is not new (for example, see the literature on constructed response tests from the early 2000s; Arthur Jr et al., 2002; Edwards & Arthur Jr, 2007), having human raters score thousands of responses is an expensive and time-consuming process (M. C. Campion et al., 2016). These costs have previously limited the practical viability of utilizing open-ended data in large-scale hiring contexts. However, recent developments in the Computer-Assisted Text Analysis (CATA) literature (E. D. Campion & Campion, 2020) suggest that computers may be able to produce ratings similar to human ratings (M. C. Campion et al., 2016; Hartwell et al., 2022) and may even help predict employee performance and turnover (Sajjadiani et al., 2019).

The excitement surrounding this new technology is palpable, but its widespread adoption has led to AI implementations in hiring contexts advancing faster in some cases than the corresponding research. Unfortunately, concerns over test security have confined some of the most significant advancements in AI selection techniques to isolated and secretive development silos, which limits their broader visibility and impact. To fix this, we need a more open dialogue regarding the challenges and opportunities of using AI in staffing contexts among researchers and practitioners. In the present paper, we contribute to this conversation by detailing the development, implementation, and evaluation of a real-world application of deep learning as a supplement to an existing test battery. By presenting an in-depth look at the methods, procedures, and challenges of using deep learning in practice, we hope this research can serve as a realistic illustration of how organizations are applying these tools in the wild.

In particular, our efforts were organized around two broad research goals. First, deep learning's viability in hiring contexts is largely contingent on the model's ability to reliably reproduce scores provided by human raters (Hartwell et al., 2022). Therefore, building on recent work by M. C. Campion et al. (2016), we sought to evaluate whether a deep learning model could consistently approximate human evaluations of candidate responses in a large-scale selection context using an assessment designed explicitly for this purpose.

> *Research question 5:* Can ratings of applicant responses to open-ended questions produced by a deep learning model approximate scores provided by human raters?

Second, the ultimate goal was to use these ratings to improve hiring decisions. Therefore, we sought to evaluate the performance of deep learning ratings of applicant open-ended text responses designed to supplement an existing selection battery. Specifically, we were interested in examining (a) the predictive validity of model ratings and (b) the potential for adverse impact associated with including deep learning ratings in the selection battery.

Research question 6: What is the predictive validity of deep learning ratings of applicant open-ended text responses?

Research question 7: To what extent do deep learning ratings of open-ended applicant responses in a selection battery contribute to the risk of adverse impact?

4.1 | Methods

4.1.1 | Setting and sample

This paper describes a recent implementation of deep learning technology to supplement an existing selection system. This system described here was designed to evaluate and select candidates for managerial positions and has been in use since 2020. The analyses we present here describe data from over 11,000 real-world applicants. In addition, we present analyses associated with a subsample of 260 supervisory ratings collected as part of a predictive validity study.

4.1.2 | Selection assessment context

The manager assessment described here is an unproctored online assessment selection battery designed to simulate the daily work of a manager in the partner organization. An external consulting firm developed customized content for each assessment based on a thorough job analysis of the managerial role.[9] The assessment begins with a realistic job preview, which presents candidates with a broad overview of the managerial position. Candidates are then allowed to navigate the assessment at their own pace and in the order they choose—a design choice intended to simulate the high autonomy of the role. The assessment includes five distinct measures (i.e., Situational Judgment, Written Responses, Problem Solving, Biodata, and Personality). Overall scores for the various assessment sections are weighted using a scheme designed to maximize prediction while minimizing group differences. This selection system is designed to be compensatory, meaning good performance on one assessment can offset poor performance on another. Performance on the deep learning live scoring composite described below informed approximately 10% of the applicant's overall assessment score.

After completing the assessment, candidates are given a banded score of 1, 2, or 3 based on candidate performance on the assessment relative to other applicants. Scoring is normed so that approximately a third of all candidates fall into each of the three bands. Hiring managers are given access to each candidate's score in the form of banded competency ratings and summaries of their respective strengths and weaknesses for use in making hiring decisions. Developmental feedback reports are also generated for each candidate based on their assessment responses.

4.1.3 | Open-ended prompts

During the assessment, three open-ended prompts are given to all applicants describing real-world scenarios a new Manager will encounter. The first question presents a scenario in which competition among managers is causing internal communication issues. The candidate is asked to describe, using 25 characters or more, how they would address these concerns and why they believe their ideas will be effective. The mean response length (in total words) to Question 1 was 77.21 (*SD* = 47.33, max = 653). The second question presents concerns regarding trends in the employment lifecycle (e.g., hiring, retention, or training) and asks candidates how they would address these trends and why they believe their ideas will be effective. The mean response length (in total words) to Question 2 was 86.40 (*SD* = 53.21, max = 662). The third question deals with employee motivation and asks candidates to provide some words to help inspire their team. The mean response length (in total words) to Question 3 was 101.37 (*SD* = 50.99, max = 734).[10]

4.1.4 | Competency development and human ratings

With the free response items in hand, our next task was training a group of human raters to evaluate a subsample of candidate responses. A multi-step process was used to determine the scoring scales for each competency the human evaluators were asked to rate. First, SMEs from the organization reviewed the questions along with candidate responses and provided insight into what differentiated good responses from bad responses. Based on this input, we developed themes (i.e., competencies) for each question and confirmed these themes with organizational SMEs in a follow-up session. A scoring rubric was created for each theme, scales for each item were established, and SMEs further confirmed their content relevance. Online Supplement C provides details on the scale-point development process and the specific benchmarks human raters used to evaluate each competency. After multiple calibration sessions, trained raters evaluated 500–1,000 candidate responses on each theme.[11]

4.1.5 | Deep learning integration

To enable comparisons between human raters and the deep learning algorithm, candidate answers to the open-ended prompts were also scored using a deep learning algorithm trained on the same sample of applicant responses. The deep learning model data comprised raw text provided by the applicants in response to the open-ended prompts. These data were preprocessed to remove specific named entities. Locations, people, and organizations were replaced with generic tokens to avoid potential biases associated with particular entities from these categories using the open-source software spaCy (Honnibal et al., 2020). For instance, the name of a person (e.g., Tammy) would be replaced with [PERS], and the name of a real-world organization would be replaced with [ORG]).[12] Once approximately 1000 candidates completed the assessment, the deep learning model was tasked with reproducing subject-matter-expert ratings of various job-relevant competencies for the open-ended items. These items were originally research-only, meaning data was collected but not scored, with formal scoring to be implemented later.

The model described here was trained on an AWS spot instance with Python and PyTorch using an Nvidia V100 16GB GPU. In this implementation of deep learning, we used the RoBERTa architecture (Y. Liu et al., 2019), which is itself a specific implementation of the transformer architecture developed by Vaswani et al. (2017) coupled with a custom multi-task regression head (described below). One advantage of the transformer architecture is that it allows the model to look at the entirety of a body of text while focusing the attention mechanism on important words and ignoring irrelevant words. For instance, the word “bank” could refer either to a financial institution or the edge of a river. The attention mechanism tells the model to focus on other words in the sentence like “loan” or “river” and ignore words like “I,” “the,” or “went” to establish which bank the sentence is referring to from the context of the surrounding words.

The highest-level building block in deep learning is called a layer, which receives information from the model, transforms it, outputs it, and passes the new values as inputs to the next layer. In this context, multi-task learning refers to a neural network process whereby a model shares the unique predictions of hidden layers between tasks while also using several task-specific output layers (Ruder, 2017). Consistent with the RoBERTa architecture's specifications, our language model consists of an embedding layer, 12 transformer encoding layers that convert inputs from prior layers into more simplistic representations, and a fully connected layer linked to all previous nodes within the neural network's hidden layers. The text input into the model is split into individual words/numbers/punctuations or subwords (i.e., tokenized) before passing this input to the embedding layer. The embedding layer converts each token to a vector representation, and these embeddings are passed through the encoders and fully connected layers to produce RoBERTa's characteristic 768-dimensional representation of the text. The custom multi-task regression head takes the language model's data inputs and processes them through a dropout layer and two fully connected layers to make ten predictions.

One challenge we faced when implementing the multi-task regression head was that our dataset did not always have all possible labels for each set of responses. We solved this problem by dividing the data into ten different datasets corresponding to each label and dropping the test responses that did not have the label in the corresponding dataset. However, this meant that the language model for each label would only get exposed to the text subset that contained the corresponding label. These transformer models can be cost prohibitive if they are in constantly in memory in a production environment. As such, we determined that it made the most sense to develop one model that accurately makes ten predictions rather than ten models that each accurately makes one. To do this, we used a Mean Square Error associated with model predictions that set the new error gradient to zero for missing labels when passed backward through the model while still adjusting the language model according to the backpropagated error. This solution enabled us to overcome the problem of sparse training data by using and putting into production one model that can make predictions on ten different competencies across three unique responses.[13]

To maximize model generalizability, we implemented a k-fold cross-validation strategy. K-fold cross-validation involves slicing the data into "k-folds," each representing an even cut of available data. For example, in a dataset with a sample size of 1000 and a five-fold cross-validation, 800 responses and labels would be used to train the algorithm, and 200 responses would be withheld for model evaluation. This process would be repeated with each set of 200 responses used as the holdout set, while the remaining 800 would be used as the training set. This results in five distinct models within each fold, making separate predictions on each holdout set within the folds. The predictions made on those holdout sets are aggregated across the entire data set to evaluate the performance of the algorithmic scoring methods.[14] The k-fold methodology described here allowed us to test multiple hyperparameters[15] without overfitting the data. We used an early stopping methodology during k-folds and our final model training. Therefore, the number of epochs was not included as a specific hyperparameter.

4.1.6 | Competency retention decisions and development of live scoring composite

Our final task was evaluating each competency's performance individually to help us decide which competencies (if any) should factor into the applicants' overall assessment score during live scoring. Two criteria were used to make this decision. First was the competency's response base rate. A low base response rate for a competency meant that the candidates were not articulating a behavior frequently enough in response to the question prompt to facilitate reliable evaluation. Attempting to score competencies with low base rates could lead to interpretational issues within the broader selection context. On this criteria, our analyses revealed that text content relevant to the Future Planning and Follow-Up competencies was provided by less than 10% of candidates. As a result, these two competencies were omitted from live scoring.

The second retention criterion was the observed convergence between the deep learning model and SME ratings. As shown in Table 5, the Encourage/Motivate and Teamwork competencies produced correlations between the deep learning ratings and the human raters below the .60 target threshold described by M. C. Campion et al. (2016). Therefore, they were marked for removal (more details on these analyses can be found in the RQ1 results section below).

This left us with six competencies that met our inclusion criteria; three for Question 1, two for Question 2, and only one for Question 3 (Appreciation/Recognition). Given constraints on applicant time and the organization's desire for shorter assessments, retained items needed to be robust enough to justify their inclusion. Due to the poor performance of two of the three competencies associated with Question 3, this item was removed (along with the otherwise successful Appreciation/Recognition competency). Thus, scores for the following five competencies were retained for subsequent analyses:

> *Influence and communication* – The ability to enact power over others and encourage open conversations among team members

TABLE 5 Deep learning correlations with mean SME rating.

			Internal model performance indices			Pre-consensus SME interrater reliabilities[a]	Convergence with SME ratings
	Deep learning competency	Scale	*MSE*	*F1* score	*AUC*	*r* between SME ratings	*r* between DL ratings and Average SME Ratings
Question 1	Influence and Communication	1 to 5	.542	–	–	.65	.81
	Collaboration	1 to 5	.560	–	–	.62	.77
	Embracing Competition	Binary	–	.870	.975	.65	.85
	Future Planning	Binary	–	.930	.982	.93	.92
Question 2	Foster Internal Talent	0 to 2	.127	.760	–	.55	.77
	Creative Recruitment	1 to 5	.348	–	–	.71	.78
	Follow Up	Binary	–	.500	.500	.55	.85
Question 3	*Appreciation (Recognition)*	0 to 2	.320	.600	–	.64	.74
	Encourage/Motivate	0 to 2	.397	.390	–	.47	.49
	Teamwork	0 to 2	.512	.260	–	.46	.56

Note. Italicized competencies were removed from the assessment (see body text for rationale).

N = 500–1000 (All competencies were evaluated using an initial set of 500 SME responses. A second round of 500 SME evaluations was then applied to the five competencies marked for retention in the final models, resulting in a total N of 1000 for these responses),.

[a] Post-consensus meeting interrater single rater reliabilities for the human ratings (i.e., correlations between the human ratings after discussion during consensus meeting) ranged from .80 (Creative recruitment) to 1.00 (Foster internal talent).

Collaboration – The ability to work with a team to create solutions, share strategies, and build morale

Embracing competition – The ability to encourage healthy competition and reframe competition in a positive and constructive manner

Foster internal talent – The ability to foster deep applicant pools of internal talent

Creative recruitment – The ability to generate novel ideas for recruiting new applicants

Of these five competencies, the partner organization selected two—Collaboration and Foster Internal Talent—for the live scoring of applicant data based on (a) the strength of their relationships with job performance and (b) their minimal contribution to adverse impact relative to the other competencies. As such, a live scoring composite was created by averaging the standardized ratings of the two competencies for each applicant. Performance on this composite informed approximately 10% of the applicant's overall assessment score.[16]

4.2 | Results

4.2.1 | Descriptive statistics

Means, standard deviations, and intercorrelations among the five deep learning competency scores are displayed in Table 6. The average absolute intercorrelation among the deep learning competency scores was .17, suggesting that each construct was reasonably independent. This value was slightly higher than the average absolute intercorrelation among the SME-rated competency scores of .11, which is typical for deep learning implementations.

4.2.2 | Research question 5: Convergence between deep learning ratings and human ratings

As noted above, our first overarching research question was whether deep learning could be used in a large-scale selection context to quickly and reliably approximate human evaluations of candidate responses. As shown in Table 5, the deep learning model ratings correlated with human ratings at a rate greater than M. C. Campion et al.'s (2016) target threshold of .60 for two of the three questions and eight of the ten competencies; in some cases as high as .92. These findings support the notion that deep learning can be used to supplement or potentially even replace human evaluations of applicant open-ended text responses within the hiring process.

Further evidence of this convergence is presented in Table 6. Specifically, we found that the general pattern of intercorrelations among the deep learning competencies was similar to that of the human ratings. For instance, competencies with high intercorrelations in the human ratings (e.g., Creative recruitment) also showed higher intercorrelations for the deep learning ratings. Conversely, competencies with lower intercorrelations in the human ratings (e.g., Foster internal talent) also showed lower intercorrelations in the deep learning ratings. In fact, the only statistically significant negative intercorrelation among the human ratings (the correlation between Fostering internal talent and Creative recruitment) was also negative for the deep learning ratings. This general convergence in the direction and magnitude of intercorrelations offers further support for the model's ability to replicate the structure of human ratings.

Finally, we conducted exploratory analyses examining the convergent and discriminant validity of the deep learning model ratings and other assessment content. These analyses served two purposes. First, they facilitated a better understanding of the broader psychological constructs represented within the deep learning model's evaluations of the applicant's responses to open-ended prompts. As shown in Table 7, the largest deep learning correlations with other assessment content were with the problem-solving simulation scores. This relationship suggests there is a

TABLE 6 Correlations among deep learning competency scores and supervisory performance ratings.

