Title
Mitigating Gender and L1 Biases in Automated English Speaking Assessment

Author
Kwako, Alexander James

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Mitigating Gender and L1 Biases in Automated English Speaking Assessment

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Education

by

Alexander James Kwako

2023

ABSTRACT OF THE DISSERTATION

Mitigating Gender and L1 Biases
in Automated English Speaking Assessment

by

Alexander James Kwako
Doctor of Philosophy in Education
University of California, Los Angeles, 2023
Professor Michael H. Seltzer, Chair

Automated assessment using Natural Language Processing (NLP) has the potential to make English speaking assessments more reliable, authentic, and accessible. Yet without careful examination, NLP may exacerbate social prejudices based on gender or native language (L1). Current NLP-based assessments are prone to such biases, yet research and documentation are scarce. Considering the high stakes nature of English speaking assessment, it is imperative that tests are fair for all examinees, regardless of gender or L1 background. Through a series of three studies, this project addresses the need for more thorough investigations of bias in English speaking assessment. Study 1 examines biases in automated transcription, a key component of automated speaking assessment. Study 2 focuses on a specific type of bias known as differential item functioning (DIF), and determines which patterns of DIF are present in human rater scores and whether or not these patterns of DIF are exacerbated by a pretrained, large language model (LLM) known as BERT. Lastly, Study 3 presents a comparison of two approaches of mitigating DIF using LLMs. Results from Study 1 indicate that there are indeed biases in automated transcription, however these do not translate into biased speaking scores. In Study 2, it is shown that BERT does exacerbate human rater biases,

although the effect size is small. Finally, Study 3 demonstrates that it is possible to debias human and automated scores; however, the two approaches have limitations, particularly when the source of DIF is unknown.

The dissertation of Alexander James Kwako is approved.

Kai-Wei Chang

Li Cai

Mark P. Hansen

Michael H. Seltzer, Committee Chair

University of California, Los Angeles

2023

To my partner, Kathryn, and my mom, Jamie,

who deserve credit for all my good fortune.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deepest gratitude and appreciation to all those who have contributed to the completion of this dissertation. I extend my heartfelt thanks to my advisor, Mike Seltzer, for accepting me into the Social Research Methodology division, and for providing endless support, reassurance, and encouragement to follow my interests. And whose great taste in music has given me the chance to hear new jazz and rock music and musicians! I am truly grateful for his mentorship and constant encouragement throughout this journey.

My thanks go out to all the members of my dissertation committee. To Mark Hansen, for encouraging me to apply for the Transdisciplinary Research Acceleration Grant, for his belief in and guidance throughout this project, and for providing thoughtful and insightful feedback at every step of the way. To Li Cai, for always being available for a statistical consult, and for sharing the breadth of his expertise and wisdom about language assessment. To Kai-Wei Chang, for his commitment to sharing his knowledge of Natural Language Processing with others outside of the Computer Science department, for embarking on this interdisciplinary project, and for generously sharing his resources with us in Education.

I would also like to extend my appreciation to John Rogers, for being a role model of what great leadership looks like, and for showing me how scholarship can make an impact in the world. And my thanks to Bill Sandoval, who introduced me to educational research, and who welcomed me onto his team in my most formative years at UCLA.

My deepest gratitude goes out to my family and my partner. To my mom, for her commitment to my education, for her unconditional love and support, and for cultivating my intellectual curiosity and openness to new ideas. To my dad, for his calm support, and his veneration for serious thinking. To my brother, for teaching me the power of tenacity, for being a vigilant friend, and for renewing my interest in math and beyond. And of course to my partner, Kathryn, to whom I owe my present happiness, who taught me that I could love writing and thinking as an academic, whose curiosity and ways of thinking are an endless source of inspiration.

To all those who have supported me that I have failed to mention, I offer my sincere appreciation. Your contributions, both big and small, have played a significant role in the completion of this dissertation. Thank you for being a part of this transformative experience.

VITA

2010 B.A. (Liberal Arts), St. John's College.

2017 M.A. (Education), UCLA.

2016–2018 Graduate Student Researcher, Developing Teachers' Capacity to Promote Argumentation in Secondary Science, UCLA. Collaborated with secondary school science teachers on implementing Next Generation Science Standards, with a focus on classroom discussion.

2018–2022 Graduate Student Researcher, Institute for Democracy, Education, and Access, UCLA. Analyzed national surveys of U.S. public high school principals. Prepared public-facing reports and articles for academic journals.

2021–2022 Graduate Student Researcher, Center for Research on Evaluation, Standards, and Student Testing, UCLA. Contributed to university efforts to understand effects of COVID on undergraduate experience. Refined survey to understand state-level approaches to English Language Learner pedagogy.

2021–2022 Researcher, Transdisciplinary Research Acceleration Grant, UCLA. Developed proposal to study the use of debiasing techniques in automated English speaking assesment.

2022–2023 Dissertation Year Fellowship, UCLA. For the study of measurement and mitigation of biases in automated English speaking assessment.

PUBLICATIONS

Patterns of classroom talk through participation in discourse-focused teacher professional development. Proceedings of the 13th Annual International Conference of the Learning Sciences, 2018

What is this thing called a mechanism? Findings from a review of realist evaluations. New Directions for Evaluation, 2020.

Do Adolescents Want More Autonomy? Testing Gender Differences in Autonomy Across STEM. Journal of Adolescence, 2021.

Do Politics in Our Democracy Prevent Schooling for Our Democracy? Civic Education in Highly Partisan Times. Democracy & Education, 2021.

Using Item Response Theory to Measure Gender and Racial Bias of a BERT-based Automated English Speech Assessment System. Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications, 2022.

Principals' Responses to Student Gun Violence Protests: Deter, Manage, or Educate for Democracy. Teacher's College Record, 2023.

CHAPTER 1

Introduction

Automated assessment of non-native (L2) English speaking proficiency is made possible by recent advances in Natural Language Processing (NLP). Researchers have shown, for instance, that pretrained large language models (LLMs) can accurately replicate human rater scores in English speaking assessment (Wang et al., 2021). Although applications of NLP present new opportunities for language assessment, recent studies have revealed that NLP can propagate and, in some cases, amplify negative stereotypes of marginalized groups (Blodgett et al., 2020). Societal biases become embedded in NLP models, which may lead to unfair scoring, e.g., where examinees of a particular racial background are systematically given lower scores than others (Wang et al., 2018). Perhaps more commonly, biases of NLP-based assessments are not examined at all (e.g. Collier and Huang, 2020; Ormerod et al., 2022)).

Considering how widespread and high stakes English speaking assessments are at both the primary and secondary education levels (Cimpian et al., 2017; Educational Testing Service, 2005), it is imperative that these assessments be fair for all students, regardless of gender or racial background. This dissertation presents a set of three studies aimed at addressing the need for deeper analyses of bias in automated assessment of non-native (L2) English speaking proficiency.

This study draws on data from the English Language Proficiency Assessment for the Twenty-First Century (ELPA21), a consortium involving 7 state education agencies in the U.S. (Huang and Flores, 2018). Both the quantity and quality of data, as well as the consortium's openness to research, make ELPA21 an ideal context in which to study biases of automated assessments.

In addition to examining gender and racial biases, I explore the use of debiasing techniques to

mitigate such biases (Sun et al., 2019). As debiasing is a relatively new area of research and has not yet been applied to language assessments, the primary goal is to examine advantages as well as disadvantages afforded by various approaches. This project addresses important gaps in English speaking assessment systems, so that NLP may be applied responsibly, in order to ensure fairness for all examinees.

1.1 Automated speaking assessments

English language assessment is mandated in the United States by the Every Student Succeeds Acts (Every Student Succeeds Act, 2015). Tests are administered to over 4 million K–12 students annually (Irwin et al., 2021) and, for university admissions, over 500,000 students take the Test of English as a Foreign Language (TOEFL; Educational Testing Service, 2005). These test are often high stakes, affecting students' graduation rates and university admissions (Cimpian et al., 2017).

In order to cut costs and meet the rising demand for English language assessments, some test developers have transitioned to fully automated assessment systems. These systems automate all aspects of English language assessment, including speech and writing, which in the past have been scored solely by human raters (Evanini et al., 2017). Several researchers, however, have questioned the fairness of automated speaking assessment systems, particularly for examinees of minority groups (Wang et al., 2018b; Collier and Huang, 2020).

1.2 Advantages of automation

While the advantages of NLP-based assessments are typically framed in terms of efficiency and affordability, NLP also has the capacity to improve reliability and even advance social justice-oriented goals. In the testing literature, it is well known that human raters are subject to cognitive and social biases (Engelhard, 2002). Psychometricians sometimes categorize these biases based on raters' tendencies toward excessive severity or leniency, or whether they are prone to halo effects

(Saal et al., 1980). These biases persist even in the face of training and monitoring (Engelhard, 1994), indicating that they operate at an unconscious level (Spencer et al., 2016), making them difficult to address.

Automated systems, in contrast to human raters, are not prone to the same kind of inconsistencies or unconscious biases. Indeed, automated scoring can promote fidelity of scores by identifying and mitigating human rater errors (Bejar, 2011). Wang et al. (2018b) have demonstrated how automated systems can be used to identify overly lenient or harsh human raters in the TOEFL exam: A similar approach, though not yet attempted, could be used to identify raters with biases against minoritized groups.

While not prone to the same sorts of errors as human raters, automated systems can still be biased. However, there are methods for identifying and mitigating such biases in NLP-based applications. Zhao et al. (2018b), for example, have shown how it is possible to debias an NLP-based application that resolves coreferences (e.g., identifying subjects of sentences when pronouns are ambiguous). Originally, the application was more likely to ascribe historically male professions (such as politician or doctor) to male subjects; however, after supplementing the corpus with less biased data, the model became more gender-neutral in its coreference resolutions. Whether social biases are present in humans or in automated systems, they can be measured and mitigated with careful planning and engineering.

1.3 Examples of bias and debiasing

There is little research on bias in automated testing of English speaking proficiency. In order to illustrate what sorts of problems might occur with these systems, I consider two hypothetical examples of how biases might become embedded in automated English language assessment systems, and how bias reduction techniques could be used to address them.

1.3.1 Insufficient training data for automated speech recognition

Automated speech recognition (ASR) systems have been shown to be less accurate for non-White speakers (Koenecke et al., 2020) and, in some cases, less accurate for women (Tatman, 2017; Tatman and Kasten, 2017). These disparities may be due to lack of representation in the data itself (e.g. Zhao et al., 2018b). In English language assessment, a less accurate ASR system will generate less accurate transcripts, which may equate to less accurate scores for language-minority examinees. In some cases, ASR systems are used to generate linguistic features (e.g., pronunciation subscores), which may be biased against language-minority examinees.

One possible debiasing solution for language assessment systems that lack sufficient data for language-minority groups is to implement a filtering model. The filtering model could flag some responses as non-scorable and send them to human raters for review. Bypassing automated scoring in some cases may help to ensure that certain examinees (e.g., those who are known to have less accurate transcripts) are not penalized by the ASR system.

A more comprehensive solution would be to use a technique known as data augmentation (e.g. Zhao et al., 2018b). A simple version of data augmentation is simply to duplicate the data of underrepresented groups to artificially make them more representative. Although requiring more effort, it is also possible to apply various transformations to audio data (e.g., changing the pitch of male and female respondents, or mixing and matching speakers with different accents). These transformations would help to prevent the model from learning construct-irrelevant features like pitch or accent. This approach would remove the ability of the model to learn differences in race or gender at the data level.

1.3.2 Implicit bias in human ratings

Scholarship on implicit bias demonstrates that human behavior is influenced unconsciously and in diverse contexts by culturally-embedded associations (Greenwald and Krieger, 2006). These associations, including negative stereotypes about underrepresented groups, typically operate

independently of individuals' explicit attitudes (Karpinski and Hilton, 2001) and they are often activated by peripheral cues (Spencer et al., 2016).

In the context of English speaking proficiency, it is possible that examinees' accents may trigger human raters' implicit biases. For instance, as a whole, the U.S. population has an implicit bias against faces that have Arabic features (Park et al., 2007); it is possible that hearing a Middle Eastern accent might prompt a rater from the U.S. to (unwittingly) give a lower score. Given the saliency of contextual cues, these biases might be more prevalent when raters hear adult voices with heavier accents, and for raters who are distracted, hungry, or otherwise on autopilot.

In addition to filtering and data augmentation, another technique that would ameliorate the propagation of implicit bias is retraining the model using adversarial neural networks. Widely applicable to a number of problems encountered in deep learning, adversarial networks actively prevent the model from learning certain information. Following the example of Zhang et al. (2018), it would be possible to ensure that the language model is unable to predict the ethnic origin of examinees' responses. In removing its ability to infer race or gender, it may also remove the source of implicit bias.

1.4 Overview of study design

The general research design involves four main components: using automated transcription services to generate text from speech, constructing a language model to score examinees' (transcribed) text, measuring gender and racial biases via analysis of differential item functioning (DIF), and exploring solutions for debiasing models. Speaking items are selected from the English Language Proficiency Assessment for the 21st Century (ELPA21), to reflect a range of ages and expected duration of response. It is hypothesized that adult voices and longer items will trigger more implicit bias, and hence will be more likely to show DIF. As a part of this analysis, I examine biases in examinees' transcripts generated by the automated transcription service.

To measure gender racial bias, I focus on a specific type of bias known as DIF, common in

educational and psychological assessment (American Educational Research Association et al., 2014). Briefly, DIF occurs when equally proficient individuals who belong to different groups (e.g., male and female) are given consistently different (i.e., higher or lower) scores for certain items. Although items in large-scale assessments, including ELPA21, are analyzed for DIF prior to operational use (Anderson, 2015), there are multiple methods of testing for DIF, and patterns of DIF may change over time. The specific approach that I use to identify DIF is specifically suited to the context of this research study. Debiasing NLP-based applications is a relatively new field of research. It will be of interest to compare multiple methods of debiasing, and to determine how well each addresses DIF.

The above research goals are organized within three interconnected studies:

1. Study 1 quantifies the biases of large-scale, automated transcription services, by comparing average WER, disaggregated by gender and L1 background.

2. Study 2 identifies patterns of DIF in human raters, and determines if an LLM-based automated scoring system introduces or exacerbates these patterns of DIF.

3. Study 3 explores two techniques to mitigate bias in automated systems, and reflects on the advantages and limitations afforded by these techniques.

These three studies are discussed in separate chapters (Sections 4, 5, and 6, respectively), which enumerate study-specific research questions and methods. However, these studies also share much in common, drawing from a shared conceptual framework and many of the same methods: The literature review (Section 2) and discussion of methods (Section 3) apply generally to all three studies.

CHAPTER 2

Literature Review

This study is situated at the intersection of (1) English speaking assessment, (2) measurement of bias, and (3) language modeling. With respect to English speaking assessment, I review (a) relevant literature on the prevalence and high stakes nature of English speaking assessment in the United States, and (b) automated speaking assessment systems currently in use. With respect to measurement of bias, I discuss (a) the impact of gender and native language (L1) on L2 English speaking proficiency, and (b) differential item functioning (DIF) both in general and as it relates to English speaking assessment. Finally, with respect to language modeling, I review research on large language models (LLMs), how these models become embedded with societal biases, and how some researchers have been able to debias LLMs for specific applications.

2.1 English speaking assessment

2.1.1 English speaking assessment in the United States

English language assessment is widespread and often holds high stakes for examinees in the United States. The Every Student Succeeds Acts (Every Student Succeeds Act, 2015) requires states to administer English language assessments to all students from non-English speaking homes, and tests are administered to over 4 million K–12 students annually (Irwin et al., 2021). English language assessments are also prevalent at the post-secondary level: Every year, over 500,000 examinees take the Test of English as a Foreign Language (TOEFL), administered by Educational Testing Service (Educational Testing Service, 2005), and 3.5 million people take the International English

Language Testing System (IELTS; International English Language Testing System, 2023). In terms of real-world impact, many universities require applicants to score above a certain threshold of proficiency on the TOEFL or IELTS as an admission requirement. For K–12 students, being labeled as an "English learner" decreases high school graduation rates and college-going behavior, even among students of similar English language proficiency (Johnson, 2019).

Nearly all language assessments include speaking as one of four language proficiency domains, along with listening, reading, and writing (Council of Chief State School Officers, 2012). At the K–12 level, speaking tasks are usually open-ended, yet have a narrow contextual focus; for instance, students are asked (via written or verbal prompt) to describe what is happening in a picture (Luoma, 2004), At the post-secondary level, examinees are given more challenging tasks, e.g., talking about a particular topic for several minutes. For most assessments, examinees speak into a microphone and responses are scored by a human rater or an automated system. Using a slightly different approach, International English Language Testing System (2023) employs trained interviewers to administer and score speaking proficiency in a face-to-face setting.

For K–12 students, the majority of states administer assessments developed by one of two consortia, the World-Class Instructional Design and Assessment (WIDA) or the English Language Proficiency Assessment for the 21st Century (ELPA21). Neither consortium currently uses automated speaking assessment. Rather, for speaking items, test-takers speak into a microphone for 5–10 seconds, whereafter their responses are passed along to human raters to score.

There are two states, however, that do use automated assessment for K–12 speaking proficiency. Both states use systems developed and administered by Pearson. The Texas English Language Proficiency Assessment System (TELPAS) uses Versant, one of the first automated speaking proficiency assessments, originally developed for large businesses and administered over the telephone (Pearson Education, Inc., 2019). Given the lack of validity studies and potential unreliability of its automated system, TELPAS' recent switch to Versant has sparked some skepticism among academics (Collier and Huang, 2020). Pearson also helped to develop the speaking assessment for the Arizona English Language Learner Assessment (AZELLA), seemingly independent of Versant (Johnston et al.,

2019). At the post-secondary level, TOEFL uses an automated speaking assessment system known as SpeechRater, developed by researchers at ETS (Chen et al., 2018b).

2.1.2 Current NLP-based automated speaking assessments

Chen et al. (2018b) identify four primary components of automated speaking assessment systems: (1) an automated speech recognition (ASR) system, which includes speech-to-text transcription, (2) the extraction of linguistic features from audio and text data, (3) a filter model to identify non-scorable responses (e.g., those with audio errors or potential plagiary), and (4) a scoring model to combine linguistic features into a single score.

Depending on the purpose or sophistication of the automated assessment system, some of the above steps may be combined or omitted. Pearson, for instance, appears to omit the filter model (step 3) for TELPAS and AZELLA. In a different vein, Chen et al. (2018a) have experimented with combining steps 1, 2, and 4 into a single end-to-end model. (Models are described as end-to-end when multiple intermediate processes get subsumed within a single architecture.) Although the performance of end-to-end models can be impressive, they are a black box—that is, the linguistic features that the model learns and uses to score examinees' responses are hidden. For this reason, it can be challenging to interpret or debug end-to-end models.

2.1.2.1 (1) Automated speech recognition

The ASR system built by ETS is comprised of multiple parts, and used for multiple purposes. Under the hood, it uses a Hidden Markov Model to parse phonemes (i.e. syllables), an n-gram model to track word dependencies, and a weighted finite state transducer to limit the search space (Qian et al., 2019). Combined, these three components are able to generate transcripts and assign probabilities to text. The transcripts are used for further data processing, while the probabilities associated with the transcripts may be used as a linguistic feature (referred to as the "Average ASR Confidence Score") in the scoring model (Zhang et al., 2019).

Training ASR systems requires a prolific amount of labeled data (i.e., audio files accompanied by accurate transcripts). The exact amount of training data required to build an ASR system depends on the complexity of the speech. For speech assessment purposes, Qian et al. (2019) suggests on the order of a million words; Evanini et al. (2017) suggests 200–300 hours of speech. Data can be supplemented with external speech corpora, e.g., the Switchboard corpus (Godfrey and Holliman, 1997), in order to improve the accuracy of speech-to-text transcription; but ensuring recording quality across data sources can be challenging (Qian et al., 2019).

Even with external data, there may yet remain a significant amount of transcription error. A common index of transcription accuracy is word error rate (WER), which is the ratio of the total number of errors (i.e. erroneous insertions, deletions, or substitutions) to the total number of words in a given transcript or corpus (Qian et al., 2019). (See Section 4.2.3 for details regarding how WER is computed.) Human-human transcription of non-native speech typically yields a WER of 15–20% (Zechner, 2009). SpeechRater comes close to human parity, with a WER of 23% for non-native spontaneous speech (Tao et al., 2016). AZELLA has a higher WER of 35% (Cheng et al., 2014). By contrast, human-human WER of native speech cam be as low as 5% (Xiong et al., 2016).

One of the significant challenges of transcribing non-native speech is faithfully capturing errors. Rather than reproduce errors, ASR systems tend to predict a similar, grammatically correct substitution. One internal study showed that SpeechRater failed to capture 70% of non-native speakers' grammatical errors (Yoon et al., 2019).

2.1.2.2 (2) Linguistic features

Once audio data has been transcribed, linguistically-relevant features are extracted from text and audio data. Linguistic features are typically defined a priori, in alignment with standards of speaking proficiency (e.g. Brown et al., 2005). For instance, to assess fluency, SpeechRater calculates (among other things) the number of pauses, the duration of pauses, and the number of words per second in examinees' speech (Hsieh et al., 2019). To assess vocabulary, SpeechRater calculates the number of

unique words that examinees use (Yoon et al., 2019). In TOEFL, over 60 linguistic features have been examined as candidates for speech assessment. (Note: In an end-to-end model, these apriori decisions about which linguistic features to use are not possible, since the features are latent.)

2.1.2.3 (3) Filter model

A filter model helps to reduce error in the automated system by identifying "non-scorable responses" (Loukina and Yoon, 2019). Such responses may have had audio issues; alternatively, they may be the result of uncooperative examinees or plagiary. In such cases, examinees' responses may bypass the automated assessment system, and instead be designated for human rating. Common technical difficulties include the presence of background noise and mechanical issues with recording devices; as described above, ASR systems can be sensitive to slight discrepancies in audio quality, which may produce inaccurate transcripts and, subsequently, inaccurate test scores. There are also instances in which a filter model can be used to identify examinees who are cheating or otherwise gaming the system.

