1-2023

Automatic scoring of speeded interpersonal assessment center exercises via machine learning: Initial psychometric evidence and practical guidelines

Louis HICKMAN

Christoph N. HERDE

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg

Louis TAY

Automatic Scoring of Speeded Interpersonal Assessment Center Exercises Via Machine Learning: Initial Psychometric Evidence and Practical Guidelines

Louis Hickman[1], Christoph N. Herde[2], Filip Lievens[2], & Louis Tay[3]

[1]Department of Psychology, Virginia Tech; The Wharton School, University of Pennsylvania

[2]Singapore Management University

[3]Department of Psychology, Purdue University

Author Note

The data in the present study come from Herde and Lievens (2022). This work was supported by the Society for Industrial and Organizational Psychology (SIOP)'s 2022 Douglas W. Bray and Ann Howard Research Grant. An earlier version of this paper was presented at the SIOP 2022 Conference.

 10.1111/ijsa.12418

Abstract

Assessment center (AC) exercises such as role-plays have established themselves as valuable approaches for obtaining insights into interpersonal behavior, but they are often considered the "Rolls Royce" of personnel assessment due to their high costs. The observation and rating process comprises a substantial part of these costs. In an exploratory case study, we capitalize on recent advances in natural language processing (NLP) by developing NLP-based machine learning (ML) models to investigate the possibility of automatically scoring AC exercises. First, we compared the convergent-related validity and contamination with word count of ML scores based on models that used different NLP methods to operationalize verbal behavior. Second, for the model that maximized convergence while minimizing contamination with word count (i.e., a model that used both *n*-grams and Universal Sentence Encoder embeddings as predictors), we investigated the criterion-related validity of its scores. Third, we examined how the interrater reliability of the AC role-play scores affects ML model convergence. To do so, we applied seven NLP methods to 96 assessees' transcriptions and trained 10 sets of ML models across 18 speeded AC role-plays to automatically score assessee performance. Results suggest that ML scores recovered most of the original variance in the overall assessment ratings, and replacing one or more human assessors with ML scores maintained criterion-related validity. Additionally, ML models seemed to exhibit higher convergence when assessors consistently detected and utilized observable behaviors to make ratings (i.e., when interrater reliability was higher). Finally, we provide a step-by-step guide for practitioners seeking to implement ML scoring in ACs.

Keywords: machine learning; artificial intelligence; validation; Assessment center exercises; Interpersonal; Natural language processing

Practitioner Points

- Natural language processing and machine learning are being adopted for personnel assessment but not yet for assessment centers
- Assessment center exercises are amenable to being automatically scored
- Natural language processing and machine learning had sufficient validity to replace one or more human raters in the assessment center, suggesting it may be possible in practice to automatically score assessment centers to reduce assessor costs
- Step-by-step guidance for developing and deploying automatic assessment center scoring is provide

Automatic Scoring of Speeded Interpersonal Assessment Center Exercises Via Machine Learning: Initial Psychometric Evidence and Practical Guidelines

In assessment center (AC) exercises, assessees participate in a series of standardized, behavioral simulation exercises wherein their behaviors are observed, recorded, and evaluated by multiple trained assessors (International Taskforce on Assessment Center Guidelines, 2015). AC exercises are used for personnel selection and development, and AC scores exhibit incremental validity beyond cognitive ability and personality (Arthur et al., 2003; Hoffman et al., 2015; Sackett et al., 2017). However, AC exercises can cost hundreds of thousands of dollars to design and administer, making them prohibitively costly for many organizations (Thornton & Rupp, 2006). Assessors represent a major cost because they must be extensively trained and then take substantial time to observe and rate assessee behavior (Guidry et al., 2013; Wirz et al., 2013).

In recent years, machine learning (ML) applications in personnel assessment have emerged, including automatically scoring accomplishment records (Campion et al., 2016) and video interviews (Hickman, Bosch, et al., 2022; Nguyen et al., 2014). Such ML assessments can save organizations money by reducing the person-hours required for personnel assessment (Campion et al., 2016) and time to hire (Langer, König, et al., 2021). To automate assessment with ML, the assessment must be based on observable behaviors—which is a key aspect of AC exercise design. Indeed, AC exercises provide "a standardized evaluation of behavior based on multiple inputs" (International Taskforce, 2015, p. 1248), where the inputs include multiple behavioral exercises and ratings from multiple trained assessors. Considering all of this, it is relevant and important to examine whether AC exercises can be automatically scored.

Apart from the practical relevance and importance of investigating automatically scoring AC exercises, there is also a pressing need to better understand when ML applications work

better (or worse). Although such applications are beginning to emerge in assessment, we know little about the conditions that affect ML model convergence. Reliability is considered key to ML (Jacobucci & Grimm, 2020), and higher interrater reliability should increase ML model convergence because if human assessors cannot reliably detect and utilize behavioral cues (Funder, 1995), then an ML model trained to replicate those ratings is similarly unlikely to detect and properly utilize relevant cues.

The present study contributes to personnel assessment in several ways. First, we contribute to a growing stream of research applying ML to personnel assessment (e.g., Campion et al., 2016; Hickman, Bosch, et al., 2022; Sajjadiani et al., 2019; Speer, 2018). We develop and test ML AC models, finding that the ML models recover most of the original variance in the AC. Such approaches are being rapidly adopted in practice, yet research on them is still in its nascent stages (Rotolo et al., 2018). Second, we advance our understanding of ML model convergence by investigating the influence of interrater reliability. Although ML is often considered an atheoretical, data-driven process (Cheng et al., 2021), ML models can be considered a special kind of rater that uses observable "behaviors to replicate human ratings" (Hickman, Bosch, et al., 2022; p. 1342). In this vein, we draw on the realistic accuracy model (RAM; Funder, 1995) to explain why interrater reliability affects ML model convergence. Third, we contribute to AC practice by providing step-by-step guidance for implementing ML AC scoring. To the extent that automated scoring can replace one or more human assessors, AC costs can be substantially reduced (cf. Campion et al., 2016). In turn, pairing automated scoring with other technologies for reducing AC costs (e.g., Tippins & Adler, 2011) then holds potential for democratizing ACs and enabling a broader swath of organizations to adopt them.

Assessment Center Exercises: Evaluations of Interpersonal Behavior

AC exercises are behavioral simulations (e.g., role-plays, group discussions, oral presentations) wherein multiple trained assessors use a standardized approach to observe and rate assessees' behavior (International Taskforce, 2015; OSS Assessment Staff, 1948). The result is a series of scores for each assessee representing their behavior-based performance on dimensions and/or exercises. One of the main advantages of ACs is that they enable observing a wide variety of actual assessee behaviors. These behaviors relate to a variety of competencies like problem solving and decision-making but are especially relevant to the gamut of interpersonal skills (Dayan et al., 2002; Sackett et al., 2017). A focus on actual (interpersonal) behavior in diverse situations explains why AC exercise scores tend to exhibit incremental validity above and beyond general mental ability (GMA) and personality test scores (Arthur et al., 2003; Hoffman et al., 2015; Meriac et al., 2008). Further, when directly compared in the same samples, ACs exhibit superior criterion-related validity compared to GMA tests (Sackett et al., 2017).

However, AC exercises are costly to design and administer (Gaugler et al., 1987; Krause & Thornton, 2009; Thornton et al., 2000; Thornton & Rupp, 2006; Thornton & Potemra, 2010): Training multiple assessors and then having them observe AC exercise performance, classify assessee behaviors, and provide ratings requires a substantial investment of person-hours (Wirz et al., 2013). Therefore, researchers and practitioners have sought to streamline the design and administration of AC exercises using technology (International Taskforce, 2015; Tippins & Adler, 2011). Yet, to our knowledge, the potential of reducing costs by automating AC exercise scoring with NLP and ML is still an unexplored, albeit potentially fruitful, option.

Automating Assessment: NLP and Machine Learning

Two relatively recent developments have enabled the automatic scoring of assessments such as AC exercises: NLP and ML. NLP (and *text mining*) involves a variety of methods for

converting unstructured, natural language text into structured, quantitative data (Kobayashi et al., 2018), which can be applied to the speech of the assessees in AC exercises. By converting unstructured, natural language text into structured, quantitative data, we can then use it in ML (Kobayashi et al., 2018). In this case, supervised ML weights different aspects of assessee speech (operationalized via NLP) to predict AC exercise scores (e.g., Chapman et al., 2016; Yarkoni & Westfall, 2017). With ML, this can scale to rapidly score many assessees.