		Deep Learning Competencies					
		Influence and Communicate	Collaboration	Embracing Co petition	Foster Internal Talent	Creative Recruitment	Live Scoring Composite
Supervisory Ratings	Overall Rating Average	.00	.05	.03	.05	.00	.07
		(.00)	(.07)	(.04)	(.08)	(.00)	(.11)
	Overall Effectiveness—Single Item Rating	–.10	.02	.01	.03	–.04	.03
		(–.15)	(.03)	(.01)	(.05)	(–.06)	(.05)
	Best Ever—Single Item Rating	.03	.06	–.02	.07	.00	.09
		(.04)	(.09)	(–.03)	(.11)	(.00)	(.14)
	Prioritization and Delegation	–.01	.07	.00	.01	–.01	.07
		(–.01)	(.10)	(.00)	(.02)	(–.02)	(.11)
	Influence and Communicate	–.02	.05	.03	.07	–.03	.08
		(–.03)	(.07)	(.04)	(.11)	(–.05)	(.13)
	Collaboration	.06	.06	.08	.05	.03	.08
		(.09)	(.09)	(.11)	(.08)	(.05)	(.13)
	Focus on Associates	–.04	.02	.01	.09	–.04	.07
		(–.06)	(.03)	(.01)	(.14)	(–.06)	(.11)
	Creativity	–.02	.06	.03	.01	.00	.05
		(–.03)	(.09)	(.04)	(.02)	(.00)	(.08)
	Deliver Results	.04	.06	.01	.05	.02	.08
		(.06)	(.09)	(.01)	(.08)	(.03)	(.13)
	Customer Centered	–.05	.02	.01	.01	.05	.02
		(–.07)	(.03)	(.01)	(.02)	(.08)	(.03)
	Lead Change	.00	.04	–.01	.06	–.04	.07

(Continues)

TABLE 6 (Continued)

	Deep Learning Competencies					
	Influence and Communicate	Collaboration	Embracing Co petition	Foster Internal Talent	Creative Recruitment	Live Scoring Composite
	(.00)	(.06)	(–.01)	(.09)	(–.06)	(.11)
Problem Solving and Decision Making	.03	.06	.02	.04	.02	.07
	(.04)	(.09)	(.03)	(.06)	(.03)	(.11)
Develop Others	–.04	–.01	.07	.03	–.02	.01
	(–.06)	(–.01)	(.10)	(.05)	(–.03)	(.02)

Note. Correlations in parentheses were corrected for unreliability and range restriction using standard estimates of .54 (Shen, Cucuina, Walmsley, and Seltzer, 2014), and the sample and population SDs. For range restriction we used the Thorndike Case 2 formula ($R = (rS/s)/ \mathrm{sqrt}(1\text{-}r^2 + r^2(\hat{S}/s^2))$).
All correlations shown in Table 6 were statistically nonsignificant ($p > .05$)$N = 260$.

TABLE 7 Subgroup differences in deeplearning (DL) competency scores and the impact of deep learning assessments on group mean differences in cumulative assessment z-scores and hiring ratios.

Effect sizes (d-scores) for majority group (top) vs. subgroup (bottom) comparisons on DL competencies				
	Gender	W–B	W–H	W–A
	Males = 7,513	White = 6,886	White = 6,886	White = 6,886
	Females = 3,670	Black = 1,846	Hispanic = 1,452	Asian = 351
Influence and Communicate	–.07	.37	.17	.27
Collaboration	.00	.11	.15	.27
Embracing Competition	.20	.18	.19	.33
Foster Internal Talent	.16	.00	–.05	–.08
Creative Recruitment	–.04	.35	.32	.38
Live Scoring Composite	.10	.08	.07	.15

Group mean (standard deviations) differences in cumulative assessment z–score w/ and w/o DL competencies			
Group	Full assessment	Deep learning removed	Deep learning replaced
Females (*n* = 3,670) vs. males	–.09 (1.07)	–.08 (1.07)	–.10 (1.07)
White (*n* = 6,866) vs. others	.06 (.97)	.06 (.97)	.08 (.95)
Black (*n* = 1,846) vs. others	–.14 (1.02)	–.13 (1.02)	–.17 (1.02)
Hispanic (*n* = 1,452) vs. others	.00 (1.03)	.02 (1.03)	.00 (1.02)
Asian (*n* = 351) vs. others	–.19 (1.22)	–.17 (1.23)	–.20 (1.21)

Estimated hiring ratios relative to with and without DL competencies			
Group and cut score	Full assessment	Deep learning removed	Deep learning replaced
Females (*n* = 3,670) vs. males			
20% cut score	.93	.95	.92
50% cut score	.91	.93	.91
80% cut score	.93	.94	.90
Black (*n* = 1,846) vs. white			
20% cut score	.93	.94	.92
50% cut score	.85	.86	.81
80% cut score	.75	.76	.67
Hispanic (*n* = 1,452) vs. white			
20% cut score	.96	.97	.95
50% cut score	.97	.98	.94
80% cut score	1.02	1.07	.95
Asian (*n* = 351) vs. white			
20% cut score	.90	.92	.89
50% cut score	.91	.93	.91
80% cut score	.92	1.00	.91

Note. The first model (i.e., full assessment) presents values associated with the full live assessment, including the deep learning competencies. The second model (i.e., deep learning removed) presents values where the deep learning score weighting is reduced to 0 and allocated equally among the remaining assessments. The third model (i.e., deep learning replaced) presents values where the deep learning score weighting is reallocated to cognitive assessments.

clear cognitive element to what is being measured by the deep learning model. The situational judgment simulation also evidenced correlations as high as .20 with some of the deep learning competencies. This finding also makes some conceptual sense as both the situational judgment simulation and open-ended responses assess some form of interpersonal skill.

Second, these analyses provided another opportunity to explore similarities and differences between the model and human ratings regarding how they related to other assessment content. Here again, Table 7 supports a general pattern of convergence among the correlations produced by the model and human raters with other assessment content. In summary, our analyses corroborate assertions by M. C. Campion et al. (2016) that algorithmic evaluations of open-ended content can be used to approximate scores assigned by human raters.

4.2.3 | Research question 6: Criterion-related validity

To examine the criterion-related validity of deep learning scores (Research Question 6), we collected 260 supervisory ratings as part of an initial predictive validation study. The criteria used in this study comprised 24 items measuring job-relevant competencies, two single-item ratings, and two control items asking whether the supervisor knew the manager's performance and whether they were confident in the accuracy of their ratings, which were used to ensure their supervisors had enough information to make these evaluations. The competencies were measured with two items each and aggregated into an Overall Rating Score ($\alpha = .96$).

As shown in Table 6, our analyses suggest that correlations between the deep learning competency scores and supervisor evaluations of subsequent manager performance were small, albeit non-zero, and positive in most cases. Specifically, small, positive correlations between the deep learning competency scores and performance criteria were observed for the Collaboration, Embracing Competition, and Foster Internal Talent ratings. Slightly larger relationships also emerged between the Live Scoring Composite and the various performance criteria ratings. In contrast, the Influence and Communication and Creative Recruitment deep learning competency scores produced a less consistent pattern of relationships with performance criteria with a mix of positive and negative effects. Incremental predictive validities for competency scores were similarly modest, particularly when evaluated next to the situational judgment and problem-solving assessment, with which several competencies shared conceptual similarities. Furthermore, it is important to note that none of the correlations between competency scores and supervisor evaluations achieved statistical significance. Hence, regarding Research Question 6, only some of the observed correlations some support to the criterion-related validity of deep learning competency ratings. However, even the largest of these relationships tended to be pretty small ($r = .07$ uncorrected and $r = .11$ corrected).

Nevertheless, it is important to understand that small effects alone do not necessarily rule out the potential practical utility of using these scores to make hiring decisions. In hiring contexts of this magnitude, even small increases in predictive capabilities can make a big difference in organizational outcomes. For instance, as shown in Online Supplement Table C3, we found that the proportion of applicants rated by their managers as “Above Average” on each of the performance criteria was consistently higher (3.5%–12.1%) among employees with live scoring composite ratings in the top 50% of applicants compared to employees with live scoring composite ratings in the bottom 50% of applicants. From the organization's perspective, this potential increase in employee performance was enough to justify the decision to include these open-ended items as a supplement to the more comprehensive hiring system.

4.2.4 | Research question 7: Adverse impact

Another practically important consideration in high-stakes selection contexts is the risk of adverse impact associated with using the tool to help make hiring decisions. Given the clear cognitive element of model scores discussed above, we were particularly interested in the possibility of subgroup differences associated with the deep learning model ratings.

Table 7 presents analyses using a large sample of over 11,000 real-world candidates who reported demographic information that (a) documents subgroup differences (in d-scores) for the five deep learning competencies individually and (b) explores the consequences of including (and omitting) deep learning scores on overall group mean assessment scores and hiring ratios.

As shown in Table 7, small gender differences emerged in the various deep learning competency scores. In some cases, these differences slightly favored women (e.g., Influence and Communication & Creative Recruitment), whereas others slightly favored men (Embracing Competition and Foster Internal Talent). Small racial differences favoring White candidates were reported as well, with d-scores that fell within ranges typically associated with other procedures thought to have low adverse impact, such as structured interviews and biodata inventories (Ployhart & Holtz, 2008). Moreover, the risk of adverse impact associated with deep learning scores was not greater than that associated with other assessments already being used. In this assessment context, adverse impact is monitored quarterly, and the weights of scores are adjusted if the risk of adverse impact is identified. Given the relatively small contribution of the deep learning competencies to group differences and adverse impact documented in Table 7, we concluded that the potential risk of hiring discrimination associated with using the deep learning competencies as part of the overall scoring composite was minimal.

4.3 | Discussion

This paper contributes to research on using deep learning in hiring contexts by describing and evaluating the development of a modern implementation of a deep learning model within a real-world hiring process specifically designed to automate the evaluation of open-ended applicant responses as a supplement to an existing test battery. In this section, we present the key lessons we learned from this implementation and discuss the implications of our findings for future researchers and practitioners working in the machine-learning space.

One of the most exciting implications of the present study is the pronounced support we found for M. C. Campion et al.'s (2016) claims that computer scoring algorithms can reliably reproduce human evaluation of applicant responses to open-ended prompts. This is no small feat, as human language and communication contain a level of complexity and nuance traditionally thought to be beyond the reach of AI. Encouragingly, our data contribute to a growing body of research (e.g., Hartwell et al., 2022), making a strong case that the algorithms can accomplish this task while producing only minimal adverse impact risk.

Using AI rather than human raters can help practitioners overcome barriers to using open-ended text responses in pre-hire assessments by reducing the time needed to collect, process, and score applicant data. As M. C. Campion et al. (2016) noted, these savings can be substantial. For instance, in the present effort, it took approximately 250 human hours across three raters to score just 500 applicants, not counting the time raters spent in scale orientation and training. It would take the same trained human evaluators three and a half months working around the clock to score the data collected by this organization in just a single month. Even with access to a large team of human raters, substantial delays in hiring decisions would be inevitable, resulting in the loss of qualified applicants (Ryan et al., 2000). In contrast, the deep learning model produces its ratings nearly instantly, producing human-like competency ratings without the time and resources required of human evaluators.

Establishing a link between human and deep learning evaluations also helps address another charge commonly leveled against deep learning: the model's interpretability. Our findings demonstrate that RoBERTa can evaluate words and phrases in open-ended text much like a human would. Establishing this linkage allows deep learning to be used for decisions where human ratings are already accepted (see Mondragon, in press). Furthermore, the algorithm allows specific words and phrases used in evaluation to be identified and explored. As such, explaining the model's functioning to executives, candidates, and potentially even litigants is more straightforward than in other implementations where model criteria are more ambiguous.

Nevertheless, this enthusiasm should be tempered by our finding that model scores underperformed in predicting supervisor evaluations of future job performance. Unfortunately, our data did not facilitate an examination of the predictive validity of human ratings with job performance. As a result, we cannot say for certain whether these small effects were a result of the deep learning model itself or limitations of the underlying human evaluations upon which the model was trained. Future research should attempt to validate human and deep learning rating side by side using sufficiently powered predictive validity studies to facilitate richer comparisons of the predictive nuances associated with human versus deep learning ratings.

It is also possible that the characteristics of the open-response items themselves were limiting the predictive potential of the deep-learning ratings. For instance, competencies associated with Question 3 generally performed worse than those associated with Questions 1 and 2, which could result from greater subjectivity in prompted responses associated with this item. In the past, organizations have not been incentivized to invest in developing highly reliable and construct valid evaluations of open-ended text due to concerns over cost. By removing these barriers, deep learning can open up a renewed interest in research on best practices for constructing and validating open-ended questions for use in pre-hire assessments.

Our data also revealed a pattern of strong, positive correlations between word count and competency scores. One possible interpretation of this finding is that the deep learning model over-valued wordy responses. Although possible, an applicant's ability and willingness to provide thoughtful and detailed descriptions may also speak to their general quality as a candidate. Indeed, the nature of open-ended response scoring necessarily places some responsibility to offer enough information to draw favorable conclusions about the candidate. This is especially true for the Influence and Communication competency, where response length was strongly correlated with response performance. A supervisor who responds to questions with only a few words is unlikely to communicate well with their subordinates, which aligns with our finding that this competency had the strongest correlation with word count for both the deep learning ($r = .61$) and human ratings ($r = .53$). Nevertheless, if other deep learning competencies produced correlations this closely related to word count, we would likely want to re-examine the items' construction. In summary, our position is that correlation with response length can certainly be a nuisance factor when evaluating open-ended responses. However, there is also a case to be made that substantive information in these relationships can benefit hiring decisions.

Moving forward, we hope that deep learning's ability to reproduce and ultimately scale human evaluations helps spark a renewed scientific focus on best practices for gathering and scoring valid and reliable assessments of applicant responses to open-ended prompts. In the meantime, it is important to emphasize that as a method, deep learning should remain conceptually separate from the constructs the model is designed to evaluate (Arthur & Villado, 2008). Positive relationships between model ratings and other assessment content might allow deep learning to substitute for a longer assessment that taps the same underlying constructs (e.g., the Problem-Solving Assessment). Until then, our results suggest that deep learning should only be used as a supplement to, not a replacement for, a well-validated assessment.

5 | STUDY 4: DEVELOPING AND VALIDATING AUTOMATED SCORING FOR AN AUDIO CONSTRUCTED RESPONSE SIMULATION[17]

Simulations are widely used in personnel selection to measure knowledge, skills, abilities, and other characteristics (KSAOs). Meta-analyses demonstrate simulation criterion-related validity across a range of fidelity, from low (e.g., situational judgement tests, or SJTs; McDaniel et al., 2007) to high (e.g., ACs; Arthur Jr et al., 2003). Research suggests that higher validity is associated with increased response fidelity, such that constructed audio responses yield higher validity than constructed written or multiple-choice responses (Funke & Schuler, 1998). However, scaling constructed response to high-volume personnel selection settings has historically been cost-prohibitive given its reliance on human scoring. Though recent technological advances show promise for automated, computer-based scoring of written constructed responses (e.g., M. C. Campion et al., 2016), no published research to date has assessed the feasibility of

TABLE 8 Human interrater reliability results (full data set, including train, test, and validation) and convergence between computer and human scores (validation data sets).

KSAO	Human ICC(1)	Human ICC(2,k)	Computer R with Human Labels (Test)	Computer R with Human Labels (Validation)
Active Listening	.28	.94	.75	.74
Assertive Communication	.21	.91	.68	.66
Oral Communication	.27	.94	.73	.70
Cooperation & Coordination	.25	.93	.77	.71
Interpersonal Adaptability	.29	.94	.73	.68
Social Influence	.29	.94	.76	.67
Decision Making	.27	.94	.78	.74
Average of KSAOs			**.82**	**.76**

Note. *N* = 1709 for Human ICC. *N* = 512 for Computer R with Human Labels (Test and Validation Data).

automated scoring of audio-based constructed responses in a personnel selection context (see [author name[s] to be added, this issue, for an exception). Our study addresses this gap by evaluating whether automatic speech recognition (ASR) and NLP-based computer scoring can replicate KSAO-based human scoring of audio responses, the prediction of job performance using computer scores, and the sub-group differences for computer scores across key demographic characteristics (i.e., gender and race/ethnicity).

