Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews

Brandon M. Booth
University of Colorado Boulder
USA
brandon.m.booth@gmail.com

Louis Hickman
Purdue University
USA
louishickman@gmail.com

Shree Krishna Subburaj
University of Colorado Boulder
USA
shree.subburaj@colorado.edu

Louis Tay
Purdue University
USA
stay@purdue.edu

Sang Eun Woo
Purdue University
USA
sewoo@purdue.edu

Sidney K. D'Mello
University of Colorado Boulder
USA
sidney.dmello@colorado.edu

ABSTRACT

We introduce the psychometric concepts of bias and fairness in a multimodal machine learning context assessing individuals' hireability from prerecorded video interviews. We collected interviews from 733 participants and hireability ratings from a panel of trained annotators in a simulated hiring study, and then trained interpretable machine learning models on verbal, paraverbal, and visual features extracted from the videos to investigate unimodal versus multimodal bias and fairness. Our results demonstrate that, in the absence of any bias mitigation strategy, combining multiple modalities only marginally improves prediction accuracy at the cost of increasing bias and reducing fairness compared to the least biased and most fair unimodal predictor set (verbal). We further show that gender-norming predictors only reduces gender predictability for paraverbal and visual modalities, while removing gender-biased features can achieve gender blindness, minimal bias, and fairness (for all modalities except for visual) at the cost of some prediction accuracy. Overall, the reduced-feature approach using predictors from all modalities achieved the best balance between accuracy, bias, and fairness, with the verbal modality alone performing almost as well. Our analysis highlights how optimizing model prediction accuracy in isolation and in a multimodal context may cause bias, disparate impact, and potential social harm, while a more holistic optimization approach based on accuracy, bias, and fairness can avoid these pitfalls.

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; • **Information systems** → *Multimedia and multimodal retrieval*; **Content analysis and feature selection**; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

bias, fairness, multimodal learning, automated video interview

ACM Reference Format:
Brandon M. Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K. D'Mello. 2021. Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3462244.3479897

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10.
https://doi.org/10.1145/3462244.3479897

1 INTRODUCTION

Much research has demonstrated that humans use multimodal information channels to make social inferences, decisions, and judgments about others [21, 25, 35, 37]. For many, the path towards improving computational models that make inferences about mental constructs from observable behaviors, especially in the wild, involves capturing and utilizing more information in context across a wide range of modalities (e.g., [14, 19, 46]). However, incautious efforts following this approach may result in harmful consequences for certain groups of people stemming from biases and unfair decisions in the machine learning and prediction process. For example, in a study designed to train a machine-learned model to make inferences about job-related social variables (e.g., communication, professionalism) of job candidates based on multimodal analysis of their prerecorded "video interviews," Muralidhar et al [38] found that the model's predictions were more accurate for men than women. The authors posited that this gender disparity in performance was due to the inclusion of multimodal predictors related to "powerful speech" behaviors for which men were rewarded and women were often penalized, consistent with observed gender stereotypes [34]. More and more, these types of multimodal models are being used to aid in measurement and assessment of psychological constructs in high-stakes contexts, such as *hireability* in pre-employment screening [8, 22, 38]. Thus, it is important to better understand the potential harmful effects of including different and multiple modalities when the model predictions are used in high-stakes decision making. In this paper, we use pre-employment screening as a case study and examine the effect of combining modalities not just on model prediction performance but also in terms of bias and fairness across genders.

Bias and fairness in machine learning are not simply a byproduct of the representativeness or correspondence of a data set to its target population. Bias and unfairness emerge as a result of human decisions made throughout the model development process. The data-driven approach of throwing all available predictors into a model and seeing what sticks is useful for maximizing accuracy, but if the most useful predictors also contain encoded information about unrelated traits (e.g., gender, age) then this resulting model may generalize poorly, leading to different prediction accuracies for different groups of people and resulting in fairness concerns.

As a simple example, consider a scenario where a model is trained to assess the hireability of potential job candidates for a warehouse restocking position where considerable strength is required for lifting objects. If all available candidate information is provided, the model may learn that sex is an important predictor because it (biologically) correlates with strength [16, 26]. This model may rate males suitable for the job more often than females and raise fairness concerns. However, if no demographic information (e.g., sex) is provided but strength is still included as a predictor, the model may still rate males suitable more often than females (a possible indicator of bias) because strength serves as a proxy predictor for sex. The fairness concern in this case is further complicated by the fact that strength *is* relevant to the job and is therefore an admissible predictor in spite of its outcomes favoring males (Title VII of the US Civil Rights Act of 1964 would allow for this difference in treatment if employers could show that strength is both job relevant and consistent with business necessity). In a multimodal scenario where many more predictors are available, each potentially encoding information about traits unrelated to the target construct, the complexity of bias and fairness concerns is much greater. These considerations have been the subject of much attention in *psychometrics* research (i.e. the science of measuring latent, psychological constructs) [51], but they are only beginning to receive attention in machine learning research [3, 36], despite being paramount concerns when considering deploying these technologies in the real world.

A prime example of the use of multimodal models in a real, high-stakes decision-making scenario is an automated video interview (AVI). In AVIs, machine-learned models are trained to make assessments about the hireability of candidates for particular jobs based on prerecorded video interviews [23]. Some organizations have already begun to incorporate this technology into their hiring workflow and are interested in adopting selection procedures that 1) help them select high performing employees, 2) are unlikely to result in lawsuits (coming from bias/fairness concerns), and 3) support diversity and inclusion initiatives [43]. One company claimed to have conducted over a million AVIs by mid-2019 [22], and the adoption of AVIs has likely accelerated during the global COVID-19 pandemic.

In this work, we use AVIs as a case study to examine differences in gender accuracy, bias, and fairness measures when using unimodal or multimodal predictors to produce model predictions of candidate hireability. We initially employ baseline models using all available predictors in unimodal and multimodal setups. Because the inclusion of all predictors within each modality combination is likely to manifest in bias and fairness concerns, we additionally explore the multimodal effects on accuracy, bias, and fairness of two model variants designed to reduce gender bias and improve fairness: predictor gender-norming (i.e., *z*-scoring) within gender groups (as discussed by [47]) and iteratively removing the predictors that contain the most information about gender. We investigate whether these approaches can reduce bias and/or improve fairness when making hireability predictions from different unimodal and multimodal predictors while still preserving enough relevant information to accurately assess hireability. To support this effort, we collected responses to AVI questions from over 700 participants in a simulated hiring study, and we asked a panel of trained annotators, acting as recruiters, to provide hireability ratings for each participant using established, psychometrically validated measures. We address the following research questions:

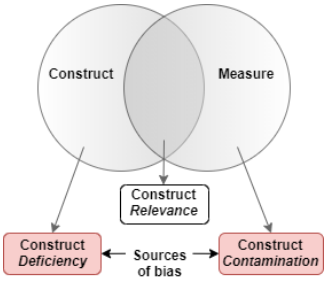

Figure 1: Sources of measurement bias in construct assessment [2, 6]

- **RQ1:** How does the inclusion of predictors from different modalities affect the accuracy, gender bias, and fairness of AVI hireability assessments?
- **RQ2a-b:** What effects do a) gender-norming features and b) iterative feature reduction to remove gender information have on accuracy, gender bias, and fairness?

