U.S. Southerners' Attitudes Towards Algorithmic Analysis of Voice Data for High-Stakes Employment and Education Evaluations

ANDREA GALLARDO*, Carnegie Mellon University, USA
LILY KLUCINEC, Carnegie Mellon University, USA
LUJO BAUER, Carnegie Mellon University, USA
LORRIE FAITH CRANOR, Carnegie Mellon University, USA

The rise of algorithmic decision-making in workplaces and schools has drawn attention to potential benefits and harms for stakeholders. Additionally, uneven performance of language technologies for non-standardized or underrepresented dialects has resulted in calls to engage speakers of those dialects. Our survey study explores the attitudes of 111 American English speakers from the southern U.S. towards voice-specific applications of algorithmic analysis in four high-stakes decision-making use cases: evaluation of candidates in hiring and college admissions interviews, employee performance evaluations based on presentations, and grading of students' pronunciation on English final exams. We elicited descriptions of potential benefits and harms with prompted consideration for dialect and speech and categorized responses inductively as well as deductively, based on existing concepts of sociotechnical harms of algorithmic systems. We contribute focused insights into an understudied but critical area: the performance of language technologies for low-resource dialects. While participants acknowledged the relevance of metrics such as pronunciation and communication skills to employers and schools, many conveyed concerns about evaluations that could negatively evaluate their voices and result in harms such as economic loss. Many participants expressed a preference for human involvement in evaluations. Building on our insights, we make recommendations for researchers and policymakers to mitigate sociotechnical harms of algorithmic evaluations that may affect various linguistic and speech communities.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**.

Additional Key Words and Phrases: natural language processing, voice assistants, automatic speech recognition, fairness, algorithmic harm

ACM Reference Format:
Andrea Gallardo, Lily Klucinec, Lujo Bauer, and Lorrie Faith Cranor. 2026. U.S. Southerners' Attitudes Towards Algorithmic Analysis of Voice Data for High-Stakes Employment and Education Evaluations. *Proc. ACM Hum.-Comput. Interact.* 10, 6, Article CSCW174 (October 2026), 42 pages. https://doi.org/10.1145/3817022

1 Introduction

The use of AI-enabled evaluations in hiring [20], job performance [96], exam grading [134] and college admissions [65] is increasingly popular. Companies targeting managers, teachers, and administrators claim that their software can algorithmically evaluate data subjects on performance metrics or eligibility standards and assess complex human traits based on data like job application

*Work completed while at Carnegie Mellon University.

Authors' Contact Information: Andrea Gallardo, Carnegie Mellon University, Pittsburgh, PA, USA, andystudent0@gmail.com; Lily Klucinec, Carnegie Mellon University, Pittsburgh, PA, USA, laklucin@andrew.cmu.edu; Lujo Bauer, Carnegie Mellon University, Pittsburgh, PA, USA, lbauer@cmu.edu; Lorrie Faith Cranor, Carnegie Mellon University, Pittsburgh, PA, USA, lorrie@cmu.edu.

This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2573-0142/2026/10-ARTCSCW174
https://doi.org/10.1145/3817022

videos [95]. While AI-enabled analysis of voice data to assess human standards of pronunciation [11, 25] and communication skills [125] may be relevant to education and employment contexts and could mitigate human biases [82], prior work shows that algorithmic evaluation of human traits can be unstable or biased [47], pseudo-scientific [95], and lead to unfair outcomes [34].

We focus on potential societal harms of algorithmic analysis of voice data in high-stakes evaluations, which may produce erroneous or biased outputs, particularly for non-standardized language, i.e., language that has not been selected, codified, elaborated, and accepted as the norm [52]. Recent work demonstrates the uneven performance of automatic speech recognition (ASR) technologies for non-standardized English dialects or accents [76, 111, 113]. In the United States (U.S.), sociolinguistic differences have historically overlapped with disparate treatment and discrimination in education and employment.

Since erroneous or biased algorithmic evaluation of voice data could worsen these discriminatory patterns, we intentionally engaged with speakers of non-standardized dialects who may be particularly affected by algorithmic analysis of voice data: Southern U.S. English speakers. Many of these speakers use dialects and linguistic features specific to this region that are distinct from standardized varieties and may be low-resource in the context of ASR technologies like voice assistants (VAs) [68, 69].

Through our survey with 111 Southern U.S. residents who self-identified as speakers of Standardized American English (SAE), Southern U.S. English (Southern), African American English (AAE), or Appalachian English (Appalachian), we investigate attitudes towards general and voice-specific algorithmic analysis. We used vignettes of algorithmic evaluations for decision-making in four high-stakes use cases: hiring interviews, presentation-based job performance evaluations, college admissions interviews, and grading on English final exams. Specifically, our research questions are:

- RQ1: How acceptable do participants find:
  (a) the general use of software for decision-making in four high-stakes use cases?
  (b) the use of voice-specific analysis in these use cases?
  (c) voice data collection by data collectors potentially involved in the four use cases?
  (d) inferences made by voice assistants about speaker attributes?
- RQ2: What benefits and harms do participants anticipate regarding the use of voice-specific analysis in high-stakes decision-making use cases?
- RQ3: How do participants relate their attitudes towards voice-specific evaluations to their experiences with non-standardized dialects, accents, and speech?

We asked participants to rate the acceptability of each use case, data collectors involved in these use cases, and algorithmic inferences based on voice data. We also solicited written responses about their overall attitudes and perceived benefits and harms of voice-specific applications.

Through qualitative analyses of these responses, we provide rich insight into the potential benefits and harms of algorithmic evaluation of voice data in employment and education. While some participants described potential benefits, such as increased objectivity or efficiency, a majority shared concerns about unfair outcomes like discrimination. Some participants also shared positive and negative experiences about their non-standardized dialect or accent, with responses conflicting in their support for or opposition to technology reinforcing or privileging the use of the dominant dialect, SAE, in workplace or educational evaluations. By deductively applying existing definitions of sociotechnical harms of algorithmic evaluations [115] to participants' responses, we provide contextualized examples of harms, such as opportunity loss and cultural erasure.

We encourage future work on accountability and agency to meet societal standards for equal opportunity in employment and education. To provide more equitable access to benefits of algorithmic evaluations, we also suggest improving the performance and robustness of language

technologies for non-standardized dialects, languages, and speech. In light of the risks of further marginalization, highlighted in responses about Appalachian, we encourage research on intentional language dataset development to counter potentially destructive effects of algorithmic evaluations based on standardized language.

2 Background and Related Work

We provide background information on AI-enabled and algorithmic evaluations in high-risk contexts, sociotechnical harms of algorithmic systems, disparate performance of ASR technologies across dialects, and marginalized American English dialects.

2.1 High-Risk Algorithmic or AI Evaluations in Education and Employment Contexts

Previous work across fields demonstrates employment [19, 46] and educational [10, 56] challenges faced by those in diverse social groups. However, the use of algorithmic or AI analysis in professional contexts [34], with its potentially invalid claims to assess human traits [120] and its unstable or biased evaluations [9, 103], has resulted in calls for enforcement of anti-discrimination laws and new regulations to prevent discriminatory evaluations. Public policy documents further support the need to address AI use in high-risk contexts, such as education, employment, health, housing, law enforcement, loan underwriting, justice, and public benefits [22]. In these contexts, decisions or evaluations made using AI or algorithmic analysis could significantly impact people's economic status, health, legal status, and educational opportunities. Furthermore, a recent representative survey of Americans found that most felt uncomfortable with AI-based or algorithmic decision-making about their lives [44].

Our work focuses on education and employment decision-making contexts where voice-specific software could algorithmically evaluate humans in hiring and college admissions interviews, presentation-based job performance evaluations, and students' pronunciation on English final exams. In these scenarios, evaluation may be based on factors such as pronunciation [11] and communication skills [125].

2.2 Sociotechnical Harms of Algorithmic Systems

Our study focuses on harms of algorithmic systems based on dialect, accent, or speech. Prior work indicates that language is an understudied characteristic in fairness literature. In a study on fairness datasets and protected attributes included therein, Simson et al. show that "language" as a feature was available for only three of 36 fairness datasets and used in only one of 233 experiments [116].

Previous work on harms of algorithmic systems conveys how marginalized communities disproportionately experience sociotechnical harms [38, 41]. Prior CSCW literature on social media and content moderation suggests that algorithmic content moderation and "shadowbanning" [27] can limit the visibility [28] and audience for marginalized creators [50] and their ability to monetize social media posts [64]. Additionally, CSCW literature highlights the shortcomings of automating data analysis that may be hard to validate or define, e.g., based on race or skin color [43]. For example, Scheuerman et al. examined facial analysis database documentation, and found that most large databases lack source material for defining race and gender categories and "do not document the process of annotating images," critiquing a data annotation method that relies on "moments of identification" in which annotators, such as crowdworkers, apply labels as though they are "obvious, static, and apolitical," when they are complex social, cultural and historical concepts [112].

Such shortcomings are important to consider as algorithmic voice analyses are used to categorize people based on nebulous and socially determined categories. For example, linguistic variables, such as accent, may be ill-defined or even defined by non-existence. Prinos et al. surveyed ASR literature and showed that accent is often not considered or described in-depth, finding few explicit indications

about "baseline" accents and misleading designations of some speakers as having no accent [93]. Additionally, Grieve et al. highlight issues with language model performance across factors like dialect, register, and time period, arguing that a language model "is inherently modeling the variety of language represented by the corpus on which it is trained, even if that variety of language is unknown and even if that corpus is a poor representation of that variety of language" [42].

To categorize the harms listed by participants in our study, we use Shelby et al.'s taxonomy of sociotechnical harms of algorithmic systems, which was based on a scoping review of 172 research articles [115]. The harms relevant to our study are defined in more detail in Appendix C.

2.3 Uneven ASR Performance for Certain Dialects and Accents

Prior work shows that ASR performs worse for non-standardized English dialects and accents [17, 111, 127], languages without enough structured data to train effective language models ("low-resource languages") [100, 101], and non-standardized dialects and accents of non-English languages [36, 93]. A common standard for measuring accuracy and performance is word error rate (WER), with a higher WER implying worse performance of an ASR model. VAs have been shown to have higher WER for speakers of certain dialects or accents, resulting in failure to understand users' commands [12, 76]. Studies also show disparate WER particularly for U.S. English dialects. Martin and Tang found that the two ASR systems they tested had a higher error rate around linguistic features of AAE [80]. Koeneke et al. found that five off-the-shelf ASR systems show disparities in transcribing identical phrases stated by Black speakers compared to White speakers in an experimental study [66], and they suggest that including AAE in training datasets could reduce disparities. Yet, work by Papakyriakopoulos et al. indicates the potential difficulties or harms of overcorrecting training data gaps [87, 88]. Other work seeks to improve model performance, including developing new forms of measurement [126], and robustness for different dialects [33, 122, 123].

Some research suggests improving ASR by going beyond a traditional computer science approach [77–79]. Markl et al. explored performance disparities in language technologies for marginalised second-language speakers and speakers of stigmatized varieties of British English and the policy implications [76, 78], arguing that "evaluation of ASR systems should … consider user experience in a broader sociolinguistic and social context" [77]. For example, Harrington et al. and Brewer et al. highlight the experiences of Black older adults in the U.S. through interviews about their use of voice assistants, with some feeling the need to "code switch" from AAVE to SAE for a Google voice assistant [16, 48, 49]. Other researchers have studied non-standardized dialect speakers' perceptions of failures [55] and preferences for model alignment [107].

2.4 Underrepresented or Marginalized American English Dialects

While it is widely acknowledged that "the typical NLP [Natural Language Processing] pipeline underrepresents speakers of most of [the 7000 human languages]," some dialects of high-resource languages are also underrepresented in NLP pipelines. This is shown in Wale et al.'s analysis of a "typical NLP pipeline," which found that "even when a language is technically supported, substantial caveats remain to prevent full participation" [131]. Prior work on toxic or hate speech detection also suggests that performance varies based on dialect [40, 110]. Thus, AI systems for high resource languages, like English, may not adequately serve all speakers of those languages. Researchers have noted that "studying speech to improve AI models is not just an engineering problem" and have called for an interdisciplinary approach by engineers, data scientists, and social scientists [114].

We selected participants in the Southern U.S. who may be familiar with non-standardized dialects, including Southern American English, Appalachian English, and African American English (AAE), which are all linguistically associated with this region [5, 7, 69–71, 98]. These dialects may be stigmatized compared to the "Northern" or standardized American dialects, resulting in

Table 1. Recruitment Groups, Inclusion Criteria, and Number of Participants Per Group

Final Group	Recruitment Group	Criteria	Count	Final
AAE	AAE-only	Speaks AAE but not SAE (can also speak Southern)	8	32
	SAE-AAE	Speaks AAE and SAE (can also speak Southern)	24	
Appalachian	Appalachian only	Speaks Appalachian but not SAE (can also speak Southern)	20	30
	SAE-Appalachian	Speaks Appalachian and SAE (can also speak Southern	10	
SAE	SAE-only	Speaks only SAE (not AAE, Appalachian, or Southern)	25	25
Southern	Southern only	Speaks only Southern (not AAE, Appalachian, or SAE)	24	24

Note: Criteria are based on combinations of self-reported dialects spoken: African American English (AAE), Appalachian English (Appalachian), Standardized American English (SAE), and Southern English (Southern).

discrimination [63, 89, 94]. However, even within the Southern U.S., dialects and accents may be considered more or less prestigious based on region [18, 23]. It is also important to recognize the diversity of language within social groups and not essentialize ethnicity, race, or regional origin as representative of a dialect. Some critics argue that grouping the speech of African Americans into one dialect ignores nuances based on many factors [106, 136]. Similar criticisms exist for the general delineation of languages [51, 117].

3 Methods

In this section we describe how we recruited participants, designed and conducted the survey, and analyzed the data. This study was approved by our institutional review board.

3.1 Recruitment

We recruited participants using Prolific's screeners and our own paid screening survey, included in Appendix A. Eligibility criteria included being at least 18, being raised and currently residing in the southern U.S., and speaking American English. Specifically, the following states where southern U.S. dialects are spoken were chosen for recruitment: West Virginia (WV), Kentucky (KY), Alabama (AL), Arkansas (AR), Georgia (GA), Louisiana (LA), Mississippi (MS), North Carolina (NC), South Carolina (SC), Tennessee (TN), and Virginia (VA) [71, 98]. A gender-balanced sample was obtained using Prolific's "Quota Sampling." A second sample with the same criteria but restricted to those over 60 years old was drawn to better balance age in our participant pool.

Participants were paid $0.80 via Prolific to complete our 4-minute screening survey. This survey ensured eligibility criteria was met, checked understanding of what voice assistants (VAs) are, and collected self-reported dialectal background. We specifically asked participants if they spoke any of the following dialects and, if so, how often: African American English (AAE), Appalachian English (Appalachian), Southern American English (Southern), and Standardized American English (SAE).

Through purposive sampling, we recruited participants in six groups: four groups composed of individuals who self-identified as speaking only one of the listed dialects often or always: AAE, Appalachian, SAE-only, and Southern-only, as well as two groups of participants who self-identified as speaking combinations of either SAE and AAE or SAE and Appalachian.