We took a KSAO-oriented approach to develop NLP models as opposed to a "black box" approach where NLP models are trained to directly predict job performance. We believe this approach offers stronger job relatedness, model explainability, and legal defensibility. We use findings and learnings from this research to identify areas in need of future research and to address practical scaling constraints.

In this study, we developed a simulation to measure *interpersonal* and *decision-making* skills (see skills in Table 8) because they are critical KSAOs across professional roles, are not measured well without more resource intensive procedures (e.g., interviews), and can be effectively evaluated via simulations (e.g., Clevenger et al., 2001). Incumbents were instructed to record audio responses to job-relevant situational judgment prompts, allowing for the direct expression of targeted KSAO behaviors. As audio-based response formats are also found in one-way behavioral interviews and ACs, the methods and results described in our study may be applicable to these assessment methods. Audio-based response formats provide several important advantages despite the greater complexity (e.g., scaling constraints, transcription errors, etc.) relative to written responses. Audio responses offer improved response fidelity, potentially contributing to stronger validity (Funke & Schuler, 1998), are associated with lower subgroup differences than written constructed responses, presumably due to the lower cognitive demands (Lievens et al., 2019), and can be more effectively delivered via mobile devices. Mobile engagement may improve accessibility to demographic segments such as younger applicants, women, Hispanics, and African Americans (Arthur et al., 2014). Finally, audio-based constructed responses could result in shorter assessment time[18], which is expected to positively impact candidates' assessment experience.

5.1 | Method and results

5.1.1 | Sample and data collection

We used a concurrent validation approach to score and validate the simulation. We recruited a diverse sample of incumbents across different roles and demographic backgrounds. The sample was composed of interns and recently graduated incumbents hired into US-based professional roles (e.g., software engineers, research scientists, product managers) at a large multinational technology company. Of the 9940 incumbents invited, 3174 (32% response rate) completed the simulation (26% female; 67% Asian, 23% White, 3% Hispanic, 3% Two or More Races, 2% African American). Range restriction on the KSAOs measures were expected as participants were recently hired using a battery of validated assessments as well as structured interviews. The majority of participants indicated they spent most of their life residing in the United States (57%), followed by China (18%), India (16%), and Canada (6%). The participants voluntarily participated in the research and were assured confidentiality. Participants were instructed to respond as if they were applying for a role similar to their current one and to record their audio responses in English. Participants who completed the study were awarded with an electronic icon on their internal company profile indicating their contribution to assessment research.

Stimuli and behaviorally anchored rating scales. The stimuli were drafted by four SMEs using detailed job analysis data, KSAO definitions, and related literature. All SMEs were Ph.D.'s in industrial-organizational (I-O) psychology with at least three years of post-graduation assessment development and validation experience. SMEs also drafted representative descriptions of good and poor responses to each scenario, which were used to create behaviorally anchored rating scales (BARS). A separate set of six SMEs with similar backgrounds and expertise reviewed the stimuli and BARS to confirm their job relatedness and ability to prompt targeted KSAO behaviors. The content development process resulted in a simulation with five interconnected, job-relevant scenarios, where each scenario elicits behaviors relevant to one or more of the targeted KSAOs. The scenarios provide a realistic, day-in-the-life experience where test takers interact with colleagues across diverse roles, backgrounds, expertise, and behavioral styles to collaborate, solve problems, and work toward team goals. The Online Supplement: Stimuli Construction provides greater details on the stimuli construction process, two modified sample items, and content validation results.

Participants were told to record a 1–2 min response (maximum = 5 min) to each of the five scenarios. The recording length was set based on I-O SME judgement and supported by MTurk pilot research where the average recorded response was 34 s (*SD* = 31 s). Participants were allowed up to three attempts to respond to each scenario. The last response recorded was scored (average recording length = 43 s, *SD* = 28 s)[19].

5.1.2 | Criterion measures

Managers were asked to complete a job performance survey of study participants. Managers were provided with background information, performance rating guidance, and tips on how to avoid rating biases (e.g., recency effects). Managers were asked to indicate their level of confidence in their rating accuracy. We removed cases where managers reported low confidence. Overall job performance was measured using four questions with a 5-point Likert-type scale (1 = Strongly Disagree; 5 = Strongly Agree; Cronbach's alpha = .94); a modified sample item is "[Name] demonstrates the required technical skills for the role." We also included domain-specific job performance dimensions tapping into KSAOs that are more closely associated with the KSAOs targeted in the simulation (see Funke & Schuler, 1998). We used eleven questions focused on interpersonal and decision-making competencies with a six-point scale (1 = Well Below Average; 6 = Best I've Ever Seen; Cronbach's alpha = .92); a modified sample item is "[Name] makes objective and well-informed decisions." All the job performance survey items were developed by the organization's I-O psychologists, validated against other performance criteria (e.g., objective indicators of performance, time-to-promotion, annual review ratings), and used in prior validation studies. We aggregated the scores to form overall and domain-specific job performance scores.

5.1.3 | Human scoring

Human SMEs scored audio responses on the targeted KSAOs and we used these human scores to train NLP scoring models. We recruited 40 I-O psychologists to serve as SME raters. We use the term "human scores" to refer to the ratings provided by the human SMEs. SMEs in this study completed at least one year of I-O psychology graduate training, and completed a graduate course in psychometric assessment or had at least five years of professional experience in the individual and/or leadership assessment domain. SMEs attended a 3-hour training session covering details about the research project, the rating task, and how to avoid rater biases (e.g., central tendency). SMEs then went through a 4-week calibration process where each week they completed ratings of audio responses, calibrated their ratings as a group, and compared those to "true" benchmark scores provided by three of the organization's internal assessment experts[20].

For the scoring task, the SMEs listened to the entire set of responses from each participant (each set = five responses to the five scenarios) before providing KSAO scores. A maximum of six randomly selected SMEs (out of the overall pool of 40 SMEs) rated each participant's responses according to the BARS. We estimated ICC(1), which represents interrater reliability if a single SME was randomly selected from the group of SMEs. In addition, we estimated ICC(2, *k*) to understand the interrater reliability when using means from a sample of SMEs drawn from a larger population. As seen in Table 8, ICC(1) ranged from .21 to .29 for *Interpersonal Skills* and was .27 for *Decision-Making Skills*. ICC(2, *k*) ranged from .91 to .94 for *Interpersonal Skills* and was .94 for *Decision-Making Skills*. These results suggest high levels of interrater reliability and are similar to those found in research based on written constructed responses (M. C. Campion et al., 2016). Table D1 in the Online Supplement includes detailed descriptive statistics and interrater reliability results for human scores. We restricted our sample to cases that were scored by at least five SMEs ($N = 1709$) based on measures of reliability.[21] KSAO scores were calculated by taking the mean of the individual human SME scores.

5.1.4 | Automated scoring

The automated scoring process consists of two parts. First, we used Amazon Transcribe ASR product to transcribe recorded audio responses to text. Second, we trained BERT NLP models to generate computer scores to replicate the (mean) human scores of the KSAOs.

We used transcribed text as input for model training as opposed to audio because research has shown that speech audio contains acoustic features that are strongly associated with demographic status (e.g., gender; Buyukyilmaz & Cibikdiken, 2016). Though human raters labeled responses by listening to respondent audio, raters received explicit training on how to minimize potential bias caused by demographic signals that may be present in audio files. By using text as the input for scoring, the model is unable to directly model the impact of accent, pitch, or other demographic signals. If audio signals were used as direct input features to ML scoring models, the models may learn superficial patterns of correlation between audio features and scores that may increase algorithmic sub-group differences.

As transcription accuracy can impact subsequent NLP model prediction accuracy, we empirically evaluated ASR performance using the Word Error Rate (WER) between expert human transcriptions and the ASR model output.[22] Lower WER scores indicate more accurate transcriptions. We used an expert human transcription service to manually transcribe 1200 mins of audio responses from a stratified random sample of 467 participants. Expert human transcribers also labeled each audio file for accent origin and strength and flagged potential issues with background noise, audio quality, or volume.

After ASR model adjustments (for additional details, see the Online Supplement: Automatic Speech Recognition Modelling), the ASR model showed less than .25 WER (a common accuracy benchmark; Peng et al., 2020) for each demographic group, an average WER of .15 for all participants, and less than a .10 difference across demographic subgroups (See Table D2 in the Online Supplement). Females had a lower WER than males (.11 vs. .16). We also found some small differences with respect to human transcribers' attribution of accent location where the WER was .11 for American-accented English, .21 for Chinese-accented English, and .18 for Indian-accented English. Taken together, our results suggest that the ASR model achieved viable accuracy levels within demographic groups and small differences across demographic groups.

Using the automatic transcriptions as the model input, we then trained NLP models to generate computer scores to replicate average human scores of the KSAOs. We used BERT as the base NLP model because it is the foundation for many state-of-the-art NLP task benchmarks (Devlin et al., 2019)[23]. As each participant's audio input consists of five responses, we used all five transcripts as inputs to the NLP model. The NLP model then predicted the seven KSAO scores simultaneously. For additional details on the NLP model architecture and training procedure, see the Online Supplement D (NLP Model Development).

A train (N = 1196; 70%), test (N = 256; 15%), and validation (N = 256; 15%) data set was built using a stratified random sampling approach. The training set was used to train the NLP model, the test set was used to select the highest performing NLP models, and the validation set was used to evaluate the performance of the selected models. The data set splits ensure an NLP model is likely to generalize to unseen data not included in model training.

To improve performance and lower sub-group differences, we used an ensemble of the top two performing NLP models. Model ensembles are a common technique in ML whereby the combination of model predictions shows better performance than any of the individual models alone (Caruana et al., 2004). The first model in the ensemble was trained using Demographic Parity Loss (DPL) to reduce potential subgroup differences (Agarwal et al., 2019), while the second model did not use this training adjustment. DPL adds a constraint during training to bias the model toward KSAO predictions that are *not* correlated with demographic status (additional details in Online Supplement: NLP Model Experiment). Our results showed that including DPL slightly reduced subgroup differences in ethnicity with minimal effects on NLP model accuracy (average computer-human score correlation loss = .04; see Table D8 in the Online Supplement: NLP Model Experiment). These results appear to be somewhat inconsistent with research suggesting that fairness-aware adjustments, like DPL, must create some amount of prediction bias or degradation (see study 1 in Zhang et al. 2023; this issue). For our research, we suspect that the observation of minimal accuracy loss may be due to the elimination of construct-irrelevant biases present for some raters, the relatively small sample sizes for most minority groups, and/or the relatively small subgroup differences in the human KSAO scores used for model training. Table D1 in the Online Supplement presents the model score descriptive statistics.

5.1.5 | Validity of the NLP model

We compared computer and human scores in the test and validation data sets to evaluate the convergent and discriminant validity of the NLP model. Table D3 in the Online Supplement present the detailed results on R, R^2, Mean Absolute Error (MAE; average absolute difference between the computer and human scores), and MSE[24]. Higher R and R^2, as well as lower MAE and MSE, indicate higher model accuracy. For *interpersonal skills*, R ranged from .66 to .74, R^2 ranged from .40 to .55, MAE ranged from .31 to .35, and MSE ranged from .14 to .20. For decision-making skills, R was .74, R^2 was .54, MAE was .33, and MSE was .17. These results suggest high levels model of convergence with human KSAO scores. The human and computer score correlations a comparable to M. C. Campion et al. (2016) results. Further, computer scores correlated with human scores better than individual human scores correlated with each other (ICC1s were below .3, Table 8), which was possible in part because the reliability of the mean of the raters was very high (ICC2s were above .9, Table 8).

We observed strong collinearity in the human KSAO scores, and computer scoring exacerbated this problem (see Table D5 in the Online Supplement). We suspect this may be an accurate reflection of the collinearity of these KSAOs in the workplace and could be driven by the fact that the simulation scenarios were designed to tap into multiple KSAOs simultaneously. However, "oral communication" seems to be distinct from the other constructs in both human and computer scoring and exhibited stronger convergent and discriminant validity both within and between scoring methods. Based on an anonymous reviewer's suggestion, we explored deriving underlying factors representing the KSAOs and using factor scores for NLP modeling (see Online Supplement: Factor-based NLP Model). Although the factor-based NLP model reduced collinearity in the computer KSAO scores, it resulted in overall worse convergent and

TABLE 9 Correlations between human scores, computer scores, and job performance (test and validation data sets).

KSAO	Overall Job Performance (H)	Overall Job Performance (C)	Domain-Specific Job Performance (H)	Domain-Specific Job Performance (C)
Active Listening	.07 (.20)	.08 (.19)	.08 (.14)	.11 (.17)
Assertive Communication	.12* (.16)	.14* (.25)	.12* (.10)	.18** (.24)
Cooperation & Coordination	.14* (.29)	.08 (.19)	.15* (.25)	.11 (.17)
Social Influence	.10 (.23)	.08 (.19)	.10 (.17)	.11 (.17)
Interpersonal Adaptability	.10 (.19)	.12* (.23)	.11 (.14)	.15* (.21)
Oral Communication	.16** (.26)	.17** (.25)	.17** (.23)	.21*** (.25)
Decision Making	.10 (.21)	.10 (.21)	.11 (.16)	.13* (.19)
Overall Score (Average of KSAOs)	.13* (.25)	.12* (.23)	.14* (.20)	.15* (.21)
Response Length (Word Count)	−.01 (.01)	−.01 (.01)	.02(.01)	.02(.01)

Note. H: Human Scores. C: Computer Scores. *N* = 290. Range restriction and unreliability corrected correlations in parentheses.**p* < .05. ***p* < .01. ****p* < .001 (2-tailed test).

criterion-related validities as well as subgroup differences. Therefore, we decided to retain the original KSAO-based NLP model.

Criterion-related validity was examined using correlations between computer scores and manager rated job performance in the test and validation data sets (Table 9 and Table D5 in the Online Supplement). The correlations with overall job performance ranged from .08 to .17 for *interpersonal skills*, was .10 for *decision-making skills*, and was .12 for an overall computer score computed by averaging the KSAO scores. The correlations with domain-specific performance ranged from .11 to .21 for *interpersonal skills*, was .13 for *decision-making skills*, and was .15 an overall computer score. Given the incumbent sample had gone through rigorous selection processes, we estimated the correlations correcting for range restriction (Case 3 in Thorndike, 1949) as well as criterion unreliability (using interrater reliability of .52; Viswesvaran et al., 1996). In the test and validation data sets, the corrected correlations with overall job performance ranged from .19 to .25 for *interpersonal skills*, was .21 for *decision-making skills*, and was .23 for an overall computer score. The correlations with domain-specific performance ranged from .17 to .25 for *interpersonal skills*, was .19 for *decision-making* skills, and was .21 for an overall computer score.

We conducted hierarchical regression analyses to estimate the incremental validity of the computer scored simulation above and beyond the existing selection assessments using the test and validation data sets. Results show the computer scored simulation explained and additional 2.8% of variance (R^2 from .024 to .052, $p = .05$, $N = 135$[25]). While the absolute effect size is small, the relative gain from the baseline suggests the practical value of including the simulation.

5.1.6 | Subgroup differences in human and computer scores

Table 10 shows our subgroup analysis of human and computer scores. As seen, there are small to medium size differences between race/ethnic groups and small differences for males-females. Specifically, Black-White results showed no substantive differences for either human or computer scores (Cohen's d = .09 and .04, respectively), Hispanic-White results showed small human and computer score differences (Cohen's d = .24 and .16, respectively, in favor of Hispanic), and the differences were at parity for Two or More Races-White for both human and computer scores (Cohen's d = -.01 and .00, respectively). We did find medium Asian-White differences for both human and computer

TABLE 10 Subgroup differences in human and computer scores (full data set, including train, test, and validation).