The filter model is helpful in improving the ASR system by acting as a manual override. Automated assessment systems are known to have certain weaknesses, and rather than build a system that is capable of handling every type of exception, it may be more cost-effective to bypass the automated assessment system in certain cases.

2.1.2.4 (4) Scoring model

The scoring model combines linguistic features to produce a summative language proficiency score for examinees (Loukina and Yoon, 2019). Finding an appropriate scoring model also requires a lot of data. In order to train a scoring model for complex constructed item types, Zechner (2019) suggests having at least 10,000–20,000 pre-scored responses. More complicated scoring models, however, might require significantly more data.

TOEFL uses linear regression, wherein features are weighted and selected using a LASSO-

based method (Loukina et al., 2015). The primary rationale for using linear regression is that it is relatively transparent and easy to communicate; Loukina and Yoon (2019) report that ETS has experimented with more complex scoring models, e.g., random forests, but they were not worth the 2–3% improvement in accuracy. Zechner (2019) reports that, at the item level, their scoring model is more reliable than human-human reliability coefficients (r = 0.65, compared to r = 0.55–0.60). In contrast to TOEFL, AZELLA combines features using a deep learning model with a single layer of latent variables (Cheng et al., 2014).

2.2 Biases in English speaking assessments

2.2.1 Issues of fairness in automated speaking assessment

During the course of test development, it is standard practice for a bias review committee to screen all items for inappropriate or content-irrelevant material, and to analyze items for differential item functioning (DIF) to ensure quantitatively that items are not biased against any particularly groups of examinees, e.g., women or racial minorities (American Educational Research Association et al., 2014). It is likely that DIF analyses were conducted by Pearson and ETS for speaking items scored by human ratings, but it is not clear whether they were repeated for automated scoring systems. Given that Versant (Pearson's ASR system for TELPAS) was not originally designed to assess K–12 students, and some evidence of inconsistency among examinees' scores (Collier and Huang, 2020), it is troubling such analyses have not been made public (or, possibly, even conducted).

2.2.2 Impact of gender and L1

In the context of bias in educational assessment, *impact* refers to the unconditional difference in scores between groups of examinees (Angoff, 1993). Most of the research on the impact of gender and L1 on language proficiency touches on myriad aspects of language proficiency, and is not specific to speaking proficiency. In general, female examinees perform better than male examinees

on language assessments (Reilly et al., 2019), and this typically including L2 English language assessments (Denies et al., 2022). These differences have early developmental roots, e.g., in word acquisition (Kaushanskaya et al., 2013) and phonetic acquisition (Dodd et al., 2003). Importantly, gender differences vary by sociocultural factors (Denies et al., 2022), and others have pointed out that some language tasks do seem to favor males (Wucherer and Reiterer, 2018).

Native language (L1) also impacts L2 English speaking proficiency. One body of literature, on "language transfer" or "cross-linguistic influence," focuses on similarities and differences between language structures (Brown et al., 2000, p. 257). One intuitive finding, for instance, states that examinees perform better on vocabulary tests when there are more L1 cognates (Leśniewska et al., 2018). Other literature examines the more complex sociocultural factors that play a role in the development of L2 English speaking. For instance, Derwing and Munro (2013) attribute differences between Slavic and Mandarin English speaking proficiency to factors like age, motivation, and conversational opportunities, which interact with L1 in complex ways.

2.2.3 Differential item functioning

This study focuses on a specific type of bias in English speaking proficiency known as differential item functioning (DIF). Historically, interest in DIF grew out of a movement in the 1960s to make standardized assessments more fair for racially underrepresented minorities by removing cultural biases from test items (Angoff, 1993). Approaches for identifying DIF continue to be used in test development to screen items for potential biases against specific groups. For example, an item with extensive references to the Bible would give Christian examinees an advantage. Such references would not be appropriate unless the explicit goal of the test was to measure knowledge of Christian theology, otherwise the item would be measuring a content-irrelevant construct (American Educational Research Association et al., 2014).

In educational assessment, a test item is said to exhibit differential item functioning (DIF) when "equally able (or proficient) individuals, from different groups, do not have equal probabilities

of answering the item correctly" (Angoff, 1993, p. 4). By definition, DIF is not a measure of unconditional mean differences (or first-order group differences), which is known as *impact*. Rather, DIF conditions on overall (and ideally unbiased) proficiency. In other words, DIF occurs when examinees who should receive similar scores (based on their performance on other test items) receive different scores on a specific item under review. In DIF terminology, overall proficiency is referred to as the *matching criterion*, the majority group is referred to as the *reference group* and the minority group is referred to as the *focal group*.

For instance, in college admissions tests of reading comprehension, males (the reference group) tend to score higher on science-related passages, whereas females (the focal group) tend to score higher on humanities-related passages, conditional on examinees' overall language proficiency (i.e., the matching criterion; Steedle et al., 2023).

There are many possible causes of DIF. The above example highlights the possibility of confounding variables (e.g. males tend to major in STEM at a higher rate; Sloane et al., 2021). One of the central concerns of this paper is that DIF might arise from implicit bias in human raters—that is, human raters might be unconsciously influenced by phonic features in examinees' voices. Also pertinent to this study, DIF could be caused by discrepancies in automated transcription of speech, or sociocultural differences rooted in examinees' gender or L1 backgrounds. Test developers have lamented that determining causes of DIF can be frustrating since studies often yield inconclusive results (Zumbo, 2007).

2.2.3.1 DIF based on gender and L1

Although there are many studies of DIF with respect to gender and L1 in large-scale English language assessment, most of these studies focus on vocabulary, listening, and writing proficiency (Kunnan, 2017). There are very few studies that focus on English speaking proficiency.

Although DIF analyses of automated speaking assessment systems are not publicly available, Wang et al. (2018b) have conducted some analyses of bias in SpeechRater. Analyzing examinees'

overall scores, they found SpeechRater to be somewhat partial towards some groups over others, based on examinees' L1 background. Although they suggested some plausible explanations for these discrepancies, e.g., an overall "[reduced number] of predicted scores at the extremes of the scoring scale" (p. 117), they concluded that more research was needed to explore these hypotheses.

2.2.3.2 Sources of DIF

Issues of fairness in automated speaking assessment are underexplored (or underreported), but examinations of sources of bias are practically non-existent. Based on literature in adjacent fields, however, it is possible to construct a list of likely sources of DIF in automated speaking assessment.

Automated transcription bias

One potential source of DIF in automated speaking assessment lies in automated transcription. Dichristofano et al. (2023) has shown that the largest providers of automated transcription services (Google, Amazon, and Microsoft) all have discrepancies in transcription accuracy based on speakers' L1 background. As text transcripts contribute the most important (and sometimes exclusive) input for most NLP-based scoring systems, it is important to consider transcription discrepancies, as they may lead to discrepancies in scores.

Human rater bias

Another potential source of DIF lies in the quality of human ratings. Although human ratings are often considered the gold standard, and the benchmark against which to evaluate the success of automated scoring systems (Zhang et al., 2019), human ratings are not ideal scores. An ideal score would be "assigned by a panel of [unbiased] expert raters and represent the true score of a spoken response" (Zechner, 2019). Although human raters receive training and are monitored over time to improve consistency (Engelhard, 2002), human-human interrater reliability typically ranges between 0.55 and 0.60 on individual items (Zechner, 2019).

Scholarship on implicit bias demonstrates that human judgment is influenced unconsciously by peripheral cues, including speakers' accents (Kang and Yaw, 2021). These biases may lead to unfair

scoring without raters even realizing it (Greenwald and Banaji, 1995). Indeed, Winke et al. (2013) reports that human raters are more lenient towards examinees who share the same L1 background. In a summary of research on the biases of raters of L2 English, Lindemann and Subtirelu (2013) report a strong disconnect between subjective evaluation of speech (e.g. using Likert scales) and more objective measures (e.g. transcription). One limitation with these studies is that they assume that L2 English speakers are of similar proficiency levels, and it is not clear whether the researchers controlled for speakers' English proficiency.

Research on implicit bias and speech suggests that, in the context of English language assessment, there may be more bias in the domain of speaking (e.g. as opposed to writing). By listening to examinees' voices, human raters may be more likely to be influenced by examinees' accents, triggering implicit bias that affects judgment during scoring.

Sociocultural factors

There are many sociocultural differences based on gender and L1 that could affect English speaking assessment. Derwing and Munro (2013), for instance, discuss how factors like age and conversational opportunities interact with L1 in nuanced ways. Gender is also a source of variation in L2 English speaking proficiency, although it varies by culture and task (Denies et al., 2022).

There may also be cultural differences that interact with item properties. Now a classic example, Freedle (2003) describes how some test items draw on cultural knowledge that disadvantage minority examinees. It is possible that certain speaking items may require an understanding of the context of schooling in the U.S.; students of some racial backgrounds may be more (or less) familiar with these practices than others.

Feature bias

One potential source of bias lies in the linguistic features and how they are combined in the scoring model. Although research has not yet examined this issue, Zhang et al. (2019) warns that "interactions of [features] and demographic groups can lead to subgroup biases" (p. 52). In broader terms, English speaking is not a monolithic construct, and without specifying these differences in

the scoring model, speakers of some language-minority backgrounds might be unfairly downgraded. For instance, speakers of some ethnic backgrounds might enunciate more slowly which may not be a marker of proficiency but simply a regional difference. Failing to specify this in the model would lead to biased scores for these examinees.

Speaking proficiency is not a simple construct. It may require non-linear combinations of multiple layers of latent variables to capture the complexity of fluency or discourse coherence. At the same time, choosing a more complex model may reduce the transparency of the linguistic features, and may make it more difficult to explore sources of bias (Loukina and Yoon, 2019).

Machine learning bias

NLP-based automated assessments often rely on machine learning methods, which are susceptible to introducing and exacerbating biases at multiple stages of development (Suresh and Guttag, 2021). Pretrained LLMs are especially susceptible to biases because they are trained on large corpora of text scraped from the web and social media (Blodgett et al., 2020). Social biases in source text resurface in downstream applications, including machine translation (Stanovsky et al., 2019) and sentiment analysis (Kiritchenko and Mohammad, 2018). In educational assessment, machine learning biases ultimately may influence examinees' scores. In section 2.3.2, I consider in greater detail how these sources of machine learning bias can occur in language modeling and in LLMs more broadly.

Other biases

Although there are too many potential sources of DIF to review in full, I highlight several miscellaneous sources of DIF that bear on the inquiry of L2 English speaking assessment. Huang et al. (2016) report that curricula vary across countries, and that these differences are a likely source of DIF in international assessment. This source of bias would also be related to the age at which examinees emigrated and entered into the U.S. schooling system. In another vein, it has been suggested that easier items might be more likely to exhibit DIF, and that this may be attributable (in part) to guessing behavior, which is in turn related to overall proficiency (Dorans and Zeller, 2004;

Santelices and Wilson, 2010). Given that speaking is a difficult aspect of L2 language acquisition (Brown et al., 2000), it is possible that examinees who are less fluent are able to guess their way through non-speaking items, but struggle with speaking items in particular.

2.3 Language modeling

2.3.1 Large language models

Language modeling refers to the construction and use of probabilistic models for language generation and inferential tasks (Jurafsky and Martin, 2023). In automated language proficiency assessment, language modeling is used in multiple ways and at multiple stages of the automated assessment process. As described above, ETS uses multiple language models which, when combined, produce speech-to-text transcriptions and linguistic features relevant to speech proficiency (Qian et al., 2019).

Language models with at least 1 layer of variables between the input and the output are classified as deep learning models (Goodfellow et al., 2016). Such models are often characterized as black boxes because the features that become embedded in the coefficients of the models are latent, and often unclear (e.g. Gretter et al., 2019). Despite the inscrutability of deep learning models, they are popular because they are easy to deploy, adaptable to a variety of tasks, and often quite powerful. Although they are not widely used in language assessment, initial trials have shown that deep learning models are moderately more accurate than the more common approach of combining manually-constructed features (Chen et al., 2018a).

A prolific amount of data is required to train most language models and, as a general rule, the more complicated the model, the more data is required. Currently there is on the order of 105 hours of speaking data publicly available (Galvez et al., 2021), and hundreds of terabytes of uncompressed text data (Iderhoff, 2023). Industry leaders like Google and AWS, however, draw from considerably more data. Sophisticated language models, whether speech or text-based, take a

lot of computational resources to train. RoBERTa, for instance, a text-based language model, is comprised of 108 parameters and was trained for five days on 1,024 × Nvidia 32GB V100 GPUs (Liu et al., 2019).

Because of the considerable costs involved in training language models, it is common to use a general language model as a foundation, and fine-tune the model (which requires significantly less data) to perform specific tasks; this technique is known as transfer learning (Jurafsky and Martin, 2023). With transfer learning, the theory is that the general language model can learn the complex semantic and grammatical dependencies of language by training on a large corpus of data. Using this general language model as a foundation, one can then train the model to perform various tasks using a much smaller dataset (sometimes only thousands of observations), and for a fraction of the computational resources (e.g., 1 GPU for 1–2 hours).

One of the most widely used text-based general language models is Bidirectional Encoder Representations from Transformers (BERT), developed by Devlin et al. (2018). BERT was trained as a masked language model, which means it learned language by trying to guess the words in a sentence. More specifically, the model was presented with a sentence in which 15% of the words were masked, and it had to predict what those correct words were. The model is constructed such that it is able to learn dependencies between the masked word(s) and every other word in the sentence using what known as an attention mechanism (Vaswani et al., 2017). Having trained the general language model, it is then fine-tuned to do a variety of tasks.

The performance of general language models, including BERT, are evaluated by their accuracy on various language tasks. One of the most widely used benchmarks is the 9 General Language Understanding Evaluation (GLUE) tasks (Wang et al., 2018a). GLUE tasks vary from identifying grammatically (in)correct sentences to assessing the (positive or negative) sentiment of movie reviews. BERT performs as well as humans on a number of these tasks.

2.3.2 Biases in large language models

There is now a large literature on bias in artificial intelligence (AI), and bias in NLP more specifically. One branch of research has focused on classifying the types of socioeconomic harms caused by biased AI (e.g. Blodgett et al., 2020). Potential harms potentially caused by biased automated assessment are discussed in Section 2.1.1. Instead, this section focuses on the technical aspects of how bias may be introduced into AI systems. Suresh and Guttag (2021) outline seven steps in the AI pipeline in which bias may be introduced. The three sources of bias most pertinent to this study include (1) a type of representation bias, (2) a type of measurement bias, and (3) learning bias. I consider each of these sources of bias in greater detail below.

Although representation bias usually refers to a mismatch between the sample and the population, there are other types of representation bias. For instance, even if the sample is representative of the population, the population itself may contain subgroups too small for the model to learn. In other words, for certain small groups, there may not be enough data for the algorithm to learn how to make accurate predictions. Along these lines, it has been shown that ASR systems are less accurate for non-White speakers (Koenecke et al., 2020) and, in some cases for women (Tatman, 2017; Tatman and Kasten, 2017). In the context of English language assessment, this may be a problem for examinees from language-minority backgrounds.

Part of the reason why representation bias is so prevalent in applications of AI is that deep learning models require a prodigious amount of data. Supplementing data with external speech corpora, e.g., Switchboard corpus (Godfrey and Holliman, 1997), improves the accuracy of speech-to-text transcription (Qian et al., 2019). Yet these external corpora may not be sufficiently large, especially for non-native speakers from minority-language backgrounds.

Measurement bias refers to systematic differences in the operationalization(s) of constructs. With respect to English speaking proficiency, measurement bias may occur when human raters consistently give lower (or less accurate) scores to individuals from a certain group. As discussed above, this may be caused by implicit biases (Spencer et al., 2016) or by interactions between

linguistic features and demographic characteristics (Zhang et al., 2019). Because NLP-based assessments are trained on human-rated data, human biases may become embedded in the model(s).

Finally, learning bias results when specifications of the model itself are susceptible to introducing or exacerbating biases. As an example, Suresh and Guttag (2021) mention how pruning, a common technique in large-scale AI models, often reduces prediction accuracy for small subgroups (Hooker et al., 2020). Such model design choices may also exacerbate existing biases. In one example, Zhao et al. (2017) found that their model predicted individuals' gender in an image by their proximity to objects in the surrounding environment (e.g. women around kitchen appliances, men around computers) at approximately twice the frequency as was found in the original training data. Likewise, for speaking assessment, certain models may be less accurate for smaller groups, or they may amplify biases by fixating on certain acoustic or textual features (e.g., different accents).

One of the principal drawbacks of deep learning models is their inscrutability, and the difficulty of characterizing which features (or which biases) are learned by the model. Although ETS' SpeechRater relies primarily on manually-constructed features, one of their most predictive features, "Average ASR Confidence," is the output of deep learning models (Qian et al., 2019). Pearson does not share details of their models in technical documentation, but they do seem to employ a scoring model that uses deep learning (Cheng et al., 2014).

In addition to the above issues, general language models come prepackaged with their own set of biases that exist even prior to fine-tuning for specific speech tasks. These biases arise from the fact that general language models are pretrained on text from Wikipedia, online news outlets, and even community forums like Reddit (Liu et al., 2019). Given the biases that exist in these spaces, it is not surprising that applications of general language models often reflect strong cultural biases against women and people of color. For example, in coreference resolution, NLP models will associate historically male occupations (such as physician or politician) to masculine pronouns, and historical female occupations to female pronouns (Zhao et al., 2018b).

2.3.3 Debiasing techniques

There are several techniques that have been proposed for mitigating biases in NLP applications. Sun et al. (2019) classify these techniques into two broad classes of methods, retraining and inference, based on the stage at which bias is addressed. In retraining, the goal is to train (or retrain) the model on an unbiased dataset or to change the way in which the model learns from data. Retraining reshapes the model at a more fundamental level, but it can also be more cumbersome or infeasible in practice. In contrast, inference-based methods address problems of bias further down the pipeline by constraining or rebalancing the models' inferences or predictions. Retraining and inference methods are not mutually exclusive, yet researchers have not explored how they might be used in conjunction.

One comprehensive yet laborious retraining solution is known as data augmentation (e.g. Zhao et al., 2018b). A simple way to augment data is to duplicate the data of underrepresented groups to artificially make them more representative, forcing the model to train on equal amounts of data from each group. A more elaborate approach for augmenting data might be to change the pitch of voices, so that male responses are made to occur in traditionally female registers, and vice versa. It would also be possible to mix accents by cutting and pasting responses from different examinees. This approach removes the ability of the model to learn differences in the pitch or accent of the speaker at the data level, and may force it to focus on more relevant aspects of speech.

Another retraining technique employs adversarial neural networks. Widely applicable to a number of problems encountered in deep learning, adversarial networks actively prevent the model from learning certain information. Following the example of Zhang et al. (2018), an adversarial network would penalize the model (during training) for predicting the gender or ethnic origin of examinees' responses. In removing its ability to infer race or gender, it may simultaneously remove the source of implicit bias. One of the benefits of adversarial neural networks is that mitigating bias may not require identifying the problematic features that the deep learning model has learned.

An example of an inference-based debiasing technique that may be relevant to the current

project is constraining predictions. In the context of English language assessment, this may involve conditioning on overall language proficiency. That is, if we know that a group of language-minority examinees is equally proficient as the reference group (based on their performance on other test items), we could constrain the algorithm to give the focal group the same distribution of scores as the reference group on each speaking item. While this approach introduces additional dependencies, it may make the algorithm fairer in its treatment of examinees who face systematic bias.

CHAPTER 3

Methods

3.1 Data

3.1.1 ELPA21

This study draws on data from the English Language Proficiency Assessment for the 21st Century (ELPA21), a consortium involving 7 state education agencies in the U.S. (Huang and Flores, 2018). Approval for this research project was granted by the ELPA21 consortium and the university institutional review board. To maintain confidentiality, certain details regarding test items and examinees are omitted.

In most frameworks for English language assessment, speaking is one of four language proficiency domains, along with listening, reading, and writing (Council of Chief State School Officers, 2012). At the K–12 level, speaking tasks are open-ended, yet have a narrow contextual focus; for instance, students are asked (via written or verbal prompt) to describe what is happening in a picture (Luoma, 2004). Examinees speak into a microphone for up to two minutes, after which their responses are passed along to human raters for scoring. All verbal responses in ELPA21 are currently scored by human raters. Human raters assign holistic integer scores based on scoring rubrics that vary by scale and item type. Consistent with best practices, human raters are trained and monitored over time to ensure consistency (Engelhard, 2002).

Analyses focused on two grade bands (2–3 and 9–12) which corresponded to two different tests. Tests were administered during the 2020–2021 school year. Although different tests were developed for each grade band, we sampled examinees and selected test items to help ensure that these two

sets of analyses were as comparable as possible.

3.1.2 Sample design and demographics

The sampling frame included all examinees in grade bands in 2–3 or 9–12 who met the following inclusion criteria: answered all three speaking items analyzed in this study; answered enough items in each of the other three domains to receive domain-specific scores; and had gender and L1 demographic information available. Furthermore, to limit the scope of the study, we excluded examinees who had an IEP or 504 Plan, examinees with non-binary gender, and examinees whose L1 was not one of the ten L1s analyzed in this study.

From the sampling frame, we sampled 15,000 students (8,377 from grade band 2–3, and 6,623 from grade band 9–12). The size of our sample was limited, in part, by the cost of automated transcription. We included all examinees whose L1 was one of our nine L1 focal groups (Table 3.1). The remainder of examinees were randomly sampled from Spanish speakers.

Demographics of grade bands 2–3 and 9–12 are presented in Table 3.1. Note that there are group differences with respect to overall language proficiency.[1] In both grade bands, male examinees scored slightly lower than female examinees. There is also heterogeneity among L1 groups.

3.1.3 L1 selection

Due to practical limitations, we focused on ten L1 groups. Spanish was the largest L1 group (constituting 82.7% of all examinees in 2020–2021) and, for this reason, served as the reference group. The other nine L1 groups were selected based on the number of examinees available, and with a view to global diversity. See Appendix 8.1 for additional details regarding L1 selection and grouping.

[1] See Section 3.3 for how language proficiency is computed for examinees.

Table 3.1: Sample descriptive statistics, in aggregate and disaggregated by gender and L1.