2 BACKGROUND AND RELATED WORK

In order to ground research on bias and fairness of multiple modalities and AVIs, we first introduce the psychometric concepts of measurement bias and fairness.

2.1 Measurement Bias and Fairness

Though bias, fairness, and ethics have recently become hot topics in affective computing and multimodal machine learning, there is a long history of research on enhancing the fairness and reducing the bias of psychological assessments [29]. This prior research focuses on measurement bias, which is distinct from other forms of bias (e.g., implicit bias, sampling bias) often covered in the artificial intelligence and machine learning (ML) literature [29, 36]. According to measurement theory, bias occurs when there is systematic (i.e., non-random) error in a measurement procedure (e.g., ML prediction) that causes a difference in measurement accuracy (as assessed by alignment with ground truth) in one group compared to another [2]. For example, gender measurement bias occurs when men and women have equal ground truth scores but the ML predictions are lower for one group compared to the other.

Many human trait modeling endeavors, including AVIs, are designed to measure latent constructs (e.g., ability, personality) that are not directly observable and instead must be inferred from measurements. The goal is to capture construct-relevant variance while minimizing *construct deficiency* (construct-relevant variance not captured), and *construct contamination* (construct-irrelevant variance that is captured). Figure 1 illustrates this, also showing that the substantive, construct-relevant variance is captured when a measure overlaps with the construct. This representation of bias comes from the psychometrics literature (e.g., [2, 6]) and is simple, useful, and helps demonstrate a risk when indiscriminately scaling the number of predictors up (i.e., the "big data" approach). Specifically, as more (imperfect) measures of a construct are used, there is more construct-irrelevant information to potentially contaminate model predictions and worsen bias (though using more predictors can also help reduce construct deficiency).

Fairness is a social construct and another important consideration when assessing a measure's validity [2]. In the context of personnel selection and AVIs, the two most relevant aspects of fairness are

procedural and distributive fairness. Procedural fairness regards the perceived fairness of the elements of the decision-making process and in a measurement context is akin to measurement bias. Distributive fairness regards whether the allocation of important resources, based on the decisions made or resulting from psychological assessment (e.g., hireability), is perceived as fair. This type of fairness is encoded in business law in the United States in the idea of *adverse impact* [1]. Adverse impact (also called disparate impact) regards whether selection ratios (i.e., the number of hires divided by the number of applicants) is substantially different across legally protected groups (e.g., race/ethnicity, gender). The most common metric for judging adverse impact is the four-fifths rule, or that the selection ratio of one group should not be less than four-fifths the selection ratio of another group [10]. Organizations considering adopting machine learning systems to aid in personnel selection are very concerned with avoiding adverse impact, because it constitutes *prima facie* evidence of discrimination that, regardless of the outcome of a court case, can cause major expense and public relations issues. However, this relevant legal statute has received little attention in previous human-centered ML research or in AVI studies.

Many measures have been proposed for quantifying manifestations of bias and unfairness in ML contexts. For example, *equal accuracy* [53] regards whether an assessment is equally accurate across groups. This measure reflects our definition of bias since a measure that is not equally accurate across groups may be contaminated by construct-irrelevant information or deficient in relevant variance. Multiple notions of distributive fairness exist, including equality, or that each group should receive equal outcomes (which is the perspective of adverse impact statutes, although some leeway is given), and equity, or that outcomes received should be proportional to their inputs [17].

In this work, we adopt *correlational accuracy* [24, 49] for assessing measurement bias, *gender predictability* for identifying potentially biased predictors, and *adverse impact* for measuring one type of distributive fairness. These metrics are defined and discussed in Section 3.7. Readers are referred to [29, 36, 50] for more information on other bias and fairness metrics.

2.2 Related Work on Bias/Fairness in AVIs

AVI research began in earnest with Nguyen et al [40] who found that their ML models could accurately assess hireability from paraverbal and visual behavioral cues. However, research using visual cues extracted from videos to infer personality traits began even earlier [5]. This work is relevant because personality– along with intelligence– are considered to be one of the most robust predictors of occupational outcomes and are hence relevant to hiring decisions [4].

Given that the job-interview process is multimodal, especially in employment interviews [27], it seems important to consider including verbal, paraverbal, and visual behaviors in AVI ML models. However, the majority of prior AVI research has used only one or two of these modalities. For example, [40] did not include verbal behavior in their AVI models even though employment interviews have been called verbal tests of ability [28]. Similarly, because the ChaLearn First Impressions data set [44] (also used as a job applicant screening challenge) used short 15-second video clips, virtually all models applied to the data have not included verbal predictors (for example, [53]).

Several recent studies such as [9, 39] have examined the accuracy of ML models constructed for personality and hireability assessment using features from each modality. Both of these works examined the accuracy of binary classification of the constructs, which is helpful in separating low and high performers but impractical when trying to select only a few of the highest ones. Rasipuram et al [45] explored multimodal modeling of communication performance in asynchronous versus face-to-face interviews, but also used binarized labels for prediction and model evaluation. Chen et al [8] constructed multimodal models to predict personality and hireability in a small (n=36) experiment and this work is perhaps most similar to the present one. These authors found 1) verbal features to be substantially more useful than the others in personality assessment, 2) both verbal and paraverbal features somewhat helpful for hireability assessments, and 3) visual features (e.g., seven facial emotions) relatively unhelpful. Each of these works focused on the accuracy of personality or hireability assessment, while the present work primarily investigates the biases stemming from these modalities as well.

Regarding bias mitigation approaches for AVIs, there was a push by some psychometricians in the 1980's towards adoption of group norming [20]. Group norming involves separately converting raw test scores to percentiles for each group, where the aim is to attenuate construct contamination (i.e., bias) by reducing the amount of construct-irrelevant (i.e., group) information [20, 26] . For example, an intelligence test would be normed separately based on test takers' race (e.g., Black and White) to address measurable group differences in test scores and equalize test performance relative to racial affiliation [20, 26]. However, group norming has been explicitly outlawed in the United States since the Civil Rights Act of 1991 due to concerns about reverse discrimination. Nevertheless, various forms of within-group normalization continue to appear in machine learning efforts (e.g., [31, 54]) and are worth exploring in low-stakes research contexts as they help develop understanding of the effectiveness of reducing group information while preserving construct-relevance. We investigate the merits of group z-norming with regards to gender and the features extracted from different modalities, and then we compare it to a feature elimination strategy where features containing gender information (i.e., gender-relevant variance) are removed prior to modeling.