		N	Human Score Mean	Computer Score Mean	Human Score Cohen's d	Computer Score Cohen's d
Race	White (Reference)	356	3.40	3.42	–	–
	Asian	1047	3.20	3.24	–.43	–.50
	Black/African American	29	3.45	3.44	.09	.04
	Hispanic/Latino	52	3.52	3.48	.24	.16
	Two or More Races	43	3.40	3.42	-.02	-.00
Gender	Male (Reference)	1138	3.23	3.26	–	–
	Female	401	3.40	3.40	.37	.36

scores (Cohen's d = -.43 and -.50, respectively). Given 61% of the Asian sample reported spending most of their life living outside English speaking countries (vs. 2% for White), we tested and found that the Asian-White differences can be partially explained by language and/or cultural differences: Asians from non-English speaking countries had lower scores than Whites (Cohen's d = -.56 and -.68 for human and computer scores, respectively), whereas Asians from English speaking countries showed small differences from Whites (Cohen's d = -.20 and -.20 for human and computer scores, respectively). Females had higher scores than males for both human and computer scores (.37 and .36, respectively). Given the observed subgroup differences, we tested whether the computer was "biased" via moderated regressions (Cleary, 1968). Results showed no significant intercept or slope differences between majority and minority subgroups, suggesting no predictive bias[26].

5.2 | Discussion

We assessed the effectiveness of ML and NLP for automatically transcribing and scoring audio-based constructed responses to a simulation. We found that computer scores trained to replicate human scoring on KSAOs predicted job performance, provided incremental predictive validity, showed similar subgroup (i.e., gender and race/ethnicity) differences to human scores and no predictive bias. We highlight practical implications and directions for future research.

First, we showed that ASR achieves viable accuracy levels across demographic groups (race, gender, and country of origin). This suggests that ASR can be used to scale assessments based on audio responses (e.g., one-way behavioral interviews, ACs) in high-volume personnel selection contexts. That said, 1.5%–3% of the responses in our sample suffered from severe audio issues (e.g., high background noise, poor audio quality, and low volume), posing challenges in both human scoring and ASR. We were able to discover low quality audio using confidence scores produced by the ASR system; such metrics could be helpful in production to monitor audio quality in real time and introduce interventions that would allow candidates to take corrective actions while taking the assessment.

Second, our uncorrected criterion validity was lower than desired. Our estimates after range restriction and criterion unreliability corrections should more accurately reflect the relationships between the simulation and job performance. However, another potential explanation is that we focused on fairly narrow and specific KSAOs (i.e., interpersonal and decision-making skills) in this simulation, and that broadening the KSAO coverage of the simulation may increase observed validity with job performance. Finally, while the absolute effect size is small, the relative gain in prediction suggests this simulation may still offer practical value in large-scale applications.

We recognize that there are several practical challenges, such as the initial investment in human scoring and the expansion of NLP model development across different stimuli, that might hinder an organization's ability to scale such assessment solutions. Future research should focus on improving human scoring quality while controlling cost (e.g., reduce the number of raters), as well as exploring stimuli-agnostic NLP models (i.e., models that can automatically score KSAOs based on varying simulation content), to help scale automated scoring of audio constructed response assessments (see study 1).

6 | STUDY 5: PRACTICAL ML ALGORITHMS FOR SELECTION ASSESSMENT SCORING: A USE CASE REPORT ON MULTI-OUTCOME PREDICTION[27]

ML techniques have drawn substantial attention from both organizational researchers and practitioners given their great promise in different aspects (e.g., increasing prediction). The approaches organizational practitioners take are slightly different from researchers. One difference is that researchers often invest a large amount of time and effort to maximize the predictability of ML models in predicting one outcome variable (e.g., Sajjadiani et al., 2019; Speer, 2018), whereas practitioners may prefer one model that achieves acceptable performance in predicting multiple outcomes and requires a reasonable level of computation power and development time. In this study, we provide an example of an organization's application of ML techniques to achieve multiple goals. These three goals are listed in the order of importance: (1) to develop a ML-based scoring algorithm that leverages pre-employment assessment data to enhance the predictability of turnover of the job candidates and their in-role performance (i.e., work productivity and quality) after/if they are hired; (2) to improve the generalizability of the assessment scoring so that it can predict multiple business outcomes (e.g., worker turnover, productivity, quality) simultaneously and across diverse candidate samples (in terms of countries of origin); and (3) to achieve better assessment efficiency by selecting items that contribute better to the prediction, thereby shortening the assessment length for future candidates. We also conducted an adverse impact analysis on the ML scoring approach and compare the results to those from the previous non-ML method.

6.1 | Methods

6.1.1 | Data overview

The organization is an international e-commerce company headquartered in China. The selection procedure concerns warehouse job placements for order fulfillment workers globally (with a projected impact of over 50,000 individuals yearly). The job candidates and incumbents for the position are from both Eastern and Western countries (e.g., China, Indonesia, the United States, Australia). The scoring algorithm developed was to be implemented on the refined assessment system to yield a predicted overall performance score for each candidate, and candidates scoring higher than the 50th percentile (i.e., a selection ratio typical of warehouse positions alike, according to the organization) were to be selected into the jobs.

The current application was built based on assessment data from 86,253 warehouse workers globally (63% male; around 60% from mainland China, around another 30% from Southeast Asia, and the rest from North America, Oceania, and the European Union) who had taken the assessment *as applicants* in 2018 or 2019 (with the system operating in their residing country's official or common language). Following a prospective validation design, these workers' job performance for *three months after* starting the job was also acquired (including *objective* work productivity and quality, as well as the turnover record at the end of their *first month*[28]).

The pre-employment assessment included two major components: a high-fidelity task simulation and a series of work history items. For the simulation, each candidate was assigned and evaluated by three separate modules (order fulfillment, cargo loading, and inventory stocking). Within each module, there were separate tasks of various complexity levels. All simulations were conducted in an augmented reality space where candidates perform the tasks mimicking the actual work. All captured information from the simulation was recorded and digitally transcribed as simulation performance markers. These performance markers (see more detailed descriptions below) were then used as predictors in the scoring model development.

For the order fulfillment module, candidates went through a few tasks of fulfilling orders (each involving multiple products) by following the order details provided (which can be repeatedly reviewed; an example detail given is the number of units for each product). Counts for each product filled, the number of times order details were reviewed, time taken from checking order details to action (averaged across checks), and total time spent on each task were recorded.

For the cargo loading module, another set of tasks was involved. In each task, candidates were presented with a handful of packaged objects (only dimensional information was available) in disarray and limited cargo space to move the packages around to fit into the cargo space. The counts of moves made were recorded for each task (the difficulty of each task was determined by the optimal number of moves).

For the inventory stocking module, several tasks were involved (each with multiple products). For each task, candidates were provided with various products where product descriptions, weights, and categories are known, as well as a shelf space with some known products in place and various spots open. Instructional rules such as how much weight each shelf level can hold, what specific categories of products can or cannot be placed together, and general rules of what categories should only be placed on a certain shelf level were given and available for repeated reviewing. For each task, the final location of each product, total moves taken to and between shelf spots, total time taken to complete the task, and the total number of times rules were reviewed were recorded.

The second component of the pre-employment assessment was the work history items. These were biodata-like Likert-type self-report questions about the candidates' past work experiences, covering perspectives such as comfort level with technology, past job natures, past turnovers, past manual labor experiences, and past performance ratings. Responses to these questions were also used as predictors in model development.

Key job outcomes were also included in the data: turnover (objective binary responses for turnover within one month of hire), overall productivity (a continuous variable based on units processed per unit time), and overall quality (a continuous variable based on defects per unit opportunity). Note that productivity and quality outcomes were calculated as scores adjusted for the specific expectations of the person (e.g., a discount for workers that are less tenured[29]).

Due to data management difficulties in organizing and integrating records across multiple regions and databases, a high degree of missingness in outcome variables was present: roughly 50%–70% on turnover, productivity, or quality. Such a level of missingness, though shocking to most researchers, is representative of the messiness of worker population data management within the organization. It was therefore decided for the current project that models were built separately for each outcome prediction first and then stacked together using subjective weights (i.e., determined by the organization) to yield an overall predictive score. This way, the model development process was based on the maximized sample for each outcome, using all available data containing values for the target outcome variables.

6.1.2 | Development

All analyses were conducted using R version 3.6.3 (R Core Team, 2020) within the *caret* package environment (version 6.0-86; *Kuhn, 2008*).

6.1.3 | Assessment shortening

Both content reviewing and predictor selection by Lasso regression were used for assessment shortening. For content reviewing, redundancies among assessment items were examined (for future item deletion to aid the goal of test shortening); several work history items (rated on Likert scales) were removed due to a high level of content similarity (e.g., different wordings of intention to leave). Additionally, Lasso regression would automatically select the best-performing predictors during model building.

6.1.4 | Data preprocessing

For each outcome-focused data subset, a 9:1 split was applied to create a training set and a testing set (i.e., 90% of the data was used to train and fine-tune the models, and 10% of the data was treated as a hold-out sample to examine model true performance). For the turnover-focused data, the split was set to be stratified (i.e., the training and testing sets have practically equal class distributions). All preprocessing was evaluated based on the training set and then applied to the testing set (N_turnover_train = 16,646, N_turnover_test = 2,073; N_productivity_train = 32,258, N_productivity_test = 3,584; N_quality_train = 28,578, N_quality_test = 3,173). Descriptive statistics along with data visualizations informed a sequence of preprocessing decisions (as suggested by Kumar, 2018): (1) all missing data on the predictors were imputed using the median (as opposed to the mean, for its robustness against outlier influences)—for the company's scalability considerations, we did not adopt more complex imputation methods[30]; (2) all categorical predictors were dummy-coded; (3) all predictors with zero or near-zero variances (i.e., when more than 95% of the values in a given variable are identical) were removed, as they would not provide enough information in differentiating outcomes—about 20% of predictors (after dummy coding) were removed, most of which were the location makers of positions practically no one had placed products on; (4) in each pair of predictors correlating above .95 with each other, the one with the largest average correlation with all other predictors was removed, as they would not contribute to incremental gains in predicting outcomes—about a dozen of predictors removed, most of which were markers that can be approximated by some other markers; (5) for all skewed numeric variables, we applied Yeo-Johnson transformation (i.e., a power transform that is natural log-based but accommodates zero and negative values; Yeo & Johnson, 2000); (6) to deal with multicollinearity (i.e., linearly dependent variables), we removed all linear combinations among the predictors (i.e., remove the predictor if it is equal to the sum of scalar multiples of other predictors), as they would contribute to spurious results in linear models—a handful of predictors were removed; and (7) due to the fact that turnover had a class imbalance issue (i.e., where turnover happened in less than 25% of cases, which may result in suboptimal class prediction performance), a mixed over- and under-sampling technique (Chawla et al., 2002) was applied to the data to yield a balanced training set. Note that all predictors were also standardized before being entered into the models. The sequence of preprocessing steps resulted in a training set of 130 predictors for further model building (including the number of simulation actions, simulation reaction time, simulation task time, scaled simulation scores, dummy-coded simulation task object locations, and scaled work history ratings).

6.1.5 | Outcome examination

We examined whether the three outcome variables (i.e., turnover, productivity, and quality) can be more effectively reflected through linkages or combinations (i.e., linear combinations, item correlations, composites, etc.). Results indicated that turnover is negatively correlated with productivity or quality (r = -.14) and that productivity and quality are not significantly correlated. Creating combinations of outcomes was, therefore, not justifiable and would not likely yield sufficient modeling advantages, further supporting our decision to build models separately to maximize prediction for each outcome.

6.1.6 | Classification algorithms

We tested the following classification models to predict whether individual warehouse worker turnover would happen (i.e., binary outcome): (1) logistic regression classifier (i.e., the "glm" method in *caret*) and its regularized version (Lasso; i.e., "glmnet"), which predicts the probability of turnover occurrence by fitting data to a logit function; (2) support vector machine (SVM) linear classifier (i.e., "svmLinear"), which finds the line that splits data between the two differently classified groups such that the distances from the closest point in each of the two groups will be farthest away;

(3) Random Forest classifier (i.e., "rf"), which ensembles decision trees and chooses the classification having the most "votes" from individual decision trees; and (4) boosted classifier, which ensembles learning algorithms that combine the prediction of several base estimators in order to improve robustness over a single estimator—specifically, we examined eXtreme Gradient Boosting (XGBoost) algorithms (i.e., "xgbTree") as they have been consistently shown to be the most robust boosting method (Chen & Guestrin, 2016; Gómez-Ríos et al., 2017).

6.1.7 | Regression algorithms

We tested the following regression models to predict productivity and quality (as they are continuous): (1) linear regression and its regularized version (Lasso), which expresses a linear relationship between the predictors and the outcome; (2) SVM linear regressor, which is an extension from SVM linear classification with analogous interpretations; (3) Random Forest regressor, which is an extension from Random Forest classification with analogous interpretations; and (4) boosted regressor, which is analogous to boosted classification but for continuous outcomes—we focused on the XGBoost algorithms.

6.1.8 | Algorithm stacking

For simultaneous prediction of both turnover and performance outcomes, we examined various manual weighting combinations for the three outcomes and stacked together the separate classification and regression prediction models to create overall assessment scores[31].

Classification model evaluations. In the context of binary turnover classifications, we used the Receiver Operating Curve (ROC) metric for model evaluation. ROC shows the tradeoff between sensitivity (i.e., the proportion of correctly identified positive cases; "true positives") and one-minus-specificity (i.e., the proportion of incorrectly identified negative cases; "false positives"). Given that it is a more comprehensive metric that takes into consideration the balance between different components contributing to accuracy, we used ROC (and specifically the area under the curve, AUC) as our main metric when selecting the final deployment model.

6.1.9 | Regression model evaluations

In the context of regression models (i.e., for productivity and quality predictions), we tracked the following two metrics to evaluate the model(s): (1) coefficient of determination (R^2), which provides a measure of how well the observed outcomes are replicated by the model—a bigger R^2 reflects more variance explained in the outcomes and indicates better model performance; and (2) Root Mean Squared Error (RMSE), which displays the plausible magnitude of the error term—a smaller RMSE indicates better model performance.

6.1.10 | Generalization

As ML models get more complex, they tend to overfit the data by capturing noise and capitalizing on chance. To ensure the model indeed captures the relationship(s) that are representative and generalizable, we used the following techniques to prevent overfitting and ensure model generalizability: (1) regularization, which aims to avoid model-data overfitting through coefficient shrinkage (i.e., the model is only as complex as it needs to be)—specifically, we adopted the Lasso regularization method as it could simultaneously facilitate our purpose of test shortening through variable selection; and (2) 10-fold cross-validation, which is a method that iteratively splits the training data into ten representative folds to train the models on nine folds and validate the model on the other fold—this prevents reliance on a single training set and helps produce more unbiased results.

6.2 | Results

6.2.1 | Classification model results for turnover prediction

In Table 11(a), we report the test set AUCs and what model hyperparameters were involved. The area under the ROC curve indicated that all models yielded similar chance-level predictive validity (with Lasso and XGBoost models only slightly outperforming the others). Determined by the comparable predictive performance, and considering the assessment simplification capabilities and the complexity of potential future modifications of the algorithm in terms of model runtime and tuning difficulties (i.e., the Lasso model could facilitate variable selection and would require less time to re-run and less effort to re-tune when the algorithm needs updating upon receiving new data), we deemed Lasso regularized logistic regression to be the best prediction model for the turnover outcome. Note that the AUCs for all classification models were barely exceeding .50, indicating that the sensitivity and specificity of the turnover prediction were both substandard (i.e., not strongly differentiating between true and false positives).