As we were not able to recruit sufficient numbers for the AAE and SAE-Appalachian groups, we consolidated the six groups into four groups corresponding to the four dialects of AAE, Appalachian, SAE, and Southern. These recruitment groups are described in Table 1. Thus, the final AAE and Appalachian groups included participants who identified as speaking AAE or Appalachian often or always, respectively, whether or not in combination with speaking SAE or Southern often or always. Note that no AAE or Appalachian speakers reported also speaking the other dialect often or always. The final SAE and Southern groups each include participants who only identified as speaking SAE or Southern often or always, respectively, and exclude participants who identified as speaking any of the other three dialects often or always.

Participants who met our eligibility criteria, selected the correct definition of a VA, and self-reported speaking at least one of the four dialects "often or always" were selected for the main survey and placed into one of four dialectal groups described above. Of 116 people who took the main survey, we removed five participants from the data sample, three due to obviously AI-generated responses that appeared to be responding to a user prompt, for example, one that began with "Certainly!" and two due to disqualifying criteria that we had accidentally overlooked.

3.2 Survey

Our survey was designed to understand participant attitudes regarding 1) the acceptability of general and voice-specific algorithmic evaluations to assist with educational and employment decision-making, including voice data collection and inferences, and 2) anticipated benefits and harms of such evaluations. Participants answered Likert and free-response questions about their attitudes and potential benefits and harms of the four use cases presented. Participants were paid $3.00 via Prolific to take this 15-minute survey. All survey questions are included in Appendix A.

Acceptability Attitudes (RQ1). We asked participants to rate acceptability on a 5-point Likert scale from extremely acceptable to extremely unacceptable. We first asked them to rate the acceptability of data collection by the following entities related to AI-enabled decision-making in education and employment contexts: companies that build and sell voice assistants (e.g., Google, Apple, Amazon), companies developing AI models (e.g., Open AI, Google, Meta), school districts, potential employers, current employers, and teachers. We then asked about acceptability of the following potential inferences made by voice technologies about speaker features: age, dialect or accent, ethnicity or race, gender, health condition, mood or emotional state, region of origin (i.e., where participants grew up), sexual orientation, and speech or voice disorder.

We then asked participants to rate the acceptability of algorithmic decision-making software in high-stakes education and employment contexts, with two use cases per context, based on prior work (Section 2.1). The order of the use cases was randomized to control for possible order effects. We avoided explicitly asking about AI to not prime them with a potentially controversial term. The four use cases are summarized in Table 2 and the complete vignettes are included in Appendix A.

Each use case consisted of a "General Scenario" vignette in which "software" was used to evaluate data subjects, which participants rated on a 5-point Likert scale from extremely acceptable to extremely unacceptable. We then presented a "Voice Specific Application" in which we described how voice-specific software would use speech data in its evaluations for the same use case. Participants were asked whether the inclusion of the voice-specific application made the scenario *less acceptable*, *more acceptable*, or if it made *no difference in acceptability*.

Following the acceptability questions for each use case, participants were presented with their general scenario answer choice and voice-specific answer choice and asked to "explain [their] overall attitude for this scenario and voice-specific application."

Table 2. Education and Employment Use Cases of Algorithmic Evaluation

Context	General Scenario	Voice-Specific Application
College Admissions Interview	A small liberal arts college uses software to determine eligibility for admission.	The software evaluates audio recordings of admissions interviews to determine suitability for small classes with interactive discussions.
English Final Exam	A high school district uses automatic grading software to grade an annual English final exam.	The software grades English word pronunciation in the spoken component of the exam.
Hiring Interview	A company uses software to rank the top job applicants throughout the job application process.	The software analyzes the job applicant's speech during a job interview to rate how understandable their speech will be to customers.
Job Performance Evaluation	A company uses software to evaluate current employees' job performance.	The software analyzes speech during presentations to evaluate public speaking and communication skills.

Note: Four education and employment use cases of AI were shown to all participants. Participants rated the acceptability of each general scenario and then rated whether they found the corresponding inclusion of a voice-specific application more or less acceptable for that use case.

Potential Benefits and Harms (RQ2). For each use case, the two final questions asked participants to list at least one potential harm and at least one potential benefit "for this use case of voice technologies," or, if they did not believe there were potential harms or benefits, to explain why.

Speech-Specific Attitudes and Experiences (RQ3). As part of our introduction to the section in which we presented participants with four use cases, we included a note, "We're especially interested in your attitudes, opinions, or experiences relating to the southern United States and dialects or accents found in this region." Through this prompt, we hoped to elicit participants' attitudes and experiences related to non-standardized dialects, accents and speech.

3.3 Data Analysis

The first and second authors qualitatively analyzed participants' acceptability explanations and potential benefits and harms using emergent coding to inductively surface themes specific to the data as well as a priori coding of algorithmic harms based on existing work. Quantitative information is provided about the themes identified for descriptive purposes only.

In our emergent or bottom-up thematic coding analysis, the first two authors double-coded responses to all free-text responses (overall attitude, benefits, and harms), and each researcher labeled responses and grouped them thematically. Through review and discussion, we resolved disagreements and iteratively refined the code categories based on emerging themes. We organized themes that arose in our analysis of participants' self-reported overall attitudes into four top-level categories of 1) negative opinion or critique, 2) positive opinion or support, 3) humans and technology, and 4) speech-related comments, each with specific subcodes developed throughout the coding process. Our emergent code book with the overall attitudes, harms, and benefits codes can be found in Table 4 in Appendix B.

Additionally, we used Shelby et al.'s taxonomy of sociotechnical harms of algorithmic systems [115] to deductively label all participants' responses listing potential harms of voice-specific

algorithmic evaluations. First, the first two authors applied the sociotechnical harm categories to the inductively coded harms, identifying relevant thematic overlap. Following this initial deductive analysis, the second author deductively labeled all participants' harm responses with corresponding sociotechnical harms, and the first author reviewed the labels, providing suggestions for changes or additions. All ambiguities and disagreements were resolved iteratively through comments and discussion. The sociotechnical harms that we used in our analysis included the Allocative Harms of *economic loss* and *opportunity loss*, Interpersonal Harms of *diminished health and well-being*, *loss of agency or control*, and *privacy violations*, Quality-of-Service Harms of *alienation* and *service or benefit loss*, and Representational Harms of *alienating social groups*, *erasing social groups*, and *stereotyping*. We provide their definitions in our deductive analysis code book in Appendix C.

3.4 Limitations

As other scholars have done [13], we consider our participants to be a non-probability sample, so their responses are not meant to generalize to broader population trends, but rather to provide insights into complex social phenomena. While our sample of 111 participants may be a relatively small sample size for a crowdsourced survey, the primary goal of this qualitative exploratory work is to better understand and characterize attitudes towards algorithmic and AI analysis of voice data from a subset of the U.S. population. We report numbers for descriptive purposes to highlight differences in the participants' responses only, as the assessment of statistical regularities for this context is beyond the scope of this work. We have little reason to believe that the outcome variable of acceptability is operationally consistent across demographic factors, such as dialect.

Participant categories were defined based on the self-identification of participants as speakers of certain dialects, and we did not "validate" their dialects. We recognize that these categories are not precise and may not fully capture participants' dialectal characteristics. We also acknowledge that Southern English may be better defined as a group or family of dialects that could overlap with some Appalachian or AAE dialects, as shown by participants who identified as speaking multiple dialects. We found it difficult to recruit speakers of Appalachian and AAE dialects who did not also report speaking SAE.

Given our focus on technology attitudes rather than linguistic analysis, we did not study the dialectal features within responses and did not observe any obvious examples of dialectal differences. Although participants did not offer language-specific insights, a few referenced specific regions and populations associated with the dialects (Section 4.5). The short, typed text responses may have contributed to the lack of obvious dialectal variation.

Additionally, our sample was relatively young, and most participants did not report having problems being understood by voice assistants or humans. Thus, our study did not capture many responses from older adults and people who report being misunderstood most of the time. We hypothesize that a study with people who have difficulty making their speech understood would be valuable for understanding potential benefits and harms of voice-specific algorithmic evaluations, as their opinions and experiences may be considerably different from those of our participants, and their insights into specific communication needs and strategies could inform the development of tools and policies to ensure their inclusion in education and employment opportunities.

Finally, to provide a positionality statement regarding dialect, our team of four researchers used SAE while conducting the described research. Our research team members speak Northeastern American English dialects (including Pittsburgh English and New York English), as well as dialects of Croatian, French, and Spanish.

4 Results

Below we summarize our survey results. We first characterize the final participant sample using self-reported demographic and prior experience information (Section 4.1). We also summarize their Likert scale responses about the acceptability of algorithmic evaluation software in each use case, data collection by entities relevant to those use cases, and potential inferences about users of voice assistants (Section 4.2). We then present our qualitative analysis of participants' free-text responses stating their attitudes and anticipated benefits and harms across all four use cases, organized into broader thematic categories of: potential benefits (Section 4.3), potential harms (Section 4.4), and speech or dialect specific concerns (Section 4.5). When quoting participants, we specify the participant number and use case. Counts of thematic codes across the four dialect groups were compared and did not indicate any stark patterns, but we provide some notable highlights in Figure 1, Section 4.2.3 and in Appendix B.

4.1 Participant Demographics and Prior Experience

All 111 participants were American adults from the southern U.S. who speak American English as their first language, as required by our recruitment criteria (Section 3.1), and Table 1 includes the final participant counts per dialect. Most participants identified as one of two ethnicities, White ($n = 56$) or Black or African American ($n = 53$). A total of 64 participants identified as female, 44 as male, three as non-binary, and two as transgender. Most participants were relatively young, with 62 participants reporting being up to 40 years old. We provide more demographic information and summary proportions in Appendix E.

Most participants had experience with voice assistants, with only three reporting having never used a voice assistant. In free-text responses, 15 participants mentioned prior experience with or knowledge about algorithmic evaluation applications similar to those in our study's four use cases, These mentioned applications included algorithmic review of resumes or videos submitted by job applicants, automatic exam grading, and ranking of college applicants. P33 wrote,

> I've had companies demand I do extremely similar stuff, even talking about myself on camera without any sort of prompts or questions, only to have my applications denied within minutes and the video never even landing on the hiring admission's radar. (P33 Hiring)

Some participants remarked on the use of software in job performance evaluations. P110 suggested that "using software to track work performance is pretty standard these days." When taking exams, P14 and P46 recalled negative experiences with their writing being "misjudged" by an algorithm.

4.2 Acceptability Attitudes (RQ1)

We summarize the 111 participants' Likert scale responses to questions about the acceptability of the use of AI-enabled software in four decision-making use cases, data collectors, and inferences made by voice assistants (VAs).

4.2.1 Unspecified "Software" Acceptable but Analysis of Voice Data Less Acceptable. Over half of participants found the general use of algorithmic evaluation software *acceptable* (somewhat or extremely) in hiring ($n = 63$), job performance evaluation ($n = 69$), and college admissions ($n = 69$) use cases. For the final exam use case, just under half ($n = 52$) found this scenario *acceptable*. For voice-specific applications across all four use cases, less than a quarter of participants rated the voice-specific application *more acceptable*, with the rest being roughly split between *less acceptable* and *makes no difference*. These results are shown in more detail in Appendix F.

4.2.2 Data Collectors: Companies More Acceptable Than Teachers, Schools or Employers. When asked to rate the acceptability of data collectors involved in the four use cases, AI companies (n=66) and VA companies (n=58) were the most commonly rated as acceptable, followed by teachers (n=48), current employers (n=44), school districts (n=31), and potential employers (n=29).

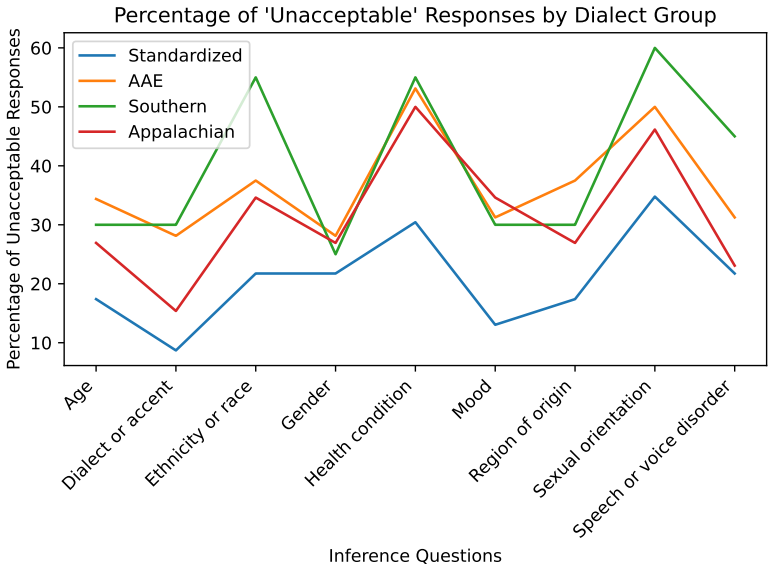

Fig. 1. Line graph showing percentage of "unacceptable" ratings for the given categories Four lines represent the four dialect groups (AAE, Appalachian, Southern, and Standardized (SAE)). The self-identified SAE group had the least number of people find it unacceptable to make inferences for all categories.

4.2.3 Some Inferences More Acceptable Than Others. When asked about the acceptability of inferences made by VAs, inferring dialect or accent was found *acceptable* by the most participants ($n = 76$), followed by gender ($n = 67$), mood or emotion ($n = 62$), age ($n = 61$), region of origin ($n = 59$) and speech disorder ($n = 58$). Less than half of participants found inferring ethnicity or race ($n = 43$), health conditions ($n = 25$) or sexual orientation ($n = 25$) *acceptable*.

For eight of nine questions about inference making, the SAE group had the least number of people and lowest percentage of their group who found the inference making unacceptable, as shown in Figure 1. However, it is important to note that these are small and non-representative samples of 32 or less people per group, with the following group sizes: Southern: 24, SAE: 25, Appalachian: 30, AAE: 32.