	Grade Band 2–3			Grade Band 9–12		
	n	**%**	**Avg. Proficiency**	**n**	**%**	**Avg. Proficiency**
All	8377	100	0.18 (0.91)	6623	100	0.16 (0.93)
Gender						
Male	4310	51.5	0.13 (0.9)	3648	55.1	0.14 (0.94)
Female	4067	48.5	0.23 (0.92)	2975	44.9	0.2 (0.92)
L1						
Spanish	4205	50.2	0.08 (0.85)	3481	52.6	0.23 (0.92)
Marshallese	692	8.3	-0.0 (0.86)	891	13.5	-0.05 (0.75)
Russian	862	10.3	0.28 (0.9)	375	5.7	0.49 (0.86)
Vietnamese	522	6.2	0.41 (0.9)	402	6.1	0.36 (0.93)
Arabic	499	6	0.33 (0.88)	414	6.3	0.06 (0.86)
Mandarin	439	5.2	0.88 (0.89)	203	3.1	0.44 (1.02)
Hindi	416	5	0.75 (0.82)	185	2.8	0.67 (0.82)
Mayan	238	2.8	-0.66 (0.88)	258	3.9	-0.84 (0.95)
Persian	295	3.5	-0.05 (1.01)	197	3	-0.07 (0.94)
Swahili	209	2.5	0.22 (0.87)	217	3.3	0.04 (0.93)

3.1.4 Item selection

Speaking items were selected to span a range of response times (i.e., length or quantity of speech). Specifically, for each grade band, we selected one speaking item that was short in duration (i.e., requiring examinees to produce a phrase or simple sentence to answer the prompt), one medium-length item (i.e., requiring 2–3 sentences or a compound sentence), and one long item (i.e., requiring 3+ sentences). Table 3.2 presents the lengths of items 1–3, based on average audio duration (in seconds) and average number of words, for both grade bands. To increase comparability between grade bands, our selection of items also took into consideration item type and item information.

Table 3.2: Item descriptive statistics.

		Grade Band 2–3			**Grade Band 9–12**		
Item #	**Length**	**Num. of Categories**	**Avg. Seconds**	**Avg. Words**	**Num. of Categories**	**Avg. Seconds**	**Avg. Words**
Item 1	Short	3	6.4 (4.9)	6.0 (6.5)	4	8.3 (5.0)	11.5 (7.1)
Item 2	Medium	5	17.2 (13.3)	25.1 (23.2)	6	14.9 (9.1)	22.8 (16.7)
Item 3	Long	6	36.9 (23.1)	51.1 (35.0)	5*	34.7 (18.9)	65.0 (38.4)

Note: Item 3 for grade band 9–12 was re-scaled from a 6-point scale to a 5-point scale. This change was made in light of the fact that one group of respondents (Hindi) did not receive any 1s. Combining 1s and 2s improved model convergence.

3.2 Transcription

3.2.1 Automated transcription

Automated transcripts were generated using Amazon Web Services. Transcript requests were sent using Amazon's API, boto3, during October 7–12 and November 14–16, 2022. Scripts were written

in Python 3.8.12 (Python Software Foundation, 2022). Default transcription settings were used, with output language set to "en-US." Amazon provides multiple transcripts by default; the most probable transcripts were selected for analyses.

3.2.2 Transcript standardization

Text was standardized in order to remove certain stylistic differences between automated transcription services. Following the example of Koenecke et al. (2020), standardization included: changing numerals to unhyphenated words (e.g. "forty two" instead of "42"); removing punctuation, with the exception of apostrophes and translatable symbols (e.g. "percent" instead of "%"); removing expressions of hesitation (such as "um" or "hm"); converting all text to lowercase; and formalizing slang (e.g. replacing "cuz" with "because").

3.3 Differential item functioning (DIF)

As discussed in Section 2.2.3, DIF occurs when there are group differences, conditioned on "unbiased" proficiency estimates. The "unbiased" proficiency estimate, θ, is referred to as the *matching criterion*. In this study, the matching criterion is examinees' non-speaking English language proficiency. By excluding speaking items, we ensured that estimates of θ were not contaminated by the same type(s) of biases under examination. To compare examinees' of similar θ, the sample was divided into ten strata based on which quantile of the standard normal distribution their non-speaking English proficiency resided.

The majority group is referred to as the *reference group*; and the minority group is referred to as the *focal group*. For gender, the reference group is male, and the focal group is female; for L1, the reference group is Spanish, and the nine focal groups are listed in Table 3.1.

3.3.1 Matching criterion

Examinees' non-speaking English proficiency was used as the matching criterion in DIF analyses. Non-speaking proficiency was inferred from examinees' responses to test items in non-speaking domains (i.e. listening, reading, and writing). Items were modeled using an Item Response Theory (IRT) framework (Cai et al., 2016), consistent with modeling choices used in production. One difference, however, was that non-speaking items were modeled as a unidimensional construct, as opposed to being modeled as three correlated dimensions, because (1) it simplified interpretation of the matching criterion, which was non-speaking proficiency as a whole, (2) it yielded smaller margins of error, and (3) model fit was acceptable for both grade bands, in terms of limited-information fit statistics and Tucker-Lewis (non-normed) fit indices (M2 RMSEA $\leq .03$ and M2 TLI $\geq .96$).

3.3.2 DIF effect sizes

As summarized by Michaelides (2008), a common method to evaluate DIF for ordinal items is based on the standardized mean difference (SMD) between reference and focal groups (Dorans and Kulick, 1986).[2] SMD is calculated as follows:

$$SMD = \sum_j \frac{N_{F.j}}{N_{F..}} \frac{\sum_u N_{Fuj} u}{N_{F.j}} - \sum_j \frac{N_{F.j}}{N_{F..}} \frac{\sum_u N_{Ruj} u}{N_{R.j}}$$

where N_{Fuj} is the number of examinees in the focal group F who have a non-speaking language proficiency score that puts them in stratum j, and who received score u (on the item in question). Multiplying this quantity by the score, u, and dividing by the number of examinees in the focal group in stratum j, yields the expected score for the focal group. A similar procedure is followed for the reference group (in the rightmost expression). Before taking the difference, the expected scores are weighted by the proportion of examinees in the focal group in stratum j.

[2] In the approach used in these studies, instead of using the Mantel test (Mantel, 1963), significance tests were based on bootstrap sampling distributions and B-H adjusted p-values, described in Sections 3.4 and 3.5.

The effect size, z, is the ratio of SMD to the standard deviation (pooled between the two groups).[3] Intuitively, z represents how much the focal group outperforms the reference group, comparing examinees of similar proficiency, in units of standard deviation.

What counts as a large or small effect size is based on a system originally proposed by Zwick et al. (1993) and is currently in use by the Educational Testing Service and other educational assessment organizations. Generalizing the system to ordinal items, Allen et al. (2001, p. 150) designates items as having strong DIF (labeled "CC") if z is greater than or equal to 0.25. Items have weak DIF ("AA") if z is less than 0.17. And items have moderate DIF ("BB") if z is between 0.17 and 0.25.

3.3.2.1 Absolute effect size

For certain research questions, the primary interest is not in determining which specific groups are (dis)advantaged, but only in quantifying the amount of DIF. In other words, we are not interested in the direction of DIF, but only the magnitude. To address these questions, we base our analyses on the absolute value of z, $z_{abs} = |z|$. We also refer to this metric as the absolute effect size or absolute DIF.

3.3.2.2 Differences between effect sizes

We also computed differences in effect sizes (i.e. differences in DIF between human and automated scores, differences between items, and differences between grades). In each of these comparisons, we were interested not in DIF itself, but in first-order differences of DIF. We refer to these quantities as $\Delta z = z_i - z_j$, and $\Delta z_{abs} = |z_{abs,i} - z_{abs,j}|$, where i and j represent two different effect sizes (e.g. DIF based on human vs. BERT scores, or DIF of item 1 vs. item 2). In research questions 2–3, we also examine second order differences, $\Delta\Delta z_{abs} = |\Delta z_{abs,i} - \Delta z_{abs,j}|$.

[3] Ormerod et al. (2022) refer to the effect size as z, a convention we follow.

3.3.3 Aggregate DIF metrics

Analyses of DIF typically revolve around pairwise comparisons at the item level. This fine-grained level of analysis is not suited for making general claims about DIF (i.e. across multiple items or multiple focal groups). Aggregating effect sizes allows us to make more general claims about DIF.

3.3.3.1 Overall DIF

To evaluate DIF across items, we computed z based on examinees' summed score (i.e. summed across all items of interest). That is, for grade bands 2–3 and 9–12, we added examinees' responses to items 1–3, and computed z according to the procedure outlined in Section 3.3.2. Since z is in units of standard deviation, it is unaffected by differences in items' scales, and thus generalizes well to a summed score.

3.3.3.2 Factor DIF

Analyses of DIF are usually localized to pairwise comparisons involving one focal group and the reference group. For factors containing more than one focal group, however, we are interested in evaluating DIF for the factor as a whole. To evaluate DIF for the entire factor, we take an unweighted stratified mean of all pairwise comparisons, $\bar{z}_{abs} = \frac{1}{p}\sum z_{abs,i}$, where p is the number of focal groups. Note that in the case where there is 1 focal group, $\bar{z}_{abs}$ reduces to z_{abs}.

3.4 Statistical estimation

To compute confidence intervals and p-values, we used a simple bootstrap procedure (Efron and Tibshirani, 1994). Examinees were resampled within grade band, gender, and L1 groups, as these characteristics were central to the study design. Statistics were calculated from 1,000 bootstrapped samples. Confidence intervals were determined from .025 and .975 quantiles for each estimate. p-values of Δz and $\Delta\Delta z$ were determined by assuming a normal distribution and taking the minimum

of a two-sided quantile of the CDF evaluated at 0.

One of the additional advantages of using bootstrap estimators is that it was easier to include variation associated with examinees' non-speaking English proficiency in estimates. That is, prior to each resampling, examinees' non-speaking English proficiency estimates were redrawn from a normal distribution with EAP as the mean standard error of EAP as the standard deviation. Examinees were then stratified based on matching criterion, as described in section 3.3.2.

3.5 Controlling false discovery rate

Because of the large number of hypothesis tests (and corresponding p-values) examined in this research project, it was necessary to control for "findings" resulting from random chance. We controlled false discovery rate at the nominal level of .05 using the Benjamini-Hochberg (B-H) technique (Benjamini and Hochberg, 1995). We use the term *statistically significant* (or simply *significant*) when an estimated p-value is below the B-H adjusted p-value. In practical terms, we are placing an upper bound of .025 on "the probability of being erroneously confident about the direction of the population comparison" (Williams et al., 1999, p. 43).

3.6 Language modeling

3.6.1 BERT models

Six separate classification models were trained for each of the items analyzed in this study. Cross-entropy served as the loss function. The maximum number of input tokens depended on the item length; the cutoff was set at two standard deviations above the mean number of tokens for each item.

I selected BERT as the focus of analysis, after exploring several possible models with differing hyperparameters as a part of a previous pilot study (Kwako et al., 2022). For off-the-shelf (OOS) BERT models, I used the pre-trained, uncased BERT base model provided by Huggingface (Wolf

et al., 2020; Devlin et al., 2018). By OOS, I specifically refer to models that (1) come pre-trained, (2) remain unmodified with respect to model architecture, and (3) use conventional loss functions. As described in the next section, however, OOS models were fine-tuned. Modified BERT models (used in Study 3) were made with Pytorch (Paszke et al., 2019) in Python 9.3.12 (Python Software Foundation, 2022).

3.6.2 BERT training

Data were split 1:1 into testing and training sets.[4] Testing and training sets were split so as to maintain equal proportions of examinees by gender and L1.

Based on prior research, I selected learning rates of 1e-6 for the BERT layers and 2e-6 for classification heads, with a batch size of 16 (Kwako et al., 2022). To slow down overfitting, all but the last attention layer and classification head were frozen during training. Models were trained for 10 epochs, and the epoch with the lowest test loss was selected as the final scoring model for each item.

3.6.3 Performance metrics

To measure performance (or *reliability*) of BERT models, I compared BERT scores to human rater scores using three different metrics: accuracy, correlation, and quadratic weighted kappa. These are common metrics to assess inter-rater reliability. Approximately 10% of items were doubly scored by human raters, which provided a baseline for evaluating the reliability of BERT models.

[4]A larger percentage of data was set aside for testing (50% as opposed to the conventional 20%) specifically for Study 3, which required a robust calculation of z and z_{abs} for the test set.

CHAPTER 4

Study 1: Gender and L1 Biases in Automated Speech Transcription

Automated speech recognition (ASR) is an integral component of automated English speaking assessment. Yet, ASR is known to underperform for certain marginalized groups. Study 1 examines gender and native language (L1) biases in automated transcription in the context of English speaking assessment. To accomplish this, I analyze word error rate (WER) of transcripts generated by Amazon. Appendix 8.2 also presents WER of transcripts generated by Microsoft and Google. Findings are validated, in part, by repeating analyses on a public dataset known as L2-ARCTIC (Zhao et al., 2018a). Results show that, for ELPA21, WER is higher for individuals whose L1 is Vietnamese, and lower for Arabic. Study 1 finds no significant differences based on gender.

4.1 Study 1 overview

Automated transcription of speech consistently underperforms for underrepresented groups (Dichristofano et al., 2023; Hutiri and Ding, 2022). This may pose a problem for automated English speaking assessment, which relies on transcripts for scoring purposes (Zechner and Evanini, 2019; Johnston et al., 2019). Discrepancies in ASR may lead to biased scores for certain groups of examinees, which may limit students' success in secondary school (Johnson, 2020) and access to higher education (Johnson, 2019).

Study 1 examines discrepancies in the accuracy of automated transcription by comparing the WER of groups of examinees disaggregated by gender and L1. This study focuses on automated

transcripts generated by Amazon, since Amazon was used for generating transcripts in Studies 2–3. However, in Appendix 8.2, I explore whether similar trends exist for Microsoft and Google and, separately, whether results are consistent with a public dataset known as L2-ARCTIC (Appendix 8.3; Zhao et al., 2018a). By identifying disparities in transcription accuracy, this study allows for analyzing transcription accuracy as a potential source of differential item functioning (DIF) in Study 2.

4.1.1 Biases in automated transcription

There is a growing body of literature showing that automated transcription is less accurate for L2 English speakers and for those with regional L1 English accents. Prior studies show that there are more transcription errors for L2 English speech than L1 English speech (Dichristofano et al., 2023; Markl, 2022; Meyer et al., 2020). Evidence suggests that accuracy is especially lower when speakers' L1 is tonal (Chan et al., 2022). Among L1 English speakers, non-hegemonic English accents are less accurate, e.g., African American Vernacular English in the U.S. (Koenecke et al., 2020) and the Belfast accent in the U.K. (Markl, 2022).

Gender disparities vary depending on the context. Some studies report lower accuracy for women (Tatman, 2017; Hutiri and Ding, 2022), whereas others report lower accuracy for men (Dichristofano et al., 2023; Zuluaga-Gomez et al., 2023; Markl, 2022), and still others report no gender differences at all (Chan et al., 2022; Tatman and Kasten, 2017). These inconsistencies may be due to differences between datasets, automated transcription services, or statistical methods (e.g. inadequate p-value adjustments for multiple hypothesis tests).

4.1.2 Research questions and study design

This study revolves around a set of pairwise comparisons of average WER, disaggregated by speakers' gender, L1, and grade band. More specifically, Study 1 addresses the following research questions:

1. Are there differences in WER, on average, between examinees with different L1 backgrounds?
2. Are there differences in WER, on average, between male and female examinees?
3. Are there differences in WER, on average, between younger (grade band 2–3) and older (grade band 9–12) examinees?

I employ a study design that addresses the above research questions efficiently and rigorously by (1) standardizing automated transcripts, (2) making statistical adjustments for more accurate comparisons, and (3) employing p-value adjustments to minimize false discovery rate.

4.2 Study 1 methods

The ELPA21 sample, described in Section 3.1, was subsampled for Study 1, according to a process described in the following section. Methods of generating automated transcripts may be found in Section 3.2.1, and p-value adjustments may be found in Section 3.5. In terms of item selection, analyses in Study 1 focused exclusively on items 1 and 2.

4.2.1 ELPA21 subsample

From the sample of 15,000 students, a subset of 1,000 students was sampled for this study. I employed stratified random sampling, selecting 25 students randomly from each 3-factor combination of gender, L1, and grade band. The size of the sample was limited in part by the time required for manual transcription. Table 4.1 presents the descriptive statistics of the subsample used in Study 1. Overall, the subsample included 6.7 hours of speech.

In contrast to the main sample (Table 3.1), which presented descriptive statistics of examinees' non-speaking English proficiency, Table 4.1 presents examinees' overall English proficiency. For Study 1, it was not critical to separate speaking items from non-speaking items. As before, there were group differences in students' overall English language proficiency. Overall English proficiency

was used for covariate adjustment, described in Section 4.2.5.

Table 4.1: Descriptive statistics of examinees subsampled for Study 1, overall and disaggregated by gender and L1.

	Grade Band 2–3				Grade Band 9–12			
	n	**Avg. Seconds**	**Avg. Words**	**Avg. Proficiency**	**n**	**Avg. Seconds**	**Avg. Words**	**Avg. Proficiency**
All	500	24 (17)	32 (29)	0.03 (1.11)	500	24 (13)	37 (23)	0.04 (1.06)
Gender								
Male	250	22 (16)	29 (31)	-0.08 (1.11)	250	23 (11)	34 (18)	0.01 (1.02)
Female	250	25 (18)	35 (28)	0.14 (1.1)	250	26 (15)	39 (27)	0.07 (1.1)
L1								
Spanish	50	23 (16)	32 (29)	-0.05 (1.12)	50	22 (8)	32 (16)	0.08 (0.94)
Marshallese	50	23 (24)	27 (32)	-0.33 (0.86)	50	20 (9)	31 (16)	-0.15 (0.94)
Russian	50	24 (16)	33 (27)	0.09 (0.91)	50	24 (9)	38 (14)	0.47 (0.86)
Vietnamese	50	22 (14)	29 (23)	0.25 (0.99)	50	29 (16)	41 (24)	0.05 (0.96)
Arabic	50	22 (12)	32 (24)	0.26 (1.01)	50	23 (11)	36 (19)	-0.05 (0.94)
Mandarin	50	25 (15)	34 (21)	0.86 (0.84)	50	33 (25)	51 (44)	0.41 (1.04)
Hindi	50	27 (15)	39 (29)	0.55 (0.88)	50	27 (10)	46 (18)	0.73 (0.78)
Mayan	50	17 (11)	16 (17)	-1.24 (1.03)	50	20 (12)	22 (17)	-0.96 (1.21)
Persian	50	25 (15)	36 (27)	-0.32 (1.1)	50	23 (10)	31 (17)	-0.24 (1.11)
Swahili	50	32 (25)	43 (47)	0.25 (0.97)	50	24 (10)	38 (16)	0.06 (0.91)

4.2.2 Manual transcription

In transcribing L2 English speech, particularly for English language learners, there is a tension between transcribing grammatical mistakes and allowing for differences in accent, as others have noted (Yoon et al., 2019). I sought to transcribe speakers' intended meaning when the context was

clear. For example, I would faithfully transcribe "a apple" or "one apples," but would not transcribe "clothing" as "closing" if the context made the speakers' intention clear.

4.2.3 Word error rate

Performance (and biases) of automated transcription services were evaluated by calculating word error rate (WER), a metric commonly used for these purposes. Under the assumption that the manual transcript is true, WER is calculated as the ratio of total number of errors (in the automated transcript) to total number of words (in the manual transcript; Qian et al. (2019). There are three types of errors: insertions, deletions, or substitutions. Qian et al. (2019, p. 65) gives the following hypothetical example to illustrate how WER is computed. Insertions are presented in **bold**, deletions in ~~strikethrough~~ (strikethrough), and substitutions in *italics*:

Table 4.2: Example of calculating word error rate, with annotations.

Human transcription:	This is *a* sample response to ~~a~~ speaking test *question*.
ASR output:	This is *the* sample response to speaking test *quest* **shun**.

$$\frac{1\ insertion + 1\ deletion + 2\ substitutions}{10\ words\ total} = .4\ or\ 40\%\ WER$$

4.2.4 Text processing

Following text standardization (Section 3.2.2), speakers' responses were concatenated into a single line of text before calculating WER. A small number of outliers (n=21 for Amazon) had WERs exceeding 100%. In the most extreme cases, these were the result of students speaking in their native languages. Since I did not want these outliers to drive findings, WER was winsorized at 100% (Wilcox, 2012).

4.2.5 Covariate adjustment

WER was adjusted for students' English language proficiency to avoid confounding bias (Elwert and Winship, 2014). Without the adjustment, it is possible that some groups would have had a lower WER not because of transcription bias, but because the group was, on average, less fluent. To make these adjustments, I used a simple linear model: $\widehat{WER}_{adj} = \beta_0 + \sum_i G_i\beta_{1i} + \theta\beta_2$, where G_i are indicator variables of group membership, θ is English proficiency, and β_j are estimated coefficients. All statistical analyses were conducted in R 3.6.3 (R Core Team, 2020).