2.3 Contribution and Novelty of Current Study

Our paper is the first to introduce and evaluate psychometric bias and fairness in a multimodal context and demonstrate the benefits of removing gendered-features when making hireability assessments in AVIs. Our contributions are as follows:

(1) We root our investigation of unimodal versus multimodal bias in psychometrics by introducing the concepts of *measurement* bias and fairness, starting a dialog in the ML community about the trade-off between accuracy, bias, and fairness with regards to predictors and modalities.
(2) (RQ1) We use a large (n=733) data set to investigate and quantify manifestations of measurement gender bias and fairness emerging from verbal, paraverbal, and visual modalities in the context of AVIs.
(3) (RQ2a-b) We explore gender-norming and iterative gendered-predictor removal as two strategies for mitigating gender bias in an AVI context and for improving fairness by reducing the amount of gender information in the feature set. We investigate whether these approaches can reduce bias while still preserving enough relevant information to accurately assess hireability.

(4) We conclude with a discussion about the legal issues surrounding bias and fairness in AVIs and the importance of adopting a holistic optimization approach–one that considers accuracy, bias, and fairness–for machine learning pipelines when deployed for use in high-stakes decision-making scenarios.

3 METHOD

We analyze a data set of prerecorded mock interviews collected as part of an AVI study. A total of 4255 videos were collected from 733 participants and were separately rated by a panel of annotators to assess hireability. Our study protocol was reviewed for informed consent, data collection, data storage, data access, among other items and approved by a university institutional review board.

3.1 Data Collection Protocol

A total of 733 participants, consisting primarily of upper-level undergraduate students, were recruited from multiple universities and the crowd-sourcing website Prolific. Students at this level were expected to be seeking employment (either as interns or regular positions) and were thereby considered to be an appropriate sample for this work. Participants were compensated with a $10 Amazon gift card for participating when recruited directly from a university or a direct payment of $7.20 in Prolific.

Participants completed a mock interview comprising six interview questions, each of which was answered in a 1-3 minute video recording. Prior to responding to the first question, participants were given an opportunity to familiarize themselves with the online video capture system by answering the faux question, "Please tell us about yourself." The six questions were displayed one-at-a-time in random order and were designed to elicit information relevant to assessing hireability for a generic managerial/team lead role. Interviews were retained for analysis as long as features could be extracted from at least four of the six videos (detailed below). The study and interview were administered asynchronously online. Figure S1 in the supplemental materials shows a sample screenshot of the video capture system interface alongside the mock interview questions.

3.2 Ground Truth Annotation by Trained Raters

Each interview was reviewed and rated by at least three (usually four; approximately 85% of participants) different research assistants. Following employment interview best practices [7], the research assistants first underwent 1-2 hours of frame-of-reference training which included the following steps: defining the construct to be rated, reviewing the scale and scale anchors, completing practice ratings, and discussing sources of (dis)agreement with other raters. Since hireability is commonly assessed in interview studies [32], two 5-point Likert-scale items were used to rate hireability: "I would recommend that this person be hired" and "If hired, I believe this person would perform well on the job" [15]. The final hireability score was taken as the average over these two items across all raters for a given participant. An inter-rater reliability ICC(1,k) = 0.67 was computed using the one-way, random, average intra-class correlation coefficient, which is considered *moderate* agreement according to [33]. Ratings of other traits were collected (not discussed in this work) and used to provide evidence of convergent validity of the hireability ratings. The hireability scores correlated positively with general mental ($r = .15$) and verbal ($r = .22$) ability test scores [13], which are among the most valid predictors of job performance across occupations [48].

3.3 Feature Extraction

A set of features was extracted from each video's visual and audio channels capturing verbal, paraverbal, and visual behaviors.

Verbal: Verbal features included *n*-gram (unigram, bigram and trigram) frequencies and Linguistic Inquiry and Word Count (LIWC) summary categories [42]. All *n*-gram features were term frequency-inverse document frequency (TF-IDF) weighted, and the bigrams and trigrams with a point-wise mutual information (PMI) less than 4.0 were dropped to remove spurious *n*-grams (per [41]).

Paraverbal: The Geneva Minimalistic Acoustic Parameter Set of features were extracted using *OpenSmile* [18]. These features included loudness, Mel-frequency cepstral coefficients, jitter and shimmer.

Visual: Visual features were extracted from facial expressions and body motion. Emotient's *FACET* was used to extract facial expression features from individual video frames where a face could be detected, which included the likelihood estimates for 20 facial action units and the size (area) of the face. Estimates of facial expressivity, positive valence and negative valence were also computed based on the facial action unit activation [12]. Additionally, face and upper body motion was estimated using the *Motion Tracker* software [52].

One set of features per participant across the entire mock interview was generated as follows. LIWC and *n*-gram features were extracted per participant across their six videos combined. For all other features, a set of statistical functionals was independently applied to each feature separately in each of the six videos, including median, standard deviation, minimum, maximum and range, and then averaged across each participant's recordings. In total, there were approximately 5653 *verbal* features after PMI filtering (depending on the words uttered), 125 *paraverbal* features, and 250 *visual* features.

3.4 Matching on Gender

We balanced the data across genders using a matching algorithm [49] to minimize source data discrepancies in gender representation and hireability. While balancing the data may have reduced the severity of gender bias, we note that it was still present in this case study and thus enabled us to focus the analysis on manifestations of gender bias introduced by the machine learning process itself. As part of the study's procedure, participants were asked to provide their gender and 727 self-affiliated as either a man (n=262) or woman (n=465). Since the non-binary gender (n=6) representation was insufficient for statistical analysis, we excluded these six individuals and matched the data on men and women. The matching method ensured that an equal number of men and women were present in the data set by down-sampling the majority class (women) to match the minority class (men). The *designmatch* package in R [55] was used to perform bipartite cardinality matching with the constraint that the mean gender difference in interview-rated hireability was no larger than $\frac{1}{20}$ of a standard deviation. After matching, 262 men and 262 women (524 total) remained.