Although ordinary least squares (OLS) regression can be used, supervised ML often uses modern prediction methods that balance model bias and variance to maximize out-of-sample accuracy (Putka et al., 2018). *Bias* refers to ML model predictions that are consistently wrong in one direction, and *variance* refers to the extent a ML model's parameters capture the patterns in the training data rather than the population parameters (Yarkoni & Westfall, 2017). High bias occurs when a ML model underfits the data, such as when important predictors are omitted. High variance occurs when a ML model overfits the data, which is likely to occur, for example, with OLS regression when sample size (*N)* to predictor (*p*) ratios are low (i.e., *N:p* ratios; Chapman et al., 2016; Putka et al., 2018). This often occurs with natural language data given that, for example, there are many different ways of operationalizing natural language (*p*) that can simultaneously be included as predictors. When $N < p$, OLS regression does not have a unique result, but modern prediction methods generally do. For example, ridge regression (Hoerl & Kennard, 1970) regularizes regression coefficients (i.e., forces them toward zero) to reduce model complexity and overfitting, while still allowing many (e.g., $N < p$) predictors to prevent underfitting (e.g., Spisak et al., 2019).

Using NLP and ML together for personnel assessment holds significant potential benefits for organizations. Campion et al. (2016) found that their supervised ML models for

automatically scoring accomplishment records exhibited convergence comparable to a single human judge. Using these ML models to replace one of the three human judges that were historically used to score the accomplishment records would save the organization at least $163,000 per year (Campion et al., 2016).

Automating Assessment Center Exercise Scoring

The first step of automatically scoring AC exercises involves designing and administering AC exercises. Then, a set of ACs must be conducted to provide data for training the ML models. To analyze the unstructured video data, assessee responses must next be transcribed. Then, NLP is used to extract textual features and convert the unstructured, natural language text into structured, quantitative data.

Those quantitative textual features are then used as predictors of the assessor ratings during ML model training. The available data must be separated into two samples to avoid capitalization on chance: the *training* and *test* samples (e.g., Raudys & Jain, 1991). When the data has a natural split (e.g., Years 1 and 2; Campion et al., 2016; Speer, 2018), one set of data is used for training and the other for testing. However, ML research that relies on a single sample of data often uses nested *k*-fold cross-validation, wherein the data is split into *k* equally sized parts (known as *folds*; Hastie et al., 2009). Then, $k - 1$ parts are used for hyperparameter tuning and training, and the ML model trained on that data is used to predict the outcome variable in the test data (i.e., the remaining fold). The process is repeated *k* times, using each fold only once for testing.

As AC exercises are behavior-based assessments that emphasize what assessees say, they should be amenable to automatic scoring with NLP and ML. Our first research question addresses how highly the ML model out-of-sample predictions converge with observed AC

scores.

Research Question 1: To what extent do the ML AC exercise scores converge with human scores?

Importantly, although convergence is necessary, it is only one element of assessment validity in personnel selection (SIOP, 2018). Supervised ML that maximizes prediction of AC scores optimizes only one aspect of validity (i.e., convergence) and does not consider discriminant validity (Simms, 2008). ML scores may be contaminated with construct-irrelevant factors, such as word count. For example, human Test of English as a Foreign Language (TOEFL) essay scores are correlated with word count (i.e., response length), yet Educational Testing Service has routinely re-designed its models for automatically scoring the essays to reduce the influence of word count (Chodorow & Burstein, 2004; Lee et al., 2008). The concern is that the ML models will score participants not based on *what* they said but merely on *how much* they said. In the present case, the observed AC exercise scores do tend to correlate with word count. However, ML model scores should not exhibit an inflated relationship with word count compared to the observed scores, because this suggests that they are contaminated with irrelevant variance associated with response length instead of response quality.

Research Question 2: How do the observed and ML AC exercise scores relate to word count?

A common question during ML research regards how high ML model convergence must be to be useful. The emerging benchmark is that they should converge at least as highly as a single human assessor. In prior research, this has been examined by comparing single rater intraclass correlations to the convergent correlations between ML scores and aggregated human scores (Campion et al., 2016; Hickman, Bosch, et al., 2022).

Research Question 3: Do the ML AC exercise scores converge as highly with the human assessors as a single rater does?

However, even this does not address how ML model scores' criterion-related validity compares to single rater's scores' criterion-related validity. Therefore, we go beyond prior research by also comparing a single rater's (in this case, the AC exercise role-playing assessor's) criterion-related validity to the ML model predictions' criterion-related validity.

Research Question 4: Is the criterion-related validity of the ML AC scores comparable to a single rater's scores?

In practice, the goal of automating scoring via ML is to save money by replacing one or more human assessors (Campion et al., 2016). Although the information for Research Questions 1, 3, and 4 provide some evidence regarding whether ML-based scores can replace a human assessor, they do not directly address whether validity is maintained by replacing some of the human raters. In the current study, the exercises were one-on-one role-plays where the role player also served as an assessor. Such a setup necessitates the use of a role player. However, the additional two to three assessors who reviewed the videotaped interactions and provided ratings represent a substantial, additional cost that could be avoided if the combined role player and ML scores exhibit validity comparable to the original, combined scores from all human assessors.

Research Question 5: Is the criterion-related validity of the average role-playing assessor and the ML AC scores comparable to the average of all human assessor scores?

Influences on ML Model Convergence

As ML models can be considered a special kind of rater that uses observable behaviors to attempt to replicate human ratings (Hickman, Bosch, et al., 2022), theories that explain the accuracy of interpersonal perception, such the RAM, may also help explain the validity of ML

models. Initially, the RAM was developed in the context of personality trait judgments (Funder, 1995), but it has since been applied to personnel judgments in employment interviews (Christiansen et al., 2005), ACs (Haaland & Christiansen, 2002; Lievens et al., 2006; Lievens et al., 2015), and LinkedIn (Roulin & Levashina, 2019). The RAM posits that accurate interpersonal judgments can occur only when *relevant* behaviors are *available* for observation and when the judge *detects* these behaviors and correctly *utilizes* that information in making judgments (Funder, 1995; Funder, 2012). Interpersonal judgments, therefore, involve a four-stage process wherein relevant behaviors occur; some are available for observation, and the judge detects some of these behaviors, and may or may not utilize them correctly for judgments.

Given that ML models are trained to replicate human assessors, they tend to replicate the properties of those human assessors (Barocas & Selbst, 2016). Therefore, ML models can only consistently detect and utilize behaviors for scoring when human assessors also do so. The key indicator of assessors' ability to reliably detect and utilize behaviors is interrater reliability. Interrater reliability is one indicator of interpersonal judgment accuracy (Funder, 1995) that is rooted in constructivism (Kruglanski, 1989). Constructivism suggests that reality is only knowable through human perceptions, so interpersonal judgmental accuracy is a function of whether judges collectively agree. When judges disagree, it suggests they did not consistently detect and utilize behaviors to rate performance. Inconsistent judgments occur for several reasons, including differences in the judges' emotional states and agreeableness (Letzring, 2008; Wood et al., 2010), as well as differences in the availability of relevant behavioral cues across different AC exercises.

Reliability is considered foundational to successful ML (Jacobucci & Grimm, 2020). If assessors do not consistently detect and utilize behaviors (causing low interrater reliability), then

an ML model trained on those humans is also unlikely to consistently detect and utilize behaviors. We expect that this affects the convergence of ML models trained on those ratings.

Hypothesis 1: Interrater reliability relates positively to the magnitude of the ML models' convergence.

Method

We describe our sampling plan, all data exclusions, and all measures in the study. Data and research materials are not available because they were gathered in collaboration with a consultancy firm. All analysis code is available at https://osf.io/2kzqd/?view_only=1ae0b05e12f745048211859f1124a8f6. Data were analyzed using multiple packages in R and Python. The study design and analysis were not preregistered.

Archival Dataset

To explore the viability of automatically scoring AC exercises, we used an archival dataset (video and audio recordings) of 18 "flash" (3-minute) AC exercises that sampled situations relevant to junior management. Table 1 describes the situation assessees encountered in each of the 18 role-plays. These multiple, speeded AC role-plays were developed and administered in collaboration with a European business school and a professional consultancy firm to assess the strengths and weaknesses of an entire MBA cohort of 96 participants (51% female, mean age = 23.63) from 19 different countries (67% Belgian, the rest 5% or fewer). All participants had at least one year of work experience and responded to the role-plays in English.

In each role-play, role players and remote assessors rated assesses' exercise performance on a nine-point scale (*1 = should clearly be improved: starters' level* to *9 = obviously strong: role model behavior*). To ensure that these ratings were based on observable and relevant

behaviors, short checklists were developed that listed behaviors indicative of (in)effective performance per exercise. To generate the Overall Assessment Rating (OAR), we averaged each assessee's scores from the 18 exercises. More information about the development, administration, and rating procedure of the multiple, speeded role-plays can be found in Herde and Lievens (2022). As detailed by Herde and Lievens (2022), seven months after the speed AC role-plays, MBA instructors provided criterion ratings by assigning percentile scores to assessees on four criterion dimensions: task performance, contextual performance, teamwork, and communication (Goffin et al., 1996, 2009). On average, the four criterion dimensions correlated $r = .65$ (all ps $< .001$). Therefore, we averaged instructor ratings into an overall performance measure ($\alpha = .87$; see Viswesvaran et al., 2005), which served as the criterion measure.

