"It feels like we're not meeting the criteria": Examining and Mitigating the Cascading Effects of Bias in Automatic Speech Recognition in Spoken Language Interfaces.

Kelechi Ezema
Institute of Cognitive Science
University of Colorado Boulder
Boulder, Colorado, USA
kelechi.ezema@colorado.edu

Chelsea Chandler
Institute of Cognitive Science
University of Colorado Boulder
Boulder, Colorado, USA
chelsea.chandler@colorado.edu

Rosy Southwell
Institute of Cognitive Science
University of Colorado Boulder
Boulder, Colorado, USA
rosy.southwell@colorado.edu

Niranjan Cholendiran
Institute of Cognitive Science
University of Colorado Boulder
Boulder, Colorado, USA
niranjan.cholendiran@colorado.edu

Sidney D'Mello
Institute of Cognitive Science
University of Colorado Boulder
Boulder, Colorado, USA
sidney.dmello@colorado.edu

Abstract

Researchers have demonstrated that Automatic Speech Recognition (ASR) systems perform differently across demographic groups (i.e. show bias), yet their downstream impact on spoken language interfaces remains unexplored. We examined this question in the context of a real-world AI-powered interface that provides tutors with feedback on the quality of their discourse. We found that the Whisper ASR had lower accuracy for Black vs. white tutors, likely due to differences in acoustic patterns of speech. The downstream automated discourse classifiers of tutor talk were correspondingly less accurate for Black tutors when presented with ASR input. As a result, although Black tutors demonstrated higher-quality discourse on human transcripts, this trend was not evident on ASR transcripts. We experimented with methods to reduce ASR bias, finding that fine-tuning the ASR on Black speech reduced, but did not eliminate, ASR bias and its downstream effects. We discuss implications for AI-based spoken language interfaces aimed at providing unbiased assessments to improve performance outcomes.

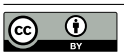
CCS Concepts

• **Computing methodologies** → Machine learning.

Keywords

automatic speech recognition, fairness, teacher discourse, racial bias

ACM Reference Format:
Kelechi Ezema, Chelsea Chandler, Rosy Southwell, Niranjan Cholendiran, and Sidney D'Mello. 2025. "It feels like we're not meeting the criteria": Examining and Mitigating the Cascading Effects of Bias in Automatic Speech Recognition in Spoken Language Interfaces.. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3706598.3714059



ACM ISBN 979-8-4007-1394-1/2025/04
https://doi.org/10.1145/3706598.3714059

1 Introduction

Educators' discourse significantly impacts student learning outcomes in the classroom [112]. For example, frameworks such as academically productive (or Accountable) talk [68] emphasize high-impact discourse moves that promote rigorous thinking and building a learning community, such as encouraging students to explain their responses or relating the content to their everyday experience. Unfortunately, a significant amount of expertise is required for teachers to successfully implement these advanced pedagogical approaches. When left to their own devices, teachers tend to default to less effective practices, such as lecturing or superficially engaging students with closed-ended questions (e.g., "yes/no", "do you understand?") rather than open-ended questions that inspire dialogic thinking [74].

High-quality professional learning can help teachers develop their teaching skills [9, 79]. Education reformers have advocated for a radical transformation in teacher professional development from one-off conference-style presentations to more job-embedded approaches [45], emphasizing the importance of feedback through classroom observation. Such observations can provide teachers with a clear and deep understanding of their performance and progress [45, 47]. Unfortunately, scaling human classroom observation is challenging due to high costs and logistical complexities. As a result, teachers often receive infrequent feedback on their instructional practices, often less than once a year [53].

In response to these challenges, researchers have explored automated AI-based systems that could help scale classroom observations towards improving teacher practice with reduced cost. Examples include talk moves [100], Teacher Talk Tool [48], and the commercial start-up TeachFX (www.teachfx.com). These systems allow teachers to record audio files of their classroom discourse and receive feedback on various dimensions of practice such as the use of academically productive talk, types of questions, and student to tutor talk ratio [16, 23, 80].

Beyond classroom teachers, automated feedback systems have also been used to improve the quality of human tutoring, an increasingly popular approach to address pandemic-related learning losses [52, 98, 113]. For example, the Human-tutor Coaching Technology (HTCT) platform provides automatic feedback based on the quality of tutorial discourse during small group tutoring sessions [12]. Because a majority of the tutors in these programs are paraprofessionals with minimal pedagogical training, there is an even more urgent need for AI-based approaches to support their ongoing professional development [15].

With the renewed interest in AI driven by the advent of tools like ChatGPT, it is likely that such automated feedback systems will be increasingly used to scale professional learning in educational settings. It is thus imperative that the systems are sufficiently accurate given their potential real-world impact on student learning outcomes [19, 95]. These systems typically use automatic speech recognition (ASR) to convert audio of classroom or tutorial recordings to text, followed by natural language processing (NLP) tools (e.g., large language models [LLMs]) that classify the text into different dimensions of pedagogical practice [49], which are visualized to provide actionable feedback (Figure 1). Providing accurate feedback on the quality of practice is an important component that underlies the success of these systems.

One major challenge to providing accurate feedback is the quality of ASR since classrooms are noisy environments with multiparty chatter, background noise, and other disruptions [24]. To this end, researchers have been contending with methods to improve the quality of ASR in classroom contexts with varied success [10, 14, 96, 97]. However, in addition to overall accuracy, the differential performance of automated feedback systems across demographic groups (i.e., bias) remains unexplored. This is particularly problematic given that bias in the accuracy of commercial ASR systems has been reported with respect to factors such as race [17, 51, 73, 108], dialect [103], and gender [42, 102]. For example, a recent study found that commercial ASR systems made about twice as many errors when transcribing speech of African American speakers compared to white speakers [51]. As a result, African American and non-native English speakers have expressed dissatisfaction and mistrust in ASR technology and attribute these issues to frequent misunderstandings and interruptions in the systems' output [18, 36, 67, 108, 111].

These growing concerns about racial bias exhibited by ASR models are exacerbated by the long history of anti-Black racism in the US. Black teachers in the US make up just 6% of the entire teacher population [72], with reports highlighting a continuous decline over the years [105]. Studies have found that Black teachers sometimes feel undervalued and disrespected in their teaching profession [21], which often pushes them to leave the profession. In the words of a Black teacher, "It feels like we're coming up short. It feels like we're not meeting the criteria, and so, we exit the field altogether" [21]. Also, African American Vernacular English (AAVE) and its regional variations [39] used by some Black speakers has been considered non-standard and inappropriate for educational use [22]. Given the delicate positions of Black teachers, it is essential that deployed educational tools do not negatively impact these groups of teachers.

Our research explored these concerns by examining racial bias in ASR and downstream automated feedback systems used in educational settings. Our key contribution lies in exploring bias throughout the computational pipeline as shown in Figure 1. We started by examining racial bias in a popular open-source ASR engine applied to real-world educational discourse (Research Question [RQ] 1), a gap that has not been explored in previous research. Because ASR in these systems serves as an input for automated feedback rather than being an end in itself, we investigated the extent to which bias in ASR has cascading impacts on bias in downstream discourse classification (RQ 2). In addition, we explored two methods to mitigate bias and its effects (RQ 3).

We conducted our research as part of a research-practice partnership [5] with ANON, a non-profit provider of high-dosage tutoring focused on low-income school districts in the US. The ANON tutoring system includes an interface that provides automated feedback on the quality of the tutorial discourse for formative improvement (i.e., not for evaluative purposes). We utilized audio recordings of authentic small group tutoring sessions from Black and white tutors to address our research questions.

2 Background and Related Work

2.1 Defining and measuring bias

Bias in the context of Machine Learning (ML) models can be defined as systematic misrepresentations or errors that favor certain groups [28]. Researchers have proposed different approaches to measure bias and fairness in ML and AI [106], with the majority focusing on establishing equitable performance for all subgroups [32, 66, 104]. Empirically, ML bias manifests when a model produces different *scores* for individuals belonging to different subgroups (e.g., race, gender) despite having the same human-verified (ground-truth) scores [104]. However, ML bias would *not* occur if the ground-truth scores were indeed different, and the model essentially reproduces that difference. Another type of bias – accuracy bias – occurs when a model's *accuracy* is different for a given subgroup aside from differences in scores. Bias would also occur if the rates of favorable (or unfavorable) outcomes differed for protected classes (groups that are legally protected from discrimination) [32, 66]. Whereas it is commonly assumed that bias in ML is solely a function of the training data, Booth et al. [11] provide a theoretical framework of how ML bias can arise as a result of contamination in different stages of a ML pipeline. For example, bias could emerge when methods used to compute features used by the ML models are themselves biased or when the features subtly encode protected features like gender or race [11].

2.2 Racial bias in Automatic Speech Recognition (ASR)

Automatic speech recognition (ASR) systems convert spoken language into text [61]. Traditionally, these systems consisted of separate acoustic and language modeling components [54, 61]. In modern end-to-end ASR systems, both components are jointly trained using end-to-end learning from large volumes of transcribed audio [101].