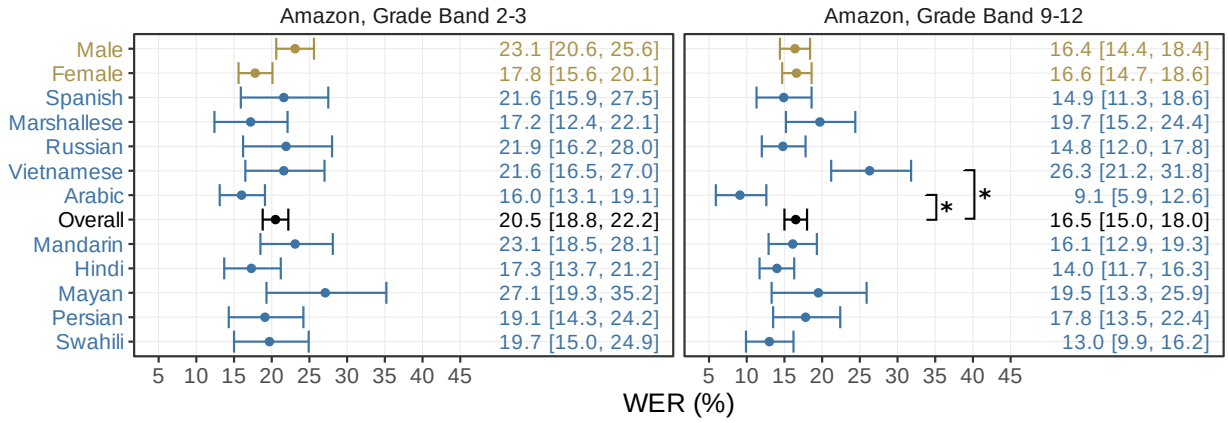
4.2.6 Statistical estimation

For gender, I conducted pairwise comparisons; for L1, I computed differences between each L1 mean and the overall (stratified) mean of all L1s. I used least square estimation with bootstrapping to calculate group WER estimates, group means difference, and p-values (Tibshirani and Efron, 1993). Bonferonni-Hochberg (B-H) adjusted p-values were used to control for false discovery rate (see Section 3.5).

4.3 Study 1 results

Figure 4.1 presents the estimated WER for examinees' speech based on Amazon's automated transcription service. For each grade band, WER is presented in aggregate (black) and disaggregated by gender (gold) and L1 (blue). Significance bars (with asterisks) indicate which pairwise comparisons were statistically significant, using B-H adjusted p-values. Results for Microsoft and Google automated transcription services (which were not used for Studies 2 or 3), are presented in Appendix 8.2.

Figure 4.1: Average WER estimates produced by Amazon's automated transcription service.



Note: Overall WER appear in black, and disaggregated WER appear in gold (gender) and blue (L1); whiskers indicate 95% confidence intervals; brackets with asterisks indicate statistically significant pairwise comparisons.

4.3.1 Gender biases

With respect to gender, results show no statistically significant differences between male and female WERs. Note that, while female speakers tended to have lower WERs, these differences were not statistically significant using B-H adjusted p-values.

4.3.2 L1 biases

Results show two statistically significant differences based on speakers' L1. Native Vietnamese speakers had a higher WER, on average, compared to other L1s. This means that automated transcription was less accurate for examinees' with Vietnamese L1 backgrounds. In contrast, native Arabic speakers had a lower WER compared to other L1s.

4.3.3 Age differences

There were also differences in WER based on speakers' grade band. Although it does not make sense to characterize these differences as biases in the same way as gender or L1, yet they may be relevant to further analyses of automated English speaking assessment. The WER of children (grades 2–3) was found to be significantly higher than the WER of young adults (grades 9–12). For Amazon, WERs were 20.5% [18.8%, 22.2%] and 16.5% [15.0%, 18.0%], respectively.

4.3.4 Differences across services and datasets

Although not the focus of Study 1, WER was computed across multiple automated transcription services (Appendix 8.2), and analyses were repeated on a public dataset known as L2-ARCTIC (Appendix 8.3). These supplemental analyses provide some evidence that the results of Study 1 are robust. Specifically, it was found that speakers with Vietnamese L1 backgrounds had a higher WER in L2-ARCTIC data, and across all three automated transcription services. Arabic speakers do not have a lower WER for L2-ARCTIC, however, and in ELPA21 the difference is not as extreme for Microsoft or Google as it is for Amazon.

4.4 Study 1 summary

The main finding of Study 1 is that there are indeed biases in automated transcription of examinees' speech for certain L1 backgrounds. Transcription was less accurate for young adults (grade band 9–12) whose L1 was Vietnamese. This finding is also consistent with a prior study of adult L2 English speech (Chan et al., 2022). It is interesting to note that this bias was not present for younger examinees (grade band 2–3). Findings from Study 1 were focused on Amazon's automated transcription service, since Amazon was selected as the main automated transcription service for Studies 2 and 3; however, similar biases were found for Microsoft and Google (Appendix 8.2).

Although our analyses did not detect gender bias, other studies have found that gender bias poses

a major problem in ASR (Hutiri and Ding, 2022). With a larger sample size, it would be possible to calculate more accurate WER estimates and to determine the size and direction of gender-based differences.

Given that I do not have access to the training data or models used in developing these services, I was unable to identify the source of the biases or attempt to mitigate them. Other researchers, however, point out that biases exists at every stage of the ASR development pipeline (Hutiri and Ding, 2022; Suresh and Guttag, 2021).

Although Study 1 highlights specific L1 biases in ASR systems, it alone is insufficient for providing evidence as to how these biases might impact examinees' test scores. In Study 2, however, I claim that these biases in automated transcription do not equate to lower (or higher) speaking scores.

CHAPTER 5

Study 2: Gender and L1 Biases in Human and Automated Scores in English Speaking Assessment

In L2 English speaking assessment, pretrained large language models (LLMs) such as BERT can score constructed response items as accurately as human raters. Less research has investigated whether LLMs perpetuate or exacerbate biases, which would pose problems for the fairness and validity of the test. Study 2 examines gender and native language (L1) biases in human and automated scores, using an off-the-shelf (OOS) BERT model. Analyses of bias focus on differential item functioning (DIF). Results show that, with respect to examinees' L1 background in grade band 9–12, there is a moderate amount of DIF, and this DIF is higher when scored by an OOS BERT model. In practical terms, the degree to which BERT exacerbates DIF is very small. Additionally, although there is more DIF for longer speaking items and for older examinees, BERT does not exacerbate these patterns of DIF.

5.1 Study 2 overview

Study 2 is designed to analyze gender and L1 biases in human and automated scores. For the automated scoring model, I use an OOS pretrained Bidirectional Encoding Representation using Transformers (BERT) because of its seminal status in the field (Devlin et al., 2018) and because it remains a focus of study in educational assessment (Wang et al., 2021). Bias is quantified using differential item functioning (DIF). I describe specific patterns of DIF in human scores, and determine whether or not BERT exacerbates DIF.

This study is designed to address four specific research questions:

1. Compared to human scores, do automated scores increase overall DIF with respect to gender or L1?
2. Does DIF increase with item length and, if so, is this exacerbated by automated scores?
3. Is DIF higher for older examinees and, if so, is this exacerbated by automated scores?
4. Which specific groups of examinees are (dis)advantaged most, and do automated scores exacerbate this (dis)advantage?

5.2 Study 2 methods

The data source and sample used in analyses were described previously (Section 3.1.2), as were automated transcription processes (Section 3.2.1), methods used to quantify DIF (Section 3.3), p-value adjustment (Section 3.5), and models and training procedures (Section 3.6). Study 2 used OOS BERT models provided by Huggingface (Wolf et al., 2020).

5.2.1 Performance metrics

In terms of performance (or reliability), OOS BERT models achieved near-parity with human raters for items 1–2 (for both grade bands), and BERT performed as well as (and, in grade band 9–12, outperformed) human raters for item 3. Table 5.1 reports the performance of each of the six BERT models in terms of accuracy, correlation, and quadratic weighted kappa, as compared to human-human agreement (which were derived from doubly-scored responses).

Table 5.1: Performance of off-the-shelf BERT scoring models for items 1–3, compared to human-human agreement, with respect to accuracy, correlation, and quadratic weighted kappa

	Grade Band 2–3						Grade Band 9–12					
	Acc.		r		QWK		Acc.		r		QWK	
Item	**H**	**B**	**H**	**B**	**H**	**B**	**H**	**B**	**H**	**B**	**H**	**B**
1	.911	.896	.793	.713	.792	.713	.929	.904	.920	.895	.920	.895
2	.756	.685	.898	.861	.898	.859	.728	.700	.911	.910	.911	.909
3	.614	.618	.834	.834	.834	.829	.694	.707	.841	.885	.609	.884

Note: "Acc." refers to accuracy, "r" to correlation, and "QWK" to quadratic weighted kappa. "H" refers to human-human comparisons (i.e. rater 2 compared to rater 1). The number of observations that were scored by two human raters ranged from 1,567–1641 for Grade Band 2–3, and from 1,254–1,293 for Grade Band 9–12. "B" refers to human-BERT comparisons (i.e. BERT compared to rater 1). The number of observations in the testing sets were 4,185 for Grade Band 2–3, and 3,306 for Grade Band 9–12.

5.3 Study 2 results

5.3.1 BERT increases DIF for L1

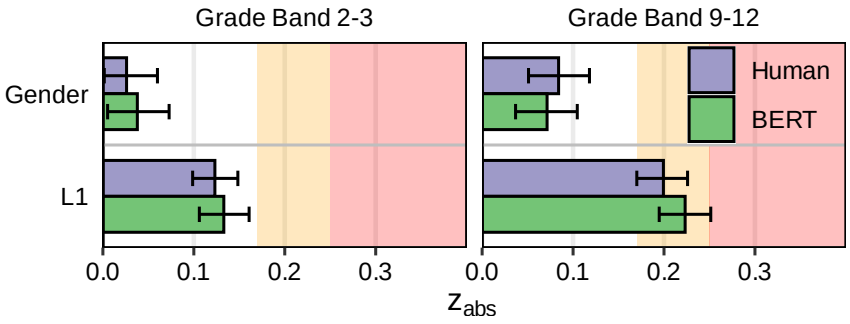
Overall, BERT-based automated scores increased DIF (to a very small degree) with respect to L1 in Grade Band 9–12. Although this difference was visible across all items in Grade Band 9–12, item 3 had the largest difference between human and automated scores.

5.3.1.1 Overall DIF of human scores

Results revealed a moderate amount of DIF in human ratings based on examinees' L1 in Grade Band 9–12. This result is visualized in Figure 5.1, which shows a gray bar (representing human scores) extending into the yellow ("moderate" DIF) region of the chart ($z_{abs} = .196, CI_{95\%} = [.170, .222], p = 5.4 \cdot 10^{-48}$). Additionally, there was non-zero DIF based on L1 in Grade Band 2–3,

and non-zero DIF based on gender in Grade Band 9–12; however, the effect sizes of these quantities were weak.

Figure 5.1: Estimates of DIF by gender and L1 over all 3 items for grade bands 2–3 and 9–12.



Note: Error bars indicate 95% confidence intervals. Yellow shaded regions correspond to moderate DIF, and red shaded regions correspond to strong DIF.

5.3.1.2 Human vs. BERT overall DIF

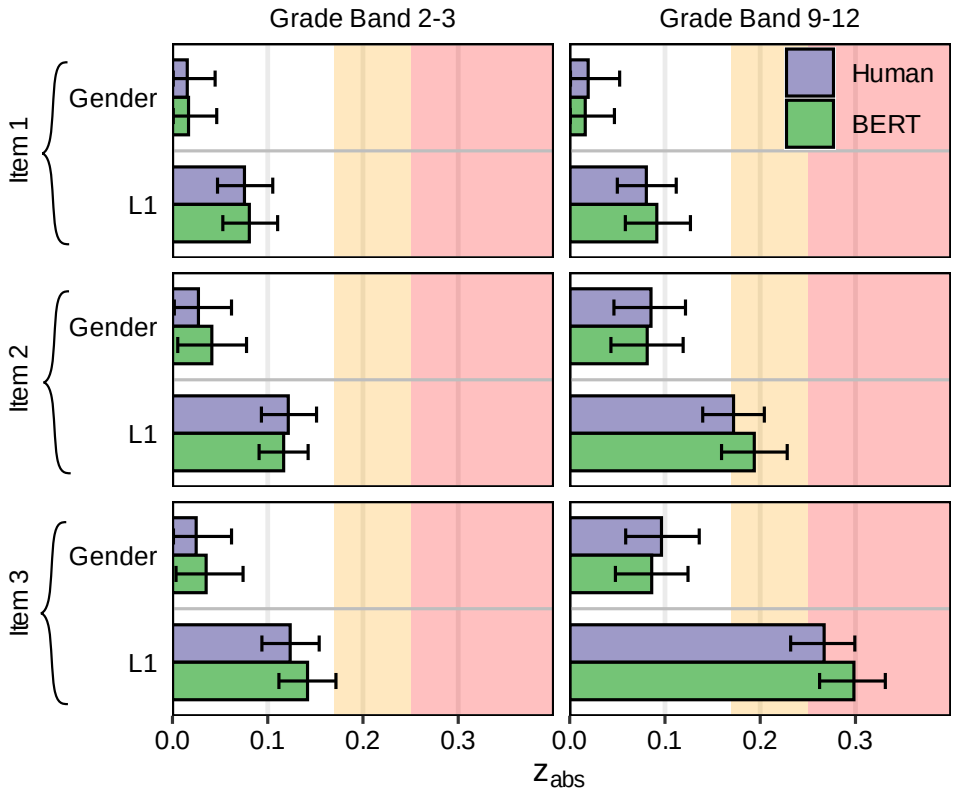
Overall DIF of automated scores was highly similar to human scores. As seen in Figure 5.1, green bars (representing BERT scores) are nearly commensurate with gray bars (representing human scores), with mostly overlapping 95% confidence intervals. Yet, there was significantly more DIF in BERT scores compared to human scores with respect to L1 in Grade Band 9–12 ($\Delta z_{abs} = .025, CI_{95\%} = [.011, .039], p = 3.3 \cdot 10^{-4}$). In practical terms, however, an effect size of 0.025 standard deviations is very small.

5.3.1.3 Human vs. BERT individual item DIF

In addition to overall DIF, I examined DIF for each individual item. Figure 5.2 presents DIF of human and automated scores, for gender and L1, across items 1–3, for each grade band. Human and automated scores are again quite consistent. For Grade Band 9–12, L1 DIF tends to be higher across all items; however, only item 3 reaches statistical significance ($\Delta z_{abs} = .032, CI_{95\%} =$

$[.010, .055], p = 3.3 \cdot 10^{-3}$). An effect size of 0.032 is very small.

Figure 5.2: Estimates of DIF by gender and L1 for each of the 3 speaking items in grade bands 2–3 and 9–12.



Note: Error bars indicate 95% confidence intervals. Yellow shaded regions correspond to moderate DIF, and red shaded regions correspond to strong DIF.

5.3.2 DIF increases with item length

Longer speaking items tended to exhibit more DIF than shorter speaking items. Automated scores, however, did not exacerbate this trend. By design, item 3 was longer than item 2, which in turn was longer than item 1. Figure 5.2 shows that, in general, item 3 had more DIF than item 2, which in turn had more DIF than item 1. Table 5.2 presents the specific values of $\Delta z_{abs,ij}$, based on human rater scores, for all three item comparisons. For example, in grade band 9–12, the difference in DIF

between items 1 and 2, based on human rater scores (i.e., the gray bars in Figure 5.2), with respect to L1, was $\Delta z_{abs,21} = .087$. That is, item 2 had .087 more standard deviations of DIF compared to item 1. Using B-H adjusted p-values, this is a statistically significant difference, as indicated by asterisks in Table 5.2.

Table 5.2: Differences in DIF between longer and shorter items, within each grade band, based on human ratings.

	Grade Band 2–3			**Grade Band 9–12**		
Factor	**2 - 1**	**3 - 1**	**3 - 2**	**2 - 1**	**3 - 1**	**3 - 2**
Gender	.012	.010	-.002	.065 *	.078 *	.013
	[-.030, .051]	[-.029, .049]	[-.042, .039]	[.021, .110]	[.031, .116]	[-.032, .055]
L1	.046 *	.053 *	.006	.087 *	.184 *	.097 *
	[.009, .085]	[.010, .093]	[-.035, .046]	[.043, .130]	[.139, .226]	[.056, .138]

Note: "*" indicates that an estimate is statistically significant using B-H adjusted p-values. 95% confidence intervals are presented in square brackets.

Although longer items tended to have more DIF, this general trend was not uniformly consistent across factors and grand bands. Specifically, the trend was less consistent for gender: There were no statistically significant differences in grade band 2–3; and in grade band 9–12, item 3 did not have more DIF than item 2 at a statistically significant level. Additionally, for grade band 2–3, item 3 did not have significantly more DIF than item 2.

In order to determine if item-item differences were exacerbated by automated scoring, we computed second-order differences, $\Delta\Delta z_{abs}$. None of these values, however, were statistically significant. We conclude that the pattern of longer items producing more DIF is consistent for both human and automated raters.

5.3.3 DIF is higher for older examinees

In general, there was more DIF for older examinees (in grade band 9–12) compared to younger examinees (in grade band 2–3), in terms of both gender and L1. Automated scores, however, did not exacerbate this trend. This trend can be seen clearly in Figure 5.1. Based on bootstrapped estimates for overall DIF with respect to gender, $\Delta z_{abs} = .059$ ($CI_{95\%} = [.011, .100], p = 4.9 \cdot 10^{-3}$); with respect to L1, $\Delta z_{abs} = 0.082$ ($CI_{95\%} = [0.047, 0.120], p = 3.8 \cdot 10^{-6}$).

When we examine individual items, this trend is present for items that are medium-length or longer (items 2 and 3) but not for short items (item 1). Visually, this can be seen in Figure 5.2. First-order differences between grade bands, Δz_{abs}, based on human ratings, are presented in Table 5.3. For example, in item 1, the difference between DIF observed in grade band 2–3 versus grade band 9–12 is $\Delta z_{abs} = .005$, with respect to L1, which is not a statistically significant difference. In items 2 and 3, however, the differences between grade band 2–3 and 9–12 are much larger ($\Delta z_{abs} = .050$ and $\Delta z_{abs} = .144$, respectively).

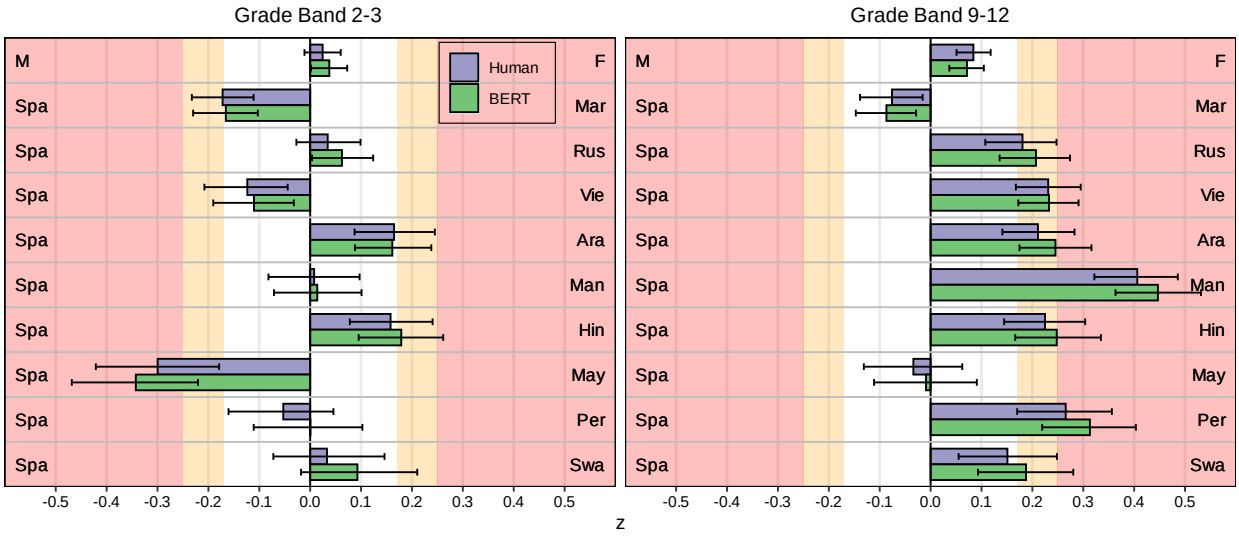
Table 5.3: Differences in DIF between grade bands for each of the 3 speaking items, based on human ratings.

Factor	**Item 1**	**Item 2**	**Item 3**
Gender	0.004 [-0.035, 0.043]	0.058 * [0.006, 0.104]	0.071 * [0.02, 0.119]
L1	0.005 [-0.037, 0.048]	0.050 * [0.006, 0.095]	0.144 * [0.098, 0.189]

Note: "*" indicates that an estimate is statistically significant using B-H adjusted p-values. 95% confidence intervals are provided in square brackets.

In order to determine if differences between grand bands were exacerbated by automated scoring, we computed second-order differences, $\Delta\Delta z_{abs}$. None of these values, however, were statistically significant. We conclude that the trend of greater DIF in older examinees was consistent for both human and automated raters.

Figure 5.3: Estimates of direction and magnitude of overall DIF.



Note: Error bars indicate 95% confidence intervals. Yellow shaded regions correspond to moderate DIF, and red shaded regions correspond to strong DIF. Reference groups are listed on the left of each chart (M = Male, Spa = Spanish); focal groups are listed on the right (L1 groups are abbreviated by the first three letters). DIF in the positive direction indicates that the focal group is favored.

5.3.4 Severity of DIF depends on L1 and grade band

The magnitude and quantity of DIF varied by L1 background, and patterns were generally not consistent across grade bands. Figure 5.3 depicts the magnitude and direction of DIF for gender and all L1 groups. For grade band 2–3, native speakers of Marshallese and Mayan languages showed evidence of moderate–strong DIF for human and BERT scores. DIF was negative for both L1 groups, indicating that these examinees fared worse on speaking items than their (equally-proficient) Spanish-speaking counterparts.

In grade band 9–12, examinees of nearly all L1 backgrounds fared better than native Spanish speakers. In this case, speaking items tended to disadvantage members of the reference group (i.e. examinees with Spanish L1 backgrounds).

As with preceding analyses, DIF based on BERT scores aligned closely with DIF based on human scores. Although results showed that BERT exacerbated DIF in L1 as a whole (Section 5.3.1), analyses of individual L1 groups did not reveal any statistically significant differences between human and BERT scores. We also did not find any statistically significant differences between human and BERT scores when examining DIF at the individual item level (Appendix 8.4).

5.4 Study 2 summary

5.4.1 Main findings

Analysis of DIF revealed specific patterns of biases in human and automated scores of English speaking assessment. With respect to human scores, there was more DIF for older examinees and for longer items. Based on commonly accepted standards regarding effect size, there was a moderate amount of overall DIF in grade band 9–12 based on examinees' L1 backgrounds. Automated scores generated by OOS BERT closely matched human scores, yet BERT was found to exacerbate overall DIF for grade band 9–12 based on examinees' L1. The degree to which BERT exacerbated bias, however, was very small.