Computer-extracted Verbal Behavior

We used a two-stage process to operationalize assessee verbal behavior. First, we paid a vendor to have humans manually transcribe the recordings of the role-plays[1]. Initially, we tested the feasibility of using computerized transcription using IBM Watson Speech-to-Text (IBM, 2019), but considerable background noise in the recordings caused computerized transcriptions to be inaccurate[2]. The AC involved simultaneously administering 18 exercises in a single, large room, causing echoes and background noise, which made it challenging for computerized transcription to identify words, unlike humans. After manual transcription, we isolated the assessee's speech (and discarded the role player's speech) for use in our analyses.

Second, we applied seven NLP techniques—including descriptive measures, a closed vocabulary (i.e., dictionary) approach, traditional open vocabulary approaches (i.e., *n*-grams),

---

1 Although the majority of recordings had video, nearly a third had only audio recordings. Therefore, we do not investigate the use of nonverbal behaviors as predictors in the ML models.

2 Due to the lower quality of some of the recordings, we could not reliably investigate the use of paraverbal behaviors (i.e., how one's voice sounds; e.g., Hickman, Bosch, et al., 2022) as predictors in the ML models.

and modern embedding methods (Eichstaedt et al., 2021)—to the transcripts to quantify assessee verbal behavior. Specifically, we: measured word count; used Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015); counted all *n*-grams where $n = 1$ & 2; calculated word embeddings with the Universal Sentence Encoder, RoBERTa (at both the document and aggregated sentence-level), and DistilBERT. We describe each in the Online Supplement.

Analytic Strategy

It is important to ensure that the ML model generalizes (or cross-validates) beyond the sample, or effect size estimates may be upwardly biased (Putka et al., 2018). Typically, in psychological research, researchers split the collected data into a calibration/training and validation/test sample (e.g., 80/20 split) to determine generalizability. However, the cross-validation results can be highly dependent on the sample split choice. Instead, in ML, the sample is split systematically into multiple *k* subsamples (i.e., *k*-folds) in order to train (or fit) the model on all but one fold and validate the model on the remaining fold. This is done multiple (i.e., *k*) times such that each fold serves once as a validation fold. Given that ML models also require hyperparameter tuning, which affects how predictor weights are assigned by the model, this tuning is done in a nested fashion (for example, nested *l* times in each of the *k* times). We used the caret R package (Kuhn, 2008) to conduct an extension of *k*-fold cross-validation: nested cross-validation. Due to the small *N* in the present study, each *k*-fold is a single participant (i.e., $k = N$), and this is known as leave-one-out-cross-validation (LOOCV). We used LOOCV to train and test a total of 10 ML models per exercise. The LOOCV process is illustrated in Figure 1. It involves conducting hyperparameter tuning on $N - 1$ participants (the "outer folds") using 3-fold cross-validation on the "inner folds" (i.e., folds created in the training data), then training a model on those participants using the optimal hyperparameter. The trained model then predicts

the AC exercise score for the holdout participant, and this process is repeated *N* times, holding out each participant only once.

We trained and tested 10 ML models that used the following predictor sets: (1) word count and word count squared; (2) 77 LIWC variables that focus on the content of speech (e.g., Analytical thinking, Social Processes, Positive Emotions); (3) the count of *n*-grams where $n = 1$ & 2; (4) 512 embeddings from USE, calculated on each sentence and then averaged across the participant's sentences in an exercise; (5) 768 embeddings from RoBERTa, calculated on each sentence and then averaged across the participant's sentences in an exercise; (6) 768 embeddings from DistilBERT, calculated on all available assessee speech in an exercise; (7) 768 embeddings from RoBERTa, calculated on all available assessee speech in an exercise; (8) a combination of the LIWC variables and *n*-grams; (9) a combination of LIWC variables, *n*-grams, and DistilBERT embeddings; and (10) a combination of *n*-grams and USE embeddings. As we trained and tested these models in each exercise, we trained and tested a total of 180 models.

For the first model that used word count as a predictor, we used ordinary least squares (OLS) regression in LOOCV. For the remaining nine models, we used ridge regression and conducted hyperparameter tuning (i.e., identifying the optimal value of λ) on $N - 1$ participants using 3-fold (the inner folds) cross-validation, then trained a model on those participants using the optimal hyperparameter. We tried 10 values of lambda generated by *caret* that ranged from .016 to .690. The model then predicted the AC exercise score for the holdout participant. Separately for each of the 18 exercises, this process was repeated *N* times, holding each participant out once for testing. Then, the *N* predictions in each exercise were combined and evaluated together to facilitate calculating convergent, discriminant, and criterion correlations.

To investigate how interrater reliability affects ML model convergence, we calculated the

correlation between 1) the correlation between predicted and observed AC exercise scores for the 18 exercises and 2) interrater reliability (*Hypothesis 1*). Finally, we conducted a sensitivity power analysis with G*Power (Erdfelder et al., 1996). We found that we had 80% power to detect an effect size $r = .50$ at $\alpha = .05$.

Results

To What Extent Do the ML Scores Converge with Human Ratings?

Table 2 reports the convergent correlations within each exercise, on average, and with the OAR (i.e., the average of each assessee's 18 exercise scores), for the 10 models trained and tested in our study to address Research Question 1. For example, the first column of Table 2, *WC* + WC^2, reports the convergence between the average human assessor scores and the ML scores for the models that used only word count and its quadratic as predictors. Only the models that used DistilBERT and the models that used a combination of *n*-grams and USE exhibited, on average, convergent correlations that exceeded the word count models. The average convergence of these models compares favorably to the average convergence across the Big Five for Park et al.'s self-report models (where $\bar{r} = .38$; 2015) and is similar to Hickman, Bosch, et al.'s interviewer-report models (where $\bar{r}_{min} = .38$ and $\bar{r}_{max} = .42$; 2022). The four models that included *n*-grams as predictors exhibited convergent correlations $r > .70$ for exercise 5.

In terms of the OAR (i.e., the average of each participant's 18 exercise scores), the human assessor OARs converged *r*s = .39 and .40 with the OAR from models that used RoBERTa embeddings calculated at the exercise level and LIWC variables as predictors, respectively. Meanwhile, the OAR from models that used word count, *n*-grams, and/or DistilBERT as predictors converged $.76 \leq r\text{s} \leq .79$, indicating very strong convergence at the overall level. Notably, across the 10 models, the correlation between the average convergence

(Average ($\bar{r}$)) and the OAR correlations r = .96, suggesting that higher convergence within the 18 exercises relates directly to the extent to which ML models capture the substantive variance in the AC.

Relationship with Word Count

The bottom row of Table 2 reports the average correlation between ML scores and assessee word count across the 18 exercises (Research Question 2). On average, the observed AC scores correlated $\bar{r}$ = .44 with word count. As can be seen, the strongest relationship between ML scores and word count is for the models that used word count and its quadratic term as predictors ($\bar{r}$ = .88). In other words, although this model exhibited convergence comparable to other models, as expected, the variance in scores is almost completely accounted for by variation in the word count. Notably, ML scores from models that used DistilBERT embeddings as predictors also correlated highly, on average, with the word count: when only DistilBERT embeddings were used, $\bar{r}$ = .73, and when DistilBERT was used together with LIWC variables and *n*-grams, $\bar{r}$ = .69. These correlations with word count are 66% and 57% larger than the correlations between observed scores and word count, suggesting that DistilBERT embeddings are contaminated with word count. The ML scores from models that used *n*-grams and USE embeddings as predictors exhibited convergence that, on average, exceeded the word count models and a weaker relationship with word count ($\bar{r}$ = .38) compared to the observed scores. On the basis of these two pieces of evidence, our remaining analyses focus on the models that used *n*-grams and USE embeddings as predictors[3].

How Does ML Score Convergence Compare to a Single Human Rater?

[3] The online supplement reports the correlations among observed and ML AC scores for these models to provide information regarding discriminant-related validity. The evidence suggests that the ML models adequately distinguish among the different AC exercises.

Table 3 reports the single and average interrater reliabilities for each exercise, correlations with sex and age for both the observed and ML AC scores, and the convergent correlations (and 95% confidence intervals) between observed and ML scores (for the models that used *n*-grams and USE embeddings as predictors) for each of the 18 AC exercises and the OAR. Single rater reliability (i.e., the average convergence of single human assessors), $G(q,1)$ (Putka et al., 2008), averaged .38—which is .03 lower than the average ML model convergence. In 16 AC role-plays, the 95% confidence interval (CI) for ML model convergence included the value of $G(q,1)$, and in the two role-plays where it did not, ML model convergence exceeded single rater reliability. Answering Research Question 3, ML scores converged somewhat more highly than a single human assessor[4].

How Does the Criterion-Related Validity of ML Scores Compare to Single Humans?

Table 4 reports the correlation between exercise scores, OARs, and the criteria for all human raters, the role player, and the ML scores for the 18 exercises. On average, the role player criterion correlations $\bar{r}$ = .25 and the ML model criterion correlations $\bar{r}$ = .19. In 14 AC role-plays, the ML models' criterion correlations' 95% CIs included the value for the role player—in one role-play, the 95% CI values exceeded the role player's criterion correlation, and in three role-plays, the role player's criterion correlation exceeded the 95% CI values. The role player OAR scores correlated r = .53 with the criterion, and the ML OAR scores correlated r = .47, 95% CI [.30, .61], with the criterion. Overall, to answer Research Question 4, the ML scores exhibited criterion-related validity comparable to, but of a somewhat lesser magnitude than, a single human assessor.