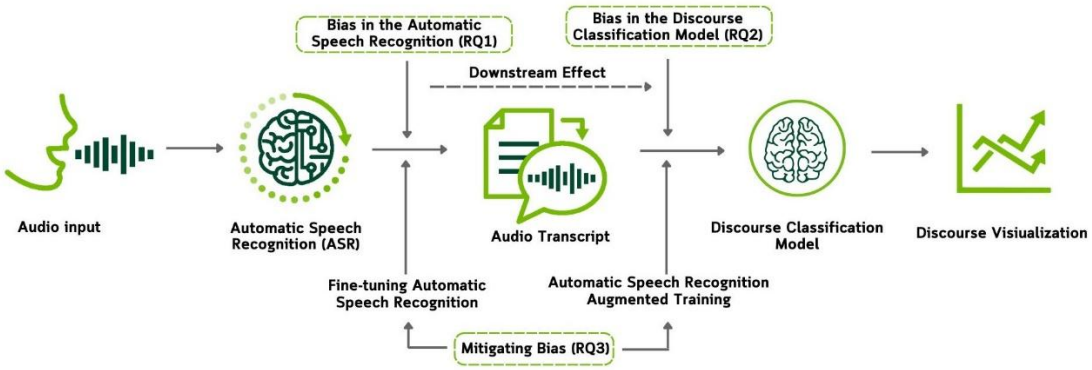


Figure 1: The automated feedback pipeline. The input is tutor speech (Audio Input), which is passed through an Automatic Speech Recognition (ASR) model, producing a text transcript (Audio Transcript). Next, the transcript is passed through a Discourse Classification Model, which classifies the texts into different discourse moves. Finally, the tutor's usage of each discourse move is presented as Discourse Visualizations. Our Research Questions (RQs) examine bias in the ASR (RQ1), its downstream effect on Discourse Classification (RQ2), and fine-tuning (in the case of the ASR system) and ASR augmented training (in the case of the classification model) as ways of mitigating bias (RQ3).

Several studies have examined racial bias in ASR, highlighting significant disparities in accuracy across various demographic factors [18, 30, 36, 51, 63, 73, 91]. Koenecke et al. [51] reported an average error rate for Black speakers (35%), nearly double that of white speakers (19%) in five commercial ASR systems. They found that acoustic factors were primarily responsible for the racial bias, as a significant performance gap persisted even when Black and white individuals spoke identical phrases [51]. A related kind of bias is accented language bias [37, 38, 58, 103], with studies indicating consistent underperformance on minority dialects and accents. For example, Slaughter et al. [91], investigated bias in the popular Whisper ASR and found that the model's embeddings exhibit bias based on race, gender, nationality, and physical disabilities. Martin and Tang [63] explored the habitual "be" of AAVE, a grammatical feature used to express regular actions or states (e.g., "She *be* tired after school," meaning "She is *usually* tired after school" [34]) and found that ASR systems struggle with these AAVE features.

Human-Computer Interaction (HCI) researchers have qualitatively analyzed the challenges faced by Black and African American speakers when interacting with commercial ASRs [13, 18, 36]. For example, Harrington et al. [36] highlighted the difficulties older Black speakers encountered in accessing health-related information via Google Home, requiring them to engage in a form of "cultural code switching." Likewise, Cunningham et al. [18] noted that African American speakers often engage in a form of "invisible labor" by adjusting their natural speech patterns to make the technology work. They attributed these challenges to a lack of inclusivity in the language model design and datasets, which fail to adequately represent the linguistic features of more diverse speakers [18].

Similarly, Feng et al. [27] highlight several reasons for bias in ASR, suggesting that a primary cause is the underrepresentation of minority groups in the training data. Other reasons could be transcription bias from human annotators, pronunciation variability within and across groups, or the quality of the recording equipment [27]. Thus, the most common approach to de-biasing ASR models involves diversifying the speech datasets [73, 93] or additional fine-tuning (i.e., adjusting model weights) using smaller datasets of minority speakers [31, 109] or synthetic data from a generative model [88], although whether these techniques suffice to eliminate bias and its harms remains an open question.

3 Novelty, Contribution, & Research Questions (RQ)

We investigate ASR bias in automatic feedback interfaces used in authentic educational settings, specifically high-impact tutoring sessions. While previous studies have independently explored bias in different components of conversational AI systems, their interconnection (Figure1 above) is yet to be examined. Our main **contribution** is to address this gap by instantiating aspects of Tay et al.'s [104] conceptual model of ML bias, which emphasizes the importance of examining bias at multiple levels of the ML pipeline. Our research was guided by three research questions (RQs).

- RQ1: To what extent are contemporary ASR systems biased against Black (compared to white) tutors, and what is the source of the bias? We addressed this question by applying Whisper ASR [82] (one of the most prominent open-source ASRs) to tutorial data. We also expanded on Koenecke et al.'s approach [51] to investigate whether the source of bias emerges from the spoken content (what was said) or the acoustics (how it was said).
- RQ2: To what extent does ASR bias impact downstream discourse classification? We investigated this question by examining differences (for Black vs. white tutors) in the accuracy and scores produced by a RoBERTa [59] encoder model trained to classify academically productive talk in tutorial discourse.
- RQ3: What are the most effective ways to mitigate ASR bias, and does this have an impact on downstream classification?

To address this question, we fine-tuned the Whisper ASR on demographically diverse datasets and contrasted it with an ASR-augmented training strategy to mitigate bias in the discourse classification models.

This current research is **novel** in three significant areas: First, to the best of our knowledge, this is the first study that links racial biases in ASR to corresponding biases on downstream classification tasks. A second novel aspect pertains to the use of different strategies to reduce bias, including debiasing the ASR itself, or making the discourse classifier more robust to ASR errors. Previous research has mainly focused on either documenting racial bias in ASR [51, 63, 103] or mitigating it [31, 109], and emphasizes a single stage of the pipeline. By examining different approaches aimed at multiple stages of the pipeline, we aim to provide more comprehensive recommendations. Our work is also novel as we examined the impacts of ASR bias in authentic educational settings. This expands evidence of ASR racial bias to a new domain, extending existing works that finds racial bias in healthcare [36], sociolinguistic interviews [51], and everyday speech [63, 103]. Lastly, our research-practice partnership with ANON ensures that our research findings have immediate real-world effects, by providing guidance on how to interpret model results (RQs 1 and 2) while also improving the underlying models (RQ3) used to enhance thousands of tutorial sessions.

4 Research context and data

4.1 Research Context

We (the research organization) partnered with ANON, a large non-profit provider of tutoring services to Title I schools in the U.S. (i.e., public schools with predominantly low-income and historically marginalized student populations). This partnership emerged from shared mutual goals of the research organization – to leverage advances in technology to address important societal needs – with those of ANON, which aims to help historically marginalized students increase their mathematics achievement scores at no cost. ANON partners with another non-profit organization to recruit recent college graduates for a "service year" of employment as tutors. Tutoring occurs in small groups of 2-5 students during the school day, where students are physically present in a classroom while tutors work remotely via a virtual tutoring interface.

The interface records video and audio from the tutor and students. The videos are primarily collected for security purposes (i.e., to ensure safety of students who are minors during the tutoring sessions with adult tutors), and secondly for quality improvement of the tutoring program (but not evaluative) purposes. To this latter point, tutors are paired with dedicated coaches who review recordings of their tutoring sessions along with AI-generated feedback for in-depth analysis of tutors' strengths and areas for improvement. Recording and feedback coaching cycles are a routine component of the program, which reduces (though does not eliminate) perceptions of being surveilled.

Policies for recording video and audio and notification of stakeholders were established by the individual school districts within federal, state, and district-specific requirements. These data were collected by ANON under agreements with the individual districts. ANON then de-identified the data and shared deidentified audio (but not video) and transcripts with the research team under a Data Usage Agreement signed by both organizations. The overall research project of developing AI-based professional learning was approved by the research organization's Institutional Review Board (IRB), of which the present study focuses on the detection and amelioration of racial bias in ASR to improve the equity of benefits for tutors and students.

4.2 Analysis of Tutorial Discourse

Tutoring discourse was analyzed using the Academically Productive Talk (APT) framework [68], which outlines six tutor discourse moves (*talk moves)* that promote student learning and equitable participation: (1): *Keeping everyone together* (e.g. "What did Eliza just say her equation was?"); (2) *Getting students to relate to another's ideas* (e.g. "Do you agree with Juan that the answer is 7/10?"); (3) *Restating* (building off a prior response, e.g. "Add two here"); (4) *Pressing for accuracy* (e.g. "Can you give an example of an ordered pair?"); (5). *Revoicing* (e.g. "Julia told us she would add two here."); and (6) *Pressing for reasoning* (e.g. "Why could I argue that the slope should be increasing?") [100].

The research team worked with ANON to develop an AI-powered interface that provides formative feedback on tutors' usage of the talk moves. The recordings are automatically transcribed with Whisper ASR [82] (detailed in Section 5) and input to a discourse classifier (detailed in Section 6), which outputs model-estimated occurrences of talk moves from each utterance, which are then presented as feedback in interfaces similar to Figure 2. Because this feedback is used to guide tutor learning, it is essential that it is both accurate and unbiased, which is the focus of the present work.

4.3 Data

We obtained 164 recordings from 65 tutors of 9^{th} grade Mathematics small group tutoring sessions, alongside tutor-provided demographic data consisting of race descriptions, personal pronouns, and graduation year. These recordings were transcribed by both human annotators and with OpenAI Whisper large-v2 [82], an ASR that is publicly available, which enables replication. The Whisper ASR is also used by the automatic feedback interface, meaning that insights gained from this study could inform practice through the creation of more inclusive models or providing guidance on the usage of the models.