5.4.2 Sources of DIF

Although our findings do not confirm any causes of DIF, they do allow us to rule out several possibilities.

5.4.2.1 Transcription (in)accuracy

Study 1 showed that there were discrepancies in word error rate (WER) of automated transcription based on L1. Specifically, automated transcription struggled with speakers of Vietnamese L1 backgrounds. Yet given the close correspondence between human and automated scores—for all examinees, not just Vietnamese examinees—it appears unlikely that transcription inaccuracies

engendered lower or higher scores.

5.4.2.2 Implicit bias

Our automated scoring system was based exclusively on transcripts of examinees' speech. No phonic information was used in the automated scoring process. It is notable, then, that there was no mitigation of DIF in automated scores using a text-based BERT model. In other words, removal of acoustic input did not reduce bias. From this, we can conclude that examinees with *identical* (transcribed) responses could not have received higher or lower scores, on average, based on gender or L1.

Although text-based automated scores did not mitigate bias, this does not necessarily imply that human raters were unaffected by implicit bias. It is possible, for instance, that examinees with different accents also had different (transcribed) responses, which yet affected human judgment.

5.4.3 Accuracy and DIF

As the performance of automated scoring improves to match (or exceed) that of human raters, one might have expected the magnitude of DIF to also match (or potentially reduce) that of human raters. Contrary to expectation, for longer speaking items, automated scores exceeded the performance of human raters yet increased DIF. More research is needed to determine the relationship between performance and DIF.

CHAPTER 6

Study 3: Mitigating Gender and L1 Biases in Automated English Speaking Assessment

As reported in Study 2, there was a moderate degree of differential item functioning (DIF) for grade band 9–12 based on examinees' native language (L1). Drawing from a new body of research in machine learning known as *debiasing*, Study 3 explores the possibility of reducing DIF. I compare two different debiasing approaches: The adversarial approach, which removes specific aspects of gender or L1 from examinees' responses; and the shrinkage approach, which averages automated scores with examinees' expected scores (ES). Results show that the adversarial approach fails to reduce DIF, producing identical scores to off-the-shelf (OOS) BERT models. Although the shrinkage approach uniformly reduces DIF, it does so at the expense of item information.

6.1 Study 3 overview

As a growing field in machine learning, debiasing is focused on making predictions more equitable with respect to protected attributes (Elazar and Goldberg, 2018). In many contexts, the prediction task (in this case, scoring English speech) may be intertwined with protected attributes (i.e., gender or L1), and it is the goal of debiasing to untangle this association and prevent protected information from affecting task predictions. Sometimes the association between the task and the protected attribute is referred to as *leakage*, in the sense that information in the prediction function leaks information about the protected class. There are numerous techniques that have been proposed and applied toward specific tasks, but because it is a relatively new field, there is little consensus as to

which techniques are most effective (Sun et al., 2019).

Debiasing techniques have not been applied to English speaking assessment, yet they may offer a solution for reducing biases (including the patterns of DIF enumerated in Study 2). It may be possible to reduce biases introduced by LLMs, and perhaps even to reduce bias below what is seen in human rater scores. Sun et al. (2019) review a number of debiasing techniques, organized by methodological approach. In Study 3, I apply two of these techniques toward mitigating gender and L1 biases in automated English speaking assessment, and discuss the relative merits and weaknesses of each.

6.1.1 The adversarial approach

The adversarial approach is characterized by predicting the protected class, and then preventing the model from using this information in task predictions. In one application of this technique, Wang et al. (2019) first trained a model to predict gender and, subsequently, during task training, reversed the optimization function so that the model unlearned its ability to predict gender.

This approach is appealing because it attempts to remove biases specific to the protected attributes. This approach is surgical in its attempt at removing only the parts of the prediction function that are related to the protected attributes. Aspects of the protected attribute that are related to the task function are actively unlearned. In image recognition, this is analogous to removing facial details or hair styles that might be related to gender. With respect to English speaking assessment, this approach would (in theory) unlearn aspects of the scoring function that are related to gender or L1.

6.1.2 The constrained prediction approach

Another approach is constrained prediction, which involves altering predictions after training is completed. One way of doing this is to alter the probability distributions so that each protected attribute has equal odds of receiving the same prediction (e.g. Yatskar et al., 2016). The method

that I employ in Study 3 is to use a shrinkage estimator (Fourdrinier et al., 2018), which is tailored specifically to shrink estimates towards zero DIF. The specific condition that shrinks predicted scores towards zero DIF is when examinees' scores are exactly equal to their expected scores, based on their responses to non-speaking items.[1]

6.1.3 Research questions and study design

Study 3 explores two approaches of debiasing gender and L1 in English speaking assessment. The specific research questions I address are:

1. Does the adversarial approach reduce overall DIF, based on gender or L1?
2. What are the causes of the adversarial model's success or failure?
3. Does a constrained prediction approach reduced overall DIF, based on gender or L1?
4. What are the causes of the constrained prediction model's success or failure?

In Section 6.4, I also discuss the advantages and limitations afforded by each approach, and I suggest possible developments and further research that could redress those limitations.

6.2 Study 3 methods

The data source and sample used in analyses were described previously (Section 3.1.2), as were automated transcription processes (Section 3.2.1), methods used to quantify DIF (Section 3.3), p-value adjustment (Section 3.5), and models and training procedures (Section 3.6). Consistent with Study 2, Study 3 uses OOS BERT models provided by Huggingface (Wolf et al., 2020), as well as modified BERT models; modifications were made with Pytorch (Paszke et al., 2019), based on original models provided by Huggingface.

[1] Expected scores are described in greater detail in Section 6.2.3.

6.2.1 Adversarial models

Adversarial models included three classification heads, corresponding to score, gender, and L1 predictions. Cross-entropy loss was calculated for each classification head. Before taking each optimization step, losses were combined in such a way as to reduce bias. Following the example of Wang et al. (2019), the final objective function is

$$L_p = \sum_{(X_i,h_i,Y_i)} [L(p(X_i), Y_i) - \lambda_g L_{c_g}(c_g(h_i), g_i) - \lambda_l L_{c_l}(c_l(h_i), l_i)]$$

where $p(X_i)$ are the predicted scores, Y_i are the actual scores, and $c_g(h_i)$ and $c_l(h_i)$ are *critics*—that is, functions that predict protected classes (i.e. gender and L1, respective) from BERT's contextual token embedding (i.e., the pooled output layer), h. λ_g and λ_l represent hyperparameters to weight the importance of the gender and L1 critic, respectively.

6.2.2 Shrinkage models

The DIF shrinkage approach computes examinees scores from a weighted average of (1) examinees' predicted scores, using regression BERT models, and (2) examinees' expected scores (ESs), based on their responses to non-speaking items. ESs are described in greater detail in Section 6.2.3.

Regression BERT models are OOS BERT models. Instead of predicting score categories using cross-entropy loss, however, these models are trained to predict a single, continuous score using mean squared error loss. Thus, scores predicted by regression BERT were non-integer, float values.

Regression scores (RSs) were combined with examinees' ESs, at varying weights (or proportions), to produce regularized scores. Study 3 includes five different weighting conditions: $1RS + 0ES$, $.75RS + .25ES$, $.5RS + .5ES$, $.25RS + .75ES$, and $0RS + 1ES$. Combined scores were rounded to the nearest whole integer, in order to remain consistent with scores produced by human raters.

6.2.3 Expected score estimates

Based on examinees' responses to non-speaking items, an expected score (ES) was computed for responses to each item in examinees' respective grade bands. That is, each examinee had three ESs, corresponding to items 1–3. ES is a function of non-speaking proficiency, θ, as well as IRT item parameters, which include the discrimination parameter, a, and the difficulty parameters, $\mathbf{b}$. More specifically, for examinee i, responding to item j, across all possible scores, u, $ES_{ij} = E[U = u|\theta_i, a_j, \mathbf{b}_j]$. For computation of θ, see Section 3.3. With respect to IRT item parameters, each item was modeled using a graded response model (Samejima, 1997) and computed using flexMIRT (Cai, 2012), based on the responses of all examinees in the sample.

6.3 Study 3 results

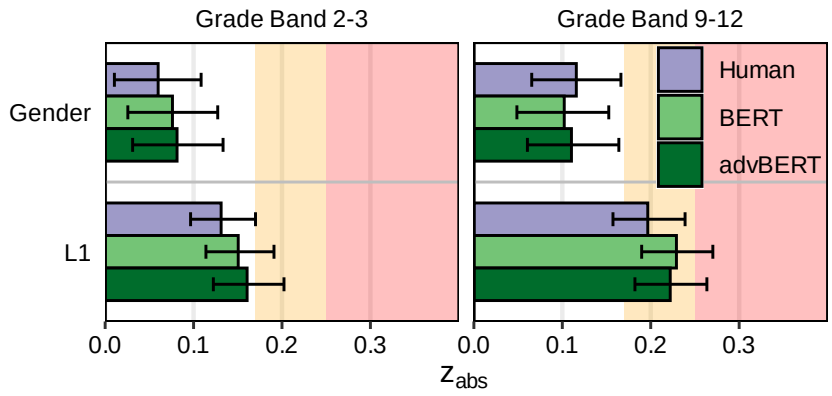
6.3.1 The adversarial approach

The adversarial model failed to remove overall DIF, based on gender or L1, in either grade band 2–3 or 9–12. Figure 6.1 compares overall DIF based on human scores (gray) to OOS BERT (green) and the adversarial BERT model (dark green). Although there were minor differences between OOS BERT and adversarial BERT, Δz_{abs}, none of these differences were statistically significant, and may be attributable to noise (e.g. from slightly different starting values).

There were also no statistically significant differences between OOS BERT and adversarial BERT when examining the direction and magnitude of overall DIF, based on gender or any individual L1 group (Figure 6.2). Again, although there were minor discrepancies, based on Δz, none of the discrepancies were statistically significant.

One of the reasons why the adversarial BERT model may not reduce DIF is that there is little to no leakage. Indeed, BERT struggled to predict the protected attributes in examinees' responses, let alone the protected attribute present in the score prediction function. Figure 6.3 presents the accuracy with which BERT predicted gender based on examinees' responses for each of the 3

Figure 6.1: Comparisons of overall DIF across human, off-the-shelf BERT, and adversarial BERT, by gender and L1 over all 3 items for grade bands 2–3 and 9–12.

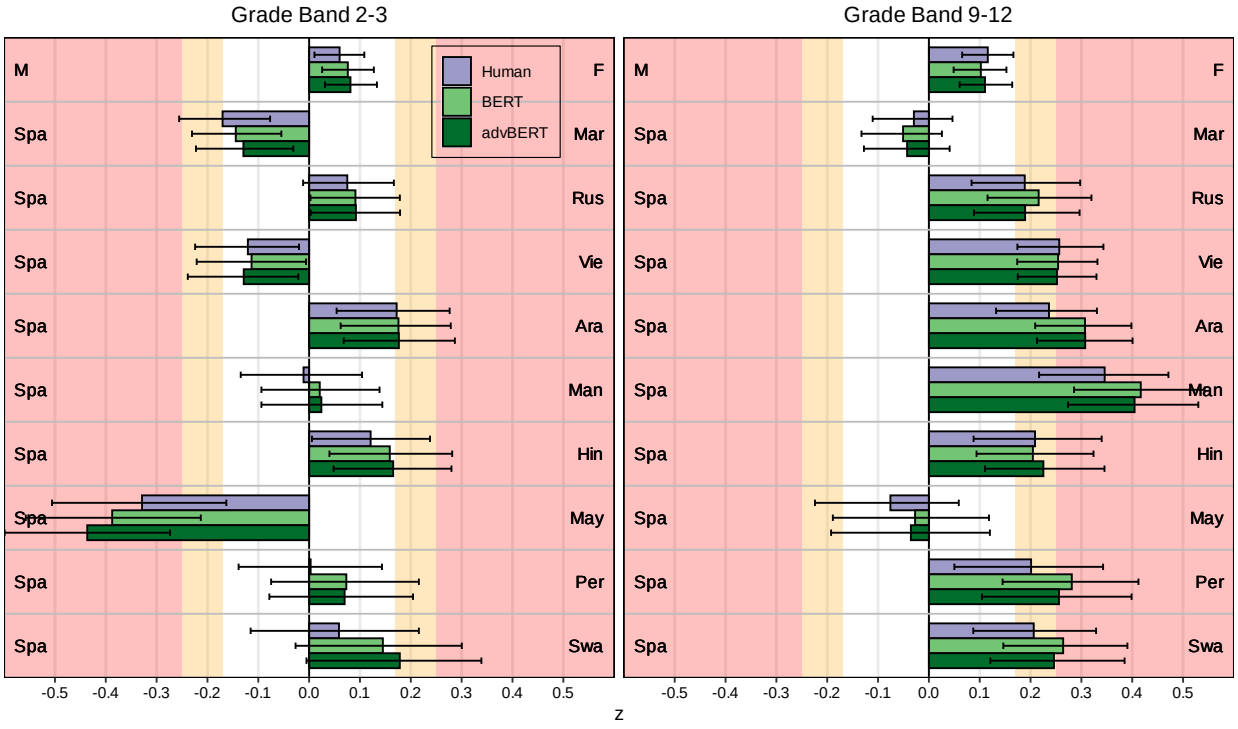


Note: BERT represents the off-the-shelf BERT model, and advBERT represents the adversarial BERT model. Error bars indicate 95% confidence intervals. Yellow shaded regions correspond to moderate DIF, and red shaded regions correspond to strong DIF. Estimates of magnitude of L1 DIF are the average z_abs for all 9 focal groups.

speaking items in grade bands 2–3 and 9–12. In both grade bands, there are many incorrect (i.e. off-diagonal) predictions. Indeed, the predictions are nearly random, suggesting that BERT identified little gender-related signal in examinees' responses. In grade band 9–12, the majority of responses were predicted as male.

With respect to predicting L1 group based on examinees' responses, results were even more extreme. For both grade bands, and across all three items, BERT predicted Spanish for nearly all examinees (Figure 6.4). Although the majority of examinees had Spanish L1 backgrounds (50.2% of examinees in grade band 2–3, and 52.6% in grade band 9–12; Table 3.1), one might have expected BERT to have predicted minority L1 groups at higher rates. Despite this limitation, BERT's prediction accuracy is higher than that of a Naïve Bayes classifier (Appendix 8.5). Note that analyses of confusion matrices do not control for examinees' scores, which precludes the possibility of identifying more subtle differences in examinees' responses.

Figure 6.2: Comparisons of direction and magnitude of overall DIF across human, off-the-shelf BERT, and adversarial BERT, by gender and each L1 group for grade bands 2–3 and 9–12.

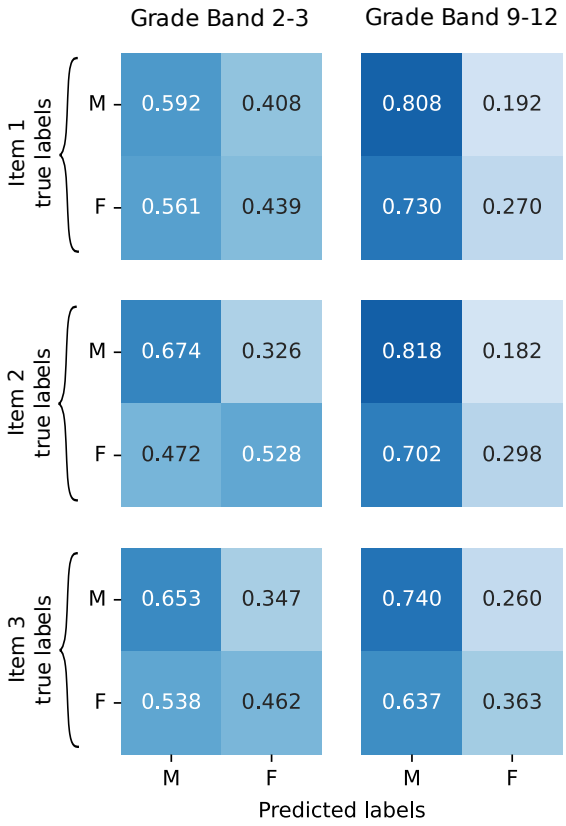


Note: BERT represents the off-the-shelf BERT model, and advBERT represents the adversarial BERT model. Error bars indicate 95% confidence intervals. Yellow shaded regions correspond to moderate DIF, and red shaded regions correspond to strong DIF. Reference groups are listed on the left of each chart (M = Male, Spa = Spanish); focal groups are listed on the right (L1 groups are abbreviated by the first three letters). DIF in the positive direction advantages the focal group.

6.3.2 The shrinkage approach

Shrinking BERT scores towards examinees' expected scores (ESs) reduces DIF and, at the extreme, eliminates DIF entirely. The degree of reduction of DIF depends on the percentage of the weighted average apportioned to ES. If zero weight is apportioned to ES, then results are identical to patterns of DIF seen from OOS BERT models. As a greater percentage of the score is apportioned to ES (i.e.

Figure 6.3: Confusion matrix of BERT predictions of gender for each of the 3 speaking items in grade bands 2–3 and 9–12.

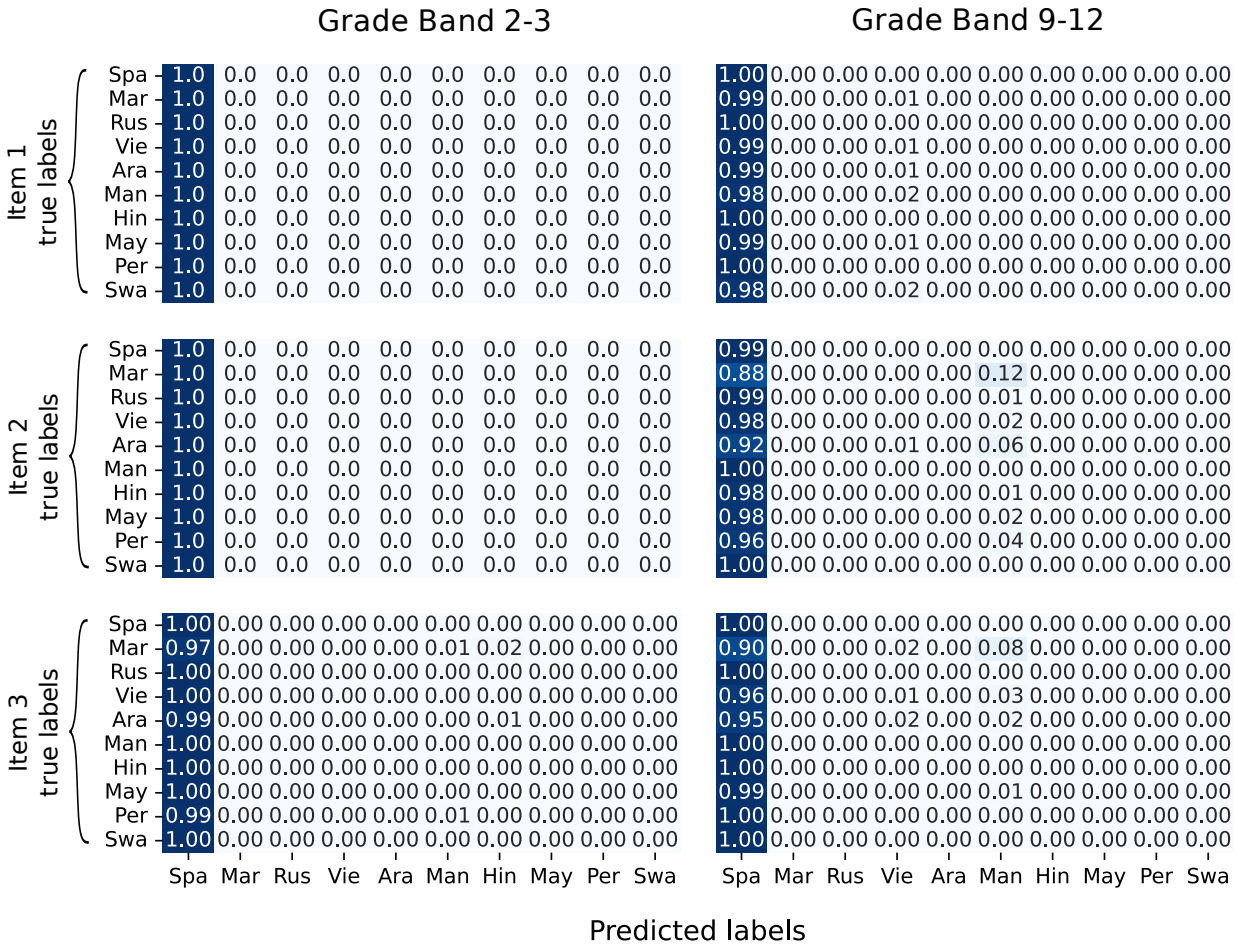


Note: accuracy is conditioned on the true number of examinees within each gender.

25%, 50%, and 75%), there is a corresponding decrease in DIF. In the extreme case, when scores are exclusively determined by ES, DIF is eliminated. Recall that ES is based on examinees' responses to non-speaking items (Section 6.2.3); thus, although DIF is reduced or eliminated, the information provided by the item is correspondingly reduced or eliminated. In the extreme case, "speaking" scores are based on non-speaking items, and examinees' responses are not taken into consideration at all. Although impractical, the extreme example is helpful in illustrating the downside of reducing DIF with the shrinkage model.

Figure 6.5 depicts the inverse relationship between DIF and the weight apportioned to ES across

Figure 6.4: Confusion matrix of BERT predictions of L1 group for each of the 3 speaking items in grade bands 2–3 and 9–12.