[4] One concern is that our results may be upwardly biased due to the use of LOOCV. As we report in Online Supplement Table S3, the results were largely consistent when we used 10-fold cross-validation, 5-fold cross-validation, or 3-fold cross-validation. For LOOCV, convergence averaged .41, whereas for 10-fold, 5-fold, and 3-fold cross-validation, convergence averaged .40.

Do ML Scores Maintain Criterion-Related Validity When Replacing Human Assessors?

The first column of Table 4 reports the criterion correlations for the average of all human assessors' ratings, and the final column of Table 4 reports the same information for the average of the role player and ML scores. On average, across the 18 exercises, the combined human assessors correlate with the criterion $\bar{r}$ = .32, and the averaged role player and ML scores correlate with the criterion $\bar{r}$ = .28. Additionally, the combined human assessor OAR scores correlated r = .57 with the criterion, and the role player plus ML OAR scores correlated r = .55, 95% CI [.39, .68], with the criterion[5]. Answering Research Question 5, the average of the role player and ML scores exhibited criterion-related validity comparable to the averaged human assessors.

Influence of Interrater Reliability on ML Model Convergence

As seen in Table 2, there was considerable variation in convergence for each model across the 18 AC role-plays. For example, for the models that used *n*-grams and USE embeddings as predictors, average convergence $\bar{r}$ = .41, r_{sd} = .16, r_{min} = .06 (exercise 8), and r_{max} = .73 (exercise 5). Hypotheses 1 posited that ML models for AC role-plays with higher interrater reliability would exhibit higher convergence. For the two NLP methods that exhibited average convergence $\bar{r}$ < .24 (i.e., those with poor validity), the relationship between convergence and interrater reliability $\bar{r}$ = .18. For the word count ML models, the relationship between convergence and interrater reliability r = .67. And for the remaining ML models, the relationship between convergence and interrater reliability $\bar{r}$ = .61. The correlation is significant at p < .01 when r > .60, and thus, Hypothesis 1 is supported.

[5] To check if the results would hold with non-role player assessors, we averaged the ML scores together with the scores from one of the raters who later observed and rated assessee performance. The assessor plus ML OAR scores correlated r = .55 with the criterion—identical to the role player plus ML OAR scores' criterion correlation.

Discussion

Main Conclusions

First, AC exercises are amenable to being automatically scored using NLP and ML. AC exercises are behavior-based assessment methods, and the present study used speeded AC role-plays where the assessees' verbal responses were the key inputs for human assessor scores. As a result, several NLP methods resulted in ML models that converged comparably to single human assessors. Importantly, however, the models that used DistilBERT embeddings as predictors were contaminated with variance associated with word count, whereas the models that used *n*-grams and USE embeddings as predictors were less contaminated with word count than the human ratings themselves. Further, several ML model scores recovered the majority of substantive variance in the AC, as evidenced by multiple OAR correlations exceeding .70.

Additionally, the ML OAR scores (from the models that used *n*-gram and USE embeddings as predictors) exhibited criterion-related validity (r = .47) comparable to a single human assessor (r = .53; in this case, the role-playing assessor), albeit of a somewhat lesser magnitude than the role player's OAR scores. Further, when the ML and role player scores were averaged together, their criterion-related validity was very similar to the average of all human assessors for the OAR scores ($r_{\text{role player+ML}}$ = .55, r_{humans} = .57).

Therefore, ML scores could be used to supplement role player ratings by replacing the additional 2-3 assessors who later observed and rated performance without sacrificing validity. In our experience, the cost of freelance assessors is often around $75 per hour. In this study, 96 assessees completed 18 AC role-plays, for a total of 1,728 videos that last about 3 minutes each. Assuming assessors review and rate 2 videos in 10 minutes, the cost of a single human assessor would be approximately $10,800. Therefore, the savings could be around $21,600 for replacing

two human raters, or $32,400 for replacing three human raters. These savings would multiply as the number of assessees grows, and pairing these savings with other cost-savings measures, such as remote administration, could enable more organizations to adopt ACs for assessment.

Second, these ML scores were valid despite the small sample size in the present study (N = 96) and the relatively brief sample of behavior. Regarding sample size, Sajjadiani et al. (2019) trained their turnover attributions classifier on 1,000 observations; Campion et al. (2016) trained their accomplishment record scorers on 41,429 observations; and Park et al. (2015) trained their system for scoring the Big Five from Facebook posts on 66,732 observations. Further, the exercises lasted only about three minutes. During that time, assessees spoke an average of 260 words per exercise, whereas Park et al. (2015) discarded all participants who had fewer than 1,000 words in their Facebook posts. Generally, NLP and supervised ML are thought to be useful only when applied to "big" data—which could include many observations or a lot of data per observation. Our results demonstrate that ML can be applied to relatively small samples (both number of participants and amount of behavior) when the assessment is rooted in observable behavior. However, we do not suggest deploying ML models trained on such small samples. We observed considerable variability in the validity of our models depending on the predictor set used, even though we would a priori expect many of these methods to perform comparably.

Third, scant knowledge and empirical evidence are currently available to explain when and under what conditions ML models are likely to be exhibit high convergence. This is an important consideration because, much like AC exercises, collecting data for and developing ML models is an expensive and arduous endeavor. Such data generally involves collecting observations and having multiple trained assessors rate each observation (e.g., Campion et al., 2016; Hickman, Bosch, et al., 2022). We hypothesized and found that ML models exhibit higher

convergence in exercises where assessors consistently detected and properly utilized behaviors because this increases interrater reliability. Interrater reliability is key for supporting the validity of any inferences made using human ratings, so research design should account for this. For example, evaluative situations should ensure numerous relevant behaviors are available for thoroughly trained assessors to detect and utilize, and ratings should be facilitated via behavioral observation aids, such as checklists of effective behaviors, behavioral observation scales, or behaviorally anchored rating scales. Behavioral observation aids help to increase interrater reliability by providing a common frame of reference for judging performance that alleviates response biases and rater idiosyncrasies (Jacobs et al., 1980; Barnardin & Smith, 1981; Roch et al., 2012).

New Frontiers: Investigating Specific Behaviors in ACs with ML

Although not a core focus of the present manuscript, NLP and ML can be used to identify the behaviors that lead to high AC exercise scores. Researchers are interested in investigating behavior in ACs (Breil et al., 2022), and many popular press books give advice on how to behave in AC exercises. NLP and ML can provide empirical evidence on how specific behaviors translate into scores—Figures 2 through 4 illustrate the stemmed *n*-grams and LIWC categories most strongly associated with the ML AC scores in the three role-plays with the highest convergent correlations: 3, 5, and 14[6]. For example, assessees should avoid responding with short phrases that contain assent words (e.g., yeah, okay, agree), particularly when presented with exaggerated criticisms or inappropriate suggestions for last-minute changes (like boycotting, rescheduling, or changing key elements of the event, as in these role-plays). Not only were assent words associated with poor performance in role-plays 5 and 14, but assent words

[6] We do not report correlations with USE embeddings because they are not interpretable.

were also negative predictors of conscientiousness in Hickman, Bosch, et al.'s (2022) automated video interview personality assessments. Similarly, when encountering hostility (as in role-play 3), it is important to maintain a positive emotional tone (LIWC categories *tone* and *posemo*) and avoid reflecting the role player's negativity back to them (LIWC categories *negemo* and *anx*). Further, the (in)effective behaviors also speak to situations beyond AC exercises. For example, in role-play 14, trying to absolve oneself of responsibility by placing the blame and responsibility on the volunteer coordinator (who was female; LIWC categories *female* and *shehe*) was associated with lower scores. Likely, trying to blame others is detrimental in most situations when a problem needs to be solved. Such findings may enhance our understanding of effective interpersonal behavior both within AC role-plays and elsewhere. Importantly, using computers to measure behavior requires much less time and labor compared to traditional, human-coding approaches to measuring behavior. As a result, NLP and ML more broadly may open up additional opportunities to investigate behavior in ACs and beyond at a scale that was previously unfeasible due to the human labor bottleneck.

Guidelines for Developing and Deploying ML AC Scoring

Table 5 reports our guidelines for developing and deploying ML AC scoring, and we describe the major steps here. The first step involves designing the AC and its exercises, and extensive guidance has been provided elsewhere on designing ACs (e.g., International Taskforce, 2015; Thornton & Rupp, 2006). To facilitate automatic scoring, the AC should be designed and administered such that video and audio of performance can be clearly recorded. A complication here is that many ACs involve exercises wherein multiple assessees interact (e.g., leaderless group discussion)—such exercises may be less amenable to automatic scoring due to difficulties in automatically transcribing and distinguishing different assessees' speech. As a result, ACs

focused on one-on-one roleplays and exercises that involve a single assessee completing tasks (e.g., in-basket exercise) may be best suited for automatic scoring.