A majority of the tutors self-reported their race as Black or African American (henceforth, Black – 30.8%), or white (26.2%) with personal pronouns of She/Her/Hers (60.0%) or He/Him/His (30.8%). There were insufficient number of tutors from other racial and pronoun categories, so we focused on Black and white tutors. Due to our methodology of matching utterances on confounding variables (Section 4.4), we excluded tutors who opted out from providing personal pronouns or did not provide their graduation year, which resulted in 88 tutoring sessions from 34 unique tutors. We extracted a total of 18,379 tutor utterances from the transcripts spanning 17.41 hours of recordings.

4.4 Propensity Matching

Koenecke et al. [51] provided a systematic method to quantify ASR bias while addressing extraneous variables, which we replicated

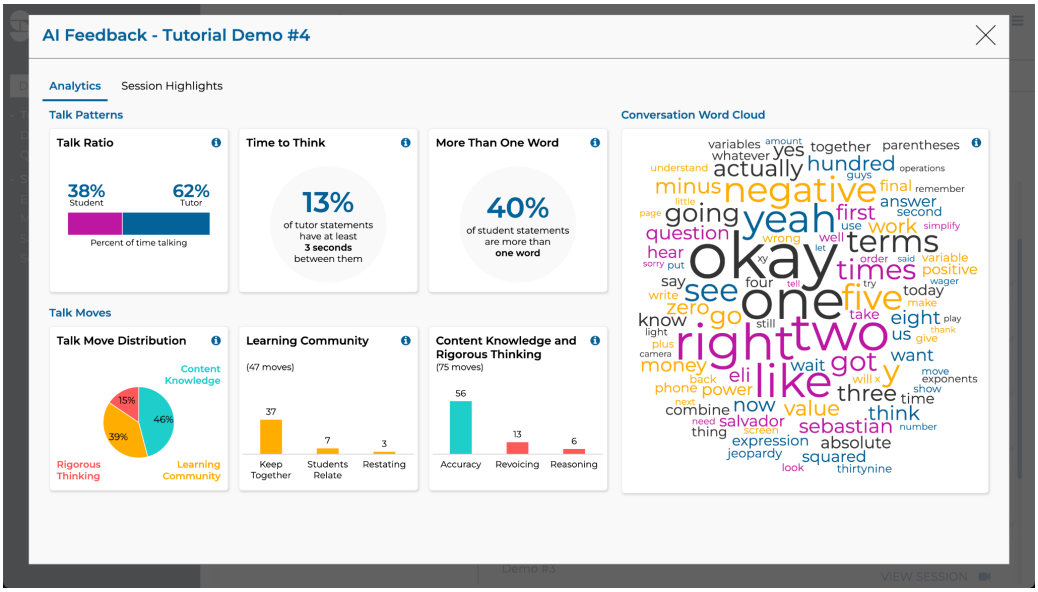


Figure 2: Screenshot of the tutor feedback interface showing visualizations of the usage of Talk Patterns, Talk Moves, and Conversation Word Cloud

here. Specifically, their method relied on propensity-score matching [62] a widely-used statistical technique that estimates the effect of an independent variable on a dependent variable by matching cases on potential confounding variables (covariates) that may predict the independent variable [85]. ASR accuracy (dependent variable) may vary due to other confounding factors such as speaker's gender [51], age [89], and noise [73], in addition to race (independent variable). Therefore, we employed propensity-score matching [62] to balance these confounding factors by creating a subset of audio snippets from white and Black tutors with comparable distributions of personal pronouns, age, duration, and audio quality (i.e., signal-to-noise ratio (SNR) in the acoustic channel), thereby enabling us to isolate the effect of racial disparities while controlling for these confounding factors.

Propensity scores were computed at the utterance level (using the Python version 3.10.9 package *PsmPy* [50]) by fitting a logistic regression model that regressed race on the following confounding factors: an indicator variable for tutor personal pronouns, graduation year (as an approximation for age), natural log (to address outliers) of the utterance length (measured in seconds), and SNR. Nearest neighbor matching without replacement was performed on the computed propensity scores, using a caliper size of 0.01 (which is 0.2 * the standard deviation of the propensity scores [1]). The matched data set comprised 12,572 utterances reflecting 11.96 hours of audio: 6,286 by 18 Black speakers (50% He/Him/His) and an equal number of utterances by 16 white speakers (38% He/Him/His). Successful matching was verified by checking the propensity logit distributions and by ensuring that the resulting distributions of covariates were approximately equal between races. This Propensity Matched dataset was the main dataset used in the analyses.

4.5 Talk Move Coding

Talk move labels were annotated by expert coders, who demonstrated high levels of inter-rater reliability (Cohen's kappa > 0.8). The coding scheme was developed with experts in math education and APT [68]. The annotations are based on the human transcript of a recording and are provided at the utterance level, though coders use the surrounding context for disambiguation. We coded a portion (11,983 utterances) of the full dataset for talk moves with 2,949 (24.6%) utterances coded as at least one talk move with the following distributions: Pressing for Accuracy (11.2%), Keeping Everyone Together (8.1%), Revoicing (3.5%), Pressing for Reasoning (0.7%), Restating (0.6%), Relating (0.5%), and None (75.4%). We used the same propensity-matching approach as above to select 7,292 total utterances[1]: 3,647 by 13 Black speakers and 3,645 utterances by 13 white speakers from the 68 recordings, totaling 7.1 hours.

4.6 Statistical modeling

We used linear mixed effects (LME) models to statistically analyze our data to account for repeated observations and nesting (i.e., clustering) of recordings within tutors. Specifically, we regressed our dependent variables (DVs - see below) on the interactions between race and personal pronouns with nested random intercepts of recordings within tutors, using the *lme4* package [2] in R version 3.6.3 [81]; more complex random effects structures resulted in convergence errors due to a lack of variance. Post-hoc analyses of significant interactions indicated different results for each race (main effects) by personal pronouns (moderator), which were explored using *emmeans* [86]. We used two-tailed tests with a $p < .05$

[1]There were two duplicate utterances from white speakers, resulting in an unequal number of utterances in this dataset.

cutoff for significance. A sample statistical model is shown below:

$$DV \sim Race$$
$$[Black|White]\times Personal\ Pronouns[SheHerHers|HeHimHis]$$
$$+(1|Tutor:Recording)+(1|Tutor)$$

5 RQ1: ASR bias and sources of bias

We investigated racial bias in the Whisper ASR, disentangling the source of the bias in the system to either the acoustic or linguistic components of the model. Generally, Word Error Rate (WER) is used to evaluate the performance of ASR systems [73], i.e. the ratio of word-level edit operations (substitutions, deletions, insertions) needed to transform a human (reference) transcript to the ASR (hypothesis) result, divided by the number of words in the reference:

$$Match\ Error\ Rate = (substitutions + deletions + insertions) \div (reference\ word\ count)$$

The alignment of reference and hypothesis was computed following text normalization to remove non-spoken annotations, spell out numbers, expand contractions, strip punctuation, and convert text to lowercase. Because WER is unbounded and can reach very large values where many insertions are present (as we see in the case of repetitive insertions of a single word or phrase, which Whisper is known to be prone to [14]), such extreme values can inflate scores. For this reason, the Match Error Rate (MER) [70], which is bounded from -0 to 1, was used as a summary error metric for ASR accuracy. However, we also include WER statistics for comparison to other work and examined bias in each type of ASR edit operation (i.e., error) separately

$$Match\ Error\ Rate = (substitutions + deletions + insertions) \div (insertions + reference\ word\ count)$$

5.1 Overall ASR Errors and Bias

Table 1 provides descriptives of WER, MER, and different error types by race on the Propensity Matched dataset (left). The linear mixed effects model that regressed MER on the race × personal pronouns interaction revealed a significantly higher error rate for Black tutors (p=0.009). There were significantly more deletion (p=0.004), and substitutions (p=0.019) errors for Black tutors, however the increase in insertions was not significant (p=0.173). In general, the ASR WER for Black tutors was higher (by 24%) than their white counterparts, which replicates prior research [51]. There were no significant differences by personal pronouns nor a significant interaction between race and personal pronouns.

Because the distributions were zero-inflated (33% of utterances were transcribed perfectly, and each error type only occurred in approximately one-third of utterances), we fit logistic regression models to examine whether the presence/absence of errors in each utterance varied by race, personal pronouns, and their interaction. We found that the odds of at least one error were significantly higher for Black tutors (odds ratio [OR] =1.87, p=0.014), and specifically, the odds for at least one deletion or at least one substitution error was significantly higher for Black tutors (deletion: OR=1.74, p=0.006; substitution: OR=1.47, p=0.019). There was also an interaction with personal pronouns for deletion errors, which indicated at least one deletion per utterance was more likely for Black than white tutors, but only for speakers preferring He/Him/His personal pronouns (OR=1.74 vs 0.93 for She/Her/Hers).

5.2 Sources of Bias

We investigated whether bias could be attributed to the linguistic (what was said) versus the acoustic (how it was said) components of the ASR system, building off the approach of [51].