4.3 Anticipated Benefits: Efficiency, Objectivity, Improving Skills (RQ2)

Participants suggested that algorithmic evaluations could increase efficiency ($n = 82$) by saving time and labor, increasing speed, or generally streamlining processes. This could relieve teachers, hiring managers, and administrators of tasks such as grading or reading applications. Participants ($n = 32$) also suggested a potential benefit of algorithmic evaluations mitigating human biases by providing objective, fair, and consistent assessments or by not showing bias or favoritism. Additionally, participants suggested that algorithmic evaluations could help employees or students improve skills ($n = 30$), such as providing feedback to students on their pronunciation. P38 suggested "it

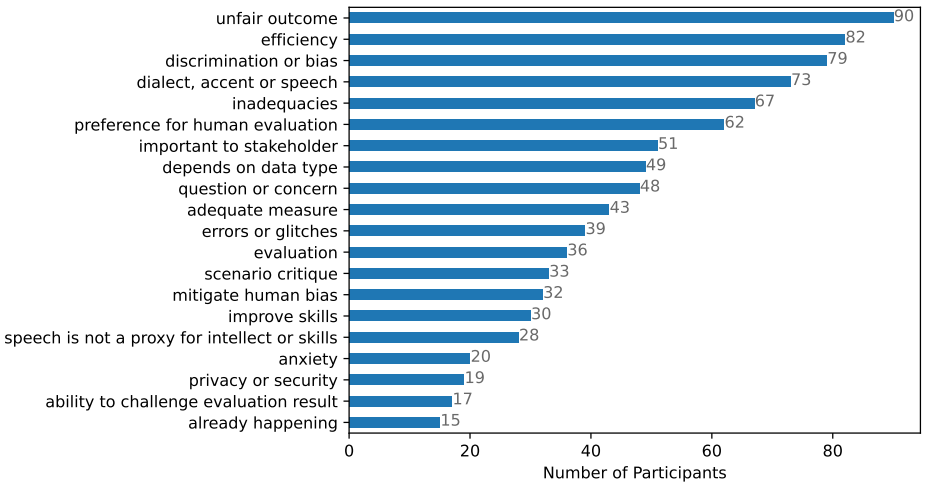

Fig. 2. Count of themes, counted once per participant across all four use cases. See definitions in Appendix B.

would be beneficial for students who are mispronouncing words. They can have it played back to them to see what they sound like" (Exam). Other participants noted how the software could broadly improve communication skills across both the employment and education contexts.

While not necessarily a benefit, we note that when discussing benefits, some participants expressed confidence in the ability of software, computers, and AI to evaluate understandability, communication skills, and pronunciation ($n = 43$) adequately or even better than a human could.

4.4 Anticipated Harms (RQ2)

We summarize the harms listed by participants using the categories of sociotechnical harms of algorithmic systems developed by Shelby et al., focusing on allocative, quality-of-service, interpersonal, and representational harms. First, we present allocative harms corresponding to *opportunity loss* and *economic loss* (Section 4.4.1). We then present quality-of-service harms of *service or benefit loss* due to uneven system performance (4.4.2), followed by interpersonal harms of *loss of agency or control*. We also present emotional and psychological harms of *alienation* and *diminished health and well-being* (Section 4.4.4), and interpersonal harms of *privacy violations* (Section 4.4.5). Finally, we present representational harms of *alienating, erasing,* and *stereotyping social groups*. Definitions and examples of each harm are provided in Appendix C. Table 3, below, lists total unique participant counts of sociotechnical harms across all scenarios, and Appendix D includes counts per scenario.

4.4.1 Allocative Harms: Opportunity Loss and Economic Loss in Employment and Education. Among all responses, a total of 90 participants mentioned potential unfair outcomes or negative decisions for each use case: rejection of job or college applicants, lower final exam grades (and thus a lower overall grade), and negative job performance evaluations resulting in the denial of pay raises. Across the four use cases, 79 participants specifically named discrimination or bias, and 67 participants indicated a basis for discrimination, such as accent or disability. The use case that garnered the most comments about discrimination or bias was hiring ($n = 52$), followed by exam grading ($n = 49$), college admissions ($n = 47$), and job performance evaluation ($n = 35$). For example, P34 expressed concern about how voice-specific evaluations may "lead to less job opportunities for people who have accents or speech impediments" (Hiring). Negative outcomes not about

Table 3. Unique Participant Counts Per Sociotechnical Harm Category and Subcategory

Harm	Harm Sub-Type	Unique Participants
Allocative Harm	Economic Loss	63
	Opportunity Loss	25
Quality-of-Service Harm	Service or Benefit Loss	46
	Alienation	8
Interpersonal Harm	Loss of Agency or Control	11
	Diminished health and well-being	3
Representational Harm	Erasing Social Groups	11
	Stereotyping	4
	Alienating Social Groups	1

discrimination included those resulting from error, malicious intent, inadequate training of the system, or simply unspecified "unfair" outcomes.

Adopting the categorization in Shelby et al., we consider the loss of job opportunities and discrimination in employment contexts to be *economic losses*, and the loss of educational opportunity and discrimination in education to be *opportunity losses*, implying loss of access to education. We also introduce additional concepts of economic loss, namely the loss of academic capital due to lower grades, the loss of labor, time or resources for an employer, and the loss of students for a university.

Economic Loss. Across all use cases, 63 participants listed harms of *economic loss*. More participants provided economic harms for the use case of hiring ($n = 47$) than for job performance evaluation ($n = 30$), pronunciation evaluation for a final exam ($n = 28$), or interviews for college admissions ($n = 11$). For both of the hiring and job performance evaluation use cases, participants listed harms of the loss of labor, time or resources for employers (hiring: $n = 16$, job eval.: $n = 3$), the loss of opportunities in employment contexts (hiring: $n = 31$, job eval.: $n = 6$), and discrimination (see Section 4.4.1). For example, P95 wrote, "There is potential harm to the company by not having the best candidates make it though the initial review process by the software" (Hiring).

For the job performance evaluation use case, 21 participants suggested potential loss of income or finances due to not receiving a raise or getting fired as a result of the evaluation. P34 wrote, "It could lead to reduced wages for employees who have heavier accents, cultural dialects, and speech disorders." For exam grading, 28 participants suggested potential harm to a student's grade due to erroneous, inaccurate, unfair, or discriminatory evaluations. P11 wrote, "Harms could massively hurt someone's grade, GPA, and future credibility if they failed this course due to having an accent or speech impediment," suggesting that harm to a student's GPA could be costly. Regarding college admissions, 11 participants suggested the college would potentially lose out on "the best," "qualified," and "highly intelligent" students. P50 wrote that the college could "develop a pattern of accepting ineligible people." Such responses suggest a potential loss in value in terms of talent.

Opportunity Loss. For college admissions interviews, 24 participants mentioned the opportunity loss of access to education (denied admission). P1 wrote, "People with speech disorders or who are not fluent in English may be unfairly excluded." P46 wrote that applicants could be "wrongly disqualified for having accents, using colloquialisms or whatever that is common in their area

but not in the wider US, not being skilled speakers despite being very intelligent in other ways, etc." Regarding algorithmic exam grading, three participants mentioned a general loss of academic opportunity, distinct from the economic loss of a lower GPA, suggesting that lower exam grades could impact a student's "future."

Note that we treat opportunity loss in employment contexts as economic loss (see paragraph above) based on prior work [115]. For both education use cases, participants also mentioned the harm of discrimination, presented as an allocative harm in Section 4.4.1.

4.4.2 Quality of Service Harms: Service or Benefit Loss. Almost half of participants ($n = 46$) wrote about how evaluations based on voice data could disproportionately fail or perform worse for certain people, e.g., those with Southern accents or speech disorders. Such disparate performance and failures correspond to the harms of *service or benefit loss*. For example, P59 suggested that uneven performance of a scoring system for job applicants could result in a loss of benefits for people with a certain dialect or accent: "the voice assistant may hear a voice that they haven't yet heard, and then give that candidate a bad score. When in reality, that candidate is great for the role, but the voice assistant just hasn't heard that specific dialect or accent before" (P59 Hiring). P34 wrote that a speech disorder or "heavy accent" should not negatively affect a student's grade, which would be a loss of the benefit of automatic grading, suggesting that skewed performance by algorithmic grading systems "may also disproportionately affect foreign students or students with different cultural background" (Exam).

4.4.3 Interpersonal Harm: Loss of Agency or Control. Regarding harms of reduced human autonomy resulting from the use of algorithmic evaluations, 11 responses about harms evoked the interpersonal harm of *loss of agency or control*, where evaluators such as teachers or managers may lack agency to apply their human discernment, modify results or errors, or conduct tasks that humans are more equipped to perform. In particular, participants wrote about the 1) loss of human authority to manage or contest evaluations, and 2) changes in speech styles to conform to expected input.

Across all responses, over half of participants (n=67) noted inadequacies and expressed skepticism about technology's ability to perform nuanced evaluations, as well as a preference for human evaluation (n=62) in the high-stakes decision-making use cases described in our survey.

Loss of Human Authority to Manage or Contest Evaluations. Six participants mentioned the harm of the potential absence of authority figures who would normally be involved in evaluations, such as teachers or managers. P55 wrote about the harm of "removing humans from the process altogether," adding, "I'm very wary of just putting programs in charge of making decisions about people without careful oversight" (Admissions).

Change in Speech to Conform to Expected Input. Five participants wrote about the harm of algorithmically informed conformity of speech, suggesting shifts in speech toward more robotic or unnatural sounding speech. P19 wrote that a "harm would be taking away more of the casual element present in some businesses that I enjoy and trying to force everything to be robotic style professional" (Job Performance). Similarly, P96 wrote that "people will speak like a robot" (Admissions). P33 suggested that job applicants might give "unnatural responses" that are scripted (Hiring). P101 suggested that "[i]f the student does not speak exactly like the computer their grade will be impacted" (Exam).

In responses about potential benefits, a few participants wrote about speakers of non-standardized dialects changing their dialect or accent to conform with SAE. P71 suggested that the English exam grading program could help students "get better at faking their accent so that in interviews and phone calls, judgmental people don't automatically assume folks from this area are uneducated" (Exam). P60 suggested that college applicants with accents would benefit from voice evaluations,

because it "would encourage such students who have deep accents to add another dialect to their portfolio, 'proper English,' which I admit is understandable to almost anyone who know[s] English" (Admissions). However, P72 provided insight into the challenge of adopting a different dialect, stating, "[I]t's likely that the way we speak would be a large part of the results. I have spent years trying not to sound like I was from southern [West Virginia]. Yet, I still do" (Job Performance). Thus, even P72's voluntary efforts to make changes conforming to SAE were unsuccessful.

Inimicable Human Capabilities. Participants named human skills that would be hard to imitate and tasks that should be human prerogatives, such as those where humans should have the upper hand. These included tolerance for differences, subjectivity, detecting nuance in pronunciation, reading body language or tone, having empathy for others, and taking additional factors into consideration. For example, P38 highlighted that "a foreign born student may technically not speak as well as a US one, but a teacher could give them the same grade taking it into account. It is relative to each student" (Exam). P27 asked, "Why use software when managers can use the power of observation to determine how well an employee is doing[?]" (Job Performance). Similarly, P36 wrote, "It can't observe things like a worker staying after hours to finish a task or one that helps others with their work" (Job Performance).

Some responses evoked the notion of dignity and a wish to be seen or heard by another person, highlighting the humanity of the act of giving each other time and consideration. P33 criticized the "societal problem of AI software across the board" and the absence of human review for applicants who submit videos for college applications, calling it "completely dehumanizing, because it shows the applicant that the admissions office can't even be bothered to speak with the applicants directly." P6 suggested that AI could not sympathize or empathize, writing, "[I]t is a cold lifeless ai that cannot make moral judgments and can only go by some algorithm" (Hiring). P87 wrote that they "think it is just unfair to a student for them to have to be graded by something without a heart, soul, or tongue to speak for itself" (Exam).

Criticisms of technology's inadequacies included an inability to accurately interpret or quantify human emotions, body language, differences in speech due to dialects and accents, acceptable differences in speaking style and presentation, or other factors. P99 suggested that human reviewers "will be more aware of biases," adding, "[I]f say you were talking to someone of your own race, it may or may not be clear what you're saying to the computer, but a human will understand more" (Hiring). P55 also expressed doubt that an algorithm would be able to understand "different ways of speaking" and that "it seems like a lot of minorities would be negatively impacted" (Exam). Algorithmic evaluation inadequacies also included the inability to understand charisma, personality, or potential for growth. For example, P38 wrote that technology used in hiring interviews "can't understand charisma. ... The way one's voice makes people feel is important." P113 also indicated that, for the hiring use case, algorithms "may overlook vital intangible qualities like personality fit or potential for growth," implying that human reviewers could account for these qualities in their hiring decisions.

4.4.4 Alienation and Diminished Health and Well-being. In their responses with potential harms of algorithmic evaluations, a few participants described negative emotions and psychological harms. The Quality-of-Service harm of *alienation* was evoked by eight participants who listed expressions of adverse emotions in response to all four scenarios, including anxiety, fear, hurt feelings, overthinking, and feeling self-conscious, put on the spot, cheated, or uncomfortable. Three participants gave examples of the Interpersonal Harm of *diminished health and well-being*, noting potential psychological effects of stress (Job Performance) and decreased self-esteem (Exam).

4.4.5 Privacy Violations. A total of 19 participants mentioned potential data privacy and security harms or concerns, such as malicious uses of voice data, nonconsensual data disclosure or collection, data breaches, and data storage concerns. For example, P108 believed that voice recordings could be"used for other purposes later that employee did not authorize" (Job Evaluation). P27 shared similar concerns for the hiring use case: "I can only imagine some higher up having some nefarious plan on how to monetize the recorded interviews." Other respondents pointed out issues related to data storage, such as P22, who asked about the college admissions interview, "What if you [don't] get admitted? Where does your voice recording go?" P108 was concerned that recordings would be stored "too long, and [be] susceptible to a breach" (Hiring). P95 highlighted a risk of unwanted disclosure and use of personal information based on voice data, writing that algorithmically grading pronunciation "could result in the person's race or ethnicity being discovered, risking possible racism from the grading of the exam."

4.4.6 Representational Harms: Alienating, Erasing, and Stereotyping Social Groups. We identified examples of the representational harms of *alienating, erasing, and stereotyping social groups*.

Alienating Social Groups. One response specified a harm of *alienating social groups*, identifying a potential failure (or opportunity) to recognize the relevance of a social group. P72 wrote that they "would hope the employer is hiring people from the area to work in that area." That is, they would hope the company would recognize relevant local employees who speak an Appalachian dialect from southern West Virginia: "I am from southern WV and can understand their speech and they can understand me, but I have met people from other parts of the country who did not understand me and ones I could not understand. I don't think an AI could get that difference."

Erasing Social Groups. A few participants ($n = 11$) wrote about the harm of *erasing social groups*, by which algorithmic evaluations would fail to recognize or parse regional accent, dialect, or speech types, and would thus create or perpetuate their systematic absence, effectively erasing them from purview of the algorithmic system. P71 wrote, "Our accents are part of who we are. Asking us to speak more of a northern or blank accent is unfair and extremely difficult and it bleaches out our culture and way of life" (Exam). P30 described a harm of "[a]ccents and regional pronunciations being overlooked or misjudged" (Exam). P114 wrote that the "software would be over generalized and most likely would not account for things like regional accent, birth nation, speech disorders, etc." (Exam).

Stereotyping Social Groups. Participants conveyed the potential for algorithmic evaluations to socially exclude or reinforce false perceptions about social groups, including speakers of Appalachian. Among responses listing potential harms, five participants wrote about the harm of stereotyping, with P17 and P26 suggesting that evaluations might reflect negative social stereotypes about Southerners and with P96 and P104 suggesting algorithmic output could reflect stereotypical pairing of attibutes, such as using voice to infer race or age. P17 said she had a 3.8 GPA in high school but shared how various personal factors, such as her gender, "would be used to assume I have lower intelligence level automatically, especially my accent" (Admissions). P72, who who self-identified as speaking Appalachian often or always, wrote, "I was once told they were surprised I wore shoes" (Admissions), suggesting that algorithmic evaluations may continue to impose existing stereotypes.