The second step involves administering the AC. The site should be prepared in advance to ensure that the audio-visual recording equipment functions properly, including checking that the audio from one exercise is cleanly captured even when other exercises are ongoing. Prior to AC administration, follow best practices for structuring rating scales and training roleplayers and assessors. Readers interested in these more general points should refer to existing resources (e.g., International Taskforce, 2015; Thornton & Rupp, 2006).

The third step involves transcribing the recorded assessee performance. Several commercial software packages are available for automatically transcribing speech. Comparing these tools can be important, as some may be systematically less accurate for racial and ethnic minorities (e.g., Koenecke et al., 2020). If multiple speakers are involved, diarization is the process of identifying multiple speakers and tracking them throughout the interactions, and it is an option that must enabled (i.e., is not a default) on many computerized transcription systems.

The fourth step involves applying NLP to the transcriptions. Although, as we showed, a variety of methods are available, word embedding methods—especially those based on transformers—have become the most popular choice for many NLP applications. Such embedding methods tend to achieve high convergence, yet they are often less interpretable than older methods, such as *n*-grams. Numerous resources are available online for working with embedding methods, and Table 5 cites some common NLP R packages and Python libraries.

After converting the unstructured text data into quantitative data, the fifth step involves training ML models to use assessee behavior to predict AC exercise scores and testing them. To use all available data and provide the most accurate estimates of out-of-sample performance, we

suggest conducting repeated *k*-fold cross-validation (Krstajic et al., 2014). This involves conducting nested *k*-fold cross-validation multiple times, with each repeat randomly shuffling the composition of the *k*-folds. *k*-fold cross-validation provides a nearly unbiased estimate of the true convergence (Varma & Simon, 2006), and repeating the process ensures that any chance variation is accounted for (Krstajic et al., 2014). At this stage, multiple models (either due to different NLP methods or algorithms) can be trained and tested.

The sixth step involves comparing the psychometric properties of these models. As we demonstrated, investigating psychometric properties beyond convergence is important to ensure that the resulting ML model scores are capturing as much intended variance as possible (i.e., minimizing construct deficiency) without including additional, irrelevant variance (i.e., construct contamination). Landers and Behrend (2022) provide additional guidance on considerations at this stage beyond the quantitative psychometric properties of the ML model scores.

The seventh step involves training the final models. Because *k*-fold cross-validation involves training and testing *k* models, none of those models should be the final ones deployed for assessment. Instead, all available data should be used to train the final models that will be deployed for personnel assessment. Although the psychometric properties of these models will not be known, the results from *k*-fold cross-validation are nearly unbiased (Varma & Simon, 2006), and additional training data is generally beneficial for ML models. For algorithms that include hyperparameters, the hyperparameters can be tuned using all available data in non-nested *k*-fold cross-validation (e.g., Hickman, Bosch, et al., 2022 used 10-fold cross-validation to identify optimal hyperparameters for models trained on entire samples).

The eighth step involves developing and evaluating ML model explanations for relevant stakeholders. Explaining the model design process, reliability, and validity to stakeholders can

improve reactions to the ML system, including fairness perceptions (Langer, Oster, et al., 2021). For example, hiring managers and applicants could be informed about how the models were developed and their psychometric properties, and hiring managers could be trained on effectively using the ML scores in decision-making. However, explanations may actually *worsen* reactions from stakeholders affected by ML decisions (Newman et al., 2020).

The ninth step involves deploying the final models to assess new AC participants. Ideally, these models would not be used initially to make selection decisions but would instead be piloted by comparing their scores to the scores assigned by human assessors. This would further increase confidence that the ML models provide valid scores, but if the ML models exhibit poor validity, then it suggests they should not be used moving forward. If they are deployed immediately for assessment, we suggest following the example set by essay scoring in standardized testing—if the ML model score and a single human disagree, then another human should provide a performance rating to resolve the difference.

The tenth step occurs after the ML models have been deployed for assessment—continuous monitoring and validation of the ML scores. This is a best practice for any selection system, because to the extent that the job or more global conditions change, predictor-criterion relations may also change. Continuously monitoring for both validity and bias are important to maximize utility and reduce the likelihood of litigation.

Limitations

Although we illustrated that NLP and supervised ML could be used on small sample sizes, this was also a limitation of the present study. We showed that validity could be maintained using nested LOOCV, but we also observed that several models that we expected to perform well exhibited worse validity relative to other methods. One reason is that the small

sample size prevented us from using some NLP methods to their full potential, such as fine-tuning end-to-end deep learning embedding methods. Additionally, the small sample size meant we could not retain a true holdout sample, and therefore, our study does not address the question of whether models trained on such a small sample would generalize to new, future groups of assessees. The small sample size may have also resulted in a more homogeneous sample than is encountered in practice, which could have inflated effect sizes. To obtain more accurate effect size estimates, future work should aim to use larger samples and test the generalizability of ML scores' validity in new, temporally lagged samples.

Conclusion

AC exercises are costly to design and administer, and recent years have witnessed the emergence of NLP and ML for reducing assessment costs. This study investigated the potential of replacing one or more human assessors with ML scores. The results suggest it may be possible to replace one or more human assessors with NLP and ML in speeded interpersonal AC exercises to save money and achieve criterion validities similar to those obtained from human assessors. Interrater reliability had a large influence on ML model convergence.

References

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.

Assessment Staff, O. S. S. (1948). *Assessment of men: Selection of personnel for the Office of Strategic Services*. New York, NY: Rinehart & Co.

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

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*(2), 190-203.

Bogaert, J., Trbovic, N., & Van Keer, E. (2005). *Ability Test Suite—Level III – Manual.* Ghent, Belgium: Hudson.

Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021). Bias and fairness in multimodal machine learning: A case study of automated video interviews. *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21)*. https://doi.org/10.1145/3462244.3479897

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology, 65*(3), 546–553. https://doi.org/10.1037/0022-3514.65.3.546

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin Slices of Behavior as Cues of Personality and Intelligence. *Journal of Personality and Social Psychology, 86*(4), 599–614. https://doi.org/10.1037/0022-3514.86.4.599

Breil, S. M., Lievens, F., Forthmann, B., & Back, M. D. (2022). Interpersonal behavior in assessment center role-play exercises: Investigating structure, consistency, and effectiveness. *Personnel Psychology,* advance online publication. https://doi.org/10.1111/peps.12507

Brüeckl, M., & Heuer, F. (2018). *irrNA: Coefficients of interrater reliability: Generalized for randomly incomplete datasets*. Available at https://cran.r-project.org/package=irrNA

Byham, W. C. (1977). Assessor selection and training. In J. L. Moses & W. C. Byham (Eds.), *Applying the assessment center method* (pp. 89–125). Pergamon Press.

Byham, W. (2016). *Assessment centers for large populations.* Presented at the International Congress on Assessment Center Methods, Bali, Indonesia.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 1, pp. 687-731). Palo Alto, CA: Consulting Psychologists Press.

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101*(7), 958–975. https://doi.org/1.1037/apl0000108

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, *41*, 1054–1072. https://doi.org/10.1016/j.jrp.2007.01.004

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., & Kurzweil, R. (2018, November). Universal sentence encoder for English. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations* (pp. 169-174).

Chapman, B. P., Weiss, A., & Duberstein, P. R. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological Methods, 21*(4), 603–620. https://doi.org/10.1037/met0000088

Cheng, M. M., & Hackett, R. D. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. *Human Resource Management Review*, *31*(1), 100698. https://doi.org/10.1016/j.hrmr.2019.100698

Christiansen, N. D., Wolcott-Burnam, S., Janovics, J. E., Burns, G. N., & Quirk, S. W. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. *Human Performance*, *18*(2), 123-149.

Dayan, K., Kasten, R., & Fox, S. (2002). Entry-level police candidate assessment center: An efficient tool or a hammer to kill a fly? *Personnel Psychology, 55*(4), 827–849. https://doi.org/10.1111/j.1744-6570.2002.tb00131.x

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181.04805*.

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V. A., Smith, L. K., Buffone, A., Iwry, J., Seligman, M. E. P., & Ungar, L. H. (2021). Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, *26*(4), 398-427.

Erdfelder, E., Faul, F., & Buchner, A. (1996). G*POWER: A general power analysis program. *Behavior research methods, instruments, & computers*, *28*(1), 1-11.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*(4), 652–67.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*(3), 177-182.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, *72*(3), 493-511.

Gebauer, J. E., Sedikides, C., Wagner, J., Bleidorn, W., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2015). Cultural norm fulfillment, interpersonal belonging, or getting ahead? A large-scale cross-cultural test of three perspectives on the function of self-esteem. *Journal of Personality and Social Psychology*, *109*(3), 526-548.

Guidry, B. W., Rupp, D. E., & Lanik, M. (2013). Tracing cognition with assessment center simulations: Using technology to see in the dark. In *Simulations for personnel selection* (pp. 231-257). Springer, New York, NY.

Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology*, *55*(1), 137-163.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-84858-7

Herde, C. N., & Lievens, F. (2020). Multiple speed assessments: Theory, practice, and research evidence. *European Journal of Psychological Assessment, 36*(2), 237–249. https://doi.org/10.1027/1015-5759/a000512

Herde, C. N., & Lievens, F. (2022). Multiple, speeded assessments under scrutiny: Underlying theory, design considerations, reliability, and validity. *Journal of Applied Psychology,* Advance online publication. https://doi.org/10.1037/apl0000603

Hickman, L. (2021). *Algorithmic ability prediction in video interviews* [Unpublished doctoral dissertation]. Purdue University. https://doi.org/10.25394/PGS.14687172.v1

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, 107*(8), 1323-1351. https://doi.org/1.1037/apl0000695

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, *25*(1), 114-146. https://doi.org/1.1177/1094428120971683

Hirevue. (2022). *Explainability statement* (White paper). Available at https://webapi.hirevue.com/wp-content/uploads/2022/04/HV_AI_Short-Form_Explainability_1pager.pdf

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Hoffman, B. J., Kennedy, C. L., LoPilato, A. C., Monahan, E. L., & Lance, C. E. (2015). A review of the content, criterion-related, and construct-related validity of assessment center exercises. *Journal of Applied Psychology, 100*(4), 1143–1168. https://doi.org/1.1037/a0038707

Hunter, J. E., Schmidt, F. L., & Judiesch, M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology*, *75*(1), 28-42.

IBM. (2019). *IBM Watson Speech to Text*. Available at https://www.ibm.com/watson/services/speech-to-text/

International Taskforce on Assessment Center Guidelines. (2015). Guidelines and ethical considerations for assessment center operations. *Journal of Management*, *41*(4), 1244-1273.

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. *Personnel Psychology*, *33*(3), 595-640.

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809-816.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality*, *61*(4), 521-551.

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

Kim, H., Di Domenico, S. I., & Connelly, B. S. (2019). Self–other agreement in personality reports: A meta-analytic comparison of self-and informant-report means. *Psychological Science*, *30*(1), 129-138.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational Research Methods*, *21*(3), 733-765.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, *117*(14), 7684-7689.

Krause, D. E., & Thornton III, G. C. (2009). A cross-cultural look at assessment center practices: Survey results from Western Europe and North America. *Applied Psychology*, *58*(4), 557-585.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1–26. https://doi.org/10.1053/j.sodo.2009.03.002

Kuncel, N. R., & Sackett, P. R. (2014). Resolving the assessment center construct validity problem (as we know it). *Journal of Applied Psychology, 99*, 38–47. https://doi.org/10.1037/a0034147

Kruglanski, A. W. (1989). The psychology of being" right": The problem of accuracy in social perception and cognition. *Psychological Bulletin*, *106*(3), 395-409.

Langer, M., König, C. J., & Busch, V. (2021). Changing the means of managerial work: effects of automated decision support systems on personnel selection tasks. *Journal of Business and Psychology*, *36*(5), 751-769.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473. https://doi.org/10.1016/j.artint.2021.103473

Leary, M. R., & Hoyle, R. H. (Eds.). (2009). *Handbook of individual differences in social behavior*. Guilford Press.

Leavitt, K., Schabram, K., Hariharan, P., & Barnes, C. M. (2021). The Machine Hums! Addressing Ontological and Normative Concerns Regarding Machine Learning Applications in Organizational Scholarship. *Academy of Management Review*, early online access.

Letzring, T. D. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality*, *42*(4), 914-932.

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91,* 247–258. http://dx.doi.org/1.1037/0021-901.91.2.247

Lievens, F., Schollaert, E., & Keen, G. (2015). The interplay of elicitation and evaluation of trait-expressive behavior: Evidence in assessment center exercises. *Journal of Applied Psychology*, *100*(4), 1169-1188.

Lievens, F., Tett, R. P., & Schleicher, D. J. (2009). Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. In J. J. Martocchio & H. Liao

(Eds.), *Research in personnel and human resources management* (pp. 99–152). Bingley, UK: JAI Press.

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 preprint arXiv:1907.11692*.

Lubinski, D. & Dawis, R. V. (1992). Aptitudes, skills and proficiencies. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, 2nd ed., pp. 1–59). Palo Alto, CA: Consulting Psychologists Press.

McCrae, R. R., & Costa, P. T. (1989). The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of Personality and Social Psychology*, *56*(4), 586-595.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*(2), 175-215.

McLaughlin, M. E., Carnevale, P., & Lim, R. G. (1991). Professional mediators' judgments of mediation tactics: Multidimensional scaling and cluster analyses. *Journal of Applied Psychology*, *76*(3), 465-472.

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology*, *93*(5), 1042-1052.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, *75*(6), 640-647.

Murphy, N. A. (2007). Appearing smart: The impression management of intelligence, person perception accuracy, and behavior in social interaction. *Personality and Social Psychology Bulletin*, *33*, 325–339. https://doi.org/10.1177/0146167206294871

Murphy, N. A., Hall, J. A., & Colvin, C. R. (2003). Accurate intelligence assessments in social interactions: Mediators and gender effects. *Journal of Personality*, *71*, 465–493. https://doi.org/10.1111/1467-6494.7103008

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, *160*, 149-167.

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.

Oliver, T., Hausdorf, P., Lievens, F., & Conlon, P. (2016). Interpersonal dynamics in assessment center exercises: Effects of role player portrayed disposition. *Journal of Management*, *42*(7), 1992-2017.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, *108*(6), 934-952.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pinchback, J. (2017, October 4). *Introducing talent auditions*. LinkedIn Talent Connect presentation. Retrieved from https://www.youtube.com/watch?v=1UwVTOqlPwI

Pinsight. (2019, June). *Shorter & faster: Recent trend in Assessment Centers*. https://www.pinsight.com/blog/2019/3/5/recent-trend-in-assessment-centers.

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

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*(5), 959-981.

Refaeilzadeh, P., Tang, L., & Liu, H. (2016). Cross-validation. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database systems*, 532-538.

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on pattern analysis and machine intelligence*, *13*(3), 252-264.

Raven, J. C. (1958). *The standard progressive matrices.* London, England: H. K. Lewis.

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*, 370–395. https://doi.org/1.1111/j.2044-8325.2011.02045.x

Rotolo, C. T., Church, A. H., Adler, S., Smither, J. W., Colquitt, A. L., Shull, A. C., Paul, K. B., & Foster, G. (2018). Putting an end to bad talent management: A call to action for the field of industrial and organizational psychology. *Industrial and Organizational Psychology*, *11*(2), 176-219.

Roulin, N., & Levashina, J. (2019). LinkedIn as a new selection method: Psychometric properties and assessment approach. *Personnel Psychology*, *72*(2), 187-211.

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology*, *102*(10), 1435-1447.

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.

Schäpers, P., Mussel, P., Lievens, F., König, C. J., Freudenstein, J. P., & Krumm, S. (2020). The role of Situations in Situational Judgment Tests: Effects on construct saturation, predictive validity, and applicant perceptions. *Journal of Applied Psychology*, *105*(8), 800-818.

Society for Industrial and Organizational Psychology. (2018). *Principles for the Validation and Use of Personnel Selection Procedures* (Fifth). American Psychological Association. https://doi.org/10.1017/iop.2018.195

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, *71*(3), 299-333.

Spisak, B. R., van der Laken, P. A., & Doornenbal, B. M. (2019). Finding the right fuel for the analytical engine: Expanding the leader trait paradigm through machine learning?. *The Leadership Quarterly*, *30*(4), 417-426.

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior*, *120*, 103445. https://doi.org/10.1016/j.jvb.2020.103445

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., & Bühner, M. (2020). Personality research and assessment in the era of machine learning. *European Journal of Personality*, *34*(5), 613-631.

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A Conceptual Framework for Investigating and Mitigating Machine-Learning Measurement Bias (MLMB) in Psychological Assessment. *Advances in Methods and Practices in Psychological Science*, advance online publication. https://doi.org/10.1177/25152459211061337

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied psychology*, *88*(3), 500-517.

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, *34*(4), 397–423. https://doi.org/1.1006/jrpe.20.2292

Tett, R. P., Toich, M. J., & Ozkum, S. B. (2021). Trait Activation Theory: A Review of the Literature and Applications to Five Lines of Personality Dynamics Research. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*, 199-233.

Thornton, G. C., Murphy, K. R., Everest, T. M., & Hoffman, C. C. (2000). Higher cost, lower validity and higher utility: Comparing the utilities of two tests that differ in validity, costs and selectivity. *International Journal of Selection and Assessment, 8*, 61-75.

Thornton, G. C. III, & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Thornton III, G. C., & Potemra, M. J. (2010). Utility of assessment center for promotion of police sergeants. *Public Personnel Management, 39*, 59-69.

Tippins, N. T., & Adler, S. (2011). *Technology-enhanced assessment of talent* (Vol. 30). John Wiley & Sons.

Vrijdags, A., Bogaert, J., Trbovic, N., & Van Keer, E. (2014). *Business Attitudes Questionnaire (Psychometric technical manual).* Ghent, Belgium: Hudson.