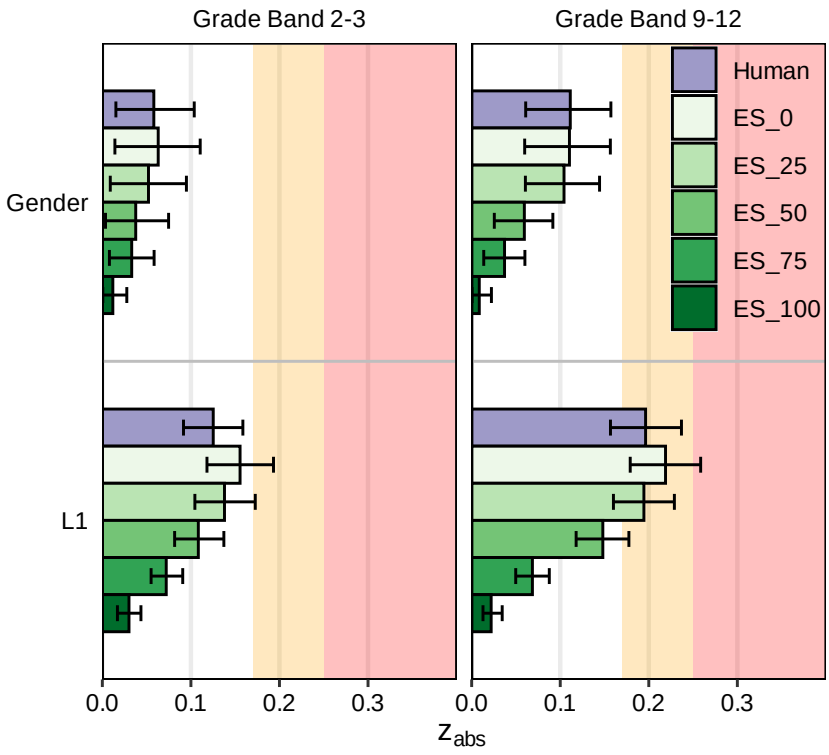


Note: accuracy is conditioned on the true number of examinees within each L1 group.

five conditions. Darker shades of green correspond to higher weights apportioned to ES. It is clear that DIF decreases monotonically as the percentage (of examinees' speaking scores) determined by ES increases.

DIF is also reduced (or eliminated) with the shrinkage model, regardless of direction or magnitude of DIF (Figure 6.6). As above, Figure 6.6 presents DIF across five shrinkage models,

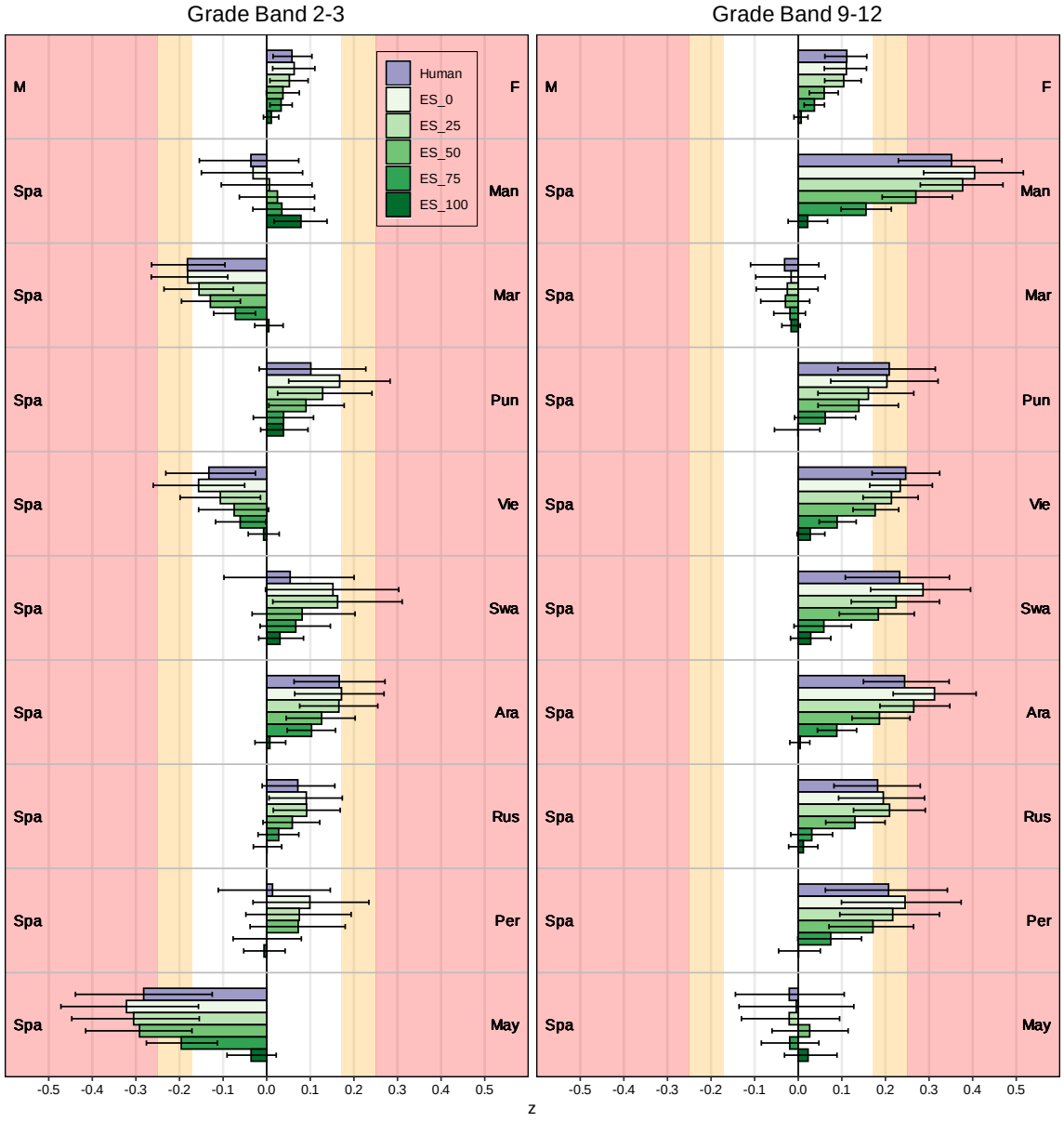
Figure 6.5: Comparisons of overall DIF across human and five shrinkage BERT models, by gender and L1 over all 3 items for grade bands 2–3 and 9–12.



Note: ES stands for "expected score," and represents the percentage of weight given to expected score (in conjunction with regression BERT). Error bars indicate 95% confidence intervals. Yellow shaded regions correspond to moderate DIF, and red shaded regions correspond to strong DIF. Estimates of magnitude of L1 DIF are the average z_abs for all 9 focal groups.

corresponding to weight apportioned to ES, by gender and L1 group for grade bands 2–3 and 9–12. Although this study examines gender and L1, the shrinkage approach would eliminate DIF across all background factors.

Figure 6.6: Comparisons of direction and magnitude of overall DIF across human and five shrinkage BERT models, by gender and each L1 group for grade bands 2–3 and 9–12.



Note: ES stands for "expected score," and represents the percentage of weight given to expected score (in conjunction with regression BERT). Error bars indicate 95% confidence intervals. Yellow shaded regions correspond to moderate DIF, and red shaded regions correspond to strong DIF. Reference groups are listed on the left of each chart (M = Male, Spa = Spanish); focal groups are listed on the right (L1 groups are abbreviated by the first three letters). DIF in the positive direction advantages the focal group.

6.4 Study 3 summary

6.4.1 Limitations and further developments for the adversarial approach

Speaking scores generated using the adversarial approach did not differ significantly from off-the-shelf BERT scores. One of the reasons the adversarial model failed to reduce DIF may be because BERT failed to detect gender or L1 differences in examinees' responses. If this is indeed the case, then there would be no leakage to mitigate, and we would not expect the adversarial approach to mitigate DIF in the first place.

Although results showed that the adversarial method was unable to reduce DIF, this could be less of a limitation of the method, and more of an indication that the method requires additional engineering. Specifically, although it was found that BERT was unable to predict gender or L1, the relationship between examinees' protected attributes and their response was not conditioned on non-speaking proficiency, which is critical in analyses of DIF. It is also possible that this lack of predictive power is due (in part) to the fact that the sample is imbalanced: There may be subtle differences in responses based on L1, but these are eclipsed by the fact that the vast majority of examinees have Spanish L1 backgrounds. More fine-grained analyses are needed to parse the relationship between responses and L1 background. Through data augmentation, it would be possible to rectify the imbalance between L1 group sizes. If these additional analyses do, in fact, reveal a relationship between examinees' responses and L1, then it might reopen the possibility of mitigating DIF through adversarial models.

Although there are other adversarial models, all would suffer from the same limitation of not being able to remove bias as long as there is no leakage. Zhang et al. (2018), for instance, suggests projecting the task-specific gradient onto the orthogonal compliment of the protect attribute gradients during training. So long as the subspace is sufficiently restricted, this could prevent the model from learning to predict gender or L1 from responses, but only if the model is able to identify a relationship between the protected attribute and the response to begin with.

6.4.2 Limitations and further developments for the shrinkage approach

Although the shrinkage approach reduces DIF (irrespective of magnitude, direction, or even background characteristic), it also reduces the information contained in speaking items. In the extreme case, examinees' responses are determined exclusively by their non-speaking English proficiency. Thus, if applied uniformly to all speaking items, the shrinkage approach precludes the possibility of measuring speaking proficiency, and confounds the very purpose of the assessment. A more judicious application of the shrinkage approach (e.g. applying it only towards those items exhibiting moderate–strong DIF), could be beneficial. Without knowing the source of DIF, however, it is impossible to know whether this more judicial application is still removing key components of speaking proficiency. This danger is plausible, especially given that longer speaking items exhibit more DIF.

One possible extension of the shrinkage approach would be to specify further constraints so as to determine algorithmically the optimal weight to apportion to examinees' expected scores. Study 3 analyzed five conditions of various weights; with additional specifications, the weight assigned would not be arbitrary, but determined a priori by some modified loss function. This addition might help clarify what is important, psychometrically. With more weight given to expected score, we reduce DIF, but we also reduce item information; it would make sense, then, to limit DIF while also maximizing some measure of item information—ideally, information that is orthogonal to non-speaking proficiency.

CHAPTER 7

Discussion

7.1 Summary of findings

7.1.1 Studies 1–3

Study 1 analyzed the accuracy of a large-scale automated transcription service by comparing average word error rate (WER) of a subsample of examinees, disaggregated by gender and L1. Results showed no difference in accuracy based on gender, for either grade band. Yet for grade band 9–12, examinees whose native language (L1) was Vietnamese tended to have less accurate transcripts, on average, while those with Arabic L1 backgrounds had more accurate transcripts.

Study 2 examined patterns of differential item functioning (DIF) with respect to scores generated by human raters and by a off-the-shelf (OOS), commonly used large language model known as BERT. Results revealed that, based on human rater scores, there was a moderate amount of overall DIF for grade band 9–12 based on examinees' L1. Across both grade bands, there was more DIF for longer speaking items, and more DIF for older examinees (grade band 9–12) compared to younger examinees (grade band 2–3). BERT exacerbated DIF by a small amount, with respect to overall DIF for grade band 9–12 based on examinees' L1. Yet BERT had identical patterns of DIF with respect to length of speaking items and age of examinees. Overall, BERT scores corresponded closely to human rater scores.

Study 3 focused on two approaches aimed at mitigating DIF, the adversarial approach and the shrinkage approach. The adversarial approach produced scores that were not significantly different from the OOS BERT model used in Study 2. One possible reason for the failure of the adversarial

model to mitigate DIF may be due to the fact that BERT struggled to identify information related to examinees' gender or L1 in examinees' responses; additional research is needed to verify if this association could be revealed by other models. In contrast to the adversarial approach, the shrinkage approach did reduce DIF for gender and L1. The drawback of the shrinkage approach is that, as DIF is reduced, information regarding examinees' speaking proficiency is also reduced. Without untangling the relationship between DIF and speaking proficiency, the shrinkage approach may undermine the very purpose of the assessment.

7.1.2 Overarching research goals

This research project was designed in part around exploring two specific construct-irrelevant drivers of DIF, automated transcription bias and implicit bias. Evidence from Studies 1–2 rule out the possibility that automated transcription is the cause of DIF. And Studies 2–3 indicate that implicit bias is likely not the main issue. Yet there is room for further exploration of implicit bias, particularly if additional data becomes available (see Section 7.2).

Another overarching goal of this research project was to explore debiasing techniques for the removal of DIF. It was especially of interest to explore adversarial methods capable of removing construct-irrelevant aspects of examinees' responses (i.e. surgically removing aspects specific to gender or L1). Such a technique could be employed regardless of the main drivers of DIF. Unfortunately, this technique did not yield fruitful results, in part because BERT did not identify any large differences in examinees' responses based on gender or L1.

7.2 Sources of DIF

Results from Studies 1–3 allow us to rule out certain sources of DIF, and suggest possibilities for examining other sources more deeply.

7.2.1 Automated transcription

Study 1 revealed that there were discrepancies in transcription accuracy that varied by L1. Yet Study 2 showed that there was very little room for these inaccuracies to affect scores. Rather, human rater scores and automated scores generated by BERT, when disaggregated by L1, were nearly identical; DIF was also nearly identical. Furthermore, we do not see DIF and accuracy trend in the same direction: examinees with Vietnamese L1 backgrounds have, on average, lower transcription accuracy, yet DIF seems to favor them over other L1 groups. Because of the tight correspondence between human and automated scores and human and automated DIF, automated transcription can be ruled out as a source of DIF.

7.2.2 Human rater bias

Research on implicit bias suggests that human raters might judge certain accents more harshly than others. That is, examinees (of different L1 backgrounds) might give the same response, yet receive different scores. For instance, Spanish speakers and Mandarin speakers who both said "she put on her shoes" may receive different scores. Study 2, however, ruled out this possibility. If examinees had received different scores, then BERT (which takes text-only input) would have reduced DIF, which was not the case. Note that this finding is consistent with scoring rubrics, which focus on the content of examinees' responses, as opposed to aspects of fluency, such as pronunciation.

Based on results from Study 2, however, the possibility still remains that examinees (of different L1 backgrounds) gave slightly different responses—that deserved the same scores—yet received different scores. For instance, consider the possibility that examinees of Spanish L1 backgrounds might have been more likely to say, "She put on her shoes," whereas examinees of Mandarin L1 backgrounds might have been more likely to say, "She was putting on her shoes." Although these responses deserve the same score, they may trigger implicit bias in human raters and be given different scores. It is possible that BERT would propagate these discrepancies.

Results from Study 3 demonstrated that BERT was unable to detect a relationship between L1

and examinees' responses.[1] That is, BERT was not able to distinguish examinees whose L1 was Spanish versus Mandarin. This finding suggests that implicit bias—and even the more nuanced type of implicit bias described above—is unlikely. An important caveat to this claim is that, in the analysis of the relationship between response and L1, Study 3 did not control for examinees' language proficiency (or, even more coarsely, human rater scores), which is critical in analyses of DIF. In examining whether human rater bias is a source of DIF, it is necessary not to compare examinees of Spanish and Mandarin L1 backgrounds as a whole, but only those examinees who deserve the same score. Unfortunately, exploring this possibility requires a great deal more data, since L1 groups would need to be partitioned by observed score, and some groups have little overlap in observed scores.

7.2.3 Sociocultural factors

Item 3 is a science-related item, and item 2 for grade band 9–12 involves some quantitative reasoning. Depending on which countries students emigrated from, these types of content-related items might have been easier for students from countries where these concepts were covered in more depth (Huang et al., 2016). However, item 2 for grade band 2–3 still exhibits some DIF with respect to L1, and it is not related to science or quantitative reasoning. Furthermore, there are non-speaking science and quantitative reasoning items. Therefore, to pursue this possibility further, it would be beneficial to examine non-speaking items related to science or quantitative reasoning to see if similar patterns of DIF emerge.

Study 2 revealed that examinees in grade band 9–12 who have Spanish L1 backgrounds tend to struggle with speaking items. There could be some sociocultural or socioeconomic cause, perhaps related to SES, migrant status, age of entry into the U.S. schooling system, or some other factor. Unfortunately, these data are not collected by all states, nor are they collected in the same way, which makes such follow-up analyses more difficult.

[1] It is possible that there is no strong relationship between L1 and examinees' responses, however this claim cannot be proven by one negative result. Other models, such as Naïve Bayes, may reveal such an association (Appendix 8.5).

There are other sociocultural factors that could be specific to speaking proficiency, such as opportunities or motivation to communicate with others (Derwing and Munro, 2013). However, exploring these factors would require data beyond what is available.

7.2.4 Feature bias

Although feature biases are less salient when using deep learning models, manual features could be helpful in determining the sources of DIF. For instance, it has been shown that embedding layers of LLMs correlate with manually specified features (Ormerod, 2022b). By examining intraclass correlation coefficients, it might be possible to ascertain which manual features are (or are not) associated with responses of each L1 group, providing additional possibilities for investigating sources of DIF.

7.2.5 Machine learning bias

Given the close correspondence between scores generated by human raters and BERT, there is little room for machine learning bias. Nevertheless, Study 2 did find that BERT exacerbated DIF with respect to L1 in grade band 9–12. Even though the effect size is small, it is worth investigating whether or not this is due to modeling or training decisions. One possible avenue of research would be to experiment with different LLMs (besides BERT). In particular, it would be useful to repeat these analyses using a smaller LLM, such as Electra, or an ensemble of models (Ormerod, 2022a).

7.2.6 Other biases

The majority of non-speaking items are relatively easy, based on their IRT parameters. Research has suggested that easier items are more likely to exhibit DIF, since they may be more likely to draw on cultural knowledge (Santelices and Wilson, 2010). If this is the case, then perhaps the non-speaking items do not constitute an appropriate set of anchor items. Although it contradicts research on implicit bias, it is possible that the speaking items are less biased than the non-speaking,

anchor items.

Dorans and Zeller (2004) suggest that there is a relationship between impact and DIF. Although they suggest that this is tied to guessing behavior, this still might be relevant for constructed response, speaking items. For instance, it has been noted that speaking is one of the most difficult aspects of L2 language acquisition (Brown et al., 2000). Thus, examinees of lower language proficiency might be able to guess their way through non-speaking items, but struggle to a far greater extent for speaking items. In this case, there may be something unique about the speaking domain that is construct-relevant, yet interacts with examinees' non-speaking language proficiency. An even more complicated case would be if speaking proficiency was also affected in a unique way by general academic proficiency, in which case there would be a mix of construct-relevant and construct-irrelevant factors associated with speaking items.

7.3 Implications

Findings revealed that there was moderate DIF, based on examinees' L1 backgrounds, specifically for medium–long speaking items in grade band 9–12 (Section 5.3.1). If certain groups of students are indeed (dis)advantaged compared to others, the test may lead to unfair assessment of students' speaking proficiency, which might potentially impact their academic trajectories (Johnson, 2019). Issues of fairness also have ramifications for validity: If the test performs differently for certain groups of students, then the test may be capturing construct-irrelevant features of students' responses. Given these potential dangers, what should be done, from a practical point of view?

7.3.1 Fairness

Unfortunately, without further analyses, there are no clear, actionable steps that can be taken to mitigate DIF in English speaking assessment. One of the main reasons for this is that the cause of DIF remains unknown. Without knowing the cause, removal of DIF runs the risk of a removal of construct-relevant aspects of the test or of examinees' responses to test items. In other words, if all

human rater scores are fair and valid—which is still a possibility—then altering examinees' scores based on DIF could make the test less fair.

Definitions of fairness dictate, to some degree, what should be done. If fairness means leveling the playing field, regardless of cause, then one can use the shrinkage approach (described in Study 3) to reduce DIF to an acceptable level. On the other hand, if fairness means ensuring that the test functions in the same way for all examinees (i.e. the test is valid), then altering scores runs the risk of deforming the construct, and should be avoided. If the cause of DIF is construct-irrelevant, however, then removing DIF would be acceptable, based on either definition of fairness.

7.3.2 Construct (ir)relevance

There are a number of possible construct-irrelevant causes of DIF which, if found to be main drivers of DIF, would promote the unambiguous policy of taking steps to remove DIF. For example, if it were clear that quantitative test items preferentially disadvantaged examinees' with Spanish L1 background in grade band 9–12, then these items should be weighted less heavily (e.g., the shrinkage approach could be used to reduce DIF to an acceptable level for these items). After all, it is not the goal of an English speaking exam to assess quantitative reasoning.

If the main driver(s) of DIF are construct-relevant, however, then removal of DIF depends on one's definition of fairness. For example, if it were clear that the source of DIF lies in the fact that L2 English speaking is a more challenging domain, and that speaking proficiency interacts with overall language proficiency in a unique way, then groups whose L1 backgrounds are lower, on average, would be expected to have lower scores on speaking items. In this case, removing DIF would change the natural properties of the speaking construct itself. Arguments about whether or not DIF should be removed in this case could be made on either side; on the whole, however, psychometricians would probably recommend not removing DIF (American Educational Research Association et al., 2014).

7.3.3 Evaluating bias

Ormerod et al. (2022) notes that the field of automated assessment (of constructed response items) has not been as rigorous as it should be when it comes to evaluating bias. Typically, studies narrowly focus on accuracy or other performance metrics, without regard to whether or not there are discrepancies by group affiliation. To ensure fairness and validity of the test, developers should examine bias alongside conventional performance metrics. However, this project also reveals the difficulties of evaluating (and responding to) test bias.

7.4 Limitations

7.4.1 Sources of DIF

Consistent with other analyses of DIF, Studies 1–3 also struggled to identify sources of DIF. Although it was possible to rule out several sources of DIF, it yet remains unknown what factors are driving DIF, and if those factors are construct relevant or irrelevant.

7.4.2 Automated scoring systems

An ideal study design would examine language models most similar to those currently in use, so as to generate the most relevant and actionable results. Practical realities, however, make it impossible to recreate systems like SpeechRater or Versant (NLP-based assessments created by ETS and Pearson, respectively); these systems were developed and refined over the course of twenty years. Despite these limitations, the automated system developed for this study does share key aspects that are similar to those of SpeechRater and Versant. Perhaps more importantly, the methods used to examine the system for bias are easily adaptable to any automated English speaking proficiency assessment.