Among all responses, about a quarter of participants (n=28) objected to the use of speech as a proxy for skills or intelligence. Participants suggested that accent or dialect could be associated with negative perceptions about intelligence, articulateness, or acceptability. For example, P10 wrote, "i am worried that they may think i am not smart just because of my accent" (Admissions). P49 wrote, "People with unacceptable dialect could be discriminated against," suggesting that algorithmic evaluation would reinforce stereotypical pairing of unacceptability (Job Performance).

Participants who objected to voice-specific evaluations to gauge skills or intellect suggested that speech does not represent the qualities most important to the evaluation. For example, P17 wrote that, for work presentations, "the actual presentation shouldn't matter, the results it presented as a result of the projects an employee worked on should be more important" (Job Performance). P82 wrote, "speech does not equal performance" (Job Performance). P72 expressed that they "have been made fun of for my accent my entire life. Yet, I have found that intelligence has nothing to do with accent. If an AI assistant judges based on accent, it needs to be scrapped" (Admissions). P19 suggested that "someone like [Stephen] Hawking," a renowned physicist and author who had a motor neuron disease that affected his speech, "would end up being declined by the software, although he's one of the smartest humans to ever live" (Admissions).

4.5 Attitudes and Experiences Related to Speech and Dialect (RQ3)

In addition to attitudes relating to language or speech, such as concerns about allocative harms (Section 4.4.1), algorithmically informed changes in speech styles (Section 4.4.3), and stereotyping (Section 4.4.6), participants also provided insights into experiences as or with speakers of non-standardized dialects, highlighting both negative and positive aspects of speaking non-standardized dialects. For example, P81 wrote that "AI voice software has a problem with my southern/Appalachian dialect. But, I've worked in customer service for 20 years and have found customers can understand me and many say they love my accent" (Hiring). P71 noted the dialectal variety of Southern English, for which there is "a strong Scottish and Irish influence in some places and more of a drawl in others. These are fine and beautiful. If it's judging on the content of their speeches and not the twang, then it's okay" (Job Performance). They also suggested that voice-specific algorithmic evaluations of hiring interviews "might be useful if you need people to take calls from a specific area of the United States," suggesting that such evaluations could promote regional dialects by inviting their usage in workplaces, as opposed to encouraging the use of SAE.

4.5.1 Support for Evaluations Based on Standardized Dialect. In contrast to responses criticizing voice-specific evaluations (Section 4.4.6), some responses expressed support for evaluations based on SAE. For example, P53 wrote that voice-specific algorithmic applications "could definitely help weed out employees with sloppy speech patterns, or ones that are too quiet to understand" (Hiring). P87 wrote that a benefit would be that "there will be people working for the company that speak clearly and are well understood by the larger population of people where this company is providing it's services" (Hiring). In terms of a company deciding raises based on job performance evaluations, P69 wrote that performance includes "communicating with proper speech," and that voice-specific analysis of presentations "should give them a more accurate view of there employees." Addressing exam grading, P99 wrote, "I don't believe there will be any harm because in those scenarios you're supposed to talk proper even if you do have a slang or accent." Similarly, P68 wrote, "if you are using words, you should know how to pronounce them" (Exam). P25 wrote that potential college students' "[s]peech should be clear and concise to be admitted" (Admissions).

5 Discussion

Prior work has suggested improving measurement of fairness-related harms [21, 37, 132] and increasing epistemological rigor by filling in gaps regarding the knowledge and values underpinning datasets and machine learning models [85]. Human-centered and participatory research can provide epistemological rigor by producing novel and culturally relevant data, including language, narratives, and values that fill knowledge gaps. While studies engaging human participants, especially those from hard-to-recruit populations, may not always result in systematically operationalizable

values, the cultural, linguistic, and epistemological exchanges they create can provide ontological awareness regarding knowledge gaps and forge paths into complex problem spaces.

We contribute to this effort to increase epistemological rigor by engaging speakers of standardized and non-standardized American English dialects and obtaining rich data about their perceptions of potential benefits and harms of algorithmic evaluation of voice data in four high-stakes use cases: college admissions, final exam grading, hiring, and job performance evaluations. Building on Shelby et al.'s taxonomy of sociotechnical harms of algorithmic systems [115], we characterized what some harms look like in employment and education contexts. By eliciting descriptions of potential harms with prompted consideration for dialect and speech, we contribute focused insights into an understudied but critical area: the performance of language technologies for non-standardized languages and speech. By categorizing these contextualized harms, we hope to assist researchers who can help measure and mitigate harms resulting from uneven performance, which may affect various linguistic and speech communities.

Below, we discuss participant responses about reinforcing standard language ideologies, improving system performance for low-resource dialects, and human accountability.

5.1 Reinforcing Versus Challenging Standard Language Ideologies

The risks of algorithmic evaluations further marginalizing speakers of non-standardized dialects through allocative harms and of alienating, erasing, and stereotyping social groups were particularly apparent in responses that evoked existing power dynamics, social stigmas, prejudices, and negative personal experiences related to speech. Some responses expressed support for evaluations based on a standardized dialect (Section 4.5), emphasized "proper" and "clear" English, assumed widespread acceptability of SAE, and depicted non-standardized language as potentially "sloppy," poorly enunciated, or stigmatized. Other responses suggested potential harms of perpetuating the stereotyping and alienation of social groups, such as Appalachian speakers, and thus reinforcing negative stigma associated with certain dialects or regions. Such responses evoke what some scholars call a "standard language ideology," or a pervasive bias for an idealized and homogeneous standard language, that perpetuates the dominance of a particular language variety and the marginalization of other varieties of the same language [74].

We found that participants imagined that "algorithmic systems' enactment of power dynamics [53, 86] can function as a minoritizing practice [24] through which unjust social hierarchies are reinforced" [115]. A few participants raised concerns about the potentially destructive effects of language standardization, including P71, who wrote that "Asking us to speak more of a northern or blank accent is unfair and extremely difficult and it bleaches out our culture and way of life" (Section 4.4.6). Such responses about the erasure of social groups due to algorithmically enforced demands to conform to normative language standards are valuable evidence of resistance to sociolinguistic homogenization. Future research investigating the ways in which standard language ideologies are embedded in language technologies [137] could help detect misalignment with speakers of regional and minoritized dialects and inform how they develop and use language technologies.

Some participants wrote about positive aspects of using non-standardized dialects, such as regional specificity and charm. In the spirit of *unmaking* literature and work that emphasizes joy, cultural assets, and "the everyday" experiences of marginalized people rather than only focusing on harms [14, 32, 128], we highlight these positive depictions and encourage future research that engages people eager to promote their language. Prior work on user-centered evaluations of language models [29, 73, 138] has emphasized the need to engage stakeholders to bring awareness to features, opportunities, and concerns that previously may not have been considered. Future work can counter the harm of cultural erasure through language documentation and preservation [91,

130, 133], as well as human-centered design that engages speakers of non-standardized dialects using participatory, ethnographic, and archival [57, 83] methods to inform the design, or rejection, of language technologies [139]. Given the dearth of data for some languages, we also look forward to research on community-led dataset creation and evaluation. We encourage researchers in this space to source and use data ethically, respecting cultural and intellectual property norms of stakeholders such as fluent speakers and linguists who developed data corpora.

5.2 Improving Performance and Robustness

Speakers of low-resource dialects may not equitably experience the benefits of algorithmic systems that use voice input, given potentially disparate error rates and unstable performance. To provide more equitable access to the benefits of algorithmic evaluations based on voice data, we encourage research that improves the performance and robustness of technologies for minoritized dialects and accents. While prior work addresses gaps in NLP performance for "low-resource" languages [2, 58] or regional accents and dialects [4, 6, 59], and harnesses voice technology to better understand and correspond with minoritized dialect groups [39, 92], NLP research can be limited by dependence on fixed datasets and limited capabilities to conduct the community outreach, trust building, and linguistic fieldwork needed for data collection. To improve performance for areas where such approaches are limited by resource constraints, we suggest long-term interdisciplinary research collaborations at the intersection of applied NLP, human-centered technology design, linguistics, and anthropology, to adequately and ethically collect and use data to make technical performance or robustness improvements. Such collaborations could develop, for example, linguistically and culturally attuned audits to evaluate how a system may fail, throw errors, or produce consistently disparate outputs based on inputs developed by a fluent speaker of a low-resource dialect, such as phonetic differences, dialect-specific morphology, or lexical semantic variations.

5.3 Accountability

Below, we highlight two key takeaways from our results regarding human accountability, with implications for research and policy.

5.3.1 Interrogating Pseudo-scientific Claims. Our results convey participants' opposition to using voice analysis to gauge skills or intellect. Some scholars have suggested that machine learning is inciting a new age of phrenology, with biometric data being used to draw conclusions about people, sometimes erroneously [45, 72]. To promote accountability for pseudo-scientific publications and products, we recommend that researchers question and investigate potentially false claims correlating biometric data with ambiguous variables [47, 120], such as personality and hireability [103, 104].

5.3.2 Human Involvement and Agency. In our study, 62 participants indicated a preference for continued human involvement in algorithmic evaluations (Section 4.4.3), and 67 participants expressed skepticism about algorithmic systems' abilities to evaluate applicants, workers, and students, A lack of human involvement could exacerbate the potential sociotechnical harms discussed above, such as economic loss and erasing social groups. Therefore, we recommend human computer interaction research on how decision makers exercise agency and accountability in algorithmic evaluation use cases. Such work could explore mental models, protocols, and technology design that could prevent a loss of agency or control. For example, studies on managers could explore how they "use and make sense of AI systems, and how this affects their discretionary decision making" [118, 119]. as well as "how hiring managers, workers, applicants, and others within an organization interact with each other and with hiring tools" [120]. To challenge potentially coercive use of algorithmic

evaluations in contexts where power imbalances are great, such as hiring, we encourage regulatory frameworks that ensure the ability to opt out without opportunity loss.

5.3.3 Policy Recommendations. We support the further development and enforcement of policies that address sociotechnical harms such as economic loss from discriminatory evaluations. Various countries and regional governments have recently passed regulations addressing high-risk AI applications [102]. In the U.S., cities such as New York City and the states of Utah and Tennessee have passed laws regarding automated employment decision-making tools [135]. In Illinois, the Illinois Human Rights Act was extended to include protections against AI-enabled discrimination based on racial data or common proxies like ZIP codes in several employment use cases, including those related to hiring and job evaluation processes [1, 35]. In Colorado, the Colorado Artificial Intelligence Act (CAIA) provides standards of use and disclosure when AI is involved in high-stakes contexts, such as employment and education, to prevent discrimination of any kind [105, 124]. The California Civil Rights Council has passed regulations regarding algorithmic decision-making systems in the workplace [67, 84, 90].

6 Conclusion

Our survey study considers the attitudes and opinions of people from the southern U.S. regarding the collection and use of their voice data by AI-enabled technologies in four high-stakes decision-making use cases. In considering four scenarios related to education and employment, participants anticipated benefits that could increase efficiency and objectivity. However, harms that could reinforce and systematize bias, result in discrimination, and pose privacy and security risks were also discussed. Thus, we recommend the development of AI applications by and for speakers of non-standardized dialects, user-centered AI auditing by these speakers, and increased legal protections.

Acknowledgments

We used Google's Gemini AI tool to generate code to help create the bar plots and tables.

References

[1] State of Illinois 103rd General Assembly. 2023. HB3773. https://www.ilga.gov/legislation/103/HB/10300HB3773.htm

[2] Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 2819–2826. https://aclanthology.org/2020.lrec-1.343

[3] Ioannis Agrafiotis, Maria Bada, Paul Cornish, Sadie Creese, Michael Goldsmith, Eva Ignatuschtschenko, Taylor Roberts, and David M Upton. 2016. Cyber Harm: Concepts, Taxonomy and Measurement. *Saïd Business School WP 2016-23* 23 (1 aug 2016), 1–45. doi:10.2139/ssrn.2828646

[4] Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How Might We Create Better Benchmarks for Speech Recognition?. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, Kenneth Church, Mark Liberman, and Valia Kordoni (Eds.). Association for Computational Linguistics, Online, 22–34. doi:10.18653/v1/2021.bppf-1.4

[5] Alëna Aksënova, Antoine Bruguier, Amanda Ritchart-Scott, and Uri Mendlovic. 2020. Algorithmic Exploration of American English Dialects. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Barcelona, Spain, 7374–7378. doi:10.1109/ICASSP40776.2020.9053751

[6] Alëna Aksënova, Zhehuai Chen, Chung-Cheng Chiu, Daan van Esch, Pavel Golik, Wei Han, Levi King, Bhuvana Ramabhadran, Andrew Rosenberg, Suzan Schwartz, and Gary Wang. 2022. Accented Speech Recognition: Benchmarking, Pre-training, and Diverse Data. arXiv:2205.08014 [eess.AS] https://arxiv.org/abs/2205.08014

[7] Rachael Meghan Allbritten. 2011. *Sounding Southern: phonetic features and dialect perceptions.* thesis. Georgetown University. https://repository.library.georgetown.edu/handle/10822/553139

[8] Julia Angwin and Terry Parris Jr. 2016. Facebook Lets Advertisers Exclude Users by Race — propublica.org. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race. [Accessed 14-01-2026].

[9] Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (San Luis Potosi, Mexico) *(EAAMO '24)*. Association for Computing Machinery, New York, NY, USA, Article 2, 18 pages. doi:10.1145/3689904.3694699

[10] Nathan Barrett, Andrew McEachin, Jonathan N Mills, and Jon Valant. 2021. Disparities and discrimination in student discipline by race and family income. *Journal of Human Resources* 56, 3 (2021), 711–748.

[11] Muzakki Bashori, Roeland van Hout, Helmer Strik, and Catia Cucchiarini. 2024. 'Look, I can speak correctly': learning vocabulary and pronunciation through websites equipped with automatic speech recognition technology. *Computer Assisted Language Learning* 37, 5-6 (2024), 1335–1363.

[12] Amanda Baughan, Xuezhi Wang, Ariel Liu, Allison Mercurio, Jilin Chen, and Xiao Ma. 2023. A Mixed-Methods Approach to Understanding User Trust after Voice Assistant Failures. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–16. doi:10.1145/3544548.3581152

[13] Scott Beach, Richard Schulz, Julie Downs, Judith Matthews, Bruce Barron, and Katherine Seelman. 2009. Disability, Age, and Informational Privacy Attitudes in Quality of Life Technology Applications: Results from a National Web Survey. *ACM Transactions on Accessible Computing* 2, 1 (May 2009), 5:1–5:21. doi:10.1145/1525840.1525846

[14] Genevieve Bell, Mark Blythe, and Phoebe Sengers. 2005. Making by making strange: Defamiliarization and the design of domestic technologies. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (jun 2005), 149–173. doi:10.1145/1067860.1067862

[15] Rena Bivens and Oliver L Haimson. 2016. Baking gender into social media design: How platforms shape categories for users and advertisers. *Social Media+ Society* 2, 4 (2016), 2056305116672486.