Wirz, A., Melchers, K. G., Lievens, F., De Corte, W., & Kleinmann, M. (2013). Trade-offs between assessor team size and assessor expertise in affecting rating accuracy in assessment centers. *Revista de Psicología del Trabajo y de las Organizaciones*, *29*(1), 13-20.

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of G-theory methods for modeling multitrait–multimethod data: Clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods*, *15*(1), 134-161.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Schleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38-45).

Wood, D., Harms, P., & Vazire, S. (2010). Perceiver effects as projective tests: What your perceptions of others say about you. *Journal of Personality and Social Psychology, 99*, 174–190.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122.

Tables

Table 1

Description of the 18 assessment center exercises

Role-play	Description
1	Role player wants to find extra volunteers but does not want to run into a conflict with other event services.
2	Role player is dissatisfied with many aspects of last year's event and threatens to take action to forbid the event.
3	Role player is hostile and wants to boycott the event because several activities threaten the attendees' safety.
4	Role player feels inexperienced and insecure because her efforts to increase sales do not pay off.
5	Role player mentions a popular sport event is scheduled on the same day and that this should be solved.
6	Role player promised a band to play at the event but the committee had already decided hosting a different band.
7	Role player wants to brainstorm about solving the problem of shortage of volunteers.
8	Role player criticizes assessee, asking to make quick decisions regarding specific entertainment issues.
9	Role player (finance coordinator) asks to make a choice among various options, while staying within the budget.
10	Role player is inexperienced, feels close to burnout, and considers resigning from her job.
11	Role player (a police inspector) is angry because the current event proposal does not meet safety regulations.
12	Role player is angry about another employee who does not meet his task expectations.
13	Role player feels disengaged and is unmotivated to switch to another catering option.
14	Role player suggests to completely change the event activities, although many preps have already been done.
15	Role player lost the registration list and has problems to acknowledge it because of potential face loss.
16	Role player (beverage supplier) mentions that a final order for beverages was never placed and his schedule is full.
17	Role player (from ICT) is furious and questions the need to take up extra IT tasks.
18	Role player mentions a double booking was made regarding the order of plates and cutlery.

Table 2

Comparison of ML model convergence, word count saturation, and relationship with interrater reliability

				Sentence-level		Document-level				
Exercise	WC + WC^2	LIWC	*n*-grams	USE	RoBERTa	DistilBERT	RoBERTa	LIWC + *n*-grams	LIWC + *n*-grams + DistilBERT	*n*-grams + USE
1	.10	-.03	.32	.38	.17	.22	.06	.31	.32	.44
2	.27	.07	.20	.39	.33	.33	.31	.19	.35	.37
3	.61	.33	.48	.30	.35	.55	.18	.60	.63	.52
4	.42	.16	.40	.39	.44	.47	.29	.37	.41	.44
5	.46	.54	.76	.61	.58	.56	.53	.76	.72	.73
6	.41	.19	.35	.35	.37	.42	.26	.35	.43	.45
7	.52	.17	.48	.28	.28	.52	.33	.44	.49	.46
8	.25	.11	.16	-.05	-.18	.20	.04	.17	.21	.06
9	.11	.23	.23	.42	.23	.08	.32	.25	.22	.37
10	.38	-.03	.32	.23	.21	.37	.01	.32	.40	.30
11	.55	.01	.47	.41	.48	.57	.29	.42	.49	.52
12	.18	.04	-.07	.21	.13	.23	.19	-.06	.06	.10
13	.47	.28	.53	.36	.18	.60	.25	.53	.61	.49
14	.55	.55	.52	.51	.59	.64	.56	.61	.66	.56
15	.58	-.24	.37	.26	.33	.61	.05	.36	.61	.40
16	.33	.26	.34	.34	.19	.46	.35	.38	.47	.44
17	.58	.16	.49	.12	.25	.45	.05	.48	.45	.37
18	.09	.27	.24	.20	.28	.31	.11	.30	.34	.27
Average ($\bar{r}$)	.38	.17	.37	.32	.29	.42	.23	.38	.44	.41
OAR	.77	.40	.76	.60	.62	.78	.39	.78	.79	.77
Word Count ($\bar{r}$)	.88	.08	.45	.18	.18	.73	.08	.47	.69	.38

Note: WC = Word Count. LIWC = Linguistic Inquiry and Word Count. USE = Universal Sentence Encoder. OAR = Overall Assessment Rating. All models used ridge regression except WC + WC^2, which used OLS regression. Sentence-level means embeddings were calculated on each sentence then averaged within participants, and document-level means embeddings were calculated on all text at once. In each exercise, the *N* predictions from LOOCV were combined to calculate correlations with other variables.

Table 3

Correlations with observed (i.e., human rated) and machine learning (n-grams and USE predictors) AC exercise scores

	Observed Scores						Machine Learning Scores			
Exercise	$G(q, 1)$	$G(q, k)$	Sex	Age	Nationality	Word Count	Observed	[95% CI]	Sex	Age	Nationality
1	.22	.63	.05	-.16	.21	.22	.44	[.26, .59]	.05	-.04	.06
2	.30	.63	-.21	-.29	.12	.37	.37	[.18, .53]	-.02	-.13	.11
3	.37	.73	.04	-.22	.20	.65	.52	[.36, .65]	.15	-.21	.20
4	.48	.75	.04	-.29	.27	.46	.44	[.26, .59]	.04	-.29	.29
5	.64	.85	-.11	-.36	.22	.50	.73	[.62, .81]	-.13	-.27	.31
6	.43	.71	-.08	-.28	.25	.47	.45	[.27, .60]	.05	-.21	.22
7	.30	.61	.13	-.22	.10	.54	.46	[.29, .60]	.03	-.34	.22
8	.26	.57	.00	-.20	.26	.30	.06	[-.14, .26]	-.01	-.13	.09
9	.23	.50	-.11	-.11	.18	.22	.37	[.18, .53]	-.13	-.17	.27
10	.45	.74	.11	-.14	.33	.43	.30	[.11, .47]	-.03	-.05	.14
11	.48	.77	.00	-.25	.26	.57	.52	[.36, .65]	-.13	-.13	.09
12	.29	.62	.00	-.17	.03	.26	.10	[-.10, .29]	.23	.00	-.09
13	.47	.77	.03	-.30	.20	.50	.49	[.32, .63]	.01	-.20	.16
14	.53	.83	.14	-.29	.31	.55	.56	[.40, .68]	.14	-.13	.26
15	.42	.79	.06	-.19	.28	.61	.40	[.22, .56]	.14	-.28	.28
16	.18	.46	-.05	-.24	.16	.35	.44	[.26, .59]	.08	-.25	.02
17	.41	.70	.07	-.28	.32	.58	.37	[.18, .53]	.06	-.22	.11
18	.34	.61	.05	-.22	.25	.26	.27	[.07, .45]	-.02	-.04	.19
OAR	--	--	.02	-.42	.40	--	.77	[.67, .84]	.06	-.42	.42

Note: $N = 96$. $P < .05$ when $r > .20$; $p < .01$ when $r > .26$. $G(q,1)$ = single rater reliability. Observed AC exercise scores are formed from the average of all available human raters (i.e., role-players and remote assessors). OAR is the average of the 18 exercise scores. We also calculated convergence with observed scores using Spearman's rank-order correlation ρ following Stachl et al.'s (2020) recommendation and found that, on average across the 18 exercises, $\bar{\rho} = .39$. These results are consistent regardless of whether predictors are standardized (i.e., mean = 0 and SD = 1).

Table 4

Comparison of human assessor and predicted AC exercise scores' criterion-related validities

Exercise	All Assessors	Role Player	ML Scores [95% CI]		Role Player and ML Mean [95% CI]	
1	.27	.20	.01	[-.19, .21]	.22	[.02, .40]
2	.36	.24	.22	[.02, .40]	.26	[.06, .44]
3	.41	.13	.24	[.04, .42]	.19	[-.01, .38]
4	.27	.26	.17	[-.03, .36]	.27	[.07, .45]
5	.40	.43	.29	[.10, .46]	.43	[.25, .58]
6	.32	.32	.31	[.12, .48]	.38	[.19, .54]
7	.22	.03	.35	[.16, .51]	.13	[-.07, .32]
8	.47	.37	.10	[-.10, .29]	.40	[.22, .56]
9	.15	.14	.12	[-.08, .31]	.14	[-.06, .33]
10	.28	.21	.28	[.08, .45]	.18	[-.02, .37]
11	.31	.30	.12	[-.08, .31]	.31	[.12, .48]
12	.31	.30	-.05	[-.25, .15]	.31	[.12, .48]
13	.27	.21	.34	[.15, .51]	.29	[.10, .46]
14	.44	.30	.23	[.03, .41]	.32	[.13, .49]
15	.43	.34	.22	[.02, .40]	.37	[.18, .53]
16	.18	.13	.16	[-.04, .35]	.20	[-.00, .39]
17	.41	.35	.13	[-.07, .32]	.38	[.19, .54]
18	.25	.28	.19	[-.01, .38]	.30	[.11, .47]
Average	.32	.25	.19	[-.01, .38]	.28	[.08, .45]
OAR	.57	.53	.47	[.30, .61]	.55	[.39, .68]

Note: Criterion correlations are Pearson's correlations between the criterion measure and a) the average of all human assessor ratings of exercise performance (All Assessors), b) the role players' ratings of exercise performance (Role Player), c) the ML predicted scores (ML scores), and d) the average of the role player and ML predicted scores in each exercise (Role Player and ML Mean). The average criterion correlation is the average convergence with the criterion across the 18 exercises. OAR is the correlation between the average of the 18 exercise scores and the criterion.