7.4.3 Mitigating DIF

Unfortunately, the adversarial approach to debiasing was not successful. Furthermore, without being able to identify the primary source(s) of DIF, using the shrinkage approach may make the test less fair. Together, these findings provide no clear policy recommendation to be taken. Additional research will need to be conducted in order to make helpful recommendations for how to handle DIF, particularly for grade band 9–12.

7.4.4 Measures of DIF

Our analyses are based around one family of metrics of uniform DIF, z. One of the benefits of z is that it is commonly used in practice, it is highly interpretable with well-established effect sizes, and it is easy to aggregate across items and focal groups. One of the drawbacks, however, is that it does not capture non-uniform DIF, and it is not ideal in terms of statistical power (Woods et al., 2013). It could be beneficial to include other metrics of DIF in future research.

7.5 Future research

Several potentially valuable extensions of the study have been proposed in Sections 6.4, 7.2, and 7.3. This section summarizes past proposals, and considers several new avenues of research.

7.5.1 Exploring other sources of bias

The source of DIF in speaking items remains unknown, yet this information is vital in determining what policy of action to take with regard to DIF. Specifically, it would be most helpful to know if the source is relevant or irrelevant to the speaking construct. If irrelevant, then mitigating DIF (e.g. using the shrinkage approach) would be warranted. One of the challenges in investigating source(s) of DIF is prioritizing research tasks, since sources must be investigated one at a time.

Perhaps of primary importance is determining if speaking is related to non-speaking language

proficiency. The advantage of using this as a starting point is that data is already available, and it is one of the few (perhaps the only) easy-to-verify construct-relevant sources of DIF.

A more laborious analysis of construct relevant sources of DIF could involve linguistic analyses. There are a number of techniques that could yield intuitive results showing the differences between examinees' responses, based on gender or L1. A simple place to start would be to examine word frequencies and bigrams. Such analyses might suggest if differences are present at the response level. Given that BERT was unable to strongly differentiate responses based on gender or L1, however, this might not yield fruitful results; prior to conducting these analyses, it might be beneficial to see if BERT can differentiate responses following data augmentation, or conditioned on score (even with the limited data available).

In a similar vein, it could be beneficial to conduct a follow-up investigation into why BERT exacerbated bias in grade band 9–12 based on examinees' L1. To explore this, it might be beneficial to borrow from some methods of explainable AI, such as integrated gradients (Ormerod, 2022b), or even something simpler such as term-frequency inverse-document-frequency (Jurafsky and Martin, 2023).

Another approach would be to examine construct-irrelevant factors. To this end, a relative easy set of analyses would involve examining migrant status, socioeconomic status, and other widely available background characteristics. It would possible to make linear adjustments to examinees' speaking scores and, based on the relationships between factors and scores, recalculate DIF.

7.5.2 Human raters as units of analysis

Instead of studying DIF with items as the central unit of analysis, it could be advantageous to study DIF with human raters as the central focus. Recent research has found that labeled data varies by background characteristics of human raters, such as race (Prabhakaran et al., 2021). Research in English language assessment has shown that raters are more favorable to speakers who share the same L1 background. Such an analysis could reveal that, even if implicit bias does not affect scores

on an aggregate level, it should still be taken into account at the individual rater level. Additional data would be required to conduct these analyses.

7.5.3 General AI scoring models

Although BERT is still a focus of research in English speaking assessment (e.g. Wang et al., 2021), it is far from state of the art. Ensembling smaller LLMs have been found to produce better results (Ormerod et al., 2021). Additionally, recent research has found that general AI models, such GPT-4, can produce accurate predictions with only a handful of training examples. It is important to continue to monitor biases for automated assessment systems built on general AI, especially since these models are changing (and supposedly improving) on a constant basis.

7.5.4 Improving the shrinkage approach

As mentioned in Study 3, the shrinkage approach could be further refined, particularly by specifying an additional component in the loss function. Doing so would allow the shrinkage model to optimize the balance between maximizing information, on the one hand, and reducing DIF, on the other.

7.5.5 Exploring other methods of DIF

It could be beneficial to include a non-uniform, and ideally more powerful, statistical method to test for DIF. The standardized mean difference (SMD) approach used in Studies 2 and 3 has the advantage of being widely deployed in practice, and it includes well-established cutpoints. However, it might be helpful to triangulate DIF (or identify other patterns of DIF) through the use of other methods, such as the IRT-based Wald Test (Woods et al., 2013) and weighted Area Between Curve (Hansen et al., 2014). These approaches would be able to identify non-uniform patterns of DIF, and perhaps have more statistical power.

CHAPTER 8

Appendices

8.1 L1 groups

In selecting L1 groups, one of our aims was to represent languages from around the globe. In some cases, this required grouping languages to reach an adequate sample size for statistical analyses. Given the constraints of sample size, we tried to ensure that L1 groups were as geo-historically related to each other as possible (Brown, 2005). The four composite L1 groups in our study were (1) Hindi, (2) Mayan languages, (3) Persian, and (4) Swahili. For simplicity, we refer to composite L1 groups by the predominate language within each group, with the exception of Hindi. It would be more accurate, however, to refer to the L1 groups as (1) Indo-Aryan, (2) Indigenous languages of Central and South America, (3) Indo-European languages of the Middle East, and (4) Niger-Congo languages.

The languages within each of the composite L1 groups are presented in Table 8.1. Note that the names of languages are derived from states' departments of education, which do not follow the same naming conventions. We made minor changes in compiling the list of languages (e.g. changing "Panjabi" to "Punjabi").

There is a great deal of heterogeneity within L1 groups, as with gender, and as with all other demographic characteristics. We note that L1 is not synonymous with cultural identity, racial identity, geographic identity, or preferred language. Despite these limitations, in the context of English speech assessment, we believe L1 is a more relevant construct than, say, conventional racial categories (e.g. White, Asian, Black).

Table 8.1: Languages of composite L1 groups by grade band.

	Grade Band 2-3		Grade Band 9-12	
Language	**n**	**%**	**n**	**%**
Hindi				
Punjabi	157	37.7	75	40.5
Hindi	124	29.8	39	21.1
Urdu	65	15.6	35	18.9
Gujarati	46	11.1	30	16.2
Marathi	24	5.8	6	3.2
Mayan languages				
Mayan languages	212	89.1	214	82.9
Q'anjob'al	24	10.1	40	15.5
Quechua	1	0.4	3	1.2
Q'eqchi	1	0.4	1	0.4
Persian				
Persian	209	70.8	97	49.2
Kurdish	76	25.8	87	44.2
Farsi	10	3.4	13	6.6
Swahili				
Swahili	89	42.6	120	55.3
Nuer	37	17.7	28	12.9
Niger-Kordofanian languages	16	7.7	16	7.4
Dinka	19	9.1	11	5.1
Kinyarwanda	7	3.3	19	8.8
Wolof	15	7.2	10	4.6
Fulah	10	4.8	5	2.3
Igbo	7	3.3	5	2.3
Yoruba	3	1.4	1	0.5
Hausa	1	0.5	1	0.5
Akan	2	1	0	0
Shona	2	1	0	0
Chichewa; Chewa; Nyanja	0	0	1	0.5
Kirundi	1	0.5	0	0

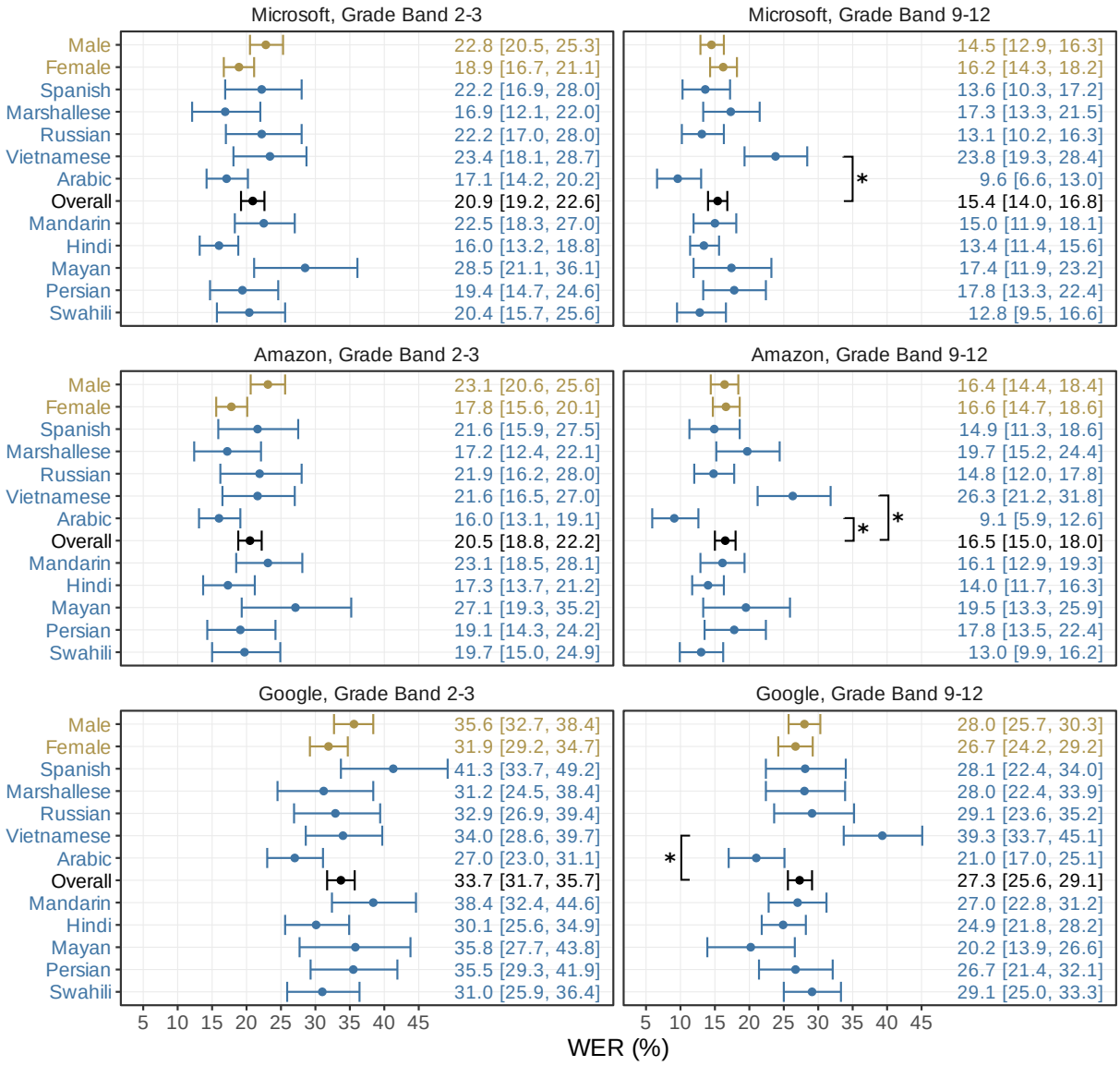
8.2 Comparison of WER across three automated transcription services

Figure 8.1 presents word error rate (WER) of examinees' speech, based on three of the largest automated transcription services (Microsoft, Amazon, and Google). WER is presented separately for younger examinees (grade band 2–3) and older examinees (grade band 9–12). Results are present in aggregate, as well as disaggregated by gender and L1.

Trends are similar across the three services. In particular, older examinees with Vietnamese L1 backgrounds have a higher WER, on average. Although WER is lower for Arabic examinees, it reaches statistical significance only for Amazon's automated transcription service.

Automated transcripts were generated from October 7–12, 2022 (for Google and Amazon) and from November 12–13, 2022 (for Microsoft). Transcript requests were sent using Microsoft, Amazon, and Google APIs for Python 3.8.12 (Python Software Foundation, 2022). Default settings were used for all services. Output language code was set to "en-US" for all three providers. Microsoft required several additional settings: The profanity filter was set to "None," and punctuation mode was set to "Automatic." Microsoft and Amazon provided multiple transcripts by default; in both cases, the most probable transcript was selected for analyses.

Figure 8.1: Average WER estimates produced by Microsoft, Amazon, and Google automated transcription services for grade bands 2–3 and 9–12.



Note: Overall WER appear in black, and disaggregated WER appear in gold (gender) and blue (L1); whiskers indicate 95% confidence intervals; brackets with asterisks indicate statistically significant pairwise comparisons.

8.3 Comparison of WER across datasets

L2-ARCTIC is a publicly available L2 English speech corpus (Zhao et al., 2018a). Similar to the subsample of ELPA21, subjects in L2-ARCTIC were sampled so as to be balanced in terms of gender and L1. Descriptive statistics of the L2-ARCTIC, alongside ELPA21, are presented in Table 8.2. L2-ARCTIC includes a total of 27.1 hours of speech from 24 adults across 6 different L1 (Table 8.2). L2-ARCTIC is comprised of scripted speech: Speakers were asked to read the ARCTIC prompts, a selection of out-of-copyright text curated to be phonetically balanced (Kominek and Black, 2003).

In the English speech assessment data (ELPA21), across all three services (Microsoft, Amazon, Google), native Vietnamese speakers had a higher WER, on average, compared to other L1s (Figure 8.2). In contrast, native Arabic speakers had a lower WER compared to other L1s. Hindi speakers also had a lower WER; however, this finding was statistically significant only for transcripts generated by Microsoft.

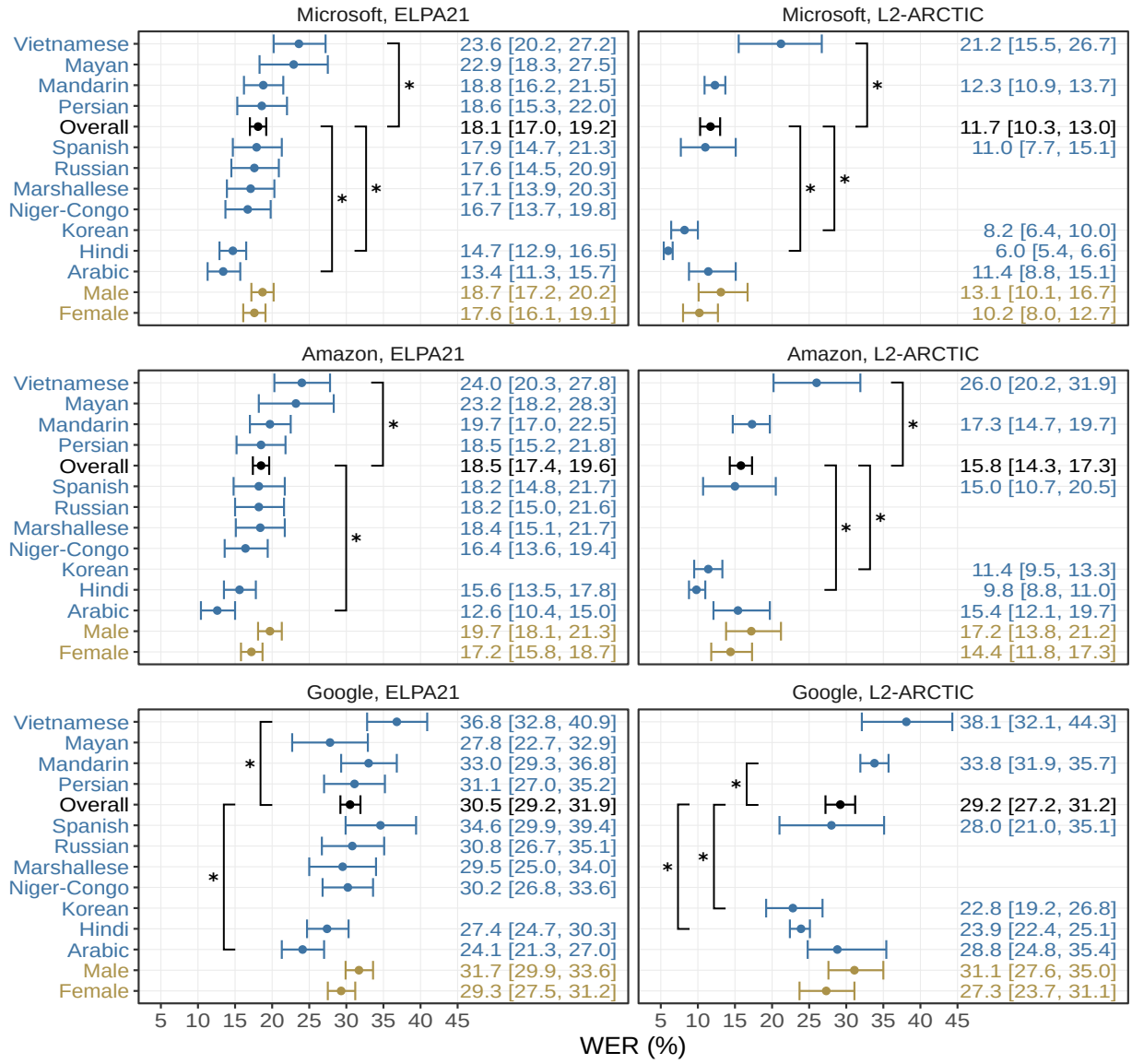
Some L1 biases were consistent with findings from L2-ARCTIC data. In particular, Vietnamese speakers still had a higher WER, on average, in L2-ARCTIC data—at least for transcripts generated by Microsoft and Amazon. Native Hindi speakers, too, were found to have lower WERs, on average, across all three services in the L2-ARCTIC data. Although neither of these comparisons was found to be statistically significant in all 6 cases, when taken together, these findings seem quite robust.

We were not able to corroborate the finding that native Arabic speakers had a lower WER. In L2-ARCTIC data, native Arabic speakers were not found to have a lower WER than other L1s, on average.

Table 8.2: Descriptive statistics of ELPA21 and L2-ARCTIC datasets, with means and standard deviations (in parentheses), overall and disaggregated by gender and L1

	ELPA21				L2-ARCTIC		
	n	Avg. Words	Avg. Seconds	Avg. Proficiency	n	Avg. Words	Avg. Seconds
All	1000	34 (27)	24 (15)	0.04 (1.09)	24	9940 (356)	4061 (518)
L1							
Vietnamese	100	35 (24)	26 (16)	0.15 (0.97)	4	10052 (1)	4320 (280)
Mayan	100	19 (17)	18 (12)	-1.10 (1.13)	–	–	–
Mandarin	100	42 (35)	29 (21)	0.63 (0.97)	4	10038 (13)	4309 (322)
Persian	100	43 (25)	27 (13)	0.64 (0.83)	–	–	–
Spanish	100	32 (24)	23 (12)	0.01 (1.03)	4	9767 (557)	4221 (649)
Russian	100	35 (22)	24 (13)	0.28 (0.90)	–	–	–
Marshallese	100	29 (25)	22 (18)	-0.24 (0.90)	–	–	–
Swahili	100	40 (35)	28 (19)	0.16 (0.94)	–	–	–
Korean	–	–	–	–	4	10049 (4)	4092 (502)
Hindi	100	34 (22)	24 (13)	-0.28 (1.10)	4	10046 (4)	3477 (334)
Arabic	100	34 (22)	23 (11)	0.10 (0.98)	4	9689 (692)	3947 (641)
Gender							
Male	500	32 (25)	23 (14)	-0.03 (1.07)	12	9950 (321)	4178 (410)
Female	500	37 (28)	26 (17)	0.11 (1.10)	12	9930 (403)	3944 (603)

Figure 8.2: Average WER estimates produced by Microsoft, Amazon, and Google automated transcription services for ELPA21 and L2-ARCTIC datasets.

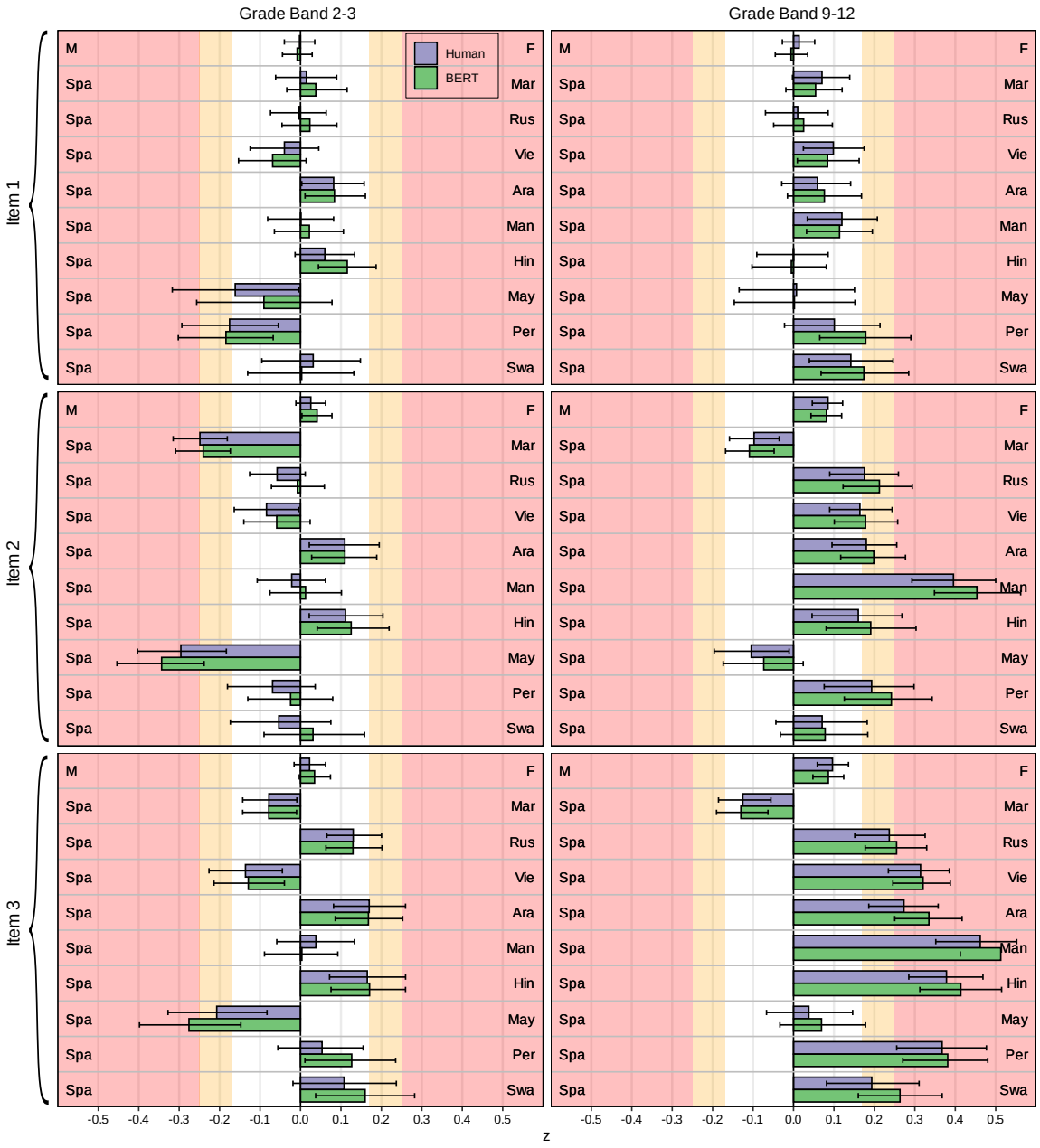


Note: Overall WER appear in black, and disaggregated WER appear in gold (gender) and blue (L1); whiskers indicate 95% confidence intervals; brackets with asterisks indicate statistically significant pairwise comparisons.