5.2.1 Acoustic Components of Bias. To isolate bias in the acoustic component of the ASR system, we selected a subset of matched n-grams (short phrases) that were spoken by both Black and white tutors. Thus, any differences in accuracy for these phrases can be attributed to differences in the acoustic component since the language is fixed. Following the approach of [51], this matching was limited to phrases of at least two words. We also ensured that the speakers of the matched n-grams had the same personal pronouns and were of similar age (approximated by graduation year). Unlike [51], we further verified that the SNR of the matched utterances was approximately equivalent, i.e., within 6dB (with the minimum discernible difference by a human listener as 3dB; [65]). Matching on the human transcripts resulted in 3,042 pairs of utterances between 2 and 9 words and 0.1 and 8.7 seconds in length.

Since the Whisper decoder implicitly contains a language model and sequence decoding makes use of the linguistic context surrounding an n-gram, it is possible that differences in ASR performance between matched n-grams still has a residual contribution from the rest of the utterance. To control for this, we isolated the audio of each n-gram within the original utterance, then transcribed the extracted n-grams with Whisper. Specifically, we used forced alignment [55] as implemented in TorchAudio, using the HuBERT ASR model [40] to extract frame-wise probability distributions over tokens in order to derive the most likely temporal alignment of each n-gram to the audio and used this to extract the corresponding n-gram audio.

As shown in Table 1 (right side), we found that the overall MER (p=0.038) and deletion rate (p<0.001) were indeed significantly higher for Black tutors even when controlling precisely for language by analyzing matched n-grams. There was no significant racial difference for the insertion and substitution rates. Post-hoc analysis on the significant race and personal pronouns interaction revealed that deletions were significantly more likely for Black tutors than for white with He/Him/His personal pronouns (21% vs 15%, p<0.001), but equivalent between races for She/Her/Hers (19% vs 18%, p=0.844 for Black and white respectively). Thus, bias was largest for Black tutors with He/Him/His personal pronouns, likely due to deletion errors.

5.2.2 Language Components of Bias. Following the approach of [51], we hypothesized that the lower ASR accuracy for the Black tutors may be a result of word usage outside of the fixed vocabulary of the ASR's language model. Using the propensity matched dataset, we collected all unique tokens (words) that were transcribed by the ASR as an approximate reconstruction of the model's vocabulary, and likewise for ground-truth human transcripts. From this set, we then computed the proportion of words in the human transcripts that had been transcribed in the reconstructed ASR vocabulary. We

Table 1: Mean error rates by race for the Propensity Matched and N-gram matched datasets.

	Propensity Matched		N-gram Matched	
Measure	*White*	*Black*	*White*	*Black*
Word error rate (WER)	0.39	0.48	0.52	0.58
Match error rate (MER)	0.24	0.31	0.41	0.46
Deletion rate	0.10	0.13	0.17	0.19
Insertion rate	0.18	0.21	0.15	0.16
Substitution rate	0.10	0.14	0.20	0.23

found that 87% of the words transcribed by humans were available in the ASR vocabulary for Black tutors with a similar rate (88%) for white tutors.

Next, we investigated the average perplexity that language models assign to the utterances for both groups. The perplexity of an utterance is defined as the exponentiated average of negative log-likelihood over tokens, each conditioned on the preceding tokens [46]. This quantifies how *surprised* the model is to encounter a given string of words. Utterances with lower perplexity are thus more easily predicted (i.e., transcribed) by the language model. Because Whisper does not have a separate language model from which we can compute perplexity based on language alone, we assumed that the distributions of other recent language models from the same OpenAI family, might align well with that used in Whisper and thus used GPT-2 to derive perplexity scores. We filtered out utterances with only a single word as we found the perplexity metric is inflated for such utterances. We found that perplexity was lower (i.e., less surprising speech based on the GPT-2 language model) for the human transcripts of Black tutors than white tutors (propensity scores of 981 and 1,125 respectively), and likewise for Whisper transcripts (866 and 968). Thus, any differences in language use among races, did not manifest in more surprising utterances for Black speakers according to a LLM trained on a broad corpus.

Overall, our results suggested that the language use of Black tutors was not more out-of-distribution or unexpected than for white tutors, indicating that language use does not explain the ASR performance gap.

6 RQ2: effects of ASR bias on downstream discourse classification

We explored the extent to which ASR bias affects downstream classification by examining how using ASR input influences the accuracy and scores generated by a RoBERTa encoder model [59] trained to classify tutor talk moves usage. We chose to use RoBERTa over more recent LLMs because this model is deployed in the automated feedback interface used for professional coaching in our application domain.

6.1 Training the Talk Moves Discourse Classifier

We completed two rounds of fine-tuning of a RoBERTa classifier to predict seven classes of talk moves (six moves plus None) in a multi-class classification setting following the approach of [100]. The training process involved standard fine-tuning of the RoBERTa model, wherein the base transformer layers were initialized from the pre-trained model, and the final layer was replaced with a fully connected layer to output probabilities for the seven classes. We used cross-entropy loss as the objective function for training. The first round of fine-tuning consisted of tuning with an out of domain dataset, comprising 567 human-annotated K-12 mathematics lesson transcripts. Similar to our data, this dataset was derived from video recordings, including lessons with either whole-class discussions and/or small group work, in addition to online lessons [99]. We fine-tuned this RoBERTa model a second time in a cross validation setting on the 11,983 labelled utterances (prior to propensity matching).

We employed 10-fold cross-validation where in each iteration, eight folds were used as a train set, one was used as a development set, and one was held out for testing. The model was trained over five epochs, saving checkpoints of the model throughout the training process. The model checkpoint with the highest accuracy on the development set was used to generate predictions on the held-out test set. The predictions from each of the 10 test sets were combined. Previously cited (better-performing) models harnessed larger context windows (e.g., seven previous and seven future utterances) and/or previous student utterances. However, since our focus is on single utterances of tutor speech, we opted for a less accurate yet more appropriate single-utterance model to investigate our research questions.

Using the talk moves Propensity Matched subset of this data, consisting of 7,292 matched utterances, we then investigated differences between races in the occurrence rate and classification errors of talk moves.

6.2 Bias in the Talk Moves Model

6.2.1 Talk Moves Model Accuracy. We first investigated whether talk moves model accuracy (left side of Table 2) differed by race when using human-transcribed input compared to ASR input. Using human-transcribed input to the talk moves model, the F1 score (a measure of classification accuracy) was 0.66 for Black tutors and 0.64 for white tutors. However, there was a larger decrease in F1 scores when moving from human to ASR transcripts for Black tutors (from .66 to .51, a 0.15 decrease) compared to white tutors (from .64 to .54, a 0.10 decrease). To test this statistically, we regressed model accuracy (1 [accurate] or 0 [inaccurate]) on the interaction between race and personal pronouns for both human and ASR transcripts via two logistic regression models. Results revealed significantly higher odds of making a classification error for Black vs. white tutor utterances for ASR transcripts (OR=1.50; p=0.013), but

Table 2: Talk Moves model error rates and occurrence rates using human and ASR transcripts

	Model Accuracy(Macro F1 Scores)		Model Predictions(% Talk Moves)	
	White	*Black*	*White*	*Black*
Ground Truth	-	-	23.3	27.1
Human Transcript	0.64	0.66	22.2	25.5
ASR Transcript	0.54	0.51	19.8	20.2

there was no equivalent difference for human-transcripts (OR=1.31, p=0.170), suggesting bias against Black tutors is likely due to the ASR accuracy bias noted in RQ1.

6.2.2 Talk Moves Occurrence. Having found lower model accuracy for the talk moves model when provided with ASR inputs, we proceeded to examine how this affected classification rate of utterances as containing a talk move (vs. None; right side of Table 2). Starting with the ground-truth human scores, we found that talk moves were actually coded at a higher rate (27%) in Black tutor's utterances, compared to 23% of the time in white tutor's utterances, though this difference only approached significance (OR=1.52; p=0.091) with a logistic mixed effects regression model that regressed the presence [1] or absence [0] of talk moves on the race × personal pronouns interaction. This relative advantage for Black tutors was maintained when the talk moves model generated predictions using the human transcript input (26% vs 22%; OR=1.56, p=0.017), but was eliminated for ASR transcript input (20% occurrence rate for both races, modeled OR=1.31, p=0.126). Thus, ASR bias adversely impacted the accuracy and scores of the underlying discourse classifiers for Black tutors.

7 RQ 3: Mitigating bias

To mitigate the racial bias observed in RQ1 and RQ2, we implemented two techniques for de-biasing the ASR and classification systems for fair performance between both race groups.

7.1 Fine-tuning the ASR Model

We hypothesized that we could address the racial performance gap in ASR by fine-tuning the Whisper ASR model with in-domain data that had an equal representation of speech from white and Black tutors. Thus, we selected an approximately equal number of utterances from Black and white tutors (~7,994 utterances each) taken from the initial dataset (prior to matching on race, personal pronouns, and graduation year). We excluded the talk move labeled subset of utterances as these were used to evaluate performance. We fine-tuned Whisper large-v2 for 10 epochs with a learning rate of 1e-5 and a batch size of 32 using the t*ransformers* Python library [110], coupled with a low-rank adaptation (LoRA) approach [41] to reduce the trainable parameters to 15 million with int8 quantization [20].