Table 5

Guidelines for developing and deploying ML AC scoring

Step	Helpful Resources
1. Design the AC	International Taskforce on Assessment Center Guidelines (2015); Thornton & Rupp (2006)
2. Administer the AC	International Taskforce on Assessment Center Guidelines (2015); Thornton & Rupp (2006)
3. Transcribe responses via software	Rev; Amazon Transcribe; IBM Watson Speech-to-Text
4. Operationalize verbal behavior	Transformers & sentence_transformers (Python libraries); text2vec & tm (R packages)
5. Train and test ML models using nested *k*-fold cross-validation	Hickman, Bosch, et al. (2022; Figure 2); Krstajic et al. (2014); Lever et al. (2016); Varma & Simon (2006); caret (R package); scikit-learn (Python library)
6. Select the model with the best psychometric properties	SIOP *Principles*; Landers & Behrend (2022)
7. Train final models on all available data	Hickman, Bosch, et al. (2022; p. 1332)
8. Develop and pilot explanations for relevant stakeholders	Landers & Behrend (2022); Langer, Oster, et al. (2021)
9. Deploy models for scoring	Rupp et al. (2020) provide guidance for piloting Pareto-optimal selection weights, which is a data-driven process like ML.
10. Ongoing monitoring of ML score validity and bias	SIOP *Principles*; Landers & Behrend (2022); Tay et al. (2022)

Figures

Figure 1

Leave-one-out cross-validation strategy

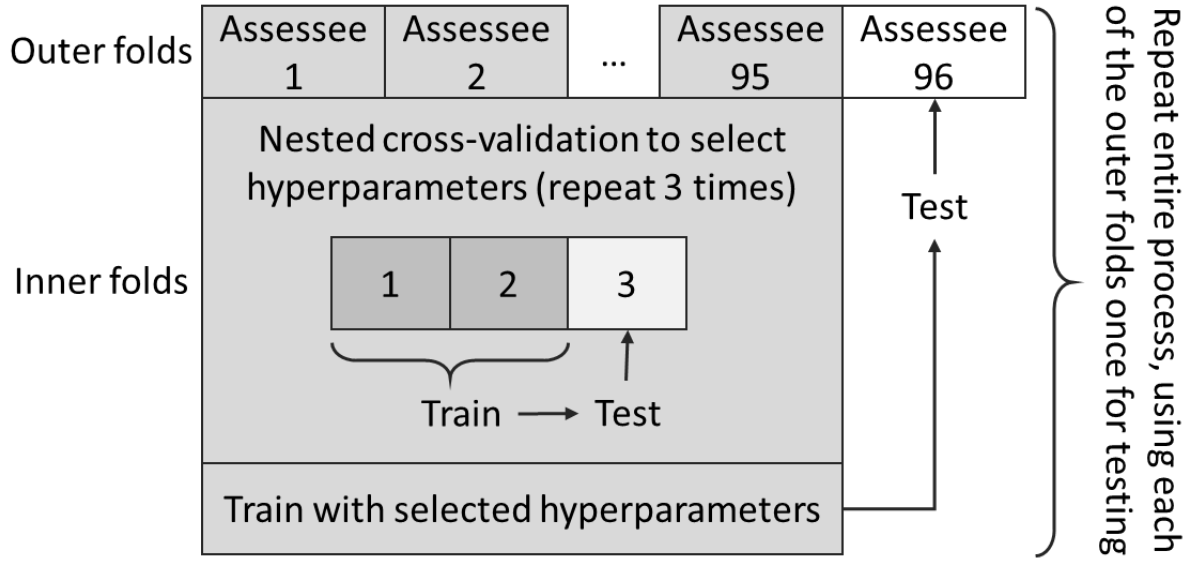
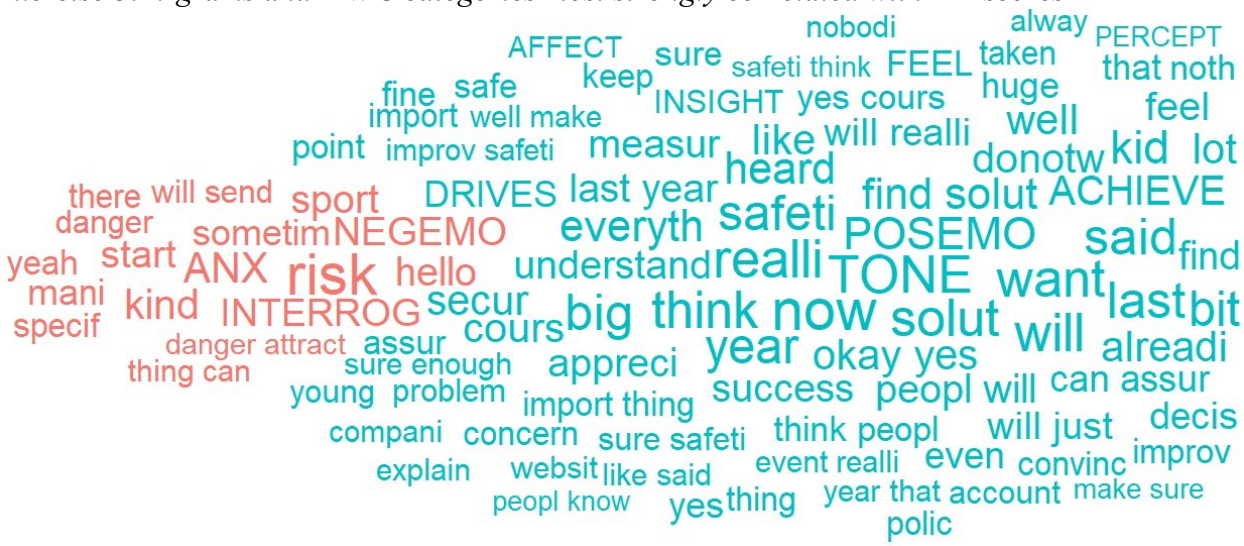


Figure 2

Exercise 3: n-grams and LIWC categories most strongly correlated with ML scores



Note: Red indicates the *n*-gram or LIWC category correlated negatively with ML scores, and blue indicates a positive correlation. LIWC categories are in ALL CAPS. Word/phrase/LIWC category size is proportional to correlation magnitude, and the negative and positive correlations are on the same scale.

Figure 3

Exercise 5: n-grams and LIWC categories most strongly correlated with ML scores

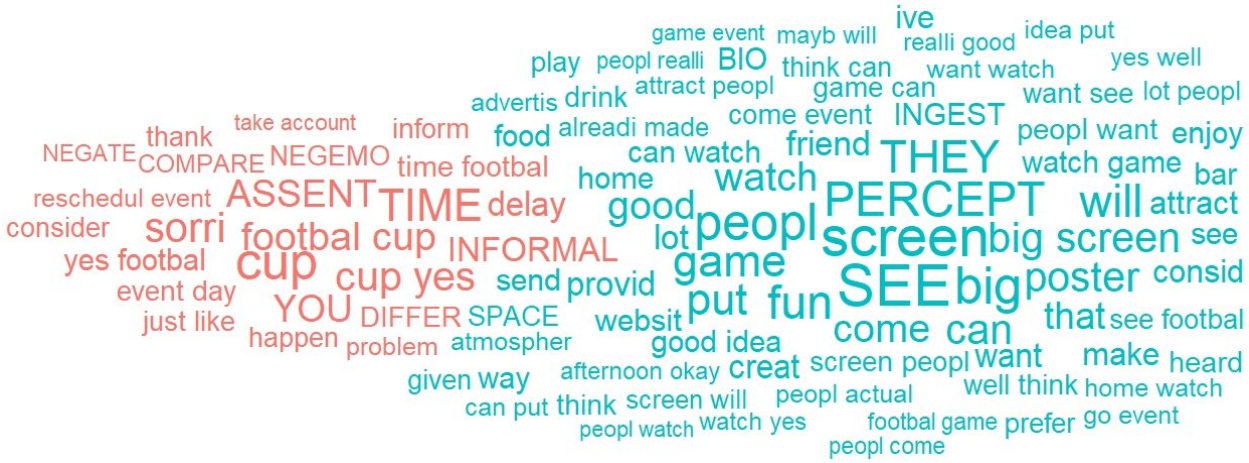


Figure 4

Exercise 14: n-grams and LIWC categories most strongly correlated with ML scores