We performed a non-parametric Fligner-Killeen test for whether the variances were equivalent across genders and found the difference to be insignificant ($\chi^2 = 0.029$, $p = 0.87$). Additionally, we performed a Kolomogorov-Smirnoff test to assess the likelihood that the data was drawn from different distributions and found the distribution difference to be insignificant ($D = 0.097$, $p = 0.76$), suggesting successful matching. Distributions of hireability scores for the matched samples of men and women are illustrated in Figure 2.

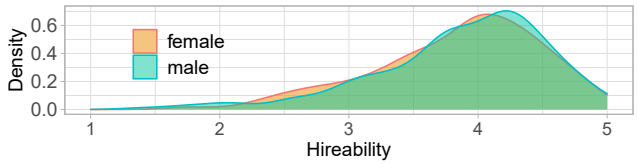

Figure 2: Distributions of hireability scores across genders

3.5 Machine Modeling (Learning)

We constrain our analysis to interpretable learning algorithms since our focus is on measuring bias and fairness with regards to legal contexts where layperson explanations of outcomes may be necessary. In our early tests, we employed elastic net and random forest (RF) regressors and found the RF achieved higher accuracy, so we selected this model as a baseline for our remaining analyses.

We separately trained three types of RF models, described as follows, on the *verbal*, *paraverbal*, and *visual* unimodal feature sets and also on their multimodal combination:

Baseline Model: All features from the chosen modality were used and were z-normalized across all participants prior to training.

Gender-normed Model: Features were z-normalized separately across men and women before training.

Reduced Features Model: A subset of the features were used for modeling, obtained via an iterative feature elimination procedure aimed at minimizing the predictability of gender. Specifically, for each modality set, we trained a model in the same manner as the *baseline* to predict *gender* (instead of hireability) using all features z-normalized across all participants. In each iteration, the 10 features with the greatest importance (i.e. feature "weights") were removed and another RF was trained in the same manner on the remaining features. A single pre-made train/test partition was used in each iteration where an equal number of men and women were present in each subset (further details in Section 3.6). In each iteration, we evaluated gender predictability using the area under the curve (AUROC) measure, and the process continued until no features remained. For each modality, a *reduced features* model was selected from the models produced at each iteration corresponding to the one with an AUROC closest to 0.5. In other words, the *reduced features* model was unique for each modality and had approximately chance-level ability to predict gender. After feature reduction, there were approximately 3760 *verbal* (34% reduction), 8 *paraverbal* (94% reduction), and 0 *visual* (100% reduction) features, an overall drop from approximately 6028 in the *combined* feature set to about 3768.

3.6 Model Training and Tuning

Premade train/test splits: To facilitate fair comparison between models, a predefined data partition was obtained for training and testing by splitting the data into five folds. Stratified sampling was used to generate five non-overlapping subsets of samples such that the distributions of hireability scores were approximately equal per subset. Since each sample corresponded to data from a single participant, this scheme resulted in a participant-independent partition. During model training, one fold at a time was held out and the remaining four used for training. Predictions were made by the trained models on the held out data and later recombined to produce the final hireability predictions for all participants.

Hyperparameter tuning: We optimized the hyperparameters in the *baseline* model separately for each modality using nested stratified five-fold cross-validation. For each training data set (from the train/test splits), the data was partitioned using stratified sampling in the same manner as described above. One subset was withheld at a time for hyperparameter validation (i.e., the inner loop) and the hyperparameter configuration with the best average performance was selected for model retraining on the entire training set to make test-set predictions. For the *paraverbal* and *visual* modalities, we tested different RF parameterizations: number of decision trees ($\{250, 500, 800\}$) and maximum depth ($\{10, 50, 80\}$). The *verbal* and *combined* modalities included additional transcript-based features, and we also tested each combination of these parameterizations: minimum document frequency ($\{0.01, 0.02, 0.03\}$) and stop words removal (none, English). The performance of the *baseline* model for *verbal* and *combined* modalities was achieved with 500 trees, max depth of 50, document frequency of 0.03 and no stop words. This combination of parameters performed near optimally as well for *paraverbal* and *visual* modalities, so to enable fair comparison of the results, we selected this set of hyperparameters to use for all models.

3.7 Evaluation Measures and Metrics

Robustness: For each set of modalities and each model variant (i.e., baseline, gender-normed, reduced features), one hundred independent trials were conducted to assess the variability in performance of the optimal tuned models. For each trial, a new five-fold train/test split was generated from a random seed, and a new RF model was trained using the best hyperparameters from the previous step. Differences in the train/test splits and in the randomized subsets of features and instances made available to each decision tree in the RF models produced variance in the output predictions. These sources of randomness allowed us to estimate the reliability of the evaluation metrics (below) relative to these sources of randomness and were more robust than considering a single instantiation.

Metrics: The predictions output by the three models for each of the four sets of modalities were evaluated in terms of accuracy, bias, and fairness. Assuming that hiring or employment screening decisions are made based on the relative rankings of participants rather than the hireability scores themselves, we used Spearman rank-based correlation (ρ) as an accuracy metric. To assess gender bias, we use correlational accuracy, or the difference in ρ between men and women, per [49]. We assessed fairness using two metrics: the adverse impact ratio (one distributive fairness metric) and the predictability of gender (a type of construct contamination bias related to procedural fairness). The AI ratio is defined as the ratio of selection ratios between two groups (in this case, men and women) where the selection ratio (SR) is the number of group members selected divided by the total number of applicants from that group. For example, let us assume 10 people are selected, 3 men and 7 women, and that there are 50 men and 50 women who applied (100 total), then $\text{SR}_{\text{men}} = \frac{3}{50}$, while $\text{SR}_{\text{women}} = \frac{7}{50}$, and the AI ratio $= \frac{3/50}{7/50} = 0.43$ (smaller value in the numerator), which would be considered quite unfair. We also measured the predictability of gender for each modality by training a RF (with the same hyperparameters as above) to predict the gender of each participant (on the held-out test sets) and then compared the output to the true gender using AUROC as the metric of interest. With these formulations, unbiased and fair models would have correlation differences close to zero, AI close to one, and AUROCs close to 0.5. Accurate models would have high Spearman correlations, with a correlation of 0.3 reflecting a medium-sized effect (Cohen's d of 0.5) [11].

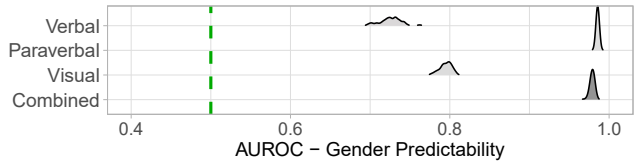

Figure 3: Distributions of gender prediction AUROCs in the *baseline* model across modalities. The green dashed line indicates the ideal value of 0.5 (gender blindness).