7.1.1 Fine-Tuning Results. Fine-tuning improved the WER on the talk moves propensity matched dataset both for Black tutors (WER of 0.44 reduced to 0.35) and for white tutors (from 0.32 to 0.28), with a reduction in the racial gap from 37% to 26%. As can be seen in Figure 3 , fine-tuning mainly decreased deletion and insertion

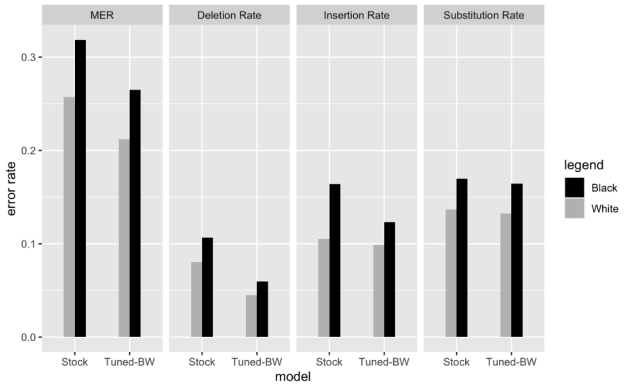


Figure 3: Error rates for the original (stock) Whisper and the model fine-tuned using equal amounts of Black and white tutor speech.

rather than substitution error rates. However, fine-tuning did not eliminate the gap entirely.

Next, we fit four LME models to predict MER and the three error rates from the three-way interaction between race, personal pronouns, and whether the model was fine-tuned or not. Of interest was whether fine-tuning reduced error overall, and if there was a significant interaction with race (i.e., whether the error rate reduction was higher for Black than white tutors). Further interaction with personal pronouns would indicate that the fine-tuning was more effective for tutors using one set of personal pronouns than the other. We found that fine-tuning significantly reduced the overall MER (estimated MER decreased by 20% from 0.24 to 0.19, p<0.001), but there was not a significant interaction with race suggesting that tutors of both races benefitted. However, when looking deeper into specific errors, the deletion rate was also significantly reduced (from 0.072 to 0.025, p<0.001) by fine-tuning and this varied by race. Post-hoc analyses indicated that differences in deletion errors between white and Black tutors, which was significant before fine-tuning (0.081 vs 0.109 respectively; p=0.015) was not statistically different for the tuned model (0.044 versus 0.060; p=0.17).

Fine-tuning the ASR improved talk moves classification accuracy on the resultant transcripts for the Black tutors and made no difference for the white tutors (left side of Table 3). For Black tutors, the F1 score increased from 0.51 to 0.58. To test significance, we regressed model accuracy for the fine-tuned transcripts on race × personal pronouns. We found the model's error rate bias reduced to

Table 3: Macro F1 scores for the talk moves classifier with and without debiasing strategies

Talk Moves Classifier				
	Original Model		ASR-augmented training	
Source of test transcript	*White*	*Black*	*White*	*Black*
Human	0.64	0.66	0.60	0.67
Baseline ASR	0.54	0.51	0.51	0.52
Fine-tuned ASR	0.54	0.58	-	-

marginal significance (OR=1.39; p=0.053) on fine-tuned data compared to the original Whisper transcripts (OR=1.50, as reported in Section 6.2.1).

7.2 ASR-Augmented Classifier Training

Our second approach to mitigate bias focused on the talk moves classification model using an ASR-augmented training technique [100]. Specifically, we trained a RoBERTa classifier on both the human and ASR transcribed versions of tutor utterances, essentially doubling the training set; all hyperparameters, parameters, and dataset splits were held constant with the human transcript-only model detailed in Section 6.1. The idea was to help the model to better understand ASR errors and become more resilient to them. As noted in Table 3 (right side), ASR-augmented training had a modest effect on racial bias in model accuracy, with a .03 decrease in F1 score for white tutors, and an F1 increase of .01 for Black tutors, but this represents less than a halving of the accuracy gap. Thus, fine-tuning the ASR appears to be a more effective strategy than ASR-augmented training of the classifier.

8 Discussion

We examined racial bias in automated feedback interfaces that aim to promote data-driven, job-embedded professional development for human tutors in a real-world context. Our focus was on ASR bias (RQ1), its downstream effects (RQ2), and efforts to mitigate it (RQ3). In the remainder of this section, we discuss our main findings, and their implications followed by a discussion of limitations and future directions.

8.1 Main findings

Our analysis revealed significant demographic disparities in ASR accuracy, with a notably higher error rate for Black tutors compared to white tutors. Additionally, we found that bias particularly manifested in higher rates of deletion errors for Black speakers. This discrepancy persisted even when controlling for language content, as evidenced by higher error rates for Black speakers in the matched n-grams. The lower ASR performance for Black speakers could not be attributed to differences in their linguistic patterns since they had lower perplexity scores compared to the white speakers. Overall, the findings for RQ1 suggests that bias may be more attributable to the ASR's deficiency in modeling the acoustic than the language component of Black tutors' speech. These findings corroborate the findings of Koenecke et al. [51], who reported a systemic underperformance in ASR systems for Black speakers which is due more to acoustic than linguistic factors, underscoring the need for improved acoustic models across demographics [75, 77].

In RQ2, we found that the downstream effects of ASR errors were evident in the decreased performance of a discourse classification model for Black tutors when using ASR input as opposed to ground-truth human transcripts. This suggests that the bias was not inherent in the discourse classifier itself but was introduced due to biased ASR inputs. The outcome of this introduced bias was impactful. While Black tutors exhibited higher-quality discourse (i.e. used more high-impact talk moves) according to expert coding, this effect was obfuscated in the automated pipeline. Specifically, the odds of detecting talk moves were higher for Black tutors than white tutors when using human transcripts, but this advantage diminished when using ASR-generated transcripts, underscoring the need for improving ASR systems and downstream models to ensure equitable performance across demographic groups.

To this point, our efforts to reduce ASR bias through fine-tuning the Whisper ASR model and using ASR-augmented training were partly successful (RQ3). Whereas fine-tuning the ASR on in-domain datasets using an equal quantity of utterances from white and Black tutors significantly reduced ASR error rates and improved talk moves classification with larger effects for Black tutors, it did not fully eliminate the bias. The second approach of ASR-augmented training of the classifier was less effective. Thus, we can conclude that fine-tuning on in-domain data appears to be a promising way to reduce ASR bias and its effects, but more work is needed to fully eliminate the bias.

8.2 Implications

In the US, racial disparities in academic performance and attainment in Mathematics and other STEM majors have been attributed to systemic inequalities [33, 43, 44, 64, 78], uneven student-teacher racial composition [56, 60], and the digital divide associated with socioeconomic status [25, 29]. Black teachers, who are underrepresented in the teaching workforce [72], could face additional challenges, including biases and ethical concerns relating to emerging AI tools [6, 107]. Research has shown that inaccurate feedback on teachers' performance can lead to stress, anxiety, burnout, diminished job satisfaction, lower self-efficacy, and limited professional growth [69, 114]. For instance, vague or inaccurate feedback can hinder professional development and reduce overall teaching effectiveness [114]. Advanced computing tools designed to support educators may have the unintended consequence of amplifying these challenges if they embed or reflect existing biases. Numerous scholars [3, 7, 35, 57, 75, 76, 83, 84, 87, 92] have recommended that educators, policymakers, and AI system developers adopt actionable strategies—such as bias audits, co-design with diverse stakeholders, implementing responsible AI principles, applying organizational

justice theory, and developing inclusive AI training datasets as part of solutions to the challenges of designing fair and unbiased AI systems [3, 7, 35, 57, 75, 76, 83, 84, 87, 92]. In the education context, when deployed with fidelity and adequate support (e.g., professional learning, technological resources, implementation support), these systems could have positive effects on teacher job satisfaction, foster professional growth, and increase retention among Black teachers, ultimately helping to address racial imbalances and systemic inequalities in US schools.

Our project takes one important step towards this goal in the context of addressing bias in technology developed to enhance the quality of high dosage tutoring for historically marginalized students. Our finding of ASR bias in the context of an automated feedback application has significant implications in the short- and long- term. First, because the system is currently in use for formative feedback and coaching, the lower accuracy for Black tutors has immediate consequences in their day-to-day work. Fortunately, because the professional development process in the current system is focused on improvement instead of evaluation and the automated feedback is only shared with their assigned coaches (not the tutors themselves), this has limited negative consequences. Nevertheless, it is imperative that this differential accuracy in the automated feedback be communicated to coaches so they can factor in how they track progress and provide support. One possibility is to consider feedback visualizations that communicate the uncertainty of the automated estimates adhering to the principle of transparency in human-centered AI design [90]. In parallel, efforts to improve accuracy, starting with replacing the default Whisper model with the fine-tuned version with lower ASR bias, should also be considered. This improved system will be more beneficial to coaches as it will provide them with more reliable evidence to better mentor and guide their assigned tutors. This will also help the professional development of all tutors, especially Black tutors because the feedback they receive will be more closely calibrated to their abilities.

The longer-term adverse consequences of bias could arise in similar automated feedback systems including commercial platforms, which may impact far more educators. Even more troubling are use cases where the outputs of the automated models are used for decision making. In the current case, the ASR-based models underpredicted the occurrence of talk moves for Black tutors, so any decision based on said models can be thought of as adversely impacting them. Even though the present stakes are much lower, the findings allude to the importance of assessing bias in any automated ML system, something that is still rarely done in practice [26] despite calls to do so [3, 4, 6, 8, 75, 83]. Further, while higher accuracy is usually desired in most AI systems, there remains the possibility of harm arising regardless of accuracy depending on how the recognized speech is used (e.g. violations of privacy facilitating increased surveillance) [71, 94, 115]. Lastly, accuracy is not the only outcome of interest, but it should be calibrated across other pertinent outcomes including reliability, fairness, privacy, generalizability, explainability, and unintended consequences.