[16] Robin N. Brewer, Christina Harrington, and Courtney Heldreth. 2023. Envisioning Equitable Speech Technologies for Black Older Adults. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 379–388. doi:10.1145/3593013.3594005

[17] Kalvin Chang, Yi-Hui Chou, Jiatong Shi, Hsuan-Ming Chen, Nicole Holliday, Odette Scharenborg, and David R. Mortensen. 2024. Self-supervised Speech Representations Still Struggle with African American Vernacular English. arXiv:2408.14262 [cs, eess]. doi:10.48550/arXiv.2408.14262

[18] Becky Childs and Christine Mallinson. 2017. *Southern Language and Linguistics* (1 ed.). Routledge, New York, New York, 54–65. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315768076-5/southern-language-linguistics-becky-childs-christine-mallinson

[19] Phoebe K Chua, Hillary Abraham, and Melissa Mazmanian. 2021. Playing the hiring game: Class-based emotional experiences and tactics in elite hiring. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–27.

[20] Jo Constantz. 2025. Job Interviews Enter a Strange New World With AI That Talks Back. Bloomberg. https://www.bloomberg.com/news/articles/2025-05-28/job-applicant-interviews-conducted-by-ai-offer-benefits-tech-glitches

[21] Emily Corvi, Hannah Washington, Stefanie Reed, Chad Atalla, Alex Chouldechova, Alex Dow, Jean Garcia-Gathright, Nick Pangakis, Emily Sheng, Dan Vann, Matthew Vogel, and Hanna Wallach. 2025. Taxonomizing Representational Harms using Speech Act Theory. https://www.microsoft.com/en-us/research/publication/taxonomizing-representational-harms-using-speech-act-theory/

[22] Council of European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). http://data.europa.eu/eli/reg/2024/1689/oj/eng

[23] Jennifer Suzanne Cramer. 2010. *The effect of borders on the linguistic production and perception of regional identity in Louisville, Kentucky.* Ph. D. Dissertation. University of Illinois at Urbana-Champaign.

[24] Roderic Crooks and Morgan Currie. 2021. Numbers will not save us: Agonistic data practices. *The Information Society* 37, 4 (2021), 201–213. doi:10.1080/01972243.2021.1920081

[25] Jonathan Dalby and Diane Kewley-Port. 1999. Explicit Pronunciation Training Using Automatic Speech Recognition Technology. *CALICO Journal* 16, 3 (1999), 425–445. arXiv:https://doi.org/10.1558/cj.v16i3.425-445 doi:10.1558/cj.v16i3.425-445

[26] Sarah Dean, Thomas Krendl Gilbert, Nathan Lambert, and Tom Zick. 2021. Axes for sociotechnical inquiry in AI research. *IEEE Transactions on Technology and Society* 2, 2 (2021), 62–70.

[27] Daniel Delmonaco, Samuel Mayworm, Hibby Thach, Josh Guberman, Aurelia Augusta, and Oliver L. Haimson. 2024. " What are you doing, TikTok?": How Marginalized Social Media Users Perceive, Theorize, and "Prove" Shadowbanning. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 154 (April 2024), 39 pages. doi:10.1145/3637431

[28] Michael Ann DeVito. 2022. How Transfeminine TikTok Creators Navigate the Algorithmic Trap of Visibility Via Folk Theorization. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 380 (Nov. 2022), 31 pages. doi:10.1145/3555105

[29] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3491102.3517441

[30] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 626, 19 pages. doi:10.1145/3491102.3517441

[31] Bryan Dosono and Bryan Semaan. 2020. Decolonizing tactics as collective resilience: Identity work of AAPI communities on Reddit. *Proceedings of the ACM on Human-Computer interaction* 4, CSCW1 (2020), 1–20.

[32] Lisa Egede, Leslie Coney, Brittany Johnson, Christina Harrington, and Denae Ford. 2024. "For Us By Us": Intentionally Designing Technology for Lived Black Experiences. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 3210–3224. doi:10.1145/3643834.3661535

[33] Jacob Eisenstein, Vinodkumar Prabhakaran, Clara Rivera, Dorottya Demszky, and Devyani Sharma. 2023. MD3: The Multi-Dialect Dataset of Dialogues. arXiv:2305.11355 [cs.CL] https://arxiv.org/abs/2305.11355

[34] Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J. Biega. 2025. Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey. *ACM Trans. Intell. Syst. Technol.* 16, 1, Article 16 (Jan. 2025), 54 pages. doi:10.1145/3696457

[35] Spencer Fane. 2024. AI In Hiring: The Impact Of Illinois' New Law On Employment Decisions And What Businesses Need To Know. https://www.mondaq.com/unitedstates/discrimination-disability-sexual-harassment/1519132/ai-in-hiring-the-impact-of-illinois-new-law-on-employment-decisions-and-what-businesses-need-to-know

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

[37] Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. 2023. FairPrism: Evaluating Fairness-Related Harms in Text Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 6231–6251. doi:10.18653/v1/2023.acl-long.343

[38] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Inf. Syst.* 14, 3 (July 1996), 330–347. doi:10.1145/230538.230561

[39] 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)*. IEEE, Seoul, Republic of Korea, 12356–12360. doi:10.1109/ICASSP48485.2024.10447116

[40] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling Bias in Toxic Speech Detection: A Survey. *Comput. Surveys* 55, 13s (July 2023), 264:1–264:32. doi:10.1145/3580494

[41] Stephen D.N. Graham. 2005. Software-sorted geographies. *Progress in Human Geography* 29, 5 (2005), 562–580. arXiv:https://doi.org/10.1191/0309132505ph568oa doi:10.1191/0309132505ph568oa

[42] Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence* Volume 7 - 2024 (2025), 18 pages. doi:10.3389/frai.2024.1472411

[43] Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. 2022. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–26.

[44] Consumer Reports Survey Group. 2024. A.I./Algorithmic Decision-Making: Consumer Reports Nationally Representative Phone and Internet Survey, May 2024. https://advocacy.consumerreports.org/wp-content/uploads/2024/07/CR-AES-AI-Algorithms-Report-7.25.24.pdf

[45] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3173574.3173582

[46] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1914–1933. doi:10.1145/2998181.2998327

[47] Elisa Harlan and Oliver Schnuck. 2021. Objective or biased: On the questionable use of Artificial Intelligence for job applications. https://interaktiv.br.de/ki-bewerbung/en/

[48] Christina N. Harrington and Lisa Egede. 2023. Trust, Comfort and Relatability: Understanding Black Older Adults' Perceptions of Chatbot Design for Health Information Seeking. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 120, 18 pages. doi:10.1145/3544548.3580719

[49] 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*. ACM, New Orleans LA USA, 1–15. doi:10.1145/3491102.3501995

[50] Camille Harris, Amber Gayle Johnson, Sadie Palmer, Diyi Yang, and Amy Bruckman. 2023. " Honestly, I Think TikTok has a Vendetta Against Black Creators ": Understanding Black Content Creator Experiences on TikTok. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–31.

[51] Roy Harris. 1981. *The language myth*. Duckworth, London.

[52] Einar Haugen. 1966. Dialect, Language, Nation. *American Anthropologist* 68, 4 (1966), 922–935. http://www.jstor.org/stable/670407

[53] Kathryn Henne, Renee Shelby, and Jenna Harb. 2021. The Datafication of #MeToo: Whiteness, Racial Capitalism, and Anti-Violence Technologies. *Big Data & Society* 8, 2 (2021), 20539517211055898. arXiv:https://doi.org/10.1177/20539517211055898 doi:10.1177/20539517211055898

[54] James L Hilton and William Von Hippel. 1996. Stereotypes. *Annual review of psychology* 47, 1 (1996), 237–271.

[55] Faye Holt, William Held, and Diyi Yang. 2024. Perceptions of Language Technology Failures from South Asian English Speakers. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 4067–4081. doi:10.18653/v1/2024.findings-acl.241

[56] Jingting Huang, Yuhan Li, and Yiran Zheng. 2022. Gender Discrimination in STEM Education. In *Proceedings of the 2022 6th International Seminar on Education, Management and Social Sciences (ISEMSS 2022)*. Atlantis Press, Chongqing, China, 1314–1323. doi:10.2991/978-2-494069-31-2_155

[57] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 306–316. doi:10.1145/3351095.3372829

[58] Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung Challenges of Building and Deploying Language Technologies for Low Resource Language Communities. arXiv:1912.03457 [cs]. doi:10.48550/arXiv.1912.03457

[59] Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the Dialect Gap and its Correlates Across Languages. arXiv:2310.15135 [cs]. doi:10.48550/arXiv.2310.15135

[60] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 158, 18 pages. doi:10.1145/3491102.3517533

[61] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic folk theories and identity: How TikTok users co-produce knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on human-computer interaction* 5, CSCW2 (2021), 1–44.

[62] Jared Katzman, Solon Barocas, Su Lin Blodgett, Kristen Laird, Morgan Klaus Scheuerman, and Hannah Wallach. 2021. Representational Harms in Image Tagging. (2021).

[63] Sharese King and Katherine D. Kinzler. 2020. Op-Ed: Bias against African American English speakers is a pillar of systemic racism. The Los Angeles Times. https://www.latimes.com/opinion/story/2020-07-14/african-american-english-racism-discrimination-speech

[64] Sara Kingsley, Proteeti Sinha, Clara Wang, Motahhare Eslami, and Jason I. Hong. 2022. "Give Everybody [..] a Little Bit More Equity"': Content Creator Perspectives and Responses to the Algorithmic Demonetization of Content Associated with Disadvantaged Groups. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 424 (Nov. 2022), 37 pages. doi:10.1145/3555149

[65] Liam Knox. 2023. Admissions Offices Deploy AI. https://www.insidehighered.com/news/admissions/traditional-age/2023/10/09/admissions-offices-turn-ai-application-reviews

[66] 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 (mar 2020), 7684–7689. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1915768117 doi:10.1073/pnas.1915768117

[67] Arsen Kourinian, Ruth Zadikany, and Remy N. Merritt. 2025. California Adopts New Employment AI Regulations Effective October 1, 2025 | Insights | Mayer Brown — mayerbrown.com. https://www.mayerbrown.com/en/insights/publications/2025/08/california-adopts-new-employment-ai-regulations-effective-october-1-2025. [Accessed 09-09-2025].

[68] William A. Kretzschmar and Charles F. Meyer. 2012. The idea of Standard American English. In *Standards of English: Codified Varieties around the World*, Raymond Hickey (Ed.). Cambridge University Press, Cambridge, England, 139–158.

[69] William Labov. 1969. *A Study of Non-Standard English*. ERIC Clearinghouse for Languages and Linguistics, Washington, D.C., USA. ERIC Number: ED024053. https://eric.ed.gov/?id=ED024053

[70] William Labov. 1991. *The three dialects of English*. Academic Press, San Diego, 1–44.

[71] William Labov, Sharon Ash, and Charles Boberg. 1997. A National Map of the Regional Dialects of American English. https://www.ling.upenn.edu/phono_atlas/NationalMap/NationalMap.html.

[72] Hao-Ping (Hank) Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks, In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA). *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Article 775, 19 pages. doi:10.1145/3613904.3642116

[73] Q. Vera Liao and Jennifer Wortman Vaughan. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review* Special Issue 5 (may 31 2024), 53 pages. https://hdsr.mit-press.mit.edu/pub/aelql9qy.

[74] Rosina Lippi-Green. 2012. *English with an Accent: Language, Ideology and Discrimination in the United States* (2nd ed.). Routledge, London; New York.

[75] Monique Mann and Tobias Matzner. 2019. Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. *Big Data & Society* 6, 2 (2019), 2053951719895805.

[76] Nina Markl. 2022. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 521–534. doi:10.1145/3531146.3533117

[77] Nina Markl and Catherine Lai. 2021. Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Online, 34–40. https://aclanthology.org/2021.hcinlp-1.6

[78] Nina Markl and Stephen Joseph McNulty. 2022. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 6328–6339. https://aclanthology.org/2022.lrec-1.680

[79] Nina Markl, Electra Wallington, Ondrej Klejch, Thomas Reitmaier, Gavin Bailey, Jennifer Pearson, Matt Jones, Simon Robinson, and Peter Bell. 2023. Automatic transcription and (de)standardisation. In *Proceedings - SIGUL 2023, 2nd Annual Meeting of the Special Interest Group on Under-resourced Languages: A Satellite Workshop of Interspeech 2023*. ISCA International Speech Communication Association, Dublin, Ireland, 93–97. doi:10.21437/SIGUL.2023-20

[80] Joshua L. Martin and Kevin Tang. 2020. Understanding Racial Disparities in Automatic Speech Recognition: The Case of Habitual "be". In *Proceedings of Interspeech 2020*. ISCA International Speech Communication Association, Shanghai, China, 626–630. doi:10.21437/Interspeech.2020-2893

[81] 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 (nov 2021), 725911. doi:10.3389/frai.2021.725911

[82] Barbara Micale. 2024. AI algorithms can add objectivity in assessing job applicants. https://news.vt.edu/content/news_vt_edu/en/articles/2024/02/science-hickman-ai-algorithms.html

[83] Hellina Hailu Nigatu, John Canny, and Sarah E. Chasins. 2024. Low-Resourced Languages and Online Knowledge Repositories: A Need-Finding Study.. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3613904.3642605

[84] Danielle Ochs, Jennifer Betts, and Zachary V. Zagger. 2025. California's Wait Is Nearly Over: New AI Employment Discrimination Regulations Move Toward Final Publication. Ogletree. https://ogletree.com/insights-resources/blog-posts/californias-wait-is-nearly-over-new-ai-employment-discrimination-regulations-move-toward-final-publication/

[85] Alexandra Olteanu, Su Lin Blodgett, Agathe Balayn, Angelina Wang, Fernando Diaz, Flavio du Pin Calmon, Margaret Mitchell, Michael Ekstrand, Reuben Binns, and Solon Barocas. 2025. Rigor in AI: Doing Rigorous AI Work Requires a Broader, Responsible AI-Informed Conception of Rigor. In *NeurIPS 2025*. Curran Associates, Inc., San Diego, USA, 21 pages. https://www.microsoft.com/en-us/research/publication/rigor-in-ai-doing-rigorous-ai-work-requires-a-broader-responsible-ai-informed-conception-of-rigor/

[86] Wanda J. Orlikowski. 2000. Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations. *Organization Science* 11, 4 (2000), 404–428. arXiv:https://doi.org/10.1287/orsc.11.4.404.14600 doi:10.1287/orsc.11.4.404.14600

[87] Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. 2023. Augmented Datasheets for Speech Datasets and Ethical Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 881–904. doi:10.1145/3593013.3594049

[88] Orestis Papakyriakopoulos and Alice Xiang. 2023. Considerations for Ethical Speech Recognition Datasets. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23)*. Association for Computing Machinery, New York, NY, USA, 1287–1288. doi:10.1145/3539597.3575793

[89] Destiny Peterson. 2023. In the Thick of Thick Accents: Employment Discrimination and the Appalachian Accent. *Appalachian Journal of Law* 22, 2 (June 2023), 19 pages. https://appalachian.scholasticahq.com/article/74188-in-the-thick-of-thick-accents-employment-discrimination-and-the-appalachian-accent

[90] Fisher Phillips. 2025. California Regulators Adopt New Discrimination Rules For Automated-Decision Systems: 3 Steps for Employers Using AI in the Workplace. https://www.jdsupra.com/legalnews/california-regulators-adopt-new-9904630/