8.4 Human vs. BERT DIF for each item

Figure 8.3 presents the magnitude and direction of DIF of items 1–3 for grade bands 2–3 and 9–12, based on gender and all nine L1 focal groups separately.

Figure 8.3: Estimates of direction and magnitude of DIF for each of the 3 speaking items in grade bands 2–3 and 9–12.



Note: Error bars indicate 95% confidence intervals. Yellow shaded regions correspond to moderate DIF, and red shaded regions correspond to strong DIF. Reference groups are listed on the left of each chart (M = Male, Spa = Spanish); focal groups are listed on the right (L1 groups are abbreviated by the first three letters). DIF in the positive direction indicates that the focal group is favored.

8.5 Predicting gender and L1

Off-the-shelf (OOS) BERT models struggled to predict examinees' gender and L1 from text responses (Figures 6.3 and 6.4). This diffculty may signal that there is little gender or L1 information in examinees' responses, or it may indicate that BERT was poorly suited for the task. To help clarify this ambiguity, the task (of predicting gender and L1) was replicated using a different model. In this research context, the Naïve Bayes (NB) classifier well-suited for comparison: NB typically performs well on smaller datasets, and it is conceptually transparent relative to BERT (Jurafsky and Martin, 2023). All n-grams of length 1–2 were used as predictors during training. Training and testing sets were identical to those used in Studies 2 and 3 (Section 3.6.2).

When it came to predicting gender, BERT outperformed NB with respect to accuracy in nearly all cases, the one exception being item 3 in grade band 9–12 (Table 8.3). With respect to F1 scores, however, NB tended to outperform BERT, particularly in grade band 9–12, where the gender imbalance was more acute (55.1% male in grade band 9–12, compared to 51.5% male in grade band 2–3; Table 3.1). NB achieved higher *macro-average F1 scores*, i.e. the mean of male and female F1 scores, as well as higher *weighted-average F1 scores*, i.e. the mean of male and female F1 scores weighted by sample size.

These differences between BERT and NB are not surprising. BERT may attain higher accuracy because it is trained to minimize cross-entropy loss against one-hot labels, which up-weights the majority group even in the absence of informative text input; information in text input needs to be strong enough to overcome class size imbalances. Although the prior in NB also up-weights the majority group, NB is also more sensitive to features (in this case 1–2 grams) unique to minority groups. Thus, NB predicts the minority group at a higher rate than BERT, resulting in higher F1 scores. It is unclear, however, why NB outperforms BERT on item 3 in grade band 9-12; further research is required to investigate this anomaly.

The patterns in BERT and NB performance for predicting L1 groups are nearly identical to those of gender. Table 8.4 shows that BERT, again, generally attains higher accuracy, whereas NB

Table 8.3: Performance of predicting gender for items 1–3, comparing off-the-shelf BERT to a Naïve Bayes classifier.

	Grade Band 2–3						Grade Band 9–12					
	Acc.		$F1_m$		$F1_w$		Acc.		$F1_m$		$F1_w$	
Item	**B**	**N**	**B**	**N**	**B**	**N**	**B**	**N**	**B**	**N**	**B**	**N**
1	.518	.513	.514	.512	.515	.511	.566	.548	.515	.538	.531	.545
2	.603	.575	.600	.575	.601	.574	.584	.560	.538	.550	.553	.557
3	.560	.559	.554	.559	.556	.559	.570	.598	.543	.588	.555	.594

Note: "Acc." refers to accuracy, "$F1_m$" to macro-average F1 score, and "$F1_w$" to weighted-average F1 score. "B" refers to BERT, and "N" refers to "Naïve Bayes." Performance was measured on the testing dataset.

attains higher F1 scores. Also consistent is the anomalous finding that NB outperform BERT on item 3 in grade band 9-12.

The question of how much gender or L1 information is embedded in examinees' responses is model-specific. Neither BERT nor NB are able to predict gender or L1 with a high degree of consistency. Yet NB does attain higher F1 scores, revealing that there are additional features in examinees' text input related to gender and L1 that are not learned by BERT. Other models may be capable of identifying additional features beyond what BERT or NB is capable of.

Table 8.4: Performance of predicting L1 group for items 1–3, comparing off-the-shelf BERT to a Naïve Bayes classifier.

	Grade Band 2–3						**Grade Band 9–12**					
	Acc.		**$F1_m$**		**$F1_w$**		**Acc.**		**$F1_m$**		**$F1_w$**	
Item	**B**	**N**	**B**	**N**	**B**	**N**	**B**	**N**	**B**	**N**	**B**	**N**
1	.502	.496	.067	.078	.336	.343	.525	.513	.071	.112	.364	.393
2	.502	.497	.067	.090	.336	.353	.523	.526	.069	.122	.365	.402
3	.502	.493	.067	.082	.337	.345	.525	.549	.071	.170	.365	.447

Note: "Acc." refers to accuracy, "$F1_m$" to macro-average F1 score, and "$F1_w$" to weighted-average F1 score. "B" refers to BERT, and "N" refers to "Naïve Bayes." Performance was measured on the testing dataset.

Bibliography

Nancy L Allen, John R Donoghue, and Terry L Schoeps. The naep 1998 technical report. *Education Statistics Quarterly*, 3(4):95–98, 2001.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for educational and psychological testing*. American Educational Research Association, 2014.

David P Anderson. Elpa21 item development process report, 2015.

William H Angoff. Perspectives on differential item functioning methodology. 1993.

Isaac I Bejar. A validity-based approach to quality control and assurance of automated scoring. *Assessment in Education: Principles, Policy & Practice*, 18(3):319–341, 2011.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

Annie Brown, Noriko Iwashita, and Tim McNamara. An examination of rater orientations and test-taker performance on english-for-academic-purposes speaking tasks. *ETS Research Report Series*, 2005(1):i–157, 2005.

H Douglas Brown et al. *Principles of language learning and teaching*, volume 4. Longman New York, 2000.

Keith Brown. *Encyclopedia of language and linguistics*, volume 1. Elsevier, 2005.

Li Cai. flexmirt: Flexible multilevel item factor analysis and test scoring [computer software]. *Seattle, WA: Vector Psychometric Group, LLC*, 2012.

Li Cai, Kilchan Choi, Mark Hansen, and Lauren Harrell. Item response theory. *Annual Review of Statistics and Its Application*, 3:297–321, 2016. Publisher: Annual Reviews.

May Pik Yu Chan, June Choe, Aini Li, Yiran Chen, Xin Gao, and Nicole Holliday. Training and typological bias in asr performance for world englishes. In *Proceedings of the 23rd Conference of the International Speech Communication Association*, 2022.

Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. End-to-end neural network based automated speech scoring. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6234–6238. IEEE, 2018a.

Lei Chen, Klaus Zechner, Su-Youn Yoon, Keelan Evanini, Xinhao Wang, Anastassia Loukina, Jidong Tao, Lawrence Davis, Chong Min Lee, Min Ma, et al. Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine. *ETS Research Report Series*, 2018(1):1–31, 2018b.

Jian Cheng, Yuan Zhao D'Antilio, Xin Chen, and Jared Bernstein. Automatic assessment of the speech of young english learners. In *Proceedings of the ninth workshop on innovative use of NLP for building educational applications*, pages 12–21, 2014.

Joseph R Cimpian, Karen D Thompson, and Martha B Makowski. Evaluating english learner reclassification policy effects across districts. *American Educational Research Journal*, 54 (1_suppl):255S–278S, 2017.

Jo-Kate Collier and Becky Huang. Test review: Texas english language proficiency assessment system (telpas). *Language Assessment Quarterly*, 17(2):221–230, 2020.

Council of Chief State School Officers. *Framework for English language proficiency development standards corresponding to the Common Core State Standards and the Next Generation Science Standards*. ERIC Clearinghouse, 2012.

Katrijn Denies, Liesbet Heyvaert, Jonas Dockx, and Rianne Janssen. Mapping and explaining

the gender gap in students' second language proficiency across skills, countries and languages. *Learning and Instruction*, 80:101618, 2022.

Tracey M Derwing and Murray J Munro. The development of l2 oral language skills in two l1 groups: A 7-year study. *Language learning*, 63(2):163–185, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alex Dichristofano, Henry Shuster, and Shefali Chandra. Global performance disparities between english-language accents in automatic speech recognition. Technical Report ASR, 2023.

Barbara Dodd, Alison Holm, Zhu Hua, and Sharon Crosbie. Phonological development: a normative study of british english-speaking children. *Clinical Linguistics & Phonetics*, 17(8):617–643, 2003.

Neil J Dorans and Edward Kulick. Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of educational measurement*, 23(4):355–368, 1986.

Neil J Dorans and Karin Zeller. Examining freedle's claims about bias and his proposed solution: Dated data, inappropriate measurement, and incorrect and unfair scoring. *ETS Research Report Series*, 2004(2):1–33, 2004.

Educational Testing Service. Test and score data summary: 2004-05 test year data test of english as a foreign language. 2005.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*, 2018.

Felix Elwert and Christopher Winship. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology*, 40:31–53, 2014.

G Engelhard. Monitoring raters in performance assessments. *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*, pages 261–287, 2002.

George Engelhard. Examining rater errors in the assessment of written composition with a many-faceted rasch model. *Journal of educational measurement*, 31(2):93–112, 1994.

Keelan Evanini, Maurice Cogan Hauck, and Kenji Hakuta. Approaches to automated scoring of speaking for k–12 english language proficiency assessments. *ETS Research Report Series*, 2017 (1):1–11, 2017.

Every Student Succeeds Act. Every student succeeds act (essa). *Pubic Law*, pages 114–95, 2015.

Dominique Fourdrinier, William E Strawderman, and Martin T Wells. *Shrinkage estimation*. Springer, 2018.

Roy Freedle. Correcting the sat's ethnic and social-class bias: A method for reestimating sat scores. *Harvard Educational Review*, 73(1):1–43, 2003.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *arXiv preprint arXiv:2111.09344*, 2021.

John J Godfrey and Edward Holliman. Switchboard-1 release 2. *Linguistic Data Consortium, Philadelphia*, 926:927, 1997.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Anthony G Greenwald and Mahzarin R Banaji. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4, 1995.

Anthony G Greenwald and Linda Hamilton Krieger. Implicit bias: Scientific foundations. *California law review*, 94(4):945–967, 2006.

Roberto Gretter, Marco Matassoni, Katharina Allgaier, Svetlana Tchistiakova, and Daniele Falavigna. Automatic assessment of spoken language proficiency of non-native children. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7435–7439. IEEE, 2019.

Mark Hansen, Li Cai, Brian D Stucky, Joan S Tucker, William G Shadel, and Maria Orlando Edelen. Methodology for developing and evaluating the promis® smoking item banks. *nicotine & tobacco research*, 16(Suppl_3):S175–S189, 2014.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

Ching-Ni Hsieh, Klaus Zechner, and Xiaoming Xi. Features measuring fluency and pronunciation. In *Automated Speaking Assessment*, pages 101–122. Routledge, 2019.

Becky H Huang and Belinda Bustos Flores. The english language proficiency assessment for the 21st century (elpa21). *Language Assessment Quarterly*, 15(4):433–442, 2018.

Xiaoting Huang, Mark Wilson, and Lei Wang. Exploring plausible causes of differential item functioning in the pisa science assessment: language, curriculum or culture. *Educational Psychology*, 36(2):378–390, 2016.

Wiebke Toussaint Hutiri and Aaron Yi Ding. Bias in automated speaker recognition. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, jun 2022. doi: 10.1145/3531146.3533089.

Nicholas Iderhoff. nlp-datasets, 2023. URL `https://github.com/niderhoff/nlp-datasets`.

International English Language Testing System. Test format: speaking, 2023. URL `https://www.ielts.org/en-us/for-test-takers/test-format`.

Véronique Irwin, Jijun Zhang, Xiaolei Wang, Sarah Hein, Ke Wang, Ashley Roberts, Christina York, Amy Barmer, Farrah Bullock Mann, Rita Dilig, et al. Report on the condition of education 2021. nces 2021-144. *National Center for Education Statistics*, 2021.

Angela Johnson. The effects of english learner classification on high school graduation and college attendance. *AERA Open*, 5(2):2332858419850801, 2019.

Angela Johnson. The impact of english learner reclassification on high school reading and academic progress. *Educational Evaluation and Policy Analysis*, 42(1):46–65, sep 2020. doi: 10.3102/0162373719877197.

Marlene Johnston, Gabriela Finn, Jonathan Wolfe, Masa Suzuki, and Hiro Fukuhara. Using an integrated approach to improve the development and scoring of technology-based speaking tests, 2019.

Daniel Jurafsky and James H Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd edition, 2023.

Okim Kang and Katherine Yaw. Social judgement of l2 accented speech stereotyping and its influential factors. *Journal of Multilingual and Multicultural Development*, pages 1–16, 2021.

Andrew Karpinski and James L Hilton. Attitudes and the implicit association test. *Journal of personality and social psychology*, 81(5):774, 2001.

Margarita Kaushanskaya, Megan Gross, and Milijana Buac. Gender differences in child word learning. *Learning and Individual Differences*, 27:82–89, 2013.

Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech

recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020. doi: /10.1073/pnas.1915768117.

John Kominek and Alan W Black. Cmu arctic databases for speech synthesis. *Carnegie Mellon University Language Technologies Institute*, 2003.

Antony John Kunnan. *Evaluating language assessments*. Taylor & Francis, 2017.

Alexander Kwako, Yixin Wan, Jieyu Zhao, Kai-Wei Chang, Li Cai, and Mark Hansen. Using item response theory to measure gender and racial bias of a bert-based automated english speech assessment system. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 1–7, 2022.

Justyna Leśniewska, François Pichette, and Sébastien Béland. First language test bias? comparing french-speaking and polish-speaking participants' performance on the peabody picture vocabulary test. *Canadian Modern Language Review*, 74(1):27–52, 2018.

Stephanie Lindemann and Nicholas Subtirelu. Reliably biased: The role of listener expectation in the perception of second language speech. *Language Learning*, 63(3):567–594, 2013.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Anastassia Loukina and Su-Youn Yoon. Scoring and filtering models for automated speech scoring. In *Automated Speaking Assessment*, pages 75–98. Routledge, 2019.

Anastassia Loukina, Klaus Zechner, Lei Chen, and Michael Heilman. Feature selection for automated speech scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–19, 2015.

Sari Luoma. *Assessing speaking*. Cambridge University Press, 2004.

Nathan Mantel. Chi-square tests with one degree of freedom; extensions of the mantel-haenszel procedure. *Journal of the American Statistical Association*, 58(303):690–700, 1963.

Nina Markl. Language variation and algorithmic bias: understanding algorithmic bias in british english automatic speech recognition. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, number disentangle. ACM, jun 2022. doi: 10.1145/3531146.3533117.

Josh Meyer, Lynn Rauchenstein, Joshua Eisenberg, and Nicholas Howell. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6462–6468 Marseille, 11–16 May 2020*, 2020.

Michalis P Michaelides. An illustration of a mantel-haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment, Research, and Evaluation*, 13(1):7, 2008.

Christopher Ormerod. Short-answer scoring with ensembles of pretrained language models. *arXiv preprint arXiv:2202.11558*, 2022a.

Christopher Ormerod, Susan Lottridge, Amy E Harris, Milan Patel, Paul van Wamelen, Balaji Kodeswaran, Sharon Woolf, and Mackenzie Young. Automated short answer scoring using an ensemble of neural networks and latent semantic analysis classifiers. *International Journal of Artificial Intelligence in Education*, pages 1–30, 2022.

Christopher M. Ormerod, Akanksha Malhotra, and Amir Jafari. Automated essay scoring using efficient transformer-based language models. *arXiv preprint arXiv:2102.13136*, 2021.

Christopher Michael Ormerod. Mapping between hidden states and features to validate automated essay scoring using deberta models. *Psychological Test and Assessment Modeling*, 64(4):495–526, 2022b.

Jaihyun Park, Karla Felix, and Grace Lee. Implicit attitudes toward arab-muslims and the moderating effects of social information. *Basic and Applied Social Psychology*, 29(1):35–45, 2007.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8026–8037, 2019.

Pearson Education, Inc. Versant english test: Test description and validation summary, 2019.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*, 2021.

Python Software Foundation. The python language reference, 2022. URL `https://docs.python.org/3.8/reference/`.

Yao Qian, Patrick Lange, and Keelan Evanini. Automatic speech recognition for automated speech scoring. In *Automated Speaking Assessment*, pages 61–74. Routledge, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

David Reilly, David L Neumann, and Glenda Andrews. Gender differences in reading and writing achievement: Evidence from the national assessment of educational progress (naep). *American Psychologist*, 74(4):445, 2019.

Frank E Saal, Ronald G Downey, and Mary A Lahey. Rating the ratings: Assessing the psychometric quality of rating data. *Psychological bulletin*, 88(2):413, 1980.

Fumiko Samejima. Graded response model. *Handbook of modern item response theory*, pages 85–100, 1997.

Maria Veronica Santelices and Mark Wilson. Unfair treatment? the case of freedle, the sat, and the standardization approach to differential item functioning. *Harvard Educational Review*, 80(1): 106–134, 2010.

Carolyn M Sloane, Erik G Hurst, and Dan A Black. College majors, occupations, and the gender wage gap. *Journal of Economic Perspectives*, 35(4):223–48, 2021.

Steven J Spencer, Christine Logel, and Paul G Davies. Stereotype threat. *Annual review of psychology*, 67:415–437, 2016.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*, 2019.

Jeffrey Steedle, Shalini Kapoor, and Shichao Wang. What's the dif? item properties associated with dif on the act. National Council on Measurement in Education, 2023.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

Harini Suresh and John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9. 2021.

Jidong Tao, Shabnam Ghaffarzadegan, Lei Chen, and Klaus Zechner. Exploring deep learning architectures for automatically grading non-native spontaneous speech. In *2016 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 6140–6144. IEEE, 2016.

Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First Workshop on Ethics in Natural Language Processing*, pages 53–59. Association for Computational Linguistics, 2017.

Rachael Tatman and Conner Kasten. Effects of talker dialect, gender and race on accuracy of bing speech and youtube automatic captions. In *Special Session: Acoustic Manifestations of Social Characteristics*, pages 934–938. International Speech Communication Assocation, INTERSPEECH, 2017. doi: .org/10.21437/Interspeech.2017-1746.

Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1), 1993.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018a.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019.

Xinhao Wang, Keelan Evanini, Yao Qian, and Matthew Mulholland. Automated scoring of spontaneous speech from young learners of english using transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 705–712. IEEE, 2021.

Zhen Wang, Klaus Zechner, and Yu Sun. Monitoring the performance of human and automated scores for spoken responses. *Language Testing*, 35(1):101–120, 2018b.

Rand R Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic press, 2012.

Valerie SL Williams, Lyle V Jones, and John W Tukey. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of educational and behavioral statistics*, 24(1):42–69, 1999.

Paula Winke, Susan Gass, and Carol Myford. Raters' l2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2):231–252, 2013.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, and Sam Shleifer. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

Carol M Woods, Li Cai, and Mian Wang. The langer-improved wald test for dif testing with multiple groups: Evaluation and comparison to two-group irt. *Educational and Psychological Measurement*, 73(3):532–547, 2013.

Barbara Verena Wucherer and Susanne Maria Reiterer. Language is a girlie thing, isn't it? a psycholinguistic exploration of the l2 gender gap. *International Journal of Bilingual Education and Bilingualism*, 21(1):118–134, 2018.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*, 2016.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542, 2016.

Su-Youn Yoon, Xiaofei Lu, and Klaus Zechner. Features measuring vocabulary and grammar. In *Automated Speaking Assessment*, pages 123–137. Routledge, 2019.

Klaus Zechner. What did they actually say? agreement and disagreement among transcribers of non-native spontaneous speech responses in an english proficiency test. In *International Workshop on Speech and Language Technology in Education*, 2009.

Klaus Zechner. Summary and outlook on automated speech scoring. In *Automated speaking assessment*, pages 192–204. Routledge, 2019.

Klaus Zechner and Keelan Evanini. *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge, 2019.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Mo Zhang, Lili Yao, Shelby J Haberman, and Neil J Dorans. Assessing scoring accuracy and assessment accuracy for spoken responses: Using human and machine scores. In *Automated speaking assessment*, pages 32–58. Routledge, 2019.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. L2-arctic: A non-native english speech corpus. In *Interspeech 2018*, 2018a.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018b.

Juan Zuluaga-Gomez, Amrutha Prasad, Iuliia Nigmatulina, Seyyed Sarfjoo, Petr Motlicek, Matthias Kleinert, Hartmut Helmke, Oliver Ohneiser, and Qingran Zhan. How does pre-trained wav2vec 2.0 perform on domain-shifted asr? an extensive benchmark on air traffic control communications. IEEE, 2023.

Bruno D Zumbo. Three generations of dif analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, 4(2):223–233, 2007.

Rebecca Zwick, John R Donoghue, and Angela Grima. Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30(3):233–251, 1993.