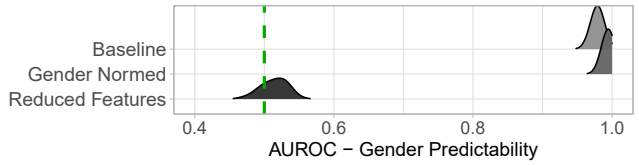

Figure 4: Distributions of gender prediction AUROCs for each model variant using the combined modality. The green dashed line indicates the ideal value of 0.5.

4 RESULTS

Table 1 shows the means and standard deviations for each metric over 100 random trials.

4.1 Gender Predictability (Procedural Fairness)

Figure 3 illustrates the gender prediction AUROC distributions for each modality across the 100 trials using the *baseline* models. The optimal AUROC is 0.5 (indicated by the dotted green line) which corresponds to the inability of the model to predict gender (i.e., gender blindness). In our matched data set, there were an equal number of men and women so the prior probability of guessing a participant's gender was 0.5, which was an ideal starting point for analyzing gender predictability.

The results indicate that all modalities contributed gender information to the *baseline* models. The *verbal baseline* and *visual baseline* models had the lowest AUROCs (0.73 and 0.79 respectively), indicating they had incomplete knowledge of gender. The baseline model's *paraverbal* and *combined* modalities were able to infer gender nearly perfectly (unsurprisingly due to the presence of vocal pitch information).

Figure 4 demonstrates how the two employed gender bias mitigation strategies affected the predictability of gender in 100 trials using the *combined* modality. Gender predictability for the *combined baseline* and *combined gender-normed* models were well above the ideal AUROC=0.5 (green dashed line), suggesting that gender-norming predictors via z-scoring did not effectively reduce the amount of gender information in a fully multi-modal setting. For the unimodal *gender-normed* models in Table 1, gender-norming did substantially reduce gender predictability for the *visual* and *paraverbal* modalities (from 0.79 to 0.58 and 0.99 to 0.69 respectively). Gender-norming was ineffective for the *verbal* modality (and by extension the *combined* modality), which is consistent with known language usage differences between men and women [30, 34]. The *reduced features* model achieved an AUROC within a few standard deviations of the ideal 0.5 for all modalities, suggesting it was successful at achieving gender blindness.

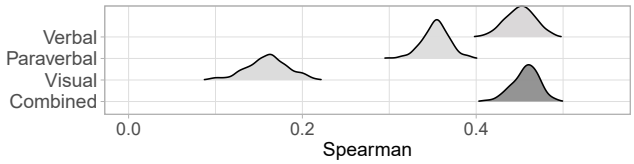

Figure 5: Distributions of ρ achieved by the *baseline* model using only the features available from each modality.

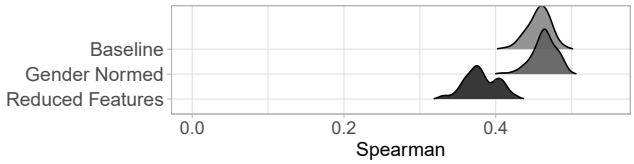

Figure 6: Distributions of ρ per model variant using the combined modality.

4.2 Accuracy (Validity)

Figure 5 illustrates the distributions of ρ over 100 trials for each modality set in the *baseline* models. Across all modalities and model variants, the standard deviations across trials was small (between 0.02 and 0.03), which indicates that each model's ρ was robust to randomness in how the train/test data was partitioned and the RF training process.

Looking first at the unimodal cases, the *paraverbal* and *visual* modalities under-performed compared to the *verbal* modality. The *verbal* feature set alone achieved $\rho = 0.45$, indicating moderate accuracy (ranked alignment between predictions and ground truth). The *combined* feature set performed negligibly better than the *verbal* modality ($\rho = 0.46$).

Figure 6 shows the performance distributions of the model variants using the *combined* modality since it is the most interesting from a multimodal perspective. We found the *combined gender-normed* model performed nearly identically to the *combined baseline* model. As expected, the *combined reduced-features* model obtained a lower ρ (also for each modality; see Table 1), ostensibly because some of the features which contained information pertinent to assessing hireability also helped predict gender and were removed. The *combined reduced-features* model's drop in ρ from the combined baseline model is significant ($p < 0.01$, using paired correlated-sample t-tests), but this model still achieved a notable $\rho = 0.38$.

4.3 Differential Correlational Accuracy (Bias)

Figure 7 shows the distributions of differences in ρ computed separately for men and women. If no bias sources were present in the data or modeling processes, then the model would be equally accurate at assessing men and women (i.e., ρ would be the same for men and women and the difference would be zero, indicated by the green dashed vertical line). In our matched data set, we observed the gender difference in mean ground truth hireability scores was -0.03, which was small in comparison to the hireability scale (1-5) and provided an ideal lower bound.

In the unimodal cases, we found upwards of 95% of the 100 trials for the *paraverbal baseline* and 100% of the trials for the *visual baseline* resulted in positive ρ differences, indicating that these modalities were consistently more accurate for women than men. The mean bias for the baseline's *verbal* modality was positive, small (.04), and lower than paraverbal (.07), which in turn, was much lower than

Table 1: Accuracy, bias, and fairness metrics per modality applied to the baseline, gender-normed, and reduced features models

Modalities	Spearman (W and M)			Gender AUROC			Spearman Diff. (W-M)			AI Ratio		
	Baseline	Gender Normed	Reduced Features	Baseline	Gender Normed	Reduced Features	Baseline	Gender Normed	Reduced Features	Baseline	Gender Normed	Reduced Features
Verbal	.45(.02)	.46(.02)	.36(.03)	.73(.01)	.99(.00)	.51(.02)	.04(.03)	.05(.03)	.01(.05)	.87(.10)	.88(.10)	.85(.11)
Paraverbal	.35(.02)	.35(.02)	.11(.03)	.99(.00)	.69(.10)	.45(.02)	.07(.03)	.10(.03)	.05(.04)	.87(.10)	.88(.09)	.85(.10)
Visual	.16(.02)	.16(.02)	.07(.02)	.79(.01)	.58(.05)	.45(.02)	.15(.05)	.16(.04)	.15(.04)	.52(.11)	.60(.12)	.80(.13)
Combined	.46(.02)	.46(.02)	.38(.02)	.98(.00)	.99(.00)	.51(.02)	.09(.03)	.10(.03)	.00(.04)	.76(.11)	.81(.12)	.87(.11)

Each entry contains the mean value across 100 independent trials followed by the standard deviation in parentheses. W = women, M = men.