[91] Claudio Pinhanez, Paulo Cavalin, Luciana Storto, Thomas Finbow, Alexander Cobbinah, Julio Nogima, Marisa Vasconcelos, Pedro Domingues, Priscila de Souza Mizukami, Nicole Grell, Majoí Gongora, and Isabel Gonçalves. 2024. Harnessing the Power of Artificial Intelligence to Vitalize Endangered Indigenous Languages: Technologies and Experiences. arXiv:2407.12620 [cs.CL] doi:10.48550/arXiv.2407.12620

[92] Claudio Santos Pinhanez, Raul Fernandez, Marcelo Carpinette Grave, Julio Nogima, and Ron Hoory. 2024. Creating an African American-Sounding TTS: Guidelines, Technical Challenges, and Surprising Evaluations. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. ACM, Greenville SC USA, 259–273. doi:10.1145/3640543.3645165

[93] Kerri Prinos, Neal Patwari, and Cathleen A. Power. 2024. Speaking of accent: A content analysis of accent misconceptions in ASR research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1245–1254. doi:10.1145/3630106.3658969

[94] Thomas Purnell, William Idsardi, and John Baugh. 1999. Perceptual and Phonetic Experiments on American English Dialect Identification. *Journal of Language and Social Psychology* 18, 1 (1999), 10–30. doi:10.1177/0261927X99018001002

[95] Katyanna Quach. 2022. AI recruitment software is "automated pseudoscience". The Register. https://www.theregister.com/2022/10/13/ai_recruitment_software_diversity/

[96] Lee Rainie, Monica Anderson, Colleen McClain, Emily A. Vogels, and Risa Gelles-Watnick. 2023. AI in Hiring and Evaluating Workers: What Americans Think. https://www.pewresearch.org/internet/2023/04/20/ai-in-hiring-and-evaluating-workers-what-americans-think/

[97] Divya Ramesh, Vaishnav Kameswaran, Ding Wang, and Nithya Sambasivan. 2022. How Platform-User Power Relations Shape Algorithmic Accountability: A Case Study of Instant Loan Platforms and Financially Stressed Users in India. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1917–1928. doi:10.1145/3531146.3533237

[98] Jeffrey Reaser, Eric Wilbanks, Karissa Wojcik, and Walt Wolfram. 2018. *Language Variety in the New South: Contemporary Perspectives on Change and Variation*. UNC Press Books, Chapel Hill, NC, USA. Google-Books-ID: l29RDwAAQBAJ.

[99] Joanna Redden, Jessica Brand, and Vanesa Terzieva. 2020. Data Harm Record - Data Justice Lab — datajusticelab.org. https://datajusticelab.org/project/data-harm-record/

[100] Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 299, 17 pages. doi:10.1145/3491102.3517639

[101] Thomas Reitmaier, Electra Wallington, Ondřej Klejch, Nina Markl, Léa-Marie Lam-Yee-Mui, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2023. Situating Automatic Speech Recognition Development within Communities of Under-Heard Language Speakers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 406, 17 pages. doi:10.1145/3544548.3581385

[102] Harvard Law Review. 2025. Resetting Antidiscrimination Law in the Age of AI. *Harvard Law Review* 138, 6 (April 2025), 1562–1584. https://harvardlawreview.org/print/vol-138/resetting-antidiscrimination-law-in-the-age-of-ai/

[103] Alene Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, and Julia Stoyanovich. 2022. Resume Format, LinkedIn URLs and Other Unexpected Influences on AI Personality Prediction in Hiring: Results of an Audit. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 572–587. doi:10.1145/3514094.3534189

[104] Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich. 2022. An external stability audit framework to test the validity of personality prediction in AI hiring. *Data Mining and Knowledge Discovery* 36, 6 (Nov. 2022), 2153–2193. doi:10.1007/s10618-022-00861-0

[105] Tatiana Rice, Kier Lamont, and Jordan Francis. 2024. The Colorado Artificial Intelligence Act FPF U.S. Legislation Policy Brief. https://leg.colorado.gov/sites/default/files/images/fpf_legislation_policy_brief_the_colorado_ai_act_final.pdf

[106] John Russell Rickford. 1997. Unequal partnership: Sociolinguistics and the African American speech community. *Language in Society* 26, 2 (1997), 161–197.

[107] Michael J. Ryan, William Held, and Diyi Yang. 2024. Unintended Impacts of LLM Alignment on Global Representation. arXiv:2402.15018 [cs.CL] https://arxiv.org/abs/2402.15018

[108] Jathan Sadowski and Evan Selinger. 2014. Creating a taxonomic tool for technocracy and applying it to Silicon Valley. *Technology in Society* 38 (2014), 161–168.

[109] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. Association for Computing Machinery, New York, NY, USA, 315–328.

[110] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1668–1678. doi:10.18653/v1/P19-1163

[111] Praveen Kumar Sattarapu, Deepti Wadera, Nguyen Phong Nguyen, Jaspreet Kaur, Sumeet Kaur, and Emmanuel Mogaji. 2024. Tomeito or Tomahto: Exploring consumer's accent and their engagement with artificially intelligent interactive voice assistants. *Journal of Consumer Behaviour* 23, 2 (2024), 278–298. doi:10.1002/cb.2195

[112] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-computer Interaction* 4, CSCW1 (2020), 1–35.

[113] Dana Serditova, Kevin Tang, and Jochen Steffens. 2025. Automatic Speech Recognition Biases in Newcastle English: an Error Analysis. In *Interspeech 2025*. ISCA International Speech Communication Association, Rotterdam, The Netherlands, 3204–3208. doi:10.21437/Interspeech.2025-1973

[114] Arathi Sethumadhavan, Joe Garvin, and Ben Noah. 2022. Speech is Human and Multifaceted: Our Approach to Studying It Should Be the Same. *Interactions* 29, 5 (Aug. 2022), 6–8. doi:10.1145/3555049

[115] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 723–741. doi:10.1145/3600211.3604673

[116] Jan Simson, Alessandro Fabris, and Christoph Kern. 2024. Lazy Data Practices Harm Fairness Research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 642–659. doi:10.1145/3630106.3658931

[117] Makoni Sinfree and Pennycook Alastair. 2006. *Disinventing and Reconstituting Languages*. Number Vol. 62 in Bilingual Education and Bilingualism. Multilingual Matters, United Kingdom. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,sso&db=e000xna&AN=181393&site=ehost-live&scope=site&custid=s8368349

[118] Mona Sloane. 2023. Automation and Recruiting: Understanding the Intersection of Algorithmic Systems and Professional Discretion in the Sourcing of Job Candidates. In *Proceedings" Automation by Design" Conference, University of Minnesota, Forthcoming*. SSRN, University of Minnesota, 19 pages.

[119] Mona Sloane. 2025. Boolean Clashes: Discretionary Decision Making in AI-Driven Recruiting. Communications of the ACM. doi:10.1145/3708596

[120] Mona Sloane, Emanuel Moss, and Rumman Chowdhury. 2022. A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns* 3, 2 (2022), 100425. doi:10.1016/j.patter.2021.100425

[121] Erin Smith. 2021. Landlords Use Secret Algorithms to Screen Potential Tenants. Find Out What They've Said About You. — propublica.org. https://www.propublica.org/article/landlords-use-secret-algorithms-to-screen-potential-tenants-find-out-what-theyve-said-about-you. [Accessed 14-01-2026].

[122] Dipankar Srirag, Aditya Joshi, and Jacob Eisenstein. 2025. Predicting the Target Word of Game-playing Conversations using a Low-Rank Dialect Adapter for Decoder Models. arXiv:2409.00358 [cs.CL] https://arxiv.org/abs/2409.00358

[123] Dipankar Srirag, Nihar Ranjan Sahoo, and Aditya Joshi. 2024. Evaluating Dialect Robustness of Language Models via Conversation Understanding. arXiv:2405.05688 [cs.CL] https://arxiv.org/abs/2405.05688

[124] General Assembly State of Colorado. 2024. Senate Bill 24-205. https://leg.colorado.gov/sites/default/files/2024a_205_signed.pdf
[125] Hung-Yue Suen, Kuo-En Hung, and Chien-Liang Lin. 2020. Intelligent video interview agent used to predict communication skill and perceived personality traits. *Human-centric Computing and Information Sciences* 10, 1 (Jan. 2020), 3. doi:10.1186/s13673-020-0208-3
[126] Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2022. Dialect-robust Evaluation of Generated Text. arXiv:2211.00922 [cs.CL] https://arxiv.org/abs/2211.00922
[127] 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*, Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach (Eds.). Association for Computational Linguistics, Valencia, Spain, 53–59. doi:10.18653/v1/W17-1606
[128] Jordan Taylor, Wesley Hanwen Deng, Kenneth Holstein, Sarah Fox, and Haiyi Zhu. 2024. Carefully Unmaking the "Marginalized User": A Diffractive Analysis of a Gay Online Community. *ACM Trans. Comput.-Hum. Interact.* 31, 6, Article 81 (Dec. 2024), 30 pages. doi:10.1145/3673229
[129] Dias Oliva Thiago, Antonialli Dennys Marcelo, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & culture* 25, 2 (2021), 700–732.
[130] Atnafu Lambebo Tonja, Fazlourrahman Balouchzahi, Sabur Butt, Olga Kolesnikova, Hector Ceballos, Alexander Gelbukh, and Thamar Solorio. 2024. NLP Progress in Indigenous Latin American Languages. arXiv:2404.05365 [cs]. doi:10.48550/arXiv.2404.05365
[131] Esma Wali, Yan Chen, Christopher Mahoney, Thomas Middleton, Marzieh Babaeianjelodar, Mariama Njie, and Jeanna Neefe Matthews. 2020. Is Machine Learning Speaking my Language? A Critical Look at the NLP-Pipeline Across 8 Human Languages. arXiv:2007.05872 [cs]. arXiv:2007.05872 [cs] doi:10.48550/arXiv.2007.05872
[132] Hanna Wallach, Meera Desai, Nick Pangakis, A. Feder Cooper, Angelina Wang, Solon Barocas, Alex Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2024. Evaluating Generative AI Systems is a Social Science Measurement Challenge. https://www.microsoft.com/en-us/research/publication/evaluating-generative-ai-systems-is-a-social-science-measurement-challenge/
[133] Azmine Toushik Wasi, Taki Hasan Rafi, and Dong-Kyu Chae. 2024. DiaFrame: A Framework for Understanding Bengali Dialects in Human-AI Collaborative Creative Writing Spaces. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose, Costa Rica) *(CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 268–274. doi:10.1145/3678884.3681862
[134] Jess Weatherbed. 2024. Texas is replacing thousands of human exam graders with AI. https://www.theverge.com/2024/4/10/24126206/texas-staar-exam-graders-ai-automated-scoring-engine
[135] Whiteford, Lisa Bauner, and Slaven Jesic. 2024. Client Alert: Avoiding Legal Pitfalls and Risks in Workplace Use of Artificial Intelligence. https://www.jdsupra.com/legalnews/client-alert-avoiding-legal-pitfalls-2004765/
[136] Walt Wolfram. 2015. The sociolinguistic construction of African American. *The Oxford handbook of African American Language* 338 (2015), 338–352.
[137] Kathryn A. Woolard. 2020. *Language Ideology*. John Wiley & Sons, Ltd, 1–21. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118786093.iela0217 doi:10.1002/9781118786093.iela0217
[138] Ziang Xiao, Wesley Hanwen Deng, Michelle S. Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q. Vera Liao. 2024. Human-Centered Evaluation and Auditing of Language Models. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3613905.3636302
[139] Jonathan Zong and J. Nathan Matias. 2024. Data Refusal from Below: A Framework for Understanding, Evaluating, and Envisioning Refusal as Design. *ACM Journal on Responsible Computing* 1, 1 (March 2024), 1–23. doi:10.1145/3630107

A Surveys

We provide details about the pilot survey as well as the final text of the screening and main surveys.

A.1 Pilot

In our pilot, we presented high-stakes use cases in a matrix question format, using the same Likert acceptability scale as in the final study (Section 3). Pilot results suggested that this kind of presentation was too brief and lacked context. For example, the education use case was described

as "teacher correcting pronunciation," which resulted in a variety of interpretations. In order to reduce ambiguity in our questions, we made all use cases more specific. We also prefaced each voice-specific application with a general scenario to allow comparisons with algorithmic or AI applications not relating to voice data.

A.2 Screening Survey

A.2.1 Voice Assistant Knowledge.

(1) What do you think a voice assistant is?
   (a) A chatbot that generates responses to typed prompts
   (b) Computer technology that responses to spoken prompts
   (c) A service provided by hotels to make reservations or offer suggestions
   (d) A large language model
   (e) I don't know

A.2.2 Voice Assistant Experience.

(1) Have you ever used a voice assistant?
(2) Do smart phone voice assistants such as Siri on iPhones or Google Assistant on Android phones have difficulty understanding you?
(3) Do home IoT voice assistants like Amazon Alexa or Google Home have difficulty understanding you?
(4) Do you change the way you talk to be better understood by voice assistants?

A.2.3 Region.

(1) What US state do you currently live in?
(2) Were you born in the United States?
(3) Did you spend most of your time before turning 18 in the U.S.?
(4) (if Yes to the previous question) In what U.S. state(s) did you spend most of your time before turning 18?

A.2.4 Language and Dialect.

(1) Is U.S. American English your first language? This can be any dialect or accent of American English, from any region in the U.S.
(2) Is American English the language that you primarily speak?

The following questions are about the frequency with which you speak the following U.S. American English dialects or accents. *[Choice of: Never, Occasionally or sometimes, or Often or Always]*

(1) How often do you speak General or Standardized U.S. American English (for example, the default American English dialect spoken by voice assistants, often used in American commercials)?
(2) How often do you speak Southern U.S. American English (also called a "southern accent")?
(3) How often do you speak African American English (also called African American Vernacular English or Black English)?
(4) How often do you speak Appalachian English (commonly spoken in central and southern Appalachia, e.g., in West Virginia and Kentucky)?
(5) Do you speak another dialect of U.S. American English not listed above? If yes, please provide the name of the dialect. If you speak multiple additional dialects, please provide the one you speak most often.
(6) (if Yes to the previous question) How often do you speak the dialect you specified above?

A.2.5 Speech. The following questions are about your speech, or how you utter words aloud.

(1) Do other English speakers have difficulty understanding you when you speak? *[Choice of: Never, Occasionally, Sometimes, About half the time, Most of the time, Always]*
(2) Do you have a speech disorder or voice disorder that could influence how well voice assistants understand you?
(3) Are there other factors besides accent, dialect, speech disorder, or voice disorder that could influence how well voice assistants understand you? If so, please specify.

A.2.6 Demographics.

(1) What is your approximate age in years?
(2) How do you describe your gender identity?
(3) How do you describe your race or ethnic identity?
(4) Do you identify as Hispanic or Latino/Latina/Latinx?
(5) What is the highest level of school you have completed or the highest degree you have received?
(6) What was your approximate household income in 2023?
(7) Do you have a formal education in a computer-related field, such as computer science or IT?
(8) Do you have a formal education in a data science related field, such as statistics or machine learning?