6.2.2 | Regression model results for productivity prediction

We present in Table 11(b) the following test set information—the R^2, RMSEs, what model hyperparameters were involved, and the correlation between the predicted productivity and the true productivity. Linear regression, linear regression with Lasso regularization, and linear SVM all yielded improved predictive results (compared to the existing baseline effects, i.e., non-ML rational scoring). The R-squared and RMSE indicated that the linear methods yielded the highest predictive performance, and the correlations between predicted and observed values are equally the highest. Considering the tradeoffs between recovered correlation, model parsimony, and next-step ensemble complexity (i.e., no difference in R-squared, RMSEs, and correlations, linear regression models being more explainable and defensible than SVM models, and the final models for the other two outcomes being Lasso regression models), we deemed Lasso regularized linear regression to be the most appropriate prediction model for the productivity outcome.

6.2.3 | Regression model results for quality prediction

We present in Table 11(c) the test set R-squared, RMSEs, what model tuning was performed, and the correlation between the predicted quality and the true quality. The R-squared indicated that Lasso predicted comparatively to the OLS regression, Random Forest, and linear SVM methods, and RMSEs indicated that Lasso yielded the smallest error and highest predictive correlation. Determined by the highest correlation between predicted quality and true quality, and considering the potential difficulty (e.g., model runtime, parameter tuning) in future model refinement, we deemed Lasso regularized linear regression to be the best prediction model for the quality outcome.

6.2.4 | Final model considerations

A final set of models was ensembled based on the chosen algorithms above. We adopted the Lasso regularized logistic/linear regression as the prediction model for each outcome separately, and assigned different sets of weights for combining the outcomes to yield a final assessment score for each individual under each weighting option[32]. The model results associated with different weighting options are shown in Table 12.

TABLE 11 Results for predicting turnover, productivity, and quality based on testing sets.

Model	ROC	Hyperparameter
Logistic regression	.50	none
Lasso regularized logistic regression	**.51**	**alpha = 1, lambda = .1**
XGBoost	.51	nrounds = 100, max_depth = 6, eta = .3, gamma = 0, colsample_bytree = 1, min_child_weight = 1, subsample = 1
Random Forest	.50	mtry = 2, ntree = 500, min.node.size = 5
SVM (linear)	-C = 1 [no convergence[a]]	

(a) Turnover model results (*N* = 2073).

Model	R-squared	RMSE	Hyperparameter	Correlation between predicted and observed values
Linear regression	.03	32.03	none	.17
Linear regression with Lasso regularization	**.03**	**32.03**	**alpha = 1, lambda = .1**	**.17**
XGBoost	.02	32.78	nrounds = 100, max_depth = 6, eta = .3, gamma = 0, colsample_bytree = 1, min_child_weight = 1, subsample = 1	.13
Random Forest	.02	32.22	mtry = 2, ntree = 500, min.node.size = 5	.14
SVM (linear)	.03	32.14	C = 1	.17

(b) Productivity model results (*N* = 3584).

Model	R-squared	RMSE	Hyperparameter	Correlation between predicted and observed values
Linear regression	.01	68.96	none	.10
Linear regression with Lasso regularization	**.01**	**68.89**	**alpha = 1, lambda = .1**	**.11**
XGBoost	.00	71.22	nrounds = 100, max_depth = 6, eta = .3, gamma = 0, colsample_bytree = 1, min_child_weight = 1, subsample = 1	.05
Random Forest	.01	68.99	mtry = 2, ntree = 500, min.node.size = 5	.10
SVM (linear)	.01	71.57	C = 1	.10

(c) Quality model results (*N* = 3,173).

Note. In regularized regression/classification, "alpha" is a mixing percentage for combining regularization penalty methods, 1 means the regularization is full Lasso, and "lambda" is the regularization parameter. In XGBoost, "nrounds" is boosting iterations, "max_depth" is maximum tree depth, "eta" is shrinkage, "gamma" is minimum loss reduction, "colsample_bytree" is subsample ratio of columns, "min_child_weight" is the minimum sum of instance weight, and "subsample" is subsample percentage. In Random Forest, "mtry" is randomly selected predictors, "ntree" is the number of trees to build, and "min.node.size" is the minimal number of samples within the terminal nodes. In SVM, "C" is cost. For more details on the algorithms, please refer to the Supplementary Materials.

[a] We recognize that our approach here is suboptimal because we did not delve into potential solutions (e.g., increase the iteration numbers, and add penalizations).

TABLE 12 Correlations between predicted outcome scores and actual outcomes.

Turnover-Productivity-Quality weighting	r with Turnover	r with Productivity	r with Quality
Non-ML rational scoring as the baseline	*–.02*	*.08*	*.03*
1, 0, 0	–.09	–.03	–.01
0, 1, 0	–.03	.16	.10
0, 0, 1	–.01	.08	.16
9, .5, .5	–.10	.00	.01
8, 1, 1	–.09	.03	.04
7, 2, 1	**–.08**	**.09**	**.06**
6, 3, 1	–.06	.12	.07
5, 3, 2	–.06	.13	.08
4, 3, 3	–.05	.14	.10
1/3, 1/3, 1/3	–.04	.15	.10

Note. N = 86,253. The "weighting" method is manual. The "baseline" refers to the originally implemented (non-ML) model, and the baseline correlations are between the respective rational criterion scores and the actual criteria. Rows below the baseline are correlations between the respective weighted blends of ML-predicted outcome scores and the actual criteria. Baseline scoring is based on composite scoring (i.e., rational scoring as determined by the organization) following traditional non-ML approaches.

As the organization's priority outcome to predict in its warehouse worker population is turnover, the final model should improve turnover prediction while maintaining or increasing prediction performance for productivity and quality. The weighting option of seven-part predicted turnover with two-part predicted production and one-part predicted quality was, therefore, considered for overall scoring. This model enhanced and maximized overall model prediction compared to the rational scoring baseline (i.e., performance from the previously in-use scoring, as shown in the first row of Table 12). The decision of choosing this weighting option over others depended on the organization's decision to greatly emphasize turnover prediction.

6.2.5 | Comparison with the baseline in validity

Compared to the previously in-use prediction baseline (i.e., the correlations between assessment score and job outcomes), our ensembled ML predictive algorithm shows significant improvements in both outcome prediction and test efficiency (i.e., achieving incremental correlation gain, using fewer items in the assessment). Compared to the baseline approach, our model yielded validity coefficients (measured in Pearson's *r*, with the point-biserial form for turnover between actual turnover and the predicted scores) that improved on turnover (-.08 for our model vs. -.02 for the baseline), largely maintained on productivity (.09 for our model vs. .08 for the baseline), and increased on quality (.06 for our model vs. .03 for the baseline).

6.2.6 | Comparison with the baseline in utility

The estimated size of the order fulfillment personnel globally for the organization is around .2 million. Our model results regarding turnover imply that implementing this new ML algorithm may potentially translate into a net cost saving of over two million U.S. dollars annually (assuming an onboarding and training cost of $4,129 per employee [SHRM Research, 2016])[33]. In addition, the improved predictions for productivity and quality could translate into

TABLE 13 Adverse impact analysis results.

Turnover-Productivity-Quality weighting	SR_female/SR_male	Cohen's *d* based on predicted scores
Non-ML baseline	.99	.01
7, 2, 1	.98	.24

Note. N = 86,253. SR = selection ratio.

roughly a total of .1 million unit productivity increase and close to one million unit defect decrease. This magnitude of projected utility is considered worthy of the resources put into the model development and the foreseen cost of implementing the new scoring algorithms.

6.2.7 | Adverse impact examination

We conducted an adverse impact analysis to examine whether the ML-based selection system has any effects on adverse impact as compared to the previous non-ML approach. We examined the four-fifth rule and calculated Cohen's *d* (see Table 13 for results and comparisons with the baseline method, i.e., non-ML rational scoring). The four-fifths rule indicates evidence of adverse impact if the selection ratio of the minority group is less than four-fifths of the selection ratio of the dominant or majority group. Cohen's *d* quantifies the effect size of the difference between predicted scores of the dominant or majority group and those of the minority group. Our adverse impact analysis results suggested that the four-fifth rule was not violated, and Cohen's *d* effect sizes showed that the gender differences in predicted scores were small[34]. Therefore, we conclude that adopting an ML approach has no effect on adverse impact as compared to the previous non-ML approach.

6.2.8 | Model maintenance considerations

In addition, because the foundation of our prediction models is Lasso regressions, which are essentially variations of logistic or linear regressions and relatively simple-in-nature, intuitive, and easy-to-explain, the model has high explainability and can facilitate identifying key variables driving the predictions. With Lasso regression, further variable examinations can be easily done to provide information on both the direction and magnitude of the relationship between each predictor and the outcome. Continuous model improvement requires re-training based on new data and examining predictor weights yielded by the model, which is easier to implement with Lasso regression models compared to other ML methods.

6.3 | Discussion

Through this organizational application of developing ML scoring algorithms for the selection assessment, we hope to inform future ML assessment scoring. Firstly, we believe placing all the emphasis on the validity coefficients of ML algorithms during their development and implementation is not sufficient. It is equally important, if not more important, to be able to provide justifications and explanations for the results of the model, rather than just improving its performance metrics. Secondly, it is important to take into account multiple business outcomes when developing predictive models, rather than concentrating on just one. By maximizing and leveraging various criteria, it is possible to achieve higher overall model efficiency and broader business impact. And thirdly, when evaluating the performance of a model, it is crucial to compare it with existing methods in terms of the utility of implementing the new model, as well as its adverse impact ratios and effect size benchmarks.

We also particularly would like to note that because the Lasso solution was deemed satisfactory in the early stage of model development for this project, we did not explore fine-tuning for other algorithms further and went largely with software default options. Proper hyperparameter tuning may result in other algorithms emerging as the best model. For demonstration and to warn readers that models should ideally be tested with hyperparameter tuning, we have performed supplementary hyperparameter tuning on the models. For instance, after fine-tuning, the most significant improvement in performance was observed in the XGBoost and Random Forest models. For productivity prediction, the correlation coefficient (*r*) increased from .13 to .18 for the XGBoost model and from .14 to .16 for the Random Forest model. For quality prediction, the *r* improved from .05 to .10 for the XGBoost model[35].

Additionally, and perhaps more importantly, our results showed that ML can increase assessment predictability, but the improvement might not be very large in the magnitude changes of the validity coefficients—the reason could be that ML might only contribute marginally when the data is highly structured and following the traditional survey or assessment format (i.e., where ML's advantage of discovering underlying patterns may not be capitalized, as the data already has rather clear patterns). Furthermore, to begin with, the validity coefficients from the baseline method were not of large magnitudes, which may be reflecting several deeper issues: (1) that the selection assessment itself is of little validity, or (2) that the measures for both the predictors and outcomes are of low quality (e.g., biodata items may be fakable and biased, and the simulations may only provide a partial—and potentially not so differentiating—coverage of job-relevant behaviors or characteristics), and (3) that more work may need to be done on improving the assessment itself, all of which may be the reason for the observed low predictive validity. Nevertheless, caution should be placed on concluding whether a small effect is meaningful to organizations, as a small effect may translate into a much bigger impact and utility when the application is scaled up.

7 | STUDY 6: NATURALISTIC EXTRACTION OF KNOWLEDGE, SKILLS, ABILITIES AND OTHER CHARACTERISTICS USING NLP WITH HUMAN-LEVEL PROFICIENCY[36]

A vital part of the hiring process is the identification of KSAOs required to complete a job. It is completed prior to the development of recruiting and personnel selection systems (Sanchez & Levine, 2010). KSAOs are also useful for outlining evaluation criteria during employee evaluations and for defining objectives during training and development (Gael, 1988). As such, determining high-quality KSAOs that cover the responsibilities of the job is critical.

However, identifying KSAOs requires a great deal of manual labor and expertise. A job analysis is generally performed by a trained industrial psychologist or a supervisor for the job. It can require creation of task lists, interviews/questionnaires, or examinations of existing job descriptions in order to understand the duties of the job (Drauden, 1988; McCormick et al., 1972). Once the duties are defined, it remains challenging to determine valuable KSAOs for the job, as they are not always straightforward to infer and can require creative thinking to discover what skills would be helpful for completing those duties (Goffin & Woycheshin, 2006). As companies grow, the process becomes more demanding, requiring KSAO ratings to be produced and maintained for sometimes thousands of jobs.

In this work, we present an approach for automatically identifying KSAOs from existing job descriptions, with the goal of alleviating the time spent by professionals on performing the task. Often these job descriptions are easier to collect because they define concrete duties performed, while the KSAOs that would be helpful in completing those duties are more difficult to obtain (Goffin & Woycheshin, 2006). Using state-of-the-art NLP techniques, a model can be trained to take an input job description and predict the KSAOs that are most relevant to the job. We collect a dataset of position descriptions and have annotators choose appropriate KSAOs for each position. We then make use of BERT (Devlin et al., 2019), a state-of-the-art NLP model, to understand the text in the job description and automatically predict KSAOs. Results show that our model achieves a prediction accuracy similar to that of humans. We discuss our findings and the challenges faced in this work.

Identifying KSAOs automatically with NLP, as presented in this work, can have a substantial positive impact on the job analysis process. We envision an AI system that can take in any document pertaining to a job detailing tasks performed in that job – a job description, performance evaluations, or resumes from existing employees – and give suggestions of KSAOs that are relevant to that job. Further, the system can make it easier to maintain KSAO lists, as the KSAOs can be automatically refreshed whenever a job description is modified (Putka et al., 2023). Though it will not replace traditional job analyses, this can greatly reduce the burden on professionals when creating KSAO ratings.

7.1 | Theory

Hickman et al., 2020 explore methodological concerns surrounding corpus construction and vocabulary inclusion-exclusion criteria within the pre-processing phase of NLP. Specific areas of concern are considered regarding how to address measurement validity when mapping research hypotheses onto applicable methods within NLP.

Recent work has also focused on the application of NLP techniques within personnel selection, recruiting, and job analysis problem domains. Work by M. C. Campion et al., 2016 achieved construct validity within grading applicant accomplishment essays based on the emulation of existing human rating data. Their work also considers the ramifications for AI in impacting workplace diversity.

Other work explores the applicability of ML as a secondary technique for uncovering trends within large or naturalistically curated data sets. Banks et al., 2019 consider the application of generative statistical models towards identifying thematic clusters within Fortune 1000 firms' job listings and other applicant-facing language content. A general observation (see Liem et al., 2018, for example) is that commercial recruiting forms and applicant tracking systems have widely been using ML methods for a number of years. However, organizational researchers have been hesitant to apply these methods (J König et al., 2020) due to a divergence in how the respective fields focus on reliability and validity of model data. Previous studies have even shown that human raters often have high variance when rating KSAOs of positions (Van Iddekinge et al., 2005).

Putka et al. (2023), concurrently with this work, applied NLP techniques to the KSAO rating task. They collected a corpus of 963 O*NET jobs to train a sparse partial least squares (SPLS) regression model, and they demonstrated the validity of their machine-generated predictions by showing strong correlation and agreement with SME ratings. Their work used more classical methods (SPLS with lemmatization and stopwords/rare words removed), while this work uses more recent NLP methods (BERT) on Army civilian job descriptions.

Our work follows a recent focus on work analysis field methods that use ML methods to extract KSAO labels from task lists (Goffin & Woycheshin, 2006). We borrow the relaxed focus on construct validity featured in M. C. Campion et al. (2016), where the focus is on reconstructing human-produced data from performing a similar task that the AI algorithm is asked to perform. Rather than adopting the typical hypothesis-testing paradigm, we instead perform an exploratory study into the ability of NLP models to identify KSAOs from text descriptions.

7.2 | Method

Collecting labeled data is a vital step in training AI models. First, we describe how we collected job descriptions and annotated them with their corresponding KSAOs. We then describe our method for training and evaluating a model for predicting KSAOs from job descriptions.