8.3 Limitations & Future Work

As with any research, our work has limitations. First, we analyzed a single ASR system, so our conclusions might not generalize to other ASR models, which is an item for future work. Likewise, we used RoBERTa for discourse classification because this is what is currently deployed in the target application, but replicating with modern LLMs and newer training approaches would be desirable.

Our second limitation pertains to the data sources. Despite the large number of utterances and statically significant findings, the number of tutors ($N = 34$) may not adequately represent data from each demographic, and the focus on two races limits generalizability. Future research with a larger sample of tutors with more diverse racial composition is recommended. There is likely bias in ASR accuracy for student speech as well, which we did not address due to a lack of data on student demographics and our focus on tutors, but this remains an important item for future work.

Methodologically, we were limited by an inability to truly isolate the language and acoustic components of the ASR, as they are inherently intertwined in the decoder. In using perplexity from GPT-2 to infer whether Whisper's coverage of language patterns differs by race, we assumed that racial biases in Whisper's language modeling aligned with other state-of-the-art language models. While the GPT-2 and Whisper models are owned by OpenAI, we are uncertain whether they share the same language components. Future work could involve using ASR models with accessible language components. Furthermore, we also did not explicitly quantify the use of AAVE speech in Black tutors as a source for the bias, which is an important future direction.

We also had limited success in debiasing the ASR model by fine tuning on in-domain speech, finding improved accuracy and reduced bias, but did not eliminate bias overall. A more robust improvement in the racial accuracy gap may require a more extensive de-biasing method and a varied dataset. Another avenue to pursue involves explicitly training the ASR with a debiasing objective, such as [88] which examined counterfactual generation of training examples from different groups and counterfactual regularization to minimize the difference in model prediction. Given our results indicate that Whisper's racial accuracy gap is driven by acoustic differences, such an approach would be particularly appropriate here.

8.4 Conclusions

Equality deems that at a minimum, automated systems perform fairly for everyone, irrespective of their demographic background. However, research has shown that this is not always achieved in AI systems, with various reports of differential performance of ASR across racial groups. Our study revealed significant ASR bias against Black tutors compared to white tutors in an AI-powered system used to provide automated feedback for professional learning. We traced the source of the bias in the ASR system primarily to acoustic factors, finding that the biased ASR negatively impacted the accuracy of the downstream classification model. As a consequence, while Black tutors had superior usage of high-quality discourse as measured by human coding of their discourse, this advantage was erased due to ASR errors in the automatic feedback. Finally, we found that the accuracy gap can be reduced, but not entirely eliminated, by fine-tuning the ASR on more diverse in-domain speech. Beyond bias, it is also likely that accuracy of the automated approaches is inherently limited due to the nature of

noisy speech in authentic environments, so it is imperative that feedback interfaces and the accompanying professional learning are robust to a modicum of inaccuracy. This would likely entail a social-technical approach that leverages the strengths of machines to sift through vast volumes of data in tandem with those of humans to draw insights from what the machine has to offer.

Acknowledgments

This research is funded by the National Science Foundation (NSF DRL 1920510) and the Learning Engineering Virtual Institutes. The opinions expressed are those of the authors and do not represent the views of the funding agencies

References

[1] Peter C. Austin. 2011. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics* 10, 2: 150–161. https://doi.org/10.1002/pst.433

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software* 67, 1. https://doi.org/10.18637/jss.v067.i01

[3] Ruha Benjamin. 2016. Catching Our Breath: Critical Race STS and the Carceral Imagination. *Engaging Science, Technology, and Society* 2: 145–156. https://doi.org/10.17351/ests2016.70

[4] Ruha Benjamin. 2019. Assessing risk, automating racism. *Science* 366, 6464: 421–422. https://doi.org/10.1126/science.aaz3873

[5] Bronwyn Bevan and William R Penuel. 2017. *Connecting research and practice for educational improvement: Ethical and equitable approaches*. Routledge.

[6] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2: 100205. https://doi.org/10.1016/j.patter.2021.100205

[7] Abeba Birhane, Vinay Uday Prabhu, and John Whaley. 2022. Auditing saliency cropping algorithms. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4051–4059.

[8] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The Forgotten Margins of AI Ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 948–958. https://doi.org/10.1145/3531146.3533157

[9] Beatrice F Birman and Georges Vernez. 2007. *State and Local Implementation of the No Child Left Behind Act: Interim Report. Teacher quality under NCLB*. US Department of Education, Office of Planning, Evaluation and Policy . . ..

[10] Nathaniel Blanchard, Michael Brady, Andrew M. Olney, Marci Glaus, Xiaoyi Sun, Martin Nystrand, Borhan Samei, Sean Kelly, and Sidney D'Mello. 2015. A Study of Automatic Speech Recognition in Noisy Classroom Environments for Automated Dialog Analysis. In *Artificial Intelligence in Education*, Cristina Conati, Neil Heffernan, Antonija Mitrovic and M. Felisa Verdejo (eds.). Springer International Publishing, Cham, 23–33. https://doi.org/10.1007/978-3-319-19773-9_3

[11] 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*, 268–277. https://doi.org/10.1145/3462244.3479897

[12] Brandon M. Booth, Jennifer Jacobs, Jeffrey B. Bush, Brent Milne, Tom Fischaber, and Sidney K. DMello. 2024. Human-tutor Coaching Technology (HTCT): Automated Discourse Analytics in a Coached Tutoring Model. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 725–735. https://doi.org/10.1145/3636555.3636937

[13] Robin N. Brewer, Christina Harrington, and Courtney Heldreth. 2023. Envisioning Equitable Speech Technologies for Black Older Adults. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 379–388. https://doi.org/10.1145/3593013.3594005

[14] Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, E. Margaret Perkoff, Michael Regan, Katharina Kann, James H. Martin, Martha Palmer, and Sidney D'Mello. 2023. A Comparative Analysis of Automatic Speech Recognition Errors in Small Group Classroom Discourse. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (UMAP '23), 250–262. https://doi.org/10.1145/3565472.3595606

[15] Ismail Celik, Muhterem Dindar, Hanni Muukkonen, and Sanna Järvelä. 2022. The Promises and Challenges of Artificial Intelligence for Teachers: a Systematic Review of Research. *TechTrends* 66, 4: 616–630. https://doi.org/10.1007/s11528-022-00715-y

[16] Chelsea Chandler, Thomas Breideband, Jason G. Reitman, Marissa Chitwood, Jeffrey B. Bush, Amanda Howard, Sarah Leonhart, Peter W. Foltz, William R. Penuel, and Sidney K. D'Mello. 2024. Computational Modeling of Collaborative Discourse to Enable Feedback and Reflection in Middle School Classrooms. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (LAK '24), 576–586. https://doi.org/10.1145/3636555.3636917

[17] Xingyu Chen, Zhengxiong Li, Srirangaraj Setlur, and Wenyao Xu. 2022. Exploring racial and gender disparities in voice biometrics. *Scientific Reports* 12, 1: 3723. https://doi.org/10.1038/s41598-022-06673-y

[18] Jay Cunningham, Su Lin Blodgett, Michael Madaio, Hal Daumé Iii, Christina Harrington, and Hanna Wallach. 2024. Understanding the Impacts of Language Technologies' Performance Disparities on African American Language Speakers. In *Findings of the Association for Computational Linguistics ACL 2024*, 12826–12833. https://doi.org/10.18653/v1/2024.findings-acl.761

[19] Dorottya Demszky, Jing Liu, Heather C. Hill, Dan Jurafsky, and Chris Piech. 2024. Can Automated Feedback Improve Teachers' Uptake of Student Ideas? Evidence From a Randomized Controlled Trial in a Large-Scale Online Course. *Educational Evaluation and Policy Analysis* 46, 3: 483–505. https://doi.org/10.3102/01623737231169270

[20] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. Retrieved September 12, 2024 from http://arxiv.org/abs/2208.07339

[21] Lambert Diana. 2024. Disrespect, low pay, lack of support keep Black teachers out of the profession. Retrieved from https://edsource.org/2024/disrespect-low-pay-lack-of-support-keep-black-teachers-out-of-the-profession

[22] Emily A. Diehm and Alison Eisel Hendricks. 2021. Teachers' Content Knowledge and Pedagogical Beliefs Regarding the Use of African American English. *Language, Speech, and Hearing Services in Schools* 52, 1: 100–117. https://doi.org/10.1044/2020_LSHSS-19-00101

[23] Sidney K. D'Mello, Nicholas Duran, Amanda Michaels, and Angela E. B. Stewart. 2024. Improving collaborative problem-solving skills via automated feedback and scaffolding: a quasi-experimental study with CPSCoach 2.0. *User Modeling and User-Adapted Interaction*. https://doi.org/10.1007/s11257-023-09387-6

[24] Sidney K. D'Mello, Andrew M. Olney, Nathan Blanchard, Borhan Samei, Xiaoyi Sun, Brooke Ward, and Sean Kelly. 2015. Multimodal Capture of Teacher-Student Interactions for Automated Dialogic Analysis in Live Classrooms. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 557–566. https://doi.org/10.1145/2818346.2830602