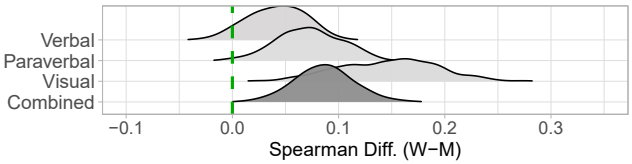

Figure 7: Distributions of bias ($\rho_{\text{women}} - \rho_{\text{men}}$) in the *baseline* model across modalities. The green dashed line indicates the ideal value of zero.

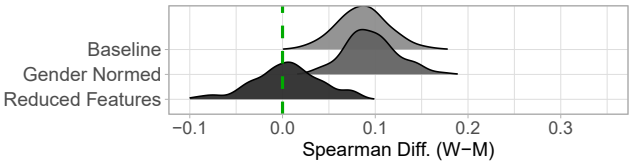

Figure 8: Distributions of bias ($\rho_{\text{women}} - \rho_{\text{men}}$) in the *baseline* model for each model variant using the combined modality. The green dashed line indicates the ideal value of zero.

visual (.15). In the *combined baseline*, 100% of the trials resulted in ρ differences greater than zero, with a magnitude in-between the three individual modalities (.04, .07, .15). In other words, for all baseline models except for the *verbal* modality, the sources of bias (whatever they may have been) were having notable differential effects on the model performance as a function of gender.

Figure 8 shows how the gender difference in ρ varied for different model variants using the combined modality (again, it is the multimodal model of interest here). Just as the *combined baseline* model, the ρ differences were positive in all of the trials for the *combined gender-normed* model (.09 and .10), also indicating the presence of upstream bias sources. The *combined reduced-features* model, however, achieved a mean ρ difference of zero, which suggests that removing the features containing gender information also substantially reduced the bias. The results in Table 1 for the *paraverbal* and *verbal* modalities also demonstrated that the *gender-normed* model did not decrease bias (in fact it increased it slightly) while the *reduced-features* model reduced bias, moderately for *paraverbal* (.07 to .05), and almost completely (.04 to .01) for the *verbal* model. No notable changes in the bias measure were apparent for the *visual* modality across model variants, which suggests that the sources of gender bias for *visual* predictors were independent of gender (and may have been due to the relatively low accuracy: $\rho = 0.16$).

4.4 Adverse Impact (Distributive Fairness)

Figure 9 shows the AI ratio distributions for each modality in the *baseline* model under the assumption that the top 10% of participants were selected. In practice the quantity of participants selected would vary

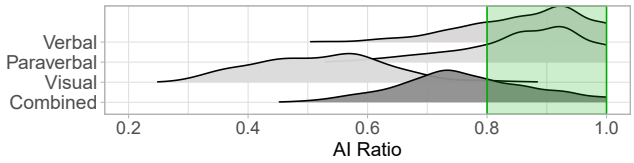

Figure 9: Distributions of adverse impact ratios in the *baseline* model across modalities. The green region is legally acceptable under the four-fifths rule.

based on organizational needs and resources, but 10% provided an illustrative snapshot of AI fairness. The optimal AI ratio is 1.0, meaning an equal proportion of men and women would be selected. Per the four-fifths rule [10], any AI ratio greater than 0.8 is considered fair by legal precedent in the US court system. If selected the top 10% of candidates with the highest ground truth hireability scores, we would observe an AI ratio of 1.0, a suitable baseline for fairness analysis.

In a majority of the random trials for *verbal baseline* (70%) and *paraverbal baseline* (72%), the AI ratio (mean AI ratio of .87 in both cases) fell within this legally acceptable range. Interestingly, the *visual baseline* fell short (mean AI ratio of .52) in 99% of the trials, which would be considered clear legal evidence of discrimination in the US. For the *combined baseline*, a majority of the trials' AI ratios (73%; mean AI = 0.76) were below the acceptable 0.8 threshold, which might also be considered legal evidence of discrimination. Defending this model in court would require evidence that the model predictions were valid/accurate (e.g., $\rho = 0.46$) and that all of the features and the target construct (e.g., hireability) used to aid in hiring decisions were job-relevant [1].

Figure 10 shows how the bias mitigation strategies impact this measure of distributive fairness when using the combined modality. Gender normalization of the features did improve the mean AI ratio (in spite of its lack of effect on bias), while the *combined reduced-features* approach improved it even further (.76 to .81 to .87, respectively). From Table 1 for the *visual* modality, the *gender-norming* model improved fairness slightly (AI of 0.60 compared to 0.52) and the *reduced-features* model helped considerably (AI up to 0.80). The AI ratio for the baseline *verbal* and *paraverbal* modality models started in an acceptable range (> 0.8) and there were no noticeable changes for the other model variants. Thus, both mitigation strategies helped the AI ratio as needed.

5 DISCUSSION

This paper has introduced the psychometric concepts of measurement bias and fairness to multimodal machine learning and examined how different sets of modalities and gender bias mitigation strategies affect measures of accuracy, bias, and fairness.

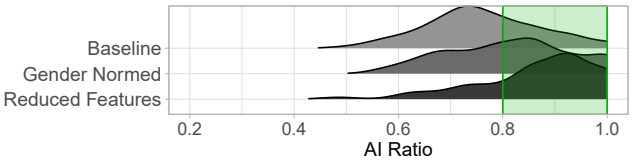

Figure 10: Distributions of adverse impact ratios for each model variant using the combined modality. The green region is legally acceptable under the four-fifths rule.

5.1 Summary of Main Findings

We demonstrated clear variation in our bias and fairness metrics based on the unimodal or multimodal predictor sets used (**RQ1**). Our results suggest that in the absence of bias mitigation, combining modalities barely improves prediction accuracy (accordant with results from a similar AVI study [8]) but often leads to increased correlational bias, group predictability, and adverse impact when compared to the least biased and most fair unimodal models (*verbal* in this case).

We have further shown that predictor bias mitigation strategies can likewise influence bias and fairness at the cost of accuracy (**RQ2a**). In particular, we showed that while predictor gender-norming seems like a plausible strategy for reducing gender predictability, it only works for certain modalities (i.e. *paraverbal* and *visual*) and it does not quite reduce it enough to reach true gender blindness (AUROCs ranged from .58 to .99). As mentioned earlier, this technique was outlawed in the US due in part to the fact that it contaminates the data with irrelevant demographic information in an attempt to improve procedural fairness. When combined with its negligible effect across all modalities on improving correlational bias and adverse impact, we do not suggest using this approach.

The *reduced features* bias mitigation approach, on the other hand, provides some benefit across modalities but further exemplifies the bias-accuracy trade-off (**RQ2b**). In terms of fairness, we found this approach successfully improved adverse impact (or avoided reducing what was already acceptable) and also achieved gender blindness across all modalities when compared to the baseline. It was additionally able to generally reduce the correlational bias (except for *visual*) and eliminate it completely for the *combined* model. However, these bias and fairness improvements were offset by its reduction in accuracy in all cases, though its accuracy for the combined model (.38) might still be acceptable.