7.2.1 | Data and sample

Job descriptions were obtained by scraping the U.S. Army FASCLASS website. The website contains over 200k job listings for Army civilian positions across the U.S. The listings are externally-facing position descriptions detailing the major duties, responsibilities, and supervisory relationships of a position[37]. We filter this dataset to include only whitecollar jobs based on the classification provided by the U.S. Office of Personnel Management.[38] Specifically, we include only jobs with an occupational series number between 0000 and 2299. The types of jobs are highly diverse, including finance, human resources, nursing, engineering, physics, and more. An average of five job descriptions from each of the 23 occupational groups were selected to be annotated, resulting in a corpus of 124 job descriptions. Figure F1 in the Online Supplement presents the number of annotated jobs in each occupational group. FASCLASS data adheres to the United States Office of Personnel Management's diversity and inclusion policies which ensure job listing data, to the largest extent possible, work towards diverse and inclusive workforce selection outcomes. Only the "Position Duties" section of each Army job description was extracted. The FASCLASS website states "Generally, major duties are those that occupy a significant portion of the employee's time. They should be only those duties currently assigned, observable, identified with the position purpose and organization, and expected to continue or recur on a regular basis." This corresponds to the "Duties and Tasks" section found in most job descriptions (Morgeson et al., 2019).

A KSAO ontology was selected from prior work by the US Army Talent Management group (KSBs; knowledge, skills, and behaviors—see Saling & Do, 2020) that reflected the requirements of Army officer positions. This ontology, described next, is extremely useful for outlining the degree of similarity between Army officer jobs and those involving leadership, management, technical, and administrative responsibilities (Borman, 1987; Dexter, 2020). Use of a common framework to describe Army jobs and civilian positions can serve as the basis for establishing the generalizability of findings generated using an Army ontology to a set of non-military job descriptions. It can highlight areas of overlap such as managerial responsibilities and job content (Dexter, 2020), though the evidence of generalizability will still be limited.

KSBs are defined hierarchically into domains, talents, and measurable KSBs. At the top level are seven domains – Cognitive, Communication, Disposition, Interpersonal, Leadership & Management, Expertise & Personal Competence, and Physical. Each domain contains several talents, and each talent contains several KSBs. Defining KSAOs hierarchically simplifies the process for annotators by allowing them to quickly find the KSAOs they want by filtering by domain and talent. The domains, talents, and KSAOs are shown in Tables F2 - F8 of the Online Supplement.

Rather than having an annotator choose several KSAOs for an entire job description, we instead split up a job description into sentences and have annotators identify KSAOs for each sentence. This makes the task more tractable for annotators by breaking up a long job description into smaller pieces. It also has the advantage of obtaining more examples with which to train the model. See the online supplement for more details.

We create a web interface for annotators to choose KSAOs that apply to a job description (see Figure F4 in the Online Supplement). The annotator is shown a job description split up by sentences. All sentences for the job description are shown together on the same web page, to ensure the annotators have sufficient context to understand each sentence. For each sentence, the annotator must (1) choose whether that sentence is a valid sentence containing KSAOs and (2) choose any number of domains, talents, and KSAOs that are present in the sentence.

A total of fifteen annotators contributed to the dataset, including 10 experts (individuals with prior job analysis experience) and five laypeople (individuals with no prior experience conducting job analyses). We make use of lay ratings because experts can be hard to find, and their time is more expensive. Laypeople, on the other hand, are cheaper and more accessible, and they can still provide valuable training data for AI models. Thus, we use both lay and expert ratings for training, but only use expert ratings for evaluation and reliability calculation.

A total of 5815 sentences were annotated taken from 124 jobs, including six jobs that were shared among all annotators and used a basis for measuring inter-rater reliability and model accuracy. Figure F1 in the Online Supplement shows the frequency of each job type. A job has on average 47 sentences, with sentences having an average of 24 words. Table F1 in the Online Supplement shows how often each KSAO was chosen by annotators. KSAOs related to expertise, program management, and communication were most prevalent in the data. Research data are not shared.

7.2.2 | Measures

Inter-rater reliability. Six jobs were selected to be shared between all annotators in order to measure inter-rater reliability. We used an Intraclass Correlation Coefficient (ICC) (Shrout & Fleiss, 1979) to compare the selected KSAOs of each job from each annotator. The final ICC score is the mean over the 10 expert annotators.

Evaluation of NLP model. We use several measures to evaluate how well our NLP model extracts KSAOs from job descriptions. We measure the correctness of machine predictions and human predictions on the six shared jobs. At a high level, a machine prediction is considered correct if it matches the human-annotated predictions. A human prediction is considered correct if it matches the predictions made by the other humans.

The measures for evaluating the effectiveness of the NLP model are as follows. Accuracy is defined as the percentage of machine predictions that matched the human predictions. In this work, we use a modified version of top-k accuracy, which makes use of the fact that our NLP model ranks the most likely KSAOs in order. Top-k accuracy is a common metric for classification, but it does not handle a sentence having multiple classes (i.e., KSAOs). Our modified top-k accuracy is defined as the percentage of sentences where at least one of the top k KSAOs chosen by the NLP model are found in the annotator-selected KSAOs. Using top-k accuracy gives the model credit even if its first choice was incorrect as long as its second or third choice was correct. We use top-1, top-3, and top-5 accuracy. Because evaluation is performed at the sentence-level, each human rater only chose at most five KSAOs for each sentence, even though they were allowed to choose any number of KSAOs. Therefore, using top-5 accuracy should be a sufficient metric for evaluating sentences from job descriptions.

Label ranking average precision (LRAP) measures what percent of the higher ranking KSAOs chosen by the model are correct. Finally, F1 score is a metric measuring the balance between precision and recall, where precision is the ratio of true positives to model-predicted positives, and recall is the ratio of true negatives to model-predicted negatives.

The same measures are computed for humans. Each annotator's predictions are compared to all of the other annotators' predictions to obtain top-k accuracies, LRAP, and F1 scores for each annotator. Then the scores are averaged over the annotators to give a single human score for each measure. A machine score that is equal to or greater than the human score shows the NLP model can predict KSAOs as well or better than humans.

7.2.3 | Training BERT to predict KSAOs

In this section, we describe our model for identifying skills from a job description. We frame this problem as a multi-label classification problem, in which an input text can be assigned multiple correct labels. Consider the following sentence, "Serves as an expert consultant to provide advice and guidance to officials, managers and other scientists and engineers within and outside the division covering a broad range of scientific or engineering activities." This sentence may have multiple valid labels: "Working With The Public," "Verbal Communication," etc.

We utilize the BERT architecture (Devlin et al., 2019) to perform the multi-label classification. BERT is a deep learning model using the Transformer architecture (Vaswani et al., 2017) for understanding natural language. It uses a mechanism called "attention" to create semantic representations of words in a piece of text by having each word "attend to" or look at each other word in the sentence when creating its own representation. This allows the model to better understand the context in which each word is used.

The BERT model has been pre-trained on large amounts of text from Wikipedia and books to learn a general understanding of English text. This was done by training it to do a simple "fill-in-the-blank" task. One or more words in each sentence are masked out, and the model must learn to predict the masked word. This simple training task, when trained on billions of words of text, results in a model that has a good understanding of general language. While this base model on its own is not very useful, it can be fine-tuned to do more specialized NLP tasks, such as detecting offensive language or predicting the emotion in a news article.

In this work, we follow previous work by fine-tuning BERT to perform classification on our job description dataset. More specifically, we add a neural network layer of randomly-initialized weights to the end of the network containing 199 outputs corresponding to the KSAOs. A sigmoid activation function is applied to each of the outputs which gives a probability for each output, representing how likely the input sentence contains each KSAO. The final prediction is obtained by selecting the KSAOs with the highest probability scores.

One problem experienced in earlier versions of the model was that it would often predict only the most common KSAOs in the training data. To solve the problem we employ two strategies. (1) We used a data augmentation technique similar to MLSMOTE (Charte et al., 2015) to generate more sentences containing the low-frequency KSAOs during training. (2) We add a module that increases the scores given to low-frequency KSAOs during evaluation using the KSAO's IRPL score. See the online supplement for details.

In addition, we compare the BERT-based models with more classical NLP approaches to determine whether more recent AI techniques can learn to better identify KSAOs than older techniques. Naive-Bayes (Hand & Yu, 2001) is a probabilistic learning method that uses the presence or absence of words to predict the class of an input text. Logistic regression using bag-of-words representations (BoW-LR) treats each input text as an unordered collection of words and then trains a logistic regression classifier on those representations. We also test logistic regression on term frequency-inverse document frequency representations (TFIDF-LR), which takes into account the relative importance of each word in a document. Finally, we test Latent Dirichlet Allocation (LDA), which can be found in the online supplement.

To evaluate the models, we perform leave-one-out cross-validation. One of six job descriptions used for evaluation is left out of the training data, while a NLP model is trained on the remaining data. Then the model is evaluated on the one job that was left out. This is repeated for all six job descriptions. The evaluation scores over the six runs are then averaged to get the final score. We experiment with the original BERT architecture and DistilBERT (Sanh et al., 2019) – a smaller and faster version of BERT that retains most of the language understanding capabilities of the original. For both models, we train both cased (i.e. takes text capitalization into account) and uncased (i.e. does not take capitalization into account) models to determine the effect of capitalization on the ability of NLP models to extract KSAOs. Implementation details are described in the Online Supplement.

7.3 | Results

7.3.1 | Inter-rater reliability for collected data

Results show that the inter-rater reliability scores for choosing KSBs (ICC = .822) was good, while ICC for rater selection of talents was slightly lower, but still moderate (ICC = .616), and ICC for selection of domain by raters was also good (ICC = .878). These rater agreement levels, while relatively high, were undoubtedly affected by several factors. There were a large number of KSBs to choose from, which makes it more likely at baseline for annotators to pick differently. Furthermore, because many of the KSAOs are extremely similar (e.g. "Even-Tempered" vs. "Emotional Control", and "Attentiveness" vs. "Focus", etc.), annotators are likely to have the same broad notion about a sentence but still choose different KSAOs. Such issues (large number of labels, label ambiguity, and annotator quality) are common in ML research and applications, and often lead to noisy datasets used for training and evaluation of models. Importantly, though, ML models can be robust to noisy data (Rolnick et al., 2017), and can often still extract meaningful patterns from the data and achieve reasonable performance.

7.3.2 | Evaluating the effectiveness of our proposed model

Table 14 shows the classification scores for each NLP model, along with annotator scores. All BERT-based models achieve top-1 accuracies above 48%, meaning the top KSAO choice for the model is a correct one according to the

TABLE 14 Classification scores for classical NLP models, BERT based NLP models, and human raters on shared jobs.

Model	Top-1 Acc	Top-3 Acc	Top-5 Acc	LRAP	F-1	Prec	Rec
Naïve-Bayes	55.01	78.46	88.06	.4991	6.12	8.02	4.95
BoW-LR	35.82	68.23	80.60	.4260	22.78	31.20	17.94
TFIDF-LR	40.30	72.07	83.58	.4402	24.38	31.08	20.06
BERT-uncased	48.94	73.83	82.77	.4314	24.34	27.32	21.94
BERT-cased	51.06	77.66	87.87	.4550	26.74	27.64	25.90
DistilBERT-uncased	48.94	73.62	82.55	.4445	24.89	25.89	23.97
DistilBERT-cased	50.43	75.96	84.26	.4463	25.56	25.65	25.47
Human	48.74	69.36	71.49	.3309	16.02	45.34	9.87

human-selected KSAOs. The best model – BERT-cased – obtained 51% top-1 accuracy. This is higher than the average human top-1 accuracy of 49%. The top-3 accuracy, top-5 accuracy, LRAP, and F1 scores for the models are all higher than the corresponding human scores. Thus, the NLP models can outperform humans for the task of extracting KSAOs from job descriptions.

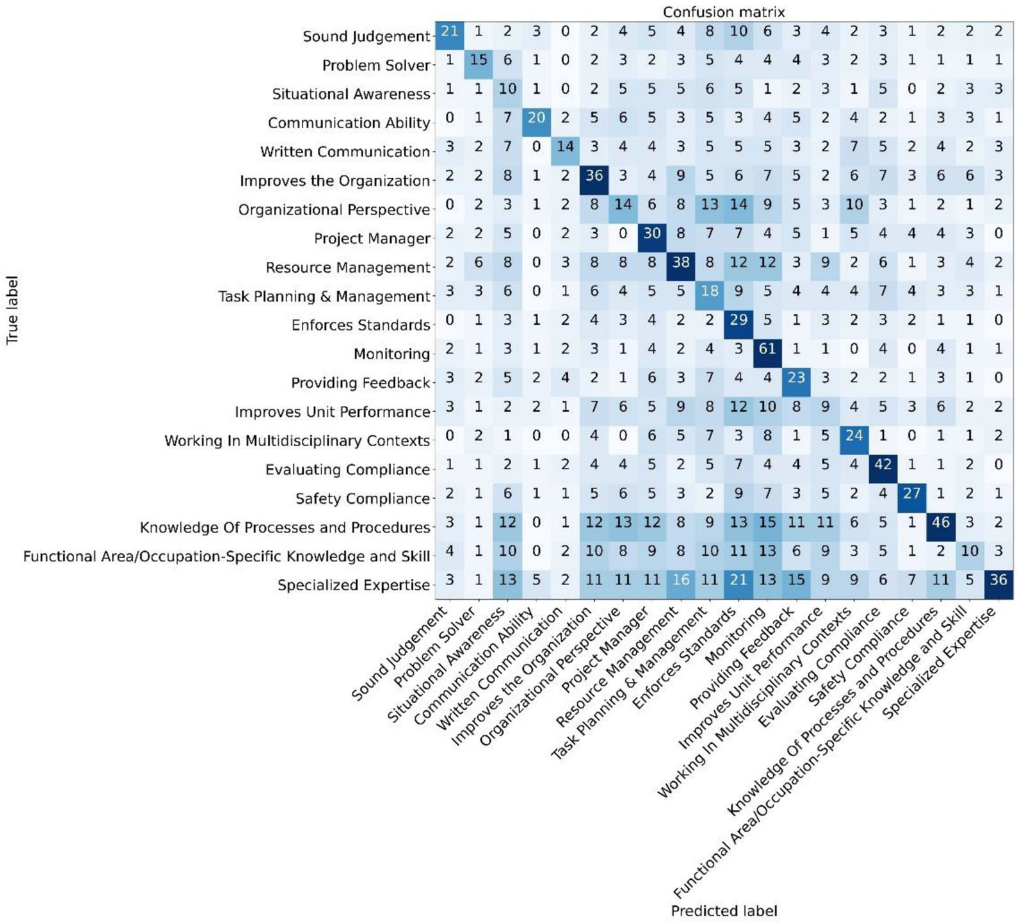
The results present a curious effect – the model can achieve higher scores than the data it was trained on. This is possible because the model was trained to recognize patterns within the overall group of raters. It can read a piece of text and predict which KSAOs would likely be chosen by most of the raters. In essence, the model knows how to best predict the majority opinion. Each individual rater, however, is not as adept at predicting the majority opinion, as evidenced by the low inter-rater agreement. This does not mean that the model only sees the ratings with optimal agreement – in fact, the model does get trained on all of the ratings. Rather, it likely learns that it can be correct more often when it emulates the human ratings with high agreement, so in a sense, it learns to "ignore" outlier ratings.

However, it should be known that the quality of the ratings is still very important and is often summarized in AI literature as "garbage in, garbage out." An AI model produces outputs that are only as good as the data that is used to train it. In our case, the AI is getting accuracy scores that are higher than each individual rater, but it's really producing outputs that are only as good as the majority opinion between all the raters.