[25] Jennifer E. Dolan. 2016. Splicing the Divide: A Review of Research on the Evolving Digital Divide Among K–12 Students. *Journal of Research on Technology in Education* 48, 1: 16–37. https://doi.org/10.1080/15391523.2015.1103147

[26] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1: eaao5580. https://doi.org/10.1126/sciadv.aao5580

[27] Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language* 84: 101567. https://doi.org/10.1016/j.csl.2023.101567

[28] Emilio Ferrara. 2023. Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*. https://doi.org/10.5210/fm.v28i11.13346

[29] Dania V. Francis and Christian E. Weller. 2022. Economic Inequality, the Digital Divide, and Remote Learning During COVID-19. *The Review of Black Political Economy* 49, 1: 41–60. https://doi.org/10.1177/00346446211017797

[30] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*: 1–83. https://doi.org/10.1162/coli_a_00524

[31] Shefali Garg, Zhouyuan Huo, Khe Chai Sim, Suzan Schwartz, Mason Chua, Alëna Aksënova, Tsendsuren Munkhdalai, Levi King, Darryl Wright, Zion Mengesha, Dongseong Hwang, Tara Sainath, Françoise Beaufays, and Pedro Moreno Mengibar. 2024. Improving Speech Recognition for African American English with Audio Classification. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12356–12360. https://doi.org/10.1109/ICASSP48485.2024.10447116

[32] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A Survey on Bias in Deep NLP. *Applied Sciences* 11, 7: 3184. https://doi.org/10.3390/app11073184

[33] Ashley Grays, Danica Moise, Erika Moore, Fanica Young, and Tahnee Wilder. 2023. Why Are We Whispering? Addressing Implicit Bias in K-12 Education. *SRATE Journal* 32, 1. Retrieved August 7, 2024 from https://eric.ed.gov/?id$=$EJ1391129

[34] Lisa J. Green. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press.

[35] Christina N. Harrington, Katya Borgos-Rodriguez, and Anne Marie Piper. 2019. Engaging Low-Income African American Older Adults in Health Discussions through Community-based Design Workshops. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3290605.3300823

[36] Christina N. Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "It's Kind of Like Code-Switching": Black Older Adults' Experiences with a Voice Assistant for Health Information Seeking. In *CHI Conference*

on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3491102.3501995

[37] Camille Harris, Chijioke Mgbahurike, and Diyi Yang. Modeling Bias in Automatic Speech Recognition. Retrieved August 7, 2024 from https://southnlp.github.io/southnlp2024/papers/southnlp2024-poster-67.pdf

[38] Drew Harwell, B. Mayes, M. Walls, and S. Hashemi. 2018. The accent gap. *The Washington Post* 19.

[39] Linette N. Hinton and Karen E. Pollock. 2000. Regional Variations in the Phonological Characteristics of African American Vernacular English. *World Englishes* 19, 1: 59–71. https://doi.org/10.1111/1467-971X.00155

[40] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. Retrieved September 4, 2024 from http://arxiv.org/abs/2106.07447

[41] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. Retrieved September 12, 2024 from http://arxiv.org/abs/2106.09685

[42] Gilhwan Hwang, Jeewon Lee, Cindy Yoonjung Oh, and Joonhwan Lee. 2019. It Sounds Like A Woman: Exploring Gender Stereotypes in South Korean Voice Assistants. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3290607.3312915

[43] Yasmiyn Irizarry. 2015. Selling students short: Racial differences in teachers' evaluations of high, average, and low performing students. *Social Science Research* 52: 522–538. https://doi.org/10.1016/j.ssresearch.2015.04.002

[44] Yasmiyn Irizarry. 2021. On Track or Derailed? Race, Advanced Math, and the Transition to High School. *Socius: Sociological Research for a Dynamic World* 7: 2378023120980293. https://doi.org/10.1177/2378023120980293

[45] Andy Jacob and Kate McGovern. 2015. The Mirage: Confronting the Hard Truth about Our Quest for Teacher Development. *TNTP*.

[46] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* 62, S1: S63–S63. https://doi.org/10.1121/1.2016299

[47] Emily Jensen, Meghan Dale, Patrick J. Donnelly, Cathlyn Stone, Sean Kelly, Amanda Godley, and Sidney K. D'Mello. 2020. Toward Automated Feedback on Teacher Discourse to Enhance Teacher Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376418

[48] Sean Kelly, Gizem Guner, Nicholas Hunkins, and Sidney K. D'Mello. 2024. High School English Teachers Reflect on Their Talk: A Study of Response to Automated Feedback with the Teacher Talk Tool. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00417-x

[49] Sean Kelly, Andrew M. Olney, Patrick Donnelly, Martin Nystrand, and Sidney K. D'Mello. 2018. Automatically Measuring Question Authenticity in Real-World Classrooms. *Educational Researcher* 47, 7: 451–464. https://doi.org/10.3102/0013189X18785613

[50] Adrienne Kline and Yuan Luo. 2022. PsmPy: A Package for Retrospective Cohort Matching in Python. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1354–1357. https://doi.org/10.1109/EMBC48229.2022.9871333

[51] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14: 7684–7689. https://doi.org/10.1073/pnas.1915768117

[52] Matthew A. Kraft and Grace T. Falken. 2021. A Blueprint for Scaling Tutoring and Mentoring Across Public Schools. *AERA Open* 7: 233285842110428. https://doi.org/10.1177/23328584211042858

[53] Matthew A. Kraft and Allison F. Gilmour. 2017. Revisiting *The Widget Effect*: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness. *Educational Researcher* 46, 5: 234–249. https://doi.org/10.3102/0013189X17718797

[54] Yogesh Kumar and Navdeep Singh. 2019. A Comprehensive View of Automatic Speech Recognition System - A Systematic Literature Review. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, 168–173. https://doi.org/10.1109/ICACTM.2019.8776714

[55] Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. CTC-Segmentation of Large Corpora for German End-to-end Speech Recognition. https://doi.org/10.48550/ARXIV.2007.09127

[56] Tamika P. La Salle, Cixin Wang, Chaorong Wu, and Jesslynn Rocha Neves. 2020. Racial Mismatch among Minoritized Students and White Teachers: Implications and Recommendations for Moving Forward. *Journal of Educational and Psychological Consultation* 30, 3: 314–343. https://doi.org/10.1080/10474412.2019.1673759

[57] Sarah Priscilla Lee, Tyler James Nanoff, Sydney Simmons, Stephanie T Jones, Vishesh Kumar, and Marcelo Worsley. 2023. Toward Co-Design with Refugee Youth: Facilitation Through a Social-emotional Framework. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, 593–597. https://doi.org/10.1145/3585088.3593870

[58] Lanna Lima, Vasco Furtado, Elizabeth Furtado, and Virgilio Almeida. 2019. Empirical Analysis of Bias in Voice-based Personal Assistants. In *Companion Proceedings of The 2019 World Wide Web Conference*, 533–538. https://doi.org/10.1145/3308560.3317597

[59] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/ARXIV.1907.11692

[60] Jasmín D. Llamas, Khoa Nguyen, and Alisia G.T.T. Tran. 2021. The case for greater faculty diversity: examining the educational impacts of student-faculty racial/ethnic match. *Race Ethnicity and Education* 24, 3: 375–391. https://doi.org/10.1080/13613324.2019.1679759

[61] Xugang Lu, Sheng Li, and Masakiyo Fujimoto. 2020. Automatic Speech Recognition. In *Speech-to-Speech Translation*, Yutaka Kidawara, Eiichiro Sumita and Hisashi Kawai (eds.). Springer Singapore, Singapore, 21–38. https://doi.org/10.1007/978-981-15-0595-9_2

[62] Alessandro Maffioli, Carolyn Heinrich, and Gonzalo Vázquez. 2010. *A Primer for Applying Propensity-Score Matching*. Inter-American Development Bank. https://doi.org/10.18235/0008567

[63] Joshua L. Martin and Kevin Tang. 2020. Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual" be". In *Interspeech*, 626–630. Retrieved August 7, 2024 from https://www.kevintang.org/Files/publications/MartinTang_2020_HabitualBe_Interspeech.pdf

[64] Ebony Omotola McGee. 2020. Interrogating Structural Racism in STEM Higher Education. *Educational Researcher* 49, 9: 633–644. https://doi.org/10.3102/0013189X20972718

[65] David McShefferty, William M. Whitmer, and Michael A. Akeroyd. 2015. The Just-Noticeable Difference in Speech-to-Noise Ratio. *Trends in Hearing* 19: 233121651557231. https://doi.org/10.1177/2331216515572316

[66] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys* 54, 6: 1–35. https://doi.org/10.1145/3457607

[67] Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "I don't Think These Devices are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans. *Frontiers in Artificial Intelligence* 4: 725911. https://doi.org/10.3389/frai.2021.725911

[68] Sarah Michaels, Mary Catherine O'Connor, Megan Williams Hall, and Lauren B Resnick. 2010. Accountable talk® sourcebook. *Pittsburg, PA: Institute for Learning University of Pittsburgh. Murphy, PK, Wilkinson, IAG, Soter, AO, Hennessey, MN, & Alexander, JF*.