A.3 Main Survey

A.3.1 Data Collectors. **Scenario:** Imagine that the following people or entities are collecting voice data from voice assistants that you encounter, for example, in a store or office. The entities will use your voice data to improve their services, product, or performance.

Question: How acceptable is it for you to give your voice data to the following data collectors? *[Choice of: Extremely unacceptable, Somewhat unacceptable, Neither acceptable nor unacceptable, Somewhat acceptable, Extremely acceptable]*

(1) Voice assistant companies like Google, Apple, and Amazon
(2) AI technology companies like Open AI, Google, Meta
(3) My school district
(4) Potential employer
(5) My current employer
(6) My teacher

A.3.2 Inferences. **Scenario:** Imagine that voice assistants can be programmed to infer or guess the following speaker attributes. That is, they could use a speaker's voice data to estimate attributes like age, gender, etc.

Question: How acceptable is it to you for voice assistants to make inferences about the following characteristics based on your voice data? *[Choice of: Extremely unacceptable, Somewhat unacceptable, Neither acceptable nor unacceptable, Somewhat acceptable, Extremely acceptable]*

(1) Age
(2) Dialect or accent
(3) Ethnicity or race
(4) Gender
(5) Health condition
(6) Mood or emotional state
(7) Region of origin (where you grew up)
(8) Sexual orientation

(9) Speech or voice disorder

Next, we will ask you about four (4) scenarios related to technologies that could be used in education and employment settings, as well as potential applications of voice technologies in those settings. For each scenario, we will ask you to provide free-text responses. We're especially interested in your attitudes, opinions, or experiences relating to the southern United States and dialects or accents found in this region. Thank you for your thoughtful responses.

A.3.3 Education Admissions.

General Scenario. A small liberal arts college uses software to determine eligibility for admission to their undergraduate program. It takes application materials and ranks students based on admissions criteria.

(1) How acceptable do you find the scenario above?
   (a) Extremely unacceptable
   (b) Somewhat unacceptable
   (c) Neither acceptable nor unacceptable
   (d) Somewhat acceptable
   (e) Extremely acceptable

Next, please consider the following voice-specific application of the technology from the scenario above.

Voice Specific Application. The software evaluates audio recordings of admissions interviews with applicants to determine their suitability for small classes with interactive discussions. Students will be evaluated based on ability to respond to questions using appropriate language and clearly articulate their opinions.

(1) Does the inclusion of this voice-specific application make the scenario more or less acceptable, or make no difference in acceptability?
   (a) Less acceptable
   (b) Makes no difference in acceptability
   (c) More acceptable

Free text responses.

(1) You selected *(general scenario answer choice)* for the general scenario and *(voice-specific answer choice)* for the voice-specific application. Please explain your overall attitude for this scenario and voice-specific application.
(2) If you believe that there are potential harms for this use case of voice technologies, please list at least one potential harm that you can imagine related to the above use case. If you do not believe that there are potential harms, please explain why.
(3) If you believe that there are potential benefits for this use case of voice technologies, please list at least one potential benefit that you can imagine related to the above use case. If you do not believe that there are potential benefits, please explain why.

A.3.4 Education Evaluation.

General Scenario. A high school district uses automatic grading software to grade students' English annual final exam, which counts for 20% of their final grade in the course for the year (grade shown on transcript).

Voice Specific Application. Exams include a spoken component, and the software grades students' English word pronunciation. The spoken component counts for 25% of the final exam grade.

A.3.5 Employment Hiring.

General Scenario. A company hiring new employees uses software to rank the top job applicants throughout the job application process. Employees are ranked based on past experience and skills. The hiring manager will review only the top 10 candidates on the list created by the software.

Voice Specific Application. The software is used during in-person job interviews. It analyzes the job applicant's speech during an exercise in which the applicant pretends to speak to a customer and rates how understandable the applicant's speech will be to customers.

A.3.6 Employment Evaluation.

General Scenario. A company uses software to evaluate current employees' job performance to determine pay raises. Employees are told ahead of time about the metrics to be considered by the software and can request re-evaluation for certain components that they can re-do, such as submitting their work product for review.

Voice Specific Application. The software considers recent presentations by employees. It analyzes speech during presentations to evaluate public speaking and communication skills.

A.3.7 Experience and Usage. Voice Assistant Experience

Thank you for your written responses. In this final part of the study, we will ask you some questions about your use of voice assistants.

(1) Which of the following devices do you use? Select all that apply and please specify the device name if you use a home voice assistant or other smart device (for example: Amazon Alexa or Google Home). *[Choice of: Home voice assistant, Smart phone, Smart TV, Smart watch, Wearable fitness tracker, Other smart device, None of the above]*
(2) Do you use a voice assistant on your smart phone, such as Siri or Google Assistant? If yes, please specify which one(s).
(3) Some voice assistants are phone-based. You might encounter them when calling a bank or an airline by phone. Have you ever encountered a phone-based voice assistant?
(4) How often do you use voice assistants or other speech recognition technology for tasks such as dictation, searching, or controlling devices? *[Choice of: Rarely or never, Once every few months, Once every few weeks, Once per week, Daily]*
(5) How often are you around someone else's voice assistant, for example, a family member or roommate's voice assistant? *[Same choices as previous question.]*
(6) Some voice assistants offer the option to improve their ability to recognize you by doing some short exercises to teach them your voice. Have you ever trained a voice assistant on your voice in this way?
(7) Where do you use voice assistants? Please select all that apply. *[Choice of: My home, Work, Car, Gaming, Home of friend(s), School, Other, NA]*

Please indicate your level of agreement or disagreement with the following statements. *[Choice of: Strongly disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Strongly agree, N/A - I do not use voice assistants, or I don't know]*

(1) Voice assistants usually understand me.
(2) I engage less with voice assistants due to past interactions with voice assistants.

(3) I am able to use voice assistants equally as well as the average or standard user they're developed for.
(4) It is important to me that speech recognition technologies, such as voice assistants, accurately understand my individual voice.
(5) It is important to me that speech recognition technologies, such as voice assistants, accurately understand voices that sound similar to mine, for example, with the same accent and dialect.
(6) (Optional) Please add anything else you would like us to know about your experiences with voice assistants.

B Appendix - Emergent Code Book

Below, we include our code book with the thematic codes we used to label participants' responses about their overall attitudes and potential benefits and harms for all four use cases in Table 4.

Counts of thematic codes across the four dialect groups were compared and did not indicate any stark patterns, but we provide some notable highlights in Figure 1, Section 4.2.3, and as follows. The potential benefits discussed by a higher ratio of self-reported SAE-speakers than any other group were: *adequate measure*, *efficiency*, and *improve skills*, and the potential harms were *errors or glitches*, and *unfair outcome*. The topics for which the SAE group ratio was the lowest of all four groups were: *dialect, accent or speech* and *speech is not a proxy for intellect or skills*. See Table 4 below for definitions of these codes.

Table 4. Inductive Code Book: Themes Emerging from Participants' Responses about Voice-specific Applications of Algorithmic Evaluations

Main Code	Subcode	Definition
Negative opinion (critique)	Scenario critique	Critique of scenario details rather than voice tech application or decision-making AI, problem with the outcome variable or the fact that such a scenario would exist
	Discrimination or bias	Unfair or negative outcome due to speaker attribute(s); mentions discrimination, bias, unfairness, prejudice, etc.
	Unfair outcome	Data subjects are treated unjustly or receive an unfair outcome (rejected, filtered out, or evaluated unfairly)
	Privacy or security	Harms related to data privacy concerns, malicious use of data, security breaches, user consent to data use, etc.
	Errors or glitches	Errors or glitches of the technology will be harmful
Positive opinion (support)	Adequate measure	Technology will be reliable, able to get the job done, capable of performing the given task adequately or even better than humans
	Efficiency	AI or software would be efficient and save time, labor or resources, e.g., reduces workload or streamlines process.
	Mitigate human bias	AI or software could help mitigate potential bias of humans involved in decision-making process. Positive view of technology related to objectivity and consistency of technology.
	Improve skills	Evaluation will help data subjects improve their skills
	Important to stakeholder	Highlights the importance or relevance of the evaluation to stakeholders (company, employee, customer, school, student) as justification for the acceptability of software/data use
	Evaluation or Finds the best	Technology will help find the best or a better selection of job candidates or students, or to evaluate communication skills or pronunciation to select the best performers
Question or concern	Already happening	Notes existence of or giving examples of similar scenarios that are already happening
	Depends on data type	Software using some data types is OK, but not others
	Question or concern	Question or concern regarding scenario that they would like more information about or a condition or variable that would influence their response
Humans & technology	Preference for human evaluation	A human should be involved, do the task or review the job done by the software. People deserve human review.
	Inadequacies	AI, software or technology in general is not able to do the given task, either at all or as well as a human could
Speech-related comment	Dialect, accent, or speech	Response mentions dialect or accent, or comments on speech attributes like "how people speak" or differences in speech patterns and pronunciation
	Speech is not a proxy for intellect or skills	Speech or voice data does not adequately represent a person's abilities or intelligence and should not be used to gauge skills

Note: We inductively developed thematic codes for participants' responses about voice-specific applications of algorithmic evaluations i.e., self-reported overall attitudes and perceptions of benefits and harms.

C Deductive Analysis Code Book: Sociotechnical Harms of Algorithmic Systems

We deductively analyzed participants' responses about potential harms for all four use cases using an a priori list of harms from Shelby et al.'s taxonomy of sociotechnical harms of algorithmic systems [115]. In Table 5 we include our "code book" listing the sociotechnical harms as they corresponded to harms that our study's participants provided in their survey responses. We also define them below based on the paper by Shelby et al. [115]. We also include examples of participant responses corresponding to these sociotechnical harms in Table 6.

The allocative harm of *opportunity loss* "occurs when algorithmic systems enable disparate access to information and resources needed to equitably participate in society" [115]. The quality of service (QoS) harm of *service or benefit loss* is the "degraded or total loss of benefits of using algorithmic systems with inequitable system performance based on identity" [81, 115]. The QoS harm of *alienation* is "self-estrangement experienced at the time of technology use, typically surfaced through interaction with systems that under-perform for marginalized individuals [81] or reinforce social alienation between humans" [15, 115]. The interpersonal harm of *privacy violation* "occurs when algorithmic systems diminish privacy, such as enabling the undesirable flow of private information [97], instilling the feeling of being watched or surveilled [99], and the collection of data without explicit and informed consent" [60, 115]. The interpersonal harm of *loss of agency or control* occurs when "algorithmic systems reduces autonomy" and when "there is insufficient ability to contest or remedy the decisions of algorithmic systems," and includes "algorithmically informed identity change," and "increased conforming of content" [8, 26, 61, 75, 109, 115, 121, 129]. The societal harm of *cultural harm* is "the development or use of algorithmic systems that affects cultural stability and safety," [3, 115] including "systemic erasure" [30, 125], "foreclos[ing] alternative ways of understanding the world" [108], and the "loss of cultural property, and harm to social values," [3, 115]. Finally, representational harms include *stereotyping social groups* through an algorithm's "beliefs about the characteristics, attributes, and behaviors of members of certain groups….and about how and why certain attributes go together" [54, 115]; *erasing social groups* through "algorithmic hegemony" [31, 115]; and *alienating social groups* by missing "the relevance of someone's membership in a specific social group" [62] during algorithmic processes [115].

Table 5. Deductive Code Book: Criteria for Applying Sociotechnical Harms of Algorithmic Systems to Harms Listed by Participants

Sociotechnical Harms	**Harm Sub-Type**	**Our Coding Criteria**
Allocative	Opportunity Loss	Mention of discrimination, unfair outcome; does not have to imply disparate impact based on group but should imply that some people lose opportunity unfairly or based on attributes.
	Economic Loss	Loss of currency or value, even if not financial. Employee loses money via loss of raises. Company loses labor/employees. University loses students. Students' GPA decreases.
Quality of Service	Service or Benefit Loss	When software has uneven performance and its scores or evaluations vary, resulting in lower scores or ranking for certain people based on factors like accent, or for unspecified reasons.
	Alienation	Indicates potential frustration or feelings of exclusion due to algorithmic evaluations.
Interpersonal	Privacy Violations	Shares concerns regarding the misuse of voice data and/or mentions concerns of data leakage.
	Loss of Agency or Control	Replacement of authority figures who could contest bad output. Increased conforming of content through intentional changes in speech styles.
	Diminished Health and Well-being	Emotional or psychological effects of algorithmic evaluation like stress or reduced self-esteem.
	General	General expression of malicious human behavior accompanying algorithmic evaluation.
Representational	Stereotyping Social Groups	Feels that evaluation would reflect social stereotypes like low intelligence; implicit stereotypes conveyed by algorithmic decisions reflecting underrepresentation; output reflects stereotypical pairing of attributes, such as using voice to determine race or age.
	Erasing Social Groups	System fails to recognize or parse certain traits, such as regional accent or dialect, which are ingrained parts of the individual. Systematic absence or underrepresentation of a social group, which effectively erases it from purview of the algorithmic system.
	Alienating Social Groups	Not acknowledging the relevance of a social group, failure to recognize strengths or challenges of certain people.

Note: We applied labels of sociotechnical harms of algorithmic systems from Shelby et al.'s taxonomy [115] as applied to our participants responses about potential harms of algorithmic evaluation of voice data in high-stakes use cases.

Table 6. Examples of Participant Responses Corresponding to Sociotechnical Harms

Sociotechnical Harms	**Harm Sub-Type**	**Examples of Harm in Employment or Education Contexts**
Allocative	Opportunity Loss	"People with speech disorders or who are not fluent in English may be unfairly excluded" (P1, Admissions)
	Economic Loss	"Good workers could get passed over for raises or promotions." (P48, Job Eval)
Quality-of-Service	Service or Benefit Loss	"One potential harm could be if the software can't pick up on nuances in speech or accents." (P24, Exam)
	Alienation	"fear in this scenario of not making the cut Over thinking the it and not exactly getting the best results" (P97, Admissions)
Interpersonal	Privacy Violations	"The only potential harm I see is one altering voice to reflect negative traits in order to make someone look bad." (P83, Job Eval)
	Diminished Health and Well-Being	"It would ... possibly hurt their self-esteem due to being hyperaware that their stutter caused them to miss points on a test that they should have been given." (P33, Exam)
	Loss of Agency or Control	"...taking away more of the casual element present in some businesses that I enjoy and trying to force everything to be robotic style professional." (P19, Job Eval)
	General	"...I would be concerned about this technology being misused for more nefarious purposes as it's initial use seems superfluous." (P27, Job Eval)
Representational	Stereotyping	"Many people not from Appalachia or from the south immediately make assumptions about intelligence and abilities based solely on how they talk." (P26, Job Eval)
	Erasing Social Groups	"The accent is part of the heritage of this area due to Scottish and Irish immigrants and other settlers. Our accents are part of who we are. Asking us to speak more of a northern or blank accent is unfair and extremely difficult and it bleaches out our culture and way of life." (P71, Exam)
	Alienating Social Groups	"I would hope the employer is hiring people from the area to work in that area. I am from southern WV and can understand their speech and they can understand me, but I have met people from other parts of the country who did not understand me and ones I could not understand. I don't think an AI could get that difference." (P72, Hiring)

Note: This table shows examples of participant responses providing potential harms of algorithmic evaluation of voice data in high-stakes use cases that correspond to sociotechnical harms of algorithmic systems from Shelby et al.'s taxonomy [115].

