Does BERT Exacerbate Gender or L1 Biases in Automated English Speaking Assessment?

Alexander Kwako
University of California, Los Angeles
akwako@ucla.edu

Yixin Wan
University of California, Los Angeles
elaine1wan@ucla.edu

Jieyu Zhao
University of Maryland, College Park
jieyuz@umd.edu

Kai-Wei Chang
University of California, Los Angeles
kwchang@cs.ucla.edu

Li Cai
University of California, Los Angeles
cai@cresst.org

Mark Hansen
University of California, Los Angeles
markhansen@ucla.edu

Abstract

In 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. This study examines gender and native language (L1) biases in human and automated scores, using an off-the-shelf (OOS) BERT model. Analyses focus on a specific type of bias known as differential item functioning (DIF), which compares examinees of similar English language proficiency. Results show that there is a moderate amount of DIF, based on examinees' L1 background in grade band 9–12. DIF is higher when scored by an OOS BERT model, indicating that BERT may exacerbate this bias; however, in practical terms, the degree to which BERT exacerbates DIF is very small. Additionally, there is more DIF for longer speaking items and for older examinees, but BERT does not exacerbate these patterns of DIF.

1 Introduction

Pretrained large language models (LLMs) present new opportunities for English speaking assessments, yet they are prone to perpetuating and, in some cases, exacerbating social prejudices (Blodgett et al., 2020). In educational assessment, researchers have shown that pretrained LLMs can replicate human scoring, including English speaking assessment, with a high degree of accuracy (Wang et al., 2021). Studies of biases of these automated scoring systems, however, is uncommon (Ormerod, 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 L1 backgrounds. This study addresses the need for deeper analyses of bias in LLM-based automated English speaking assessments.

1.1 Bias in English speaking assessment

There are many potential sources of bias in English speaking assessment. We highlight four sources that we believe are most pertinent to the study of gender and L1 biases.

Human rater bias Scholarship on implicit bias demonstrates that human judgment is influenced unconsciously by peripheral cues, including speakers' accents (Kang and Yaw, 2021). In the context of English speaking assessment, 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) reports a strong disconnect between subjective evaluation of speech (e.g. using Likert scales) and more objective measures (e.g. transcription). Although unexplored, implicit bias could also affect examinees based on gender vocal cues.

Research on implicit bias and speech suggests that there may be more bias in the speaking domain, as opposed to other domains, such as writing. By listening to examinees' voices, human raters may be more likely to be influenced by examinees'

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 668–681

accents, triggering implicit bias that affects their judgment during scoring.

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

Additionally, cultural differences may interact with item properties. In one highly-publicized study, Freedle (2003) describes how verbal items draw on cultural knowledge that disadvantage minority examinees. It is possible, then, that certain speaking items require an understanding of the context of schooling in the United States, which may be more or less familiar to examinees of different cultural backgrounds, and particularly for those who recently emigrated.

Curricular differences Huang et al. (2016) report that curricula vary across countries, and that these differences are a likely source of bias in comparative studies of international assessment. Curricular differences between countries would be particularly salient for examinees who entered into the United States schooling system at a later age.

Item difficulty Dorans and Zeller (2004) and Santelices and Wilson (2010) suggest that item difficulty might be related to guessing behavior, which in turn produces bias related to examinees' overall proficiency. 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, yet struggle with speaking items.

1.2 LLMs may exacerbate social biases

Studies have revealed that pretrained LLMs can propagate and, in some cases, amplify negative stereotypes of marginalized groups (Blodgett et al., 2020). Because LLMs are pretrained on large corpora of text largely scraped from the web, societal biases in these texts become embedded in the LLMs. These biases may surface in downstream applications, such as machine translation (Stanovsky et al., 2019) and sentiment analysis (Kiritchenko and Mohammad, 2018).

In English speaking assessment, LLMs are not yet in widespread use. Yet researchers who are exploring their use typically focus on performance metrics (e.g. accuracy) to the exclusion of biases (e.g. Wang et al., 2021). Even in the broader field of NLP-based English speaking assessment, analyses of bias are rarely conducted or reported (e.g. Collier and Huang, 2020). In one rare study, however, Wang et al. (2018) found that their automated scoring system diverged from human raters for several L1 groups.