The precision, recall, and F1 scores for the BERT models are between 21% and 28%, which may seem low. However, the scores are reasonable given the task being performed. As stated previously in the inter-rater reliability results section, there are a large number of KSAOs to choose from (199), which makes it very difficult for an AI (or human rater) to choose a KSAO that matches what the human raters chose. Further, because some of the KSAOs are similar in meaning to each other, the AI often chooses an KSAO that was incorrect, but still likely a reasonable choice qualitatively. See Table F10 of the Online Supplement for examples.

The uncased models receive significantly lower scores than the cased models. This shows that taking capitalization into account can be helpful for this task. We hypothesize that cased models are able to detect organization names (e.g. USAAC), and this can better inform the models that certain KSAOs are more likely (e.g. Improves the Organization). Of the BERT-based models, BERT-cased achieves the highest scores.

We believe the models can detect patterns in the creative thinking done in humans by recognizing that certain keywords or phrases in the job description often lead to certain KSAOs. For example: for the following sentence "Ensures coordination with USAASC G8 for funding," raters had chosen (among others) the KSAO "Interpersonal Tact." It takes creativity and higher-level thinking for a human to understand that coordination with another organization (USAASC G8) over time will necessitate having good interpersonal relationship skills to reliably obtain funding. BERT-cased predicted (among others) the KSAO "Interpersonal Relationship Building" for the sentence, which is very similar to the one chosen by raters. The model successfully detected the keyphrase pattern "coordination with [organization] for [goods]" to require interpersonal relationship KSAOs, and thus correctly predicted the KSAO.



	Sound Judgement	Problem Solver	Situational Awareness	Communication Ability	Written Communication	Improves the Organization	Organizational Perspective	Project Manager	Resource Management	Task Planning & Management	Enforces Standards	Monitoring	Providing Feedback	Improves Unit Performance	Working In Multidisciplinary Contexts	Evaluating Compliance	Safety Compliance	Knowledge Of Processes and Procedures	Functional Area/Occupation-Specific Knowledge and Skill	Specialized Expertise
Sound Judgement	21	1	2	3	0	2	4	5	4	8	10	6	3	4	2	3	1	2	2	2
Problem Solver	1	15	6	1	0	2	3	2	3	5	4	4	4	3	2	3	1	1	1	1
Situational Awareness	1	1	10	1	0	2	5	5	5	6	5	1	2	3	1	5	0	2	3	3
Communication Ability	0	1	7	20	2	5	6	5	3	5	3	4	5	2	4	2	1	3	3	1
Written Communication	3	2	7	0	14	3	4	4	3	5	5	5	3	2	7	5	2	4	2	3
Improves the Organization	2	2	8	1	2	36	3	4	9	5	6	7	5	2	6	7	3	6	6	3
Organizational Perspective	0	2	3	1	2	8	14	6	8	13	14	9	5	3	10	3	1	2	1	2
Project Manager	2	2	5	0	2	3	0	30	8	7	7	4	5	1	5	4	4	4	3	0
Resource Management	2	6	8	0	3	8	8	8	38	8	12	12	3	9	2	6	1	3	4	2
Task Planning & Management	3	3	6	0	1	6	4	5	5	18	9	5	4	4	4	7	4	3	3	1
Enforces Standards	0	1	3	1	2	4	3	4	2	2	29	5	1	3	2	3	2	1	1	0
Monitoring	2	1	3	1	2	3	1	4	2	4	3	61	1	1	0	4	0	4	1	1
Providing Feedback	3	2	5	2	4	2	1	6	3	7	4	4	23	3	2	2	1	3	1	0
Improves Unit Performance	3	1	2	2	1	7	6	5	9	8	12	10	8	9	4	5	3	6	2	2
Working In Multidisciplinary Contexts	0	2	1	0	0	4	0	6	5	7	3	8	1	5	24	1	0	1	1	2
Evaluating Compliance	1	1	2	1	2	4	4	5	2	5	7	4	4	5	4	42	1	1	2	0
Safety Compliance	2	1	6	1	1	5	6	5	3	2	9	7	3	5	2	4	27	1	2	1
Knowledge Of Processes and Procedures	3	1	12	0	1	12	13	12	8	9	13	15	11	11	6	5	1	46	3	2
Functional Area/Occupation-Specific Knowledge and Skill	4	1	10	0	2	10	8	9	8	10	11	13	6	9	3	5	1	2	10	3
Specialized Expertise	3	1	13	5	2	11	11	11	16	11	21	13	15	9	9	6	7	11	5	36

FIGURE 1 Confusion matrix for the BERT-cased model with MLSMOTE.

Note. It shows the most common errors associated with the model. The rows in the matrix represent the true number of sentences belonging to each KSAO, while columns represent the number of sentences predicted by the model for each KSAO.

Naive Bayes has higher accuracy and LRAP than BERT-based models but lower F-1 score. We explain the cause in the Error Analysis below. BoW-LR receives scores that are lower than the BERT-based models. TFIDF-LR achieves scores similar to some of the BERT-based models, but they are lower than the best model BERT-cased. Thus, classical methods based on TFIDF can reasonably be used for identifying KSAOs. These methods train faster and are simpler to implement. To get the best accuracy, however, BERT-cased should be used. For results on all 124 jobs, see Table F9 in the Online Supplement.

7.3.3 | Error analysis

Figure 1 shows the most common errors associated with BERT-cased as a confusion matrix. The rows in the matrix represent the true number of sentences belonging to each KSAO, while columns represent the number of sentences predicted by the model for each KSAO. The values that lie on the diagonal represent correct predictions (e.g., 21 of the sentences that had the true label of "Sound Judgement" were correctly predicted by the model as "Sound Judgement").

Values not on the diagonal represent errors (e.g., 1 of the sentences that had the true label of “Sound Judgement” was incorrectly predicted by the model as “Problem Solver”).

We introduced strategies including MLSMOTE to prevent the model from choosing the most common KSAOs too frequently. Qualitatively, these changes caused the model to give more varied answers that still would be reasonable. For example: for the sentence mentioned above “Ensures coordination with USAASC G8 for funding,” the model previously predicted only “Resource Management,” “Coordination,” and “Working In Multidisciplinary Contexts” – three very common KSAOs. After the changes, the model predicted “Interpersonal Relationship Building”, “Coordination,” “Motivating Others”, “Project Manager,” and “Resource Management,” which contains less common KSAOs that are still relevant.

For further validation that the model predicts KSAOs with greater variety, compare the confusion matrix to that in Figure F2 of the Online Supplement, showing the errors when not using MLSMOTE. It is clear that without MLSMOTE, the model was over-predicting certain common KSAOs such as “Monitoring,” “Specialized Expertise,” and “Resource Management.” Rather, the opposite effect seems to be apparent in the model with MLSMOTE. The common KSAOs are often incorrectly predicted as a different KSAO. We would argue that this is a more desirable consequence than over-predicting the most common KSAOs. An AI system that only suggests the same few KSAOs is not as useful as one that suggests more varied, but still relevant KSAOs.

Naive Bayes has higher accuracy and LRAP than BERT-based models but lower F-1 score. We believe Naive Bayes is better at ranking KSAOs by simply choosing the most common KSAOs in the training set. While we alleviated the problem using techniques such as MLSMOTE, Naive Bayes was not as affected by the techniques as the BERT-based models. We can see from the Naive Bayes’ confusion matrix in the Figure F3 of the Online Supplement that the problem remains. Most of the errors are concentrated in the common KSAO columns like “Monitoring.” We believe this is because Naive Bayes cannot generalize as well as BERT to words and phrases it has not seen before in training. This causes it to resort to picking the most likely KSAOs based on the class priors. This behavior is not desirable in an AI recommendation system since a professional will likely already know if the most common KSAOs are relevant or not.

7.4 | Discussion

7.4.1 | Practical implications

The purpose of this study is to investigate the possibility of using NLP models to extract KSAOs from text descriptions and whether models could be as proficient as humans at the task. Our results appear to show that NLP models can indeed perform the task as well as humans. This is a surprising finding considering the dataset’s noisy labels and its relatively small size. Deep learning models such as BERT perform best with thousands to millions of training examples, and our models were trained on the lower end of that spectrum (5815 examples). This shows that practitioners may be able to train their own models to accurately extract KSAOs with relatively little manual effort to collect training data.

These models are not limited to analyzing job descriptions. They may also be used to extract KSAOs from other forms of unstructured text or spoken language. For example, interviews of current employees or supervisors about job duties could be transcribed into text using advances in speech recognition. The text could then be analyzed using NLP models to extract KSAOs. Similarly, KSAOs can be identified from open-response questions in surveys. This would be valuable for human resources officers in performing job analyses with less effort.

Our approach can be used as part of a selection process as well. Applicant resumes could be screened for KSAOs and a score could be calculated based on how well the KSAOs in the resume match the KSAOs found in the desired job description. Sajjadiani et al., 2019 use a similar approach; however, only the job titles were used to determine KSAOs, while the benefit of our approach is the ability to glean the employee competencies from expository text. A possible criticism of our method is that KSAOs needed for a job may be more easily obtained by drawing directly from a database such as O*NET. While this is true for job descriptions, the KSAOs present in an applicant’s resume are not as easy to collect, instead requiring manual effort to analyze each resume. An NLP model, on the other hand, can be used to automate this process.

There is a well-demonstrated relationship between KSAOs and job performance (Ployhart & Bliese, 2006) However, this relationship is significantly mediated by individual factors such as adaptability (Tucker et al., 2009). Extracting KSAOs from naturalistic sources of occupational text (e.g., performance evaluations, self-reported incumbent descriptions) may allow organizations to help explain the variance in job performance across employees and therefore better understand their antecedents of success.

7.4.2 | Limitations

A limitation of our approach is that we identify KSAOs on a sentence-level basis, while many professionals in real-world organizations identify KSAOs holistically at the job level. While our method of breaking up job descriptions into sentences may not completely match the process in organizations, we still believe our method results in KSAOs that will be relevant to the job. As an example, OPM's job analysis process is to choose competencies based on the job tasks, then to link those competencies to specific tasks to verify that there is a clear relationship between the tasks that are part of the job and the competencies required to complete those tasks (Office of Personnel Management, 2003). While not exactly the same as our process, it validates our method of splitting up the job. Each KSAO should be associated with a specific task, similar to how raters labelled the data in our work. If KSAOs are needed for the job as a whole, the KSAOs that were associated with the most sentences can be chosen. A limitation of this process is that it may miss high-level KSAOs associated with a job that a human would have chosen. For this reason, we believe an AI system based on our work should be used to recommend KSAOs to humans, rather than the AI system completing the task alone.

ORCID

Georgi Yankov https://orcid.org/0000-0003-2942-6446
Jay H. Hardy III https://orcid.org/0000-0002-1064-2985
Carter Gibson https://orcid.org/0000-0003-4353-6350
Mengqiao Liu https://orcid.org/0000-0001-8426-3091
John Capman https://orcid.org/0000-0002-7704-4656
Tianjun Sun https://orcid.org/0000-0002-3655-0042
Feng Guo https://orcid.org/0000-0002-5054-1839
Bo Zhang https://orcid.org/0000-0002-6730-7336
Logan Lebanoff https://orcid.org/0000-0001-7079-0210

ENDNOTES

[1] Study 1 authored by Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., & Koohifar, F.

[2] Study 2 authored by Yankov, G., & Speer, A.

[3] The limited range of the assessor ratings might also be yielding limited variance and attenuating our final results although the exercise scores which we modelled were sums of the assessor ratings.

[4] We explored other ML models beyond XGBoost. However, the best accuracy was achieved with XGBoost, and this seems to be common with most tabular ML tasks (Nielsen, 2016).

[5] We selected five parameters critical for a simple, yet robust implementation of the XGBoost algorithm for the needs and shape of our text data: learning rate (four rates linearly spaced between .1 and .6), maximum depth (a random number between 1 and 10), minimal child weight (a random number between 1 and 10), number of estimators (300, 350, 400, and 450), and regularization alpha (0, 0.1, 0.5, 0.75). We ran the RandomizedSearchCV method from the sklearn Python library to find the optimal hyper-parameters values.

[6] Training was performed on a 6-core Linux UBUNTU virtual machine with 56 GB memory and 1 GPU. Following Devlin et al.'s (2019) guidelines, we experimented (Online Table 5) tuning the following hyperparameters: batch size (8, 16, and 24), epoch (3 vs. 4), and learning rate (3e-5, 5e-5, and 1e-4).

[7] Note these were observed correlations and therefore attenuated as a function of range restriction and criterion unreliability. We did not have adequate data to correct for range restriction. If we corrected for criterion-unreliability using the Viswesvaran et al. (1996) .52 estimate, the average AC dimension validity coefficient for assessors was .24, and it was .34 for NLP scores.

[8] Study 3 authored by Hardy, J., Gibson, C., Koenig, N., & Frost, C.

[9] The job analysis began with a review of existing job descriptions followed by in-depth interviews with subject matter experts (SMEs), visionary interviews, focus groups, and a job analysis questionnaire. Information uncovered during this process was foundational in the design of all selection assessments, including the open-ended prompts underlying the deep learning integration.

[10] The RoBERTa architecture max token length of 512. During implementation, only 4-5 cases (0.04%) exceeded this limit. As such, we did not feel it was necessary to explore methods for increasing the max token length in this case.

[11] Inter-rater reliabilities at the competency level are provided in Table 5.

[12] The use of the transformer architecture (described below) meant that beyond this pre-scrubbing of the data, no additional data cleaning (e.g., stemming, n-grams, removal of stop-words, etc.) was required.

[13] Support for the viability of this modeling strategy can be found in a recent paper by de la Vega de Leon et al. (2018), which shows that deep neural networks trained with moderate sparsity can maintain a reasonable performance compared to dense labels with less than 20% performance degradation with losses in model accuracy that are on par with other methods of multi-tasking prediction.

[14] A predefined data partition was obtained using a stratified sampling approach. In other words, within the k-folds and our final model training data, applicant responses were always represented in one set or the other, creating an equal distribution of responses in each subset. This approach to structuring the data helps minimize contaminating covariation associated with similarities in writing styles and halo effects contained within each response set.

[15] Learning Rates Tested: 5e-06, 1e-05,2e-06,1e-06; Dropout Rates Tested: 0.05, 0.08, 0.10, 0.15.

[16] We acknowledge that the decision to include only two competencies raises construct deficiency concerns. However, the purpose of the deep learning assessment scores was only to supplement information within an existing validated assessment, not to replace it. As such, the partner organization determined that maximizing the added predictive value of assessment scores was more important than maximizing coverage of the construct because many of these constructs were already well represented within the broader assessment. Nevertheless, we recognize that readers would want a more complete understanding of the model's performance independent of this decision. As a result, we provide results for the live scoring composite along with each of the five individual competency ratings to enable a more thorough evaluation of the model's potential performance.

[17] Study 4 authored by Liu, M., McNeney, D., Capman, J. F., Lowery, S. B., Kitching, M., Nimbkar, A., & Boyce, A. S.

[18] Based on a prior MTurk pilot study we conducted, audio constructed format resulted in a 13% time reduction compared to written constructed format, while controlling for the assessment content.

[19] We offered three recording attempts to promote a positive candidate experience and to buffer potential technical difficulties getting in the way of recording the audio responses.

[20] To develop benchmark scores, we trained three internal I-O psychologists with more than five years of experience working in the assessment domain. The I-O psychologists independently provided ratings on the audio responses and met as a group to derive consensus on the ratings.

[21] We evaluated reliability based on different numbers of raters ($k = 3, 4, 5$, or 6) and found that ICC(2, k) dropped by .083 or more when using fewer than five raters.

[22] WER estimates the difference between machine and human (true) transcription; a high WER indicates a larger number of word substitutions, deletions, or insertions in the automatic transcript as compared to human-transcribed text for the same audio clip (WER = (Substitutions + Insertions + Deletions) / Number of Words Spoken).