D Deductive Coding Counts: Sociotechnical Harms

We provide additional code counts for sociotechnical harms listed by participants, which we deductively coded using Shelby et al.'s taxonomy of sociotechnical harms of algorithmic systems, to supplement Table 3, which includes the total unique counts of sociotechnical harms per participant across all four scenarios (Section 4.4).

Below, in Table 7, we provide counts per scenario. Note that two themes, privacy harms and discrimination, were coded as part of our inductive or emergent coding, listed in Figure 2.

Table 7. Distribution of Sociotechnical Harms: Total Counts by Scenario

Harm	Harm Sub-Type	College	Exam	Hiring	Job Eval	Total
Allocative Harm	Economic Loss	11	28	47	30	**116**
	Opportunity Loss	24	3	0	0	**27**
Quality-of-Service Harm	Service or Benefit Loss	13	17	13	21	**64**
	Alienation	3	1	2	2	**8**
Interpersonal Harm	Loss of Agency or Control	3	5	2	2	**12**
	Diminished Health/Well-being	0	2	0	1	**3**
Representational Harm	Erasing Social Groups	5	6	1	1	**13**
	Stereotyping	1	1	1	1	**4**
	Alienating Social Groups	0	0	1	0	**1**

Note: Total counts of harms provided per participant for each use case correspond to sociotechnical harm categories ("Harm") and subcategories ("Harm Sub-Type") drawn from Shelby et al.'s taxonomy [115].

Since opportunity loss in employment contexts is categorized as economic loss in Shelby et al.'s taxonomy, in our deductive coding analysis, we broke down economic loss into the following themes that aligned with the four specific use cases we presented to participants: loss of students for universities or colleges, loss of academic currency (grade point average) for students, loss of labor, time or resources for employers, loss of opportunities for job candidates and workers, and loss of income or finances for workers. The counts for these is shown below in Table 8.

Table 8. Economic Loss Subcodes Across Four Use Cases.

Economic Loss	College	Exam	Hiring	Job Eval
Loss of students for university	11			
Loss of academic currency (grade point average)		28		
Loss of labor, time or resources for employer			16	3
Loss of opportunities in employment context			31	6
Loss of income or finances				21

E Appendix - Demographics

Participants self-reported their dialect, age, gender, race or ethnicity, and income, as shown by percentage in Table 9. Note that multiple gender and racial or ethnic identities could be selected, making their respective column total more than 100%. Participants mostly identified as one of two ethnicities, with 56 identifying as White, 53 as Black or African American, one as Asian, and one as American Indian or Alaska Native. A total of 64 participants identified as female, 44 as male, three as non-binary, and two as transgender. Most participants were relatively young, with 62 participants reporting being up to 40 years old.

Most participants reported their highest level of education to include at least some college education, with 52 reporting having obtained an Associate's or Bachelor's degree, 30 having attended college with no degree, 15 obtaining a high school degree, 11 obtaining a master's degree, two having less than a high school degree, and one obtaining a trade school degree. In terms of income, 56 participants reported that they made less than $50,000 per year, 51 between $50,000-$150,000, and three above $150,000.

Table 9. Summary of Participant Demographics

Age (Years)		Gender Identity		Household Income	
21 to 30	25.2%	Female	57.7%	Less than $50,000	50.5%
31 to 40	30.6%	Male	39.6%	$50,000 to $99,999	36.9%
41 to 50	23.4%	Non-binary	2.7%	$100,000 to $149,999	9.0%
51 to 60	18.0%	Transgender	1.8%	Above $150,000	2.7%
61 to 70	2.7%			No response	0.9%

Ethnic or Racial Identity		Dialect Group		Education	
White	51.4%	SAE-Only	22.5%	Less than high school	1.8%
Black or African American	48.6%	APP-Only	18.0%	Graduated high school	13.5%
Asian	0.9%	AAE-Only	7.2%	Some college education	27.0%
American Indian or Alaska Native	0.9%	Southern-Only	21.6%	Associate's degree	14.4%
		SAE-Appalachian	9.0%	Bachelor's degree	32.4%
		SAE-AAE	21.6%	Master's degree	9.9%
				Trade school diploma	0.9%

State (Current Residence)		State of Origin	
Alabama	1.8%	Alabama	1.8%
Arkansas	1.8%	Arkansas	1.8%
Georgia	8.1%	Georgia	7.2%
Kentucky	25.2%	Kentucky	23.4%
Louisiana	9.9%	Louisiana	11.7%
Mississippi	0.9%	Maryland	0.9%
North Carolina	9.9%	Missouri	0.9%
South Carolina	2.7%	North Carolina	6.3%
Tennessee	19.8%	South Carolina	2.7%
Virginia	9.0%	Tennessee	21.6%
West Virginia	10.8%	Texas	0.9%
Virginia	9.9%		
West Virginia	10.8%		

When reporting speaking dialects often or always, participants reported various combinations of dialects, as shown in Figure 3, which shows the number of participants for ten self-reported dialects and dialect combinations.

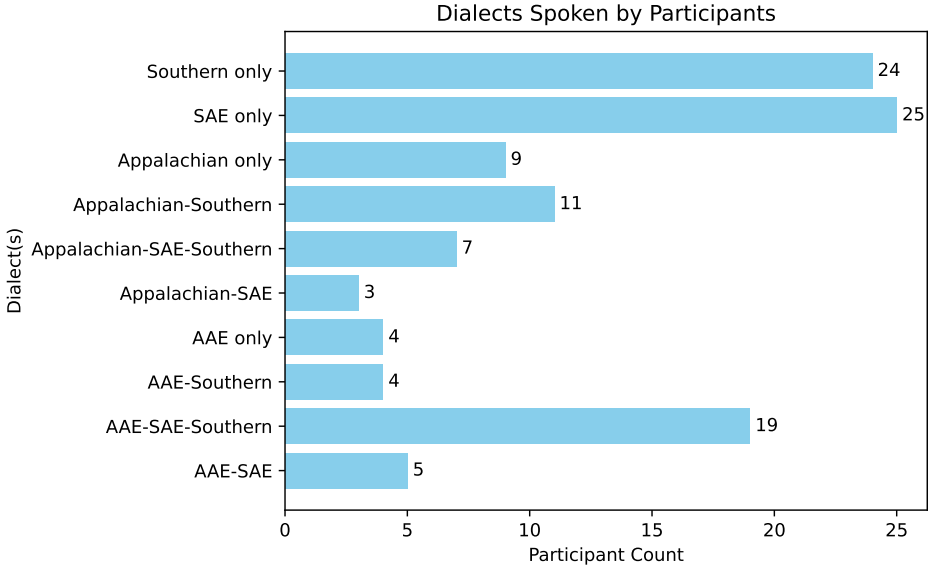

Fig. 3. Counts of dialects spoken by participants, including dialect combinations and counts.

We also gathered background information about participants' experience with and their attitudes towards voice technologies they used most often. Most participants had experience with voice assistants, with only three participants reporting having never used a voice assistant. Most participants reported using a voice assistant on their mobile phone ($n = 91$) or having encountered a phone-based voice assistant ($n = 93$). When asked if they had ever trained a VA on their voice to improve the VA's ability to recognize them, 60 participants responded Yes, and 44 responded No.

E.1 Difficulty Being Understood by Technology or Humans

When asked how frequently other English speakers, smart phones, or home IoT devices have difficulty understanding them when they speak, most participants responded "Never," "Occasionally," or "Sometimes," as shown in Figure 4.

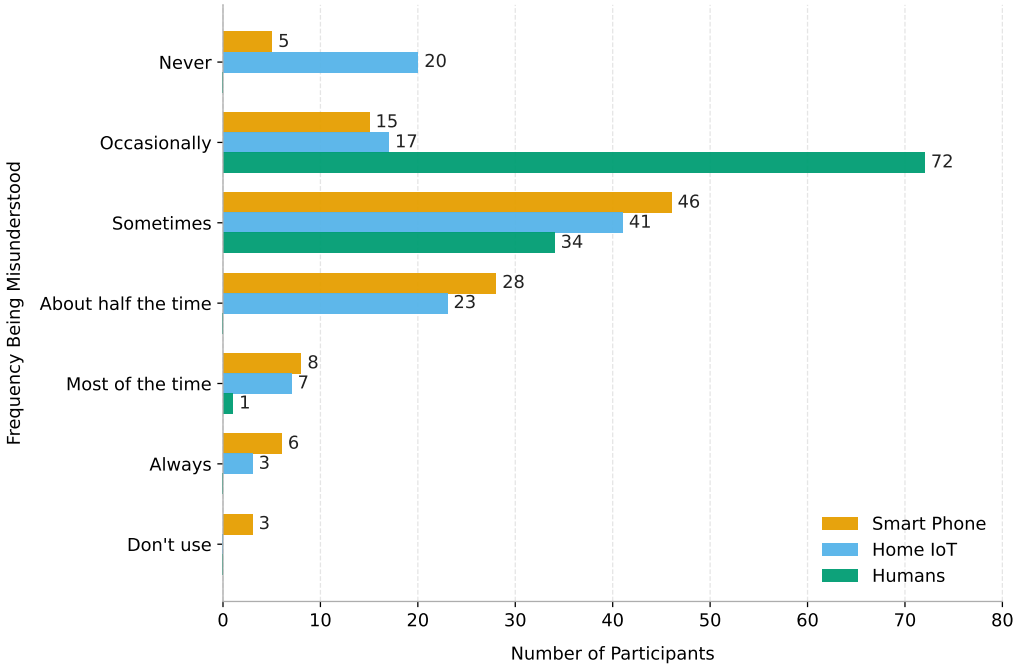

Fig. 4. Counts of participants' ratings of difficulty being understood by other English speakers, smart phones, or home IoT devices, by frequency.

When asked how much they agreed with the statement, "Voice assistants usually understand me," 40 participants selected Strongly Agree, 54 Somewhat Agree, 11 Somewhat disagree, and one Strongly disagree. For the statement, "I am able to use voice assistants equally as well as the average or standard user they're developed for," 55 participants selected Strongly Agree, 33 Somewhat Agree, eight Somewhat disagree, and three Strongly disagree.

For the following statement, "It is important to me that speech recognition technologies, such as voice assistants, accurately understand my individual voice." 51 selected Strongly Agree, 42 Somewhat Agree, five Somewhat Disagree, and three Strongly Disagree. For the statement, "It is important to me that speech recognition technologies, such as voice assistants, accurately understand voices that sound similar to mine, for example, with the same accent and dialect," 53 participants selected Strongly Agree, 37 Somewhat Agree, seven Somewhat Disagree, and two Strongly Disagree.

F Charts of Acceptability Results

We provide visual representation of the acceptability likert ratings presented in Section 4.2, i.e., participants' acceptability attitudes about general software evaluations and changes in acceptability attitudes for voice-specific applications, in two figures showing percentages of acceptability attitudes, Figure 5 and Figure 6, as well as in tables showing the numbers of participants, Table 10 and Table 11.

F.1 Acceptability of Algorithmic Evaluation "Software"

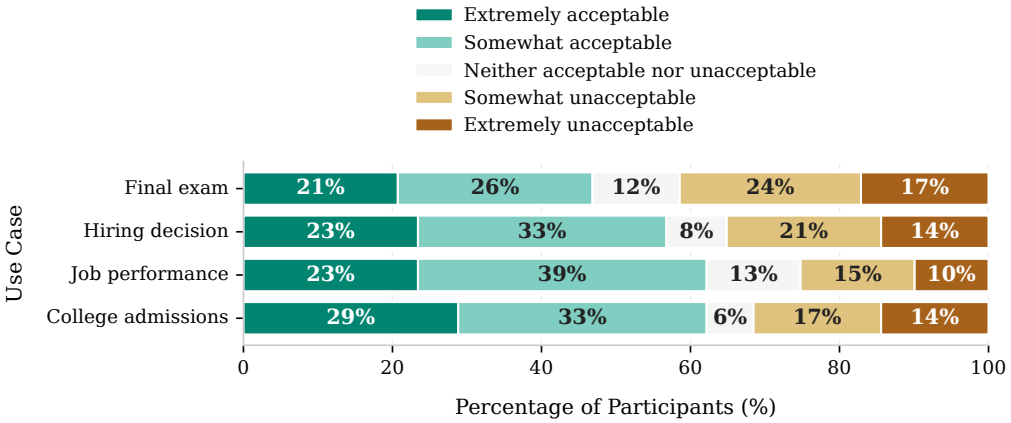

Fig. 5. Acceptability of the use of "software" for algorithmic evaluation, by percentage of participants, in four high-stakes use cases: college admissions interviews, final exam grading, job performance evaluation, and hiring interviews.

Table 10. Acceptability of Algorithmic Evaluation "Software" by Number of Participants

Scenario	Acceptable	Neither	Unacceptable
College Admission	69	7	35
Job Evaluation	69	14	28
Hiring	63	9	39
Final Exam	52	13	46

F.2 Relative Acceptability of Voice-Specific Applications of Evaluations

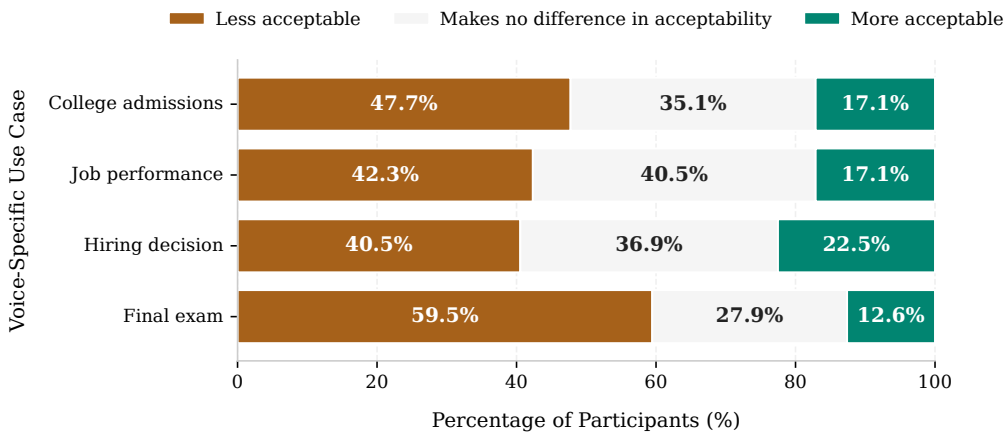

Fig. 6. Change in acceptability for voice-specific applications, by percentage of participants, after we specifically asked about the acceptability of algorithmic analysis of voice data in four use cases: college admissions, final exam grading, job performance evaluation, and hiring.

Table 11. Participant Acceptance of Voice-Specific Algorithmic Evaluations

Use Case	More Acceptable	Less Acceptable	No Difference
Hiring	25	45	41
Job Evaluation	19	47	45
College Admission	19	53	39
Final Exam	14	66	31

Received 13 May 2025; revised 13 January 2026; accepted 17 March 2026