1.3 Differential item functioning

Differential item functioning (DIF) is a specific type of bias commonly examined in educational and psychological assessment (American Educational Research Association et al., 2014). DIF occurs when "equally able (or proficient) individuals, from different groups, do not have equal probabilities of answering the item correctly" (Angoff, 1993, p. 4).

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

1.4 Study overview and research questions

This study is designed to analyze gender and L1 biases in L2 English speaking assessment, and to determine if these biases are exacerbated by a pretrained LLM-based automated scoring system. Our data come from a large-scale K-12 English language assessment known as the English Language Proficiency Assessment for the 21st Century (ELPA21; Huang and Flores, 2018). For our automated scoring model, we use an off-the-shelf pretrained Bidirectional Encoding Representation using Transformers (BERT) model (Devlin et al., 2018). We focus on BERT because of its seminal status in language modeling, and because it remains a focus of study in English speaking assessment (Wang et al., 2021). We quantify the amount of bias in human and automated scores by measuring DIF. We first describe specific patterns of DIF in human scores, and then determine whether or not BERT exacerbates DIF.

2 Methods

2.1 Data

This study draws on data from the English Language Proficiency Assessment for the 21st Century

	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)

Table 1: Sample descriptive statistics in aggregate ("All") and disaggregated by gender and L1.

(ELPA21), a consortium involving 7 state education agencies in the U.S. (Huang and Flores, 2018). To maintain confidentiality, certain details regarding test items and examinees are omitted.

Analyses focused on two grand bands (2–3 and 9–12) which corresponded to two tests administered during the 2020–2021 school year. For items in the speaking domain, examinees spoke into a microphone for up to two minutes, after which their responses were sent to human raters who assigned holistic integer scores based on item-specific scoring rubrics. All verbal responses in ELPA21 are currently scored by human raters. Consistent with best practices, raters are trained and monitored over time to ensure consistency (Engelhard, 2002).

2.2 Sample design and demographics

The sampling frame included all examinees in grade bands 2–3 or 9–12 who met the following inclusion criteria: answered all three speaking items included in this study; answered at least one item in each of the other three domains; and had gender and L1 demographic information available. To limit the scope of the study, we excluded examinees with disabilities, examinees with non-binary gender, and examinees whose L1 was other than one of the ten L1s analyzed in this study.

From the sampling frame, we sampled 15,000 students.[1] We included all examinees whose L1 was one of the nine L1 focal groups selected for study (Table 1). The remainder of examinees were randomly sampled from Spanish speakers.

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

2.3 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 A for additional details regarding L1 selection and grouping.

2.4 Item selection

Speaking items were selected to span a range of response times (i.e., length or quantity of speech). Specifically, for each grand 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

[1]The size of our sample was limited, in part, by the cost of automated transcription.

[2]See Section 2.6 for how language proficiency was computed for examinees.

Item #	Length	Grade Band 2-3			Grade Band 9-12		
		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)

Table 2: Item descriptive statistics. Item 3 for grand band 9–12 was re-scaled from a 6-point scale to a 5-point scale. This change was made due to the fact that one group of respondents (Hindi) did not receive any 1s. Combining 1s with 2s helped to improve model convergence.

sentence), and one long item (i.e., requiring 3+ sentences). Table 2 presents the lengths of items 1–3, based on average audio duration (in seconds) and average number of words, for both grand bands. To increase comparability between grand bands, our selection of items also took into consideration item type and item information.

2.5 Automated Transcription

Automated transcripts were generated using Amazon Web Services, during October 7–12 and November 14–16, 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.

We conducted an analysis of transcription accuracy and bias of Amazon's automated transcription service, reported in detail in Kwako (2023). Findings pertinent to the present study are reproduced in Appendix B

2.6 Differential item functioning

As discussed in Section 1.3, DIF occurs when there are group differences, conditional 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 (see Section 2.9 for how non-speaking English proficiency was computed). By excluding speaking items, we ensured that estimates of θ were not contaminated by the same type(s) of bias 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 was male (and the focal group was female); for L1, the reference group was Spanish (and the nine focal groups are listed in Table 1).