[23] Open source code for BERT can be found on: https://huggingface.co/docs/transformers/model_doc/bert

[24] Unless otherwise specified, all correlation coefficients are uncorrected.

[25] The sample size of 135 for the incremental validity analysis is smaller than the full test and validation data set size because the study participants took different assessments for different roles in the organization, so we selected the job with the largest N in our sample and used that subsample for this particular analysis. Results are uncorrected for range restriction and criterion unreliability.

[26] Note that the moderated regression analyses were conducted on the test and validation datasets, with small sample sizes and statistical power for detecting a small effect (Asian: $N = 238$, power = .41; Black: $N = 70$, power = .14; Hispanic: $N = 76$, power = .15; Two or More Races: $N = 69$, power = .14).

[27] Study 5 authored by Sun, T., Guo, F., Min, H., & Zhang, B.

[28] The organization, unfortunately, does not have information on the turnover types. We hereby recognize this as a limitation in this study and application, and advocate organizations to collect and examine turnover details, as they may provide valuable information for predicting turnover behaviors.

[29] Per suggestion by the reviewing team, we also explored more of the relationship between productivity and quality, and we found that there seems to be a curvilinear relationship between them: there is an overall small positive association between productivity and quality, but people with extremely high productivity would show lower quality.

[30] We recognize that this is a limited way of imputing for missingness. Our adoption of this substandard pre-processing approach was motivated by the organization's concern about computation time.

[31] We also provided more algorithm details in Online Supplement E.

[32] An alternative (and potentially more ideal) weighting approach may be to adopt empirical weighting to combine algorithm scores using generalized linear models done on an independent sample separated from the training sample(s).

[33] Though the number here is not specific to the organization and may not apply to the Asia-focused context, the SHRM statistics were believed to be applicable to all HR contexts around the world, as the methodology for the SHRM study stated that the data were collected from its members all over the world (i.e., more than 275,000 members in over 160 countries) though with a response rate of only 6%.

[34] This is a larger effect size observed than the traditional non-ML approach. We think it might be possible that this is due to the new inclusion of the work history items (and not the ML method), as biodata measures have been criticized due to their potential to elicit adverse impacts. We have examined the gender differences on biodata items and reported in the Supplementary Materials the predicted score gender differences in terms of the item-level Cohen's *d*'s. No items stood out as potentially problematic (i.e., beyond a "small" degree) as judged by the benchmarks provided by Cohen (1988), while some displayed similar effect sizes as the overall ML-based gender differences.

[35] Updated model performance with hyperparameter tuning details can be found in Online Supplement Table E2.

[36] Study 6 authored by Lebanoff, L., Phillips, H., & Newton, C.

[37] https://acpol2.army.mil/ako/fasclass/search_fs/search_fasclass.asp

[38] https://www.opm.gov/policy-data-oversight/classification-qualifications/classifying-general-schedule-positions/occupationalhandbook.pdf

REFERENCES

Agarwal, A., Dudík, M., & Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. *arXiv preprint arXiv*.

Albritton, B. H., & Tonidandel, S. (2020). How can big data science transform the psychological sciences? *The Spanish Journal of Psychology*, *23*, e44. https://doi.org/10.1017/SJP.2020.45

Arthur Jr, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, *56*(1), 125–153. https://doi.org/10.1111/j.1744-6570.2003.tb00146.x

Arthur Jr, W., Doverspike, D., Muñoz, G. J., Taylor, J. E., & Carr, A. E. (2014). The use of mobile devices in high-stakes remotely delivered assessments and testing. *International Journal of Selection and Assessment*, *22*(2), 113–123. https://doi.org/10.1111/ijsa.12062

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*(4), 985–1008.

Arthur Jr, W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*(2), 435–442. https://doi.org/10.1037/0021-9010.93.2.435

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

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*, 478–494. https://doi.org/10.1037/0021-9010.74.3.478

Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an army officer sample. *Organizational Behavior and Human Decision Processes*, *40*(3), 307–322. https://doi.org/10.1016/0749-5978(87)90018-5

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://arxiv.org/pdf/2005.14165.pdf

Buyukyilmaz, M., & Cibikdiken, A. O. (2016). Voice gender recognition using deep learning. *2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA 2016)*, *58*, 409–411.

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*, 285–325.

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. http://dx.doi.org/10.1037/apl0000108

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*. https://doi.org/10.1145/1015330.1015432

Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, *89*, 385–397. https://doi.org/10.1016/j.knosys.2015.07.019

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785–794). ACM, 10, 2939672.2939785.

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

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology*, *86*(3), 410–417. https://doi.org/10.1037/0021-9010.86.3.410

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*: Routledge Academic. https://doi.org/10.4324/9780203771587

Connelly, B. S., Ones, D. S., Ramesh, A., & Goff, M. (2008). A pragmatic view of assessment center exercises and dimensions. *Industrial and Organizational Psychology*, *1*(1), 121–124. https://doi.org/10.1111/j.1754-9434.2007.00022.x

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*: Cengage Learning.

de la Vega de León, A., Chen, B., & Gillet, V. J. (2018). Effect of missing data on multi-task prediction methods. *Journal of Cheminformatics*, *10*(1), 1–12.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 4171–4186). https://doi.org/10.48550/arXiv.1810.04805

Dexter, J. C. (2020). Human resources challenges of military to civilian employment transitions. *Career Development International*, *25*(5), 481–500. https://doi.org/10.1108/CDI-02-2019-0032

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

Drauden, G. M. (1988). Task inventory analysis in industry and the public sector. *The job analysis handbook for business, industry, and government*, *2*, 1051–1071.

Edwards, B. D., & Arthur Jr, W. (2007). An examination of factors contributing to a reduction in subgroup differences on a constructed-response paper-and-pencil test of scholastic achievement. *Journal of Applied Psychology*, *92*(3), 794–801.

Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment*, *6*(2), 115–123. https://doi.org/10.1111/1468-2389.00080

Gael, S. (1988). *The job analysis handbook for business, industry, and govt*. John Wiley & Sons.

Goffin, R. D., & Woycheshin, D. E. (2006). An empirical method of determining employee competencies/ksaos from task-based job analysis. *Military Psychology*, *18*(2), 121–130. https://doi.org/10.1207/s15327876mp1802_2

Gokaslan, A., & Cohen, V. (2019). OPENWEBTEXT corpus. URl: https://skylion007.github.io/OpenWebTextCorpus

Gómez-Ríos, A., Luengo, J., & Herrera, F. (2017). A study on the noise label influence in boosting algorithms: AdaBoost, GBM and XGBoost. *In International Conference on Hybrid Artificial Intelligence Systems* (pp. 268–280). https://doi.org/10.1007/978-3-319-59650-1_23

Hand, D. J., & Yu, K. (2001). Idiot's bayes—not so stupid after all? *International Statistical Review*, *69*(3), 385–398.

Hartwell, C., Liff, J., Gardner, C., & Mondragon, N. (2022, April 28–30). Development and validation of asynchronous competency-based structured interview scoring algorithms. In J. Lavashina & S. Baumgartner (Chairs), New developments in structured interviews: From AI to technical interviews [Symposium]. SIOP 2022 Convention.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction, second edition*: Springer.

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*, Advance online publication. https://doi.org/10.1037/apl0000695

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2020). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, *25*(1), 114–146. https://doi.org/10.1177/1094428120971683

Hoffman, C. J., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011). Exercises and dimensions are the currency of assessment centers. *Personnel Psychology*, *64*(2), 351–395. https://doi.org/10.1111/j.1744-6570.2011.01213.x

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL https://doi.org/10.5281/zenodo, 1212303(6)

International Task Force on Assessment Center Guidelines. (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment*, *17*(3), 243–253. https://doi.org/10.1111/ijsa.2009.17.issue-310.1111/j.1468-2389.2009.00467.x

Jansen, P. C. W. (2012). How to apply a dimension-based assessment center. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 121–140): Routledge/Taylor & Francis Group.

König, C. J., Demetriou, A. M., Glock, P., Hiemstra, A. M. F., Iliescu, D., Ionescu, C., Langer, M., Liem, C. C. S., Linnenbürger, A., Siegel, R., & Vartholomaios, I. (2020). Some advice for psychologists who want to work with computer scientists on big data. *Personnel Assessment and Decisions*, *6*(1), 17–23. https://doi.org/10.25035/pad.2020.01.002

Kobrin, J. L., Deng, H., & Shaw, E. J. (2011). The association between SAT prompt characteristics, response features, and essay scores. *Assessing Writing*, *16*(3), 154–169. https://doi.org/10.1016/j.asw.2011.01.001

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kumar, D. (2018, December 25). Introduction to Data Preprocessing in Machine Learning. Medium. Retrieved May 2nd, 2021 from https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d

Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, *34*(7), 3265–3315. https://doi.org/10.1093/rfs/hhaa079

Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Wicaksana, A. S., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. *Explainable and interpretable models in computer vision and machine learning* (pp. 197–253). Springer. https://doi.org/10.1007/978-3-319-98131-4_9

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology*, *18*(1), 102–121. https://doi.org/10.1080/13594320802058997

Lievens, F., Sackett, P. R., Dahlke, J. A., Oostrom, J. K., & De Soete, B. (2019). Constructed response formats and their effects on minority–majority differences and validity. *Journal of Applied Psychology*, *104*(5), 715–726. https://doi.org/10.1037/apl0000367

Liu, M. (2019, April). Machine Learning for I-O: Techniques and Real-World Applications. Symposium presented for the 34th Annual Meeting of the Society for Industrial and Organizational Psychology (SIOP): National Harbor, MD.

Liu, P., Qiu, X., Chen, X., Wu, S., & Huang, X. J. (2015). Multi-timescale long short-term memory neural network for modelling sentences and documents. *Proceedings of the EMNLP Conference* (pp. 2326–2335). https://doi.org/10.18653/v1/D15-1280

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. ArXiv:1907.11692 [Cs]. http://arxiv.org/abs/1907.11692

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics and job dimensions as based on the position analysis questionnaire (paq). *Journal of Applied Psychology*, *56*(4), 347. https://doi.org/10.1037/h0033099

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & GRUBB III, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, *60*(1), 63–91. https://doi.org/10.1111/j.1744-6570.2007.00065.x

Min, H., Peng, Y., Shoss, M., & Yang, B. (2021). Using machine learning to investigate the public's emotional responses to work from home during the COVID-19 pandemic. *Journal of Applied Psychology*, *106*(2), 214–229. https://doi.org/10.1037/apl0000886

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Morgeson, F. P., Brannick, M. T., & Levine, E. L. (2019). *Job and work analysis: Methods, research, and applications for human resource management*. Sage Publications.

Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE transactions on multimedia*, *16*(4), 1018–1031.

Nielsen, D. (2016). Tree Boosting with XGBoost Why Does XGBoost Win "Every" Machine Learning Competition? NTNU Tech Report. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2433761

Office of Personnel Management. (2003). Office of personnel management government handbook. https://www.opm.gov/policy-data-oversight/hiring-information/competitivehiring/deo_handbook.pdf#page=230

Peng, Y., Zhang, J., Zhang, H., Xu, H., Huang, H., & Siong Chng, E. (2020). A multilingual approach to joint speech and accent recognition with DNN-HMM framework. arXiv preprint arXiv: 2010.11483.

Ployhart, R. E., & Bliese, P. D. (2006). *Individual adaptability (i-adapt) theory: Conceptualizing the antecedents, consequences, and measurement of individual differences in adaptability*. Understanding adaptability: A prerequisite for effective performance within complex environments. Emerald Group Publishing Limited. https://doi.org/10.1016/S1479-3601(05)06001-7

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

Putka, D. J., Oswald, F. L., Landers, R. N., Beatty, A. S., McCloy, R. A., & Yu, M. C. (2023). Evaluating a natural language processing approach to estimating ksa and interest job analysis ratings. *Journal of Business and Psychology*, *38*, 385–410. https://doi.org/10.1007/s10869-022-09824-0

R Core Team. (2020). R: A language and environment for statistical computing, Vienna, Austria. http://www.R-project.org/

Rivlin, O. (2019). Generalizability in reinforcement learning. https://towardsdatascience.com/generalization-in-deep-reinforcement-learning-a14a240b155b

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rolnick, D., Veit, A., Belongie, S., & Shavit, N. (2017). Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694. https://doi.org/10.48550/arXiv.1705.10694

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.

Ryan, A. M., Sacco, J. M., McFarland, L. A., & Kriska, S. D. (2000). Applicant self-selection: Correlates of withdrawal from a multiple hurdle process. *Journal of Applied Psychology*, *85*(2), 163–189.

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

Saling, K. C., & Do, M. D. (2020). Leveraging people analytics for an adaptive complex talent management system. *Procedia Computer Science*, *168*, 105–111. https://doi.org/10.1016/j.procs.2020.02.269

Sanchez, J., & Levine, E. (2010). The rise and fall of job analysis and the future of work analysis. *Annual review of psychology*, *63*, 397–425.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. https://doi.org/10.48550/arXiv.1910.01108

Shen, W., Cucina, J. M., Walmsley, P. T., & Seltzer, B. K. (2014). When correcting for unreliability of job performance ratings, the best estimate is still. 52. *Industrial and Organizational Psychology*, *7*(4), 519–524.

SHRM Research. (2016). Human Capital Benchmarking Report. Society for Human Resource Management. https://www.shrm.org/about-shrm/press-room/press-releases/pages/human-capital-benchmarking-report.aspx

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420. https://doi.org/10.1037//0033-2909.86.2.420

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

Speer, A. B. (2020). Scoring dimension-level job performance from narrative comments: Validity and generalizability when using natural language processing. *Organizational Research Methods*, *24*(3), 572–594. https://doi.org/10.1177/1094428120930815

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2021). A conceptual framework for investigating and mitigating machine learning measurement bias (MLMB) in psychological assessment. PsyArXiv. https://doi.org/10.31234/osf.io/mjph3

Thompson, I., Koenig, N., Mracek, D., & Tonidandel, S. (2023). Deep learning in employee selection: Evaluation of algorithms to automate the scoring of open-ended assessments. *Journal of Business and Psychology*, *38*, 509–527. https://doi.org/10.1007/s10869-023-09874-y

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. Wiley.

Tracy, S. J. (2013). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact*: Wiley-Blackwell.

Tucker, J. S., Pleban, R. J., & Gunther, K. M. (2009). The mediating effects of adaptive skill on values-performance relationships. *Human Performance*, *23*(1), 81–99. https://doi.org/10.1080/08959280903400275

Van Iddekinge, C. H., Putka, D. J., Raymark, P. H., & Eidson Jr, C. E. (2005). Modeling error variance in job specification ratings: The influence of rater, job, and organization-level factors. *Journal of Applied Psychology*, *90*(2), 323. https://doi.org/10.1037/0021-9010.90.2.323

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574. https://doi.org/10.1037/0021-9010.81.5.557

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2019). HuggingFace's transformers: State-of-the-art natural language processing. http://arxiv.org/abs/1910.03771

Xu, J., Chen, D., Qiu, X., & Huang, X. (2016). Cached long short-term memory neural networks for document-level sentiment classification. arXiv. doi.org/10.48550/arXiv.1610.04989

Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). *Big Bird: Transformers for Longer Sequences*. NeurIPS Proceedings.

Zedeck, S. (1986). A process analysis of the assessment center method. *Research in Organizational Behavior*, *8*, 259–296.

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

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 19–27).

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: Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., Speer, A., Hardy, J. H., Gibson, C., Frost, C., Liu, M., McNeney, D., Capman, J., Lowery, S., Kitching, M., Nimbkar, A., Boyce, A., Sun, T., Guo, F., … Newton, C. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. *Personnel Psychology*, 1–63. Advance online publication. https://doi.org/10.1111/peps.12608