So, which model and which set of modalities are the best given the scope of this AVI case study? Top contenders in terms of accuracy are the *baseline* and *gender-normed* models using either the *verbal* or *combined* modalities. As mentioned, the *gender-normed* version does not achieve gender blindness for either modality, nor does the *combined baseline* version. This leaves the unimodal *verbal baseline* which also exhibits diminished gender predictability (AUROC=0.73), moderate correlational accuracy ($\rho = .45$), small correlational bias ($\rho_{\text{diff}} = 0.04$), and acceptable adverse impact (AI=0.87). In some contexts, it may be desirable to trade-off accuracy to improve bias and enhance fairness. In this case, the *combined reduced-features* model provides the next best accuracy ($\rho = 0.38$), better gender blindness (AUROC=0.51), no correlational bias ($\rho_{\text{diff}} = 0.00$), and equivalently acceptable adverse impact (AI=0.87). The *verbal reduced-features* model achieves a similar accuracy ($\rho = .36$), gender blindness (AUROC=.51), low bias ($\rho_{\text{diff}} = 0.01$), acceptable adverse impact (AI=.85), and it avoids the risk of including features with irrelevant information when making hireability assessments (such as visual ones) when compared to the *combined reduced-features* model. Based on the metrics here, we would consider the *reduced-features* approach with either the *verbal* or *combined* modalities to be better choices than the *verbal* or *combined* baseline model.

5.2 Limitations and Future Work

One important point worth highlighting is related to the amount of variance we observed for each measure of accuracy, bias, and fairness. The standard deviations computed over 100 trials in Table 1 assess the reliability of each measure relative to randomness *only* in the training and modeling processes, where many other factors were held constant (e.g., the data set, learning algorithm, hyperparameters, predictors). These standard deviations therefore serve as lower bounds on the amount of variance that we would expect to observe in other scenarios or other studies where these factors may vary. Fortunately, the standard deviations for many of the accuracy, gender predictability, and correlational bias metrics reported were relatively low, which gives us some confidence in the reliability of our findings. In contrast, the standard deviations for the AI ratios were higher (approximately 0.10), revealing that this measure is highly sensitive to experimental perturbations (even with a 10% selection ratio). Future work may seek to find ways of reducing this sensitivity to help improve distributive fairness in more of the randomized trials.

We have only begun to scratch the surface of understanding manifestations of bias according to the choice of modalities and according to bias mitigation strategies. There are many ways of measuring bias and fairness [29, 36, 50], and many more protected groups (e.g., age, gender, religion) that would need to be considered. Our analysis only considered bias and fairness effects relative to a random forest model on matched data, where bias and fairness effects may be different and perhaps worse when using black-box models or unmatched data.

Finally, our study also involved low-stakes mock interviews and the use of trained raters rather than expert raters. Though we aimed to increase ecological validity by emulating real-world conditions, extension to more ecologically valid scenarios is warranted.

6 CONCLUSION

As scientists interested in enabling machines to process and understand humans and human-produced data, it is tempting to throw as much data as possible into machine inference systems to enable them to discover unintuitive predictor relationships and to help teach us more about ourselves. This is the auspicious prospect of the "big data" revolution. This type of scientific inquiry can indeed help us uncover some unique associations that explain human behavior, but it can also be the cause of lasting harm to certain groups of people, especially when these systems are deployed in high-stakes decision-making scenarios. Our results from the AVI case study echo this point: predictors should be justifiably relevant to the target construct or else irrelevant patterns may be discovered in the noise, introducing systemic bias and causing social harms.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (IIS 1921087 and IIS 1921111) and the NSF National AI Institute for Student-AI Teaming (iSAT) (DRL 2019805). The opinions expressed are those of the authors and do not represent views of the NSF.

REFERENCES

[1] 1964. *Title VII of the Civil Rights Act of 1964.* https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964

[2] American Educational Research Association, American Psychological Association, Joint Committee on Standards for Educational, Psychological Testing (US), National Council on Measurement in Education, et al. 2014. *Standards for educational and psychological testing.* American Educational Research Association.

[3] Fernando Ávila, Kelly Hannah-Moffat, and Paula Maurutto. 2020. The seductiveness of fairness: Is machine learning the answer?–Algorithmic fairness in criminal justice systems. In *The Algorithmic Society*. Routledge, 87–103.

[4] Murray R Barrick and Michael K Mount. 1991. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology* 44, 1 (1991), 1–26.

[5] Joan-Isaac Biel, Oya Aran, and Daniel Gatica-Perez. 2011. You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5.

[6] John F Binning and Gerald V Barrett. 1989. Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology* 74, 3 (1989), 478.

[7] Michael A Campion, David K Palmer, and James E Campion. 1997. A review of structure in the selection interview. *Personnel psychology* 50, 3 (1997), 655–702.

[8] Lei Chen, Gary Feng, Michelle P Martin-Raugh, Chee Wee Leong, Christopher Kitchen, Su-Youn Yoon, Blair Lehman, Harrison Kell, and Chong Min Lee. 2016. Automatic Scoring of Monologue Video Interviews Using Multimodal Cues.. In *INTERSPEECH*. 32–36.

[9] Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. 2017. Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 504–509.

[10] Department of Justice Civil Service Commission, Department of Labor. 1978. *Uniform Guidelines on Employee Selection Procedures.* https://www.govinfo.gov/content/pkg/CFR-2017-title29-vol4/xml/CFR-2017-title29-vol4-part1607.xml

[11] Jacob Cohen. 1992. A power primer. *Psychological bulletin* 112, 1 (1992), 155.

[12] Jeffrey F. Cohn, Laszlo A. Jeni, Itir Onal Ertugrul, Donald Malone, Michael S. Okun, David Borton, and Wayne K. Goodman. 2018. Automated Affect Detection in Deep Brain Stimulation for Obsessive-Compulsive Disorder: A Pilot Study. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 40–44. https://doi.org/10.1145/3242969.3243023

[13] David M Condon and William Revelle. 2014. The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence* 43 (2014), 52–64.

[14] Vedant Das Swain, Koustuv Saha, Hemang Rajvanshy, Anusha Sirigiri, Julie M Gregg, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, Shayan Mirjafari, Raghu Mulukutla, et al. 2019. A multisensor person-centered approach to understand the role of daily activities in job performance with organizational personas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–27.

