Developing and Validating Automated Scoring for an Audio Constructed Response Simulation

Mengqiao (MQ) Liu, Denver McNeney, John F. Capman, Shane B. Lowery, Matthew Kitching,

Anjali Nimbkar, and Anthony S. Boyce

Amazon, Inc.

This article is In Press at Personnel Psychology. Full citation:

Koenig, N., Tonidandel, T., Thompson, I., Albritton, B., Koohifar, F., Yankov, S., Speer, A., Hardy, J., Gibson, C., Frost, C., Liu, M., McNeney, D., Capman, J. F., Lowery, S. B., Kitching, M., Nimbkar, A., Boyce, A., Sun, T., Guo, F., Min, H., Zhang, B., Lebanoff, L, & Newton, C. (In Press). Improving Measurement and Prediction in Personnel Selection through the Application of Machine Learning. In Press at *Personnel Psychology*.

Author Note

Correspondence concerning this article should be addressed to Mengqiao (MQ) Liu, Amazon Inc., 1600 7th Ave, Seattle, WA 98101, United States. Email: mengqiao.liu7@gmail.com

Abstract

We evaluated the effectiveness of machine learning (ML) and natural language processing (NLP) for automatically scoring a simulation requiring audio-based constructed responses. We administered the simulation to 3,174 recent new professional-level hires working in a large multinational technology company. Human subject matter experts (SMEs) scored each response using behaviorally anchored rating scales of interpersonal and decision-making skills which we then used to train Bidirectional Encoder Representations from Transformers (BERT) NLP models to generate computer scores of those skills. Results demonstrate evidence of convergent validity between human and computer scores (correlations ranging from .66 to .74), as well as criterion related-validity (uncorrected correlations ranging from .08 to .21; corrected correlations ranging from .17 to .25) and incremental validity (2.8% additional variance) above and beyond the existing assessments against supervisor ratings of incumbent job performance. Computer scores showed similar subgroup differences to human scores and exhibited no predictive bias.

Keywords: simulation, constructed response, natural language processing, validity, selection

Developing and Validating Automated Scoring for an Audio Constructed Response Simulation

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., assessment centers, or 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., Campion et al., 2016), no published research to date has assessed the feasibility of automated scoring of audio-based constructed responses in a personnel selection context. Our study addresses this gap by evaluating whether automatic speech recognition (ASR) and natural language processing (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 1) 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[1], which is expected to positively impact candidates' assessment experience.

Method and Results

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

[1] 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.

professional roles (e.g., software engineers, research scientists, product managers) at a large multinational technology company. Of the 9,940 incumbents invited, 3,174 (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 subject matter experts (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 minute response (maximum = 5 minutes) 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 seconds (*SD* = 31 seconds). Participants were allowed up to three attempts to respond to each scenario. The last response recorded was scored (average recording length = 43 seconds, *SD* = 28 seconds)[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 6-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

[2] 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.

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.

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[3]. We continued using benchmark cases (random selection of five cases) on a weekly basis to monitor and coach SME raters to maintain or improve accuracy.

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

[3] 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.

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 1, 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 (Campion et al., 2016). Table 1 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.[4] KSAO scores were calculated by taking the mean of the individual human SME scores.

[TABLE 1 ABOUT HERE]

Automated Scoring

The automated scoring process consists of two parts. First, we used Amazon Transcribe automatic speech recognition (ASR) product to transcribe recorded audio responses to text. Second, we trained Bidirectional Encoder Representations from Transformers (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,

[4] 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.

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.[5] Lower WER scores indicate more accurate transcriptions. We used an expert human transcription service to manually transcribe 1,200 minutes 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 2 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.

[5] 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).

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., 2018)[6]. 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 Appendix (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 = 0.04; see Table 8 in the Online Supplement:

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

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 1 in the Online Supplement presents the model score descriptive statistics.

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 1 above and Table 3 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[7]. 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 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 1), which was possible in part because the reliability of the mean of the raters was very high (ICC2s were above .9, Table 1).

We observed strong collinearity in the human KSAO scores, and computer scoring exacerbated this problem (see Table 5 in the Online Supplement). We suspect this may be an

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

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 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 2 and Table 5 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 0.52; 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.

[TABLE 2 ABOUT HERE]

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[8]). While the absolute effect size is small, the relative gain from the baseline suggests the practical value of including the simulation.

Subgroup Differences in Human and Computer Scores

Table 3 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 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

[8] 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.

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[9].

[TABLE 3 ABOUT HERE]

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

[9] 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).

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.

Data Availability Statement

Research data are not shared.

Tables

Table 1

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	0.28	0.94	0.75	0.74
Assertive Communication	0.21	0.91	0.68	0.66
Oral Communication	0.27	0.94	0.73	0.70
Cooperation & Coordination	0.25	0.93	0.77	0.71
Interpersonal Adaptability	0.29	0.94	0.73	0.68
Social Influence	0.29	0.94	0.76	0.67
Decision Making	0.27	0.94	0.78	0.74
Average of KSAOs			**0.82**	**0.76**

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

Table 2

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	0.07 (0.20)	0.08 (0.19)	0.08 (0.14)	0.11 (0.17)
Assertive Communication	0.12* (0.16)	0.14* (0.25)	0.12* (0.10)	0.18** (0.24)
Cooperation & Coordination	0.14* (0.29)	0.08 (0.19)	0.15* (0.25)	0.11 (0.17)
Social Influence	0.10 (0.23)	0.08 (0.19)	0.10 (0.17)	0.11 (0.17)
Interpersonal Adaptability	0.10 (0.19)	0.12* (0.23)	0.11 (0.14)	0.15* (0.21)
Oral Communication	0.16** (0.26)	0.17** (0.25)	0.17** (0.23)	0.21*** (0.25)
Decision Making	0.10 (0.21)	0.10 (0.21)	0.11 (0.16)	0.13* (0.19)
Overall Score (Average of KSAOs)	0.13* (0.25)	0.12* (0.23)	0.14*(0.20)	0.15*(0.21)
Response Length (Word Count)	-0.01 (0.01)	-0.01 (0.01)	0.02(0.01)	0.02(0.01)

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)

Table 3

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	-0.43	-0.50
	Black/African American	29	3.45	3.44	0.09	0.04
	Hispanic/Latino	52	3.52	3.48	0.24	0.16
	Two or More Races	43	3.40	3.42	-0.02	-0.00
Gender	Male (Reference)	1138	3.23	3.26	-	-
	Female	401	3.40	3.40	0.37	0.36

Select References

Agarwal, A., Dudík, M., & Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. *arXiv preprint arXiv*: 1905.12843.

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

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

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101*(7), 958–975. https://doi.org/10.1037/apl0000108

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

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

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

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

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT press.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

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, P., Qiu, X, & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv: 1605.05101

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

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

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929-1958.

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

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. 10.1037/0021-9010.81.5.557