[69] Rebeca Mireles-Rios and John A. Becchio. 2018. The Evaluation Process, Administrator Feedback, and Teacher Self-Efficacy. *Journal of School Leadership* 28, 4: 462–487. https://doi.org/10.1177/105268461802800402

[70] Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, 2765–2768. https://doi.org/10.21437/Interspeech.2004-668

[71] Pranav Narayanan Venkit, Christopher Graziul, Miranda Ardith Goodman, Samantha Nicole Kenny, and Shomir Wilson. 2024. Race and Privacy in Broadcast Police Communications. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2: 1–26. https://doi.org/10.1145/3686921

[72] National Center for Education Statistics. National Teacher and Principal Survey (NTPS) Dashboard. *NTPS State Dashboard*. Retrieved from https://nces.ed.gov/surveys/ntps/ntpsdashboard/Dashboard/US

[73] Mikel K. Ngueajio and Gloria Washington. 2022. Hey ASR System! Why Aren't You More Inclusive?: Automatic Speech Recognition Systems' Bias and Proposed Bias Mitigation Techniques. A Literature Review. In *HCI International 2022 – Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*, Jessie Y. C. Chen, Gino Fragomeni, Helmut Degen and Stavroula Ntoa (eds.). Springer Nature Switzerland, Cham, 421–440. https://doi.org/10.1007/978-3-031-21707-4_30

[74] Martin Nystrand, Adam Gamoran, Robert Kachur, and Catherine Prendergast. 1997. *Opening dialogue*. New York: Teachers College Press.

[75] Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376392

[76] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2024. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. https://doi.org/10.48550/arXiv.2402.17861

[77] Chinasa T. Okolo and Hongjin Lin. 2024. "You can't build what you don't understand": Practitioner Perspectives on Explainable AI in the Global South. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–10. https://doi.org/10.1145/3613905.3651080

[78] Francis A. Pearman and Ebony O. McGee. 2022. Anti-Blackness and Racial Disproportionality in Gifted Education. *Exceptional Children* 88, 4: 359–380. https://doi.org/10.1177/00144029211073523

[79] May Britt Postholm. 2012. Teachers' professional development: a theoretical review. *Educational Research* 54, 4: 405–429. https://doi.org/10.1080/00131881.2012.734725

[80] Samuel L. Pugh, Arjun Rao, Angela E.B. Stewart, and Sidney K. D'Mello. 2022. Do Speech-Based Collaboration Analytics Generalize Across Task Contexts? In *LAK22: 12th International Learning Analytics and Knowledge Conference* (LAK22), 208–218. https://doi.org/10.1145/3506860.3506894

[81] R Core Team. 2020. R: A Language and Environment for Statistical Computing. Retrieved from https://www.R-project.org

[82] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. https://doi.org/10.48550/ARXIV.2212.04356

[83] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

[84] Nadra Rasberry, Joshua Essandoh, Ethan Do, and Ihudiya Finda Ogbonnaya-Ogburu. 2024. Designing Technology to Support the Hospital Classroom: Preliminary Findings. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, 228–232. https://doi.org/10.1145/3678884.3681856

[85] Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1: 41–55. https://doi.org/10.1093/biomet/70.1.41

[86] Russell V. Lenth. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. Retrieved from https://CRAN.R-project.org/package$=$emmeans

[87] Soheila Sadeghi and Chunling Niu. 2024. Augmenting Human Decision-Making in K-12 Education: The Role of Artificial Intelligence in Assisting the Recruitment and Retention of Teachers of Color for Enhanced Diversity and Inclusivity. *Leadership and Policy in Schools*: 1–21. https://doi.org/10.1080/15700763.2024.2358303

[88] Leda Sarı, Mark Hasegawa-Johnson, and Chang D. Yoo. 2021. Counterfactually Fair Automatic Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29: 3515–3525. https://doi.org/10.1109/TASLP.2021.3126949

[89] Majdi Sawalha and Mohammad Abu Shariah. 2013. The effects of speakers' gender, age, and region on overall performance of Arabic automatic speech recognition systems using the phonetically rich and balanced Modern Standard Arabic speech corpus

[90] Ben Shneiderman. 2022. *Human-centered AI.* Oxford University Press.

[91] Isaac Slaughter, Craig Greenberg, Reva Schwartz, and Aylin Caliskan. 2023. Pre-trained Speech Processing Models Contain Human-Like Biases that Propagate to Speech Emotion Recognition. https://doi.org/10.48550/arXiv.2310.18877

[92] Angela D. R. Smith, Alex A. Ahmed, Adriana Alvarado Garcia, Bryan Dosono, Ihudiya Ogbonnaya-Ogburu, Yolanda Rankin, Alexandra To, and Kentaro Toyama. 2020. What's Race Got To Do With It? Engaging in Race in HCI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3334480.3375156

[93] Julie M. Smith. 2024. "I'm Sorry, but I Can't Assist": Bias in Generative AI. In *Proceedings of the 2024 on RESPECT Annual Conference*, 75–80. https://doi.org/10.1145/3653666.3656065

[94] Daniel J Solove. 2005. A taxonomy of privacy. *U. Pa. l. Rev.* 154: 477.

[95] Mengli Song, Andrew J. Wayne, Michael S. Garet, Seth Brown, and Jordan Rickles. 2021. Impact of Providing Teachers and Principals with Performance Feedback on Their Practice and Student Achievement: Evidence from a Large-Scale Randomized Experiment. *Journal of Research on Educational Effectiveness* 14, 2: 353–378. https://doi.org/10.1080/19345747.2020.1868030

[96] Rosy Southwell, Samuel Pugh, M Perkoff, Charis Clevenger, Jeffrey Bush, Rachel Lieber, Wayne Ward, Peter Foltz, and Sidney D'Mello. 2022. Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. *International Educational Data Mining Society.*

[97] Rosy Southwell, Wayne Ward, Viet Anh Trinh, Charis Clevenger, Clay Clevenger, Emily Watts, Jason Reitman, Sidney D'Mello, and Jacob Whitehill. 2024. Automatic Speech Recognition Tuned for Child Speech in the Classroom. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12291–12295. https://doi.org/10.1109/ICASSP48485.2024.10447428

[98] Zaidee Stavely. 2022. Education secretary urges more tutoring, mental health support. *Education secretary urges more tutoring, mental health support.* Retrieved from https://edsource.org/updates/education-secretary-urges-more-tutoring-mental-health-support

[99] Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. The TalkMoves Dataset: K-12 Mathematics Lesson Transcripts Annotated for Teacher and Student Discursive Moves. Retrieved August 29, 2024 from http://arxiv.org/abs/2204.09652

[100] Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. Fine-tuning Transformers with Additional Context to Classify Discursive Moves in Mathematics Classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 71–81. https://doi.org/10.18653/v1/2022.bea-1.11

[101] Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2020. End-to-end ASR: from Supervised to Semi-Supervised Learning with Modern Architectures. Retrieved September 12, 2024 from http://arxiv.org/abs/1911.08460

[102] Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59. https://doi.org/10.18653/v1/W17-1606

[103] Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Interspeech 2017*, 934–938. https://doi.org/10.21437/Interspeech.2017-1746

[104] Louis Tay, Sang Eun Woo, Louis Hickman, Brandon M. Booth, and Sidney D'Mello. 2022. A Conceptual Framework for Investigating and Mitigating Machine-Learning Measurement Bias (MLMB) in Psychological Assessment. *Advances in Methods and Practices in Psychological Science* 5, 1: 251524592110613. https://doi.org/10.1177/25152459211061337

[105] The Economist Newspaper Limited. Why America lost so many of its black teachers. Retrieved July 31, 2024 from https://www.economist.com/democracy-in-america/2019/07/08/why-america-lost-so-many-of-its-black-teachers

[106] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, 1–7. https://doi.org/10.1145/3194770.3194776

[107] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. https://doi.org/10.1145/3531146.3533088

[108] Kimi Wenzel, Nitya Devireddy, Cam Davison, and Geoff Kaufman. 2023. Can Voice Assistants Be Microaggressors? Cross-Race Psychological Responses to Failures of Automatic Speech Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3544548.3581357

[109] Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, Peng Xu, and Pascale Fung. 2020. Learning Fast Adaptation on Cross-Accented Speech Recognition. https://doi.org/10.48550/arXiv.2003.01901

[110] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. Retrieved September 12, 2024 from http://arxiv.org/abs/1910.03771

[111] Yunhan Wu, Daniel Rough, Anna Bleakley, Justin Edwards, Orla Cooney, Philip R. Doyle, Leigh Clark, and Benjamin R. Cowan. 2020. See What I'm Saying? Comparing Intelligent Personal Assistant Use for Native and Non-Native Language Speakers. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–9. https://doi.org/10.1145/3379503.3403563

[112] Fei Xie and Ali Derakhshan. 2021. A Conceptual Review of Positive Teacher Interpersonal Communication Behaviors in the Instructional Context. *Frontiers in Psychology* 12: 708490. https://doi.org/10.3389/fpsyg.2021.708490

[113] Stephanie Yang, Amreen Amin Poonawala, Tian-Shun Allan Jiang, and Bertrand Schneider. 2023. Can Synchronous Code Editing and Awareness Tools Support Remote Tutoring? Effects on Learning and Teaching. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2: 1–30. https://doi.org/10.1145/3610177

[114] Zeina Zrien and Solomon David. 2023. The Impact of School Leaders' Feedback in Enhancing Teachers' Performance towards School Improvement: A Single Case Study among Teachers in a Private School in Dubai.

[115] Privacy & Racial Justice – EPIC – Electronic Privacy Information Center. Retrieved from https://epic.org/issues/democracy-free-speech/privacy-and-racial-justice