[15] Wendy S Dunn, Michael K Mount, Murray R Barrick, and Deniz S Ones. 1995. Relative importance of personality and general mental ability in managers' judgments of applicant qualifications. *Journal of Applied Psychology* 80, 4 (1995), 500.

[16] Alice H Eagly and Wendy Wood. 1999. The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American psychologist* 54, 6 (1999), 408.

[17] Oscar Espinoza. 2007. Solving the equity–equality conceptual dilemma: A new model for analysis of the educational process. *Educational Research* 49, 4 (2007), 343–363.

[18] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. Association for Computing Machinery, New York, NY, USA, 1459–1462. https://doi.org/10.1145/1873951.1874246

[19] Tiantian Feng, Brandon M Booth, Brooke Baldwin-Rodríguez, Felipe Osorno, and Shrikanth Narayanan. 2021. A multimodal analysis of physical activity, sleep, and work shift in nurses with wearable sensor data. *Scientific reports* 11, 1 (2021), 1–12.

[20] Linda S Gottfredson. 1994. The science and politics of race-norming. *American Psychologist* 49, 11 (1994), 955.

[21] Boukje Habets, Sotaro Kita, Zeshu Shao, Asli Özyurek, and Peter Hagoort. 2011. The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of cognitive neuroscience* 23, 8 (2011), 1845–1854.

[22] Drew Harwell. 2019. A face-scanning algorithm increasingly decides whether you deserve the job. *The Washington Post* (2019).

[23] Louis Hickman, Nigel Bosch, Vincent Ng, Rachel Saef, Louis Tay, and Sang Eun Woo. 2021. Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology* (2021).

[24] Louis Hickman, Louis Tay, Sang E Woo, and Sidney D'Mello. 2021. An empirical test of machine learning measurement bias mitigation strategies. In M. Liu & L. Hickman (Chairs), Machine Learning for I-O 3.0. Symposium conducted at the 2021 Annual Conference of the Society for Industrial and Organizational Psychology.

[25] Judith Holler and Stephen C Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences* 23, 8 (2019), 639–652.

[26] Leatta M Hough, Frederick L Oswald, and Robert E Ployhart. 2001. Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment* 9, 1-2 (2001), 152–194.

[27] Allen I Huffcutt. 2011. An empirical review of the employment interview construct literature. *International Journal of Selection and Assessment* 19, 1 (2011), 62–81.

[28] John E Hunter and Hannah Rothstein Hirsh. 1987. Applications of meta-analysis. (1987).

[29] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 49–58.

[30] Priyanka D Joshi, Cheryl J Wakslak, Gil Appel, and Laura Huang. 2020. Gender differences in communicative abstraction. *Journal of personality and social psychology* 118, 3 (2020), 417.

[31] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *Aea papers and proceedings*, Vol. 108. 22–27.

[32] Donald H Kluemper, Benjamin D McLarty, Terrence R Bishop, and Anindita Sen. 2015. Interviewee selection test and evaluator assessments of general mental ability, emotional intelligence and extraversion: Relationships with structured behavioral and situational interview performance. *Journal of Business and Psychology* 30, 3 (2015), 543–563.

[33] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* 15, 2 (2016), 155–163.

[34] Campbell Leaper and Rachael D Robnett. 2011. Women are more likely than men to use tentative language, aren't they? A meta-analysis testing for gender differences and moderators. *Psychology of Women Quarterly* 35, 1 (2011), 129–142.

[35] Stephen C Levinson. 2016. Turn-taking in human communication–origins and implications for language processing. *Trends in cognitive sciences* 20, 1 (2016), 6–14.

[36] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[37] Lorenza Mondada. 2018. Multiple temporalities of language and body in interaction: Challenges for transcribing multimodality. *Research on Language and Social Interaction* 51, 1 (2018), 85–106.

[38] Skanda Muralidhar, Laurent Son Nguyen, Denise Frauendorfer, Jean-Marc Odobez, Marianne Schmid Mast, and Daniel Gatica-Perez. 2016. Training on the job: Behavioral analysis of job interviews in hospitality. In *Proceedings of the 18th acm international conference on multimodal interaction*. 84–91.

[39] Iftekhar Naim, M Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2015. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, Vol. 1. IEEE, 1–6.

[40] Laurent Son Nguyen, Denise Frauendorfer, Marianne Schmid Mast, and Daniel Gatica-Perez. 2014. Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE transactions on multimedia* 16, 4 (2014), 1018–1031.

[41] Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology* 108, 6 (2015), 934.

[42] James Pennebaker, Martha Francis, and Roger Booth. 1999. Linguistic inquiry and word count (LIWC). (01 1999).

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

[44] Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. 2016. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European conference on computer vision*. Springer, 400–418.

[45] Sowmya Rasipuram and Dinesh Babu Jayagopi. 2016. Asynchronous video interviews vs. face-to-face interviews for communication skill measurement: a systematic study. In *Proceedings of the 18th ACM international conference on multimodal interaction*. 370–377.

[46] Vinesh Ravuri, Projna Paromita, Karel Mundnich, Amrutha Nadarajan, Brandon M Booth, Shrikanth S Narayanan, and Theodora Chaspari. 2020. Investigating Group-Specific Models of Hospital Workers' Well-Being: Implications for Algorithmic Bias. *International Journal of Semantic Computing* 14, 04 (2020), 477–499.

[47] Paul R Sackett and Steffanie L Wilk. 1994. Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist* 49, 11 (1994), 929.

[48] Frank L Schmidt and John E Hunter. 1998. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin* 124, 2 (1998), 262.

[49] Louis Tay, Sang E Woo, Louis Hickman, Brandon M Booth, and Sidney D'Mello. 2021. A conceptual framework for investigating and mitigating machine learning bias for psychological assessment. *Manuscript submitted for publication* (2021).

[50] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 1–7.

[51] John D Wasserman and Bruce A Bracken. 2012. Fundamental psychometric considerations in assessment. *Handbook of Psychology, Second Edition* 10 (2012).

[52] Jacqueline Westlund, Sidney D'Mello, and Andrew Olney. 2015. Motion Tracker: Camera-Based Monitoring of Bodily Movements Using Motion Silhouettes. *PloS one* 10 (06 2015), e0130293. https://doi.org/10.1371/journal.pone.0130293

[53] Shen Yan, Di Huang, and Mohammad Soleymani. 2020. Mitigating Biases in Multimodal Personality Assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 361–369.

[54] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. 2020. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 18–19.

[55] Jose R. Zubizarreta, Cinar Kilcioglu, and Juan P. Vielma. 2018. designmatch: Matched Samples that are Balanced and Representative by Design. https://CRAN.R-project.org/package=designmatch R package version 0.3.1.