2.7 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).[3] SMD is calculated as follows:

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

where N_{Fuj} is the number of examinees in the focal group F whose θ puts them in stratum j, and who received score u on the item in question. Multiplying this quantity by 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. Before taking the difference, the expected scores are weighted by the proportion of examinees in the focal group in stratum j.

The effect size, z, is the ratio of SMD to the standard deviation (pooled between the two groups).[4] Intuitively, z represents how much the focal group outperforms the reference group, among 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 used by the Educational Testing

[3] Instead of using the Mantel test (Mantel, 1963), our significance tests were based on bootstrap sampling distributions and B-H adjusted p-values, described in Sections 2.10 and 2.11, respectively.

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

Service and other educational assessment organizations. Generalizing the system to ordinal items, Allen et al. (2001) designate items as having strong DIF if z is greater than or equal to 0.25. Items have weak DIF if z is less than 0.17. And items have moderate DIF if z is between 0.17 and 0.25.

Absolute effect size For certain research questions, the primary interest was not in determining the *direction* of DIF (i.e., which groups are advantaged or disadvantaged), but only in quantifying the *magnitude* of DIF. To address these questions, we based 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.

Differences between effect sizes We also computed differences in effect sizes (i.e. between human and automated scores, between items, and between grade bands). In each of these comparisons, we were interested not in z or z_{abs}, but in first-order differences. 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. In research questions 2 and 3, we also examined second order differences, $\Delta\Delta z_{abs} = |\Delta z_{abs,i} - \Delta z_{abs,j}|$.

2.8 Aggregate DIF metrics

Aggregating DIF effect sizes allowed us to make more general claims about DIF. Analysis of DIF typically revolves around pairwise comparisons at the item level. This fine-grained level of analysis, however, is not suited for making general claims about DIF. To make more general claims (e.g., across multiple items or focal groups) we report *overall* DIF and *factor* DIF.

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 grand bands 2–3 and 9–12, we added examinees' responses to items 1–3, and computed z according to the procedure outlined in Section 2.7. Since z is in units of standard deviation, it is unaffected by differences in items' scales, and thus generalizes well to summed score.

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 were interested in evaluating DIF for the factor as a whole. To evaluate DIF for the entire factor, we took 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}.

2.9 Non-speaking English proficiency

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 we modeled non-speaking items as a unidimensional construct because (1) it simplified interpretation of the matching criterion, since we were interested in non-speaking proficiency as a whole rather than individual domains, (2) it yielded smaller margins of error, and (3) model fit was in an acceptable range for both grade bands, based on limited-information fit statistics and Tucker-Lewis (non-normed) fit indices (M2 RMSEA $\leq$ .03 and M2 TLI $\geq$ .96).

2.10 Statistical Estimation

To compute confidence intervals and p-values, we used a simple bootstrap procedure (Efron and Tibshirani, 1994). Examinees were resampled within grand band, gender, and L1 groups, as these characteristics were central to our study design. Statistics were calculated from 1,000 bootstrapped samples. Confidence intervals were determined from .025 and .975 quantiles for each estimate. p-values were determined by assuming a normal distribution and taking the minimum of a two-sided quantile of the CDF evaluated at 0.

2.11 p-value adjustments

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, statistical significant means that we place 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).

2.12 BERT modeling

Six separate classification models were trained for each of the items analyzed in this study. Cross-entropy served as the loss function. The maxi-

mum number of input tokens depended on the item length: We set the cutoff at two standard deviations above the mean number of tokens for each item. We used the pre-trained uncased BERT base model provided by Huggingface (Wolf et al., 2020). Modeling and training were scripted using Pytorch (Paszke et al., 2019) in Python 9.3.12 (Python Software Foundation, 2022). We explored several possible models with differing hyperparameters as a part of a previous pilot study (Kwako et al., 2022).

2.13 BERT training

Data were split 1:1 into testing and training sets.[5] Testing and training sets were split so as to maintain equal proportions of examinees by gender and L1.

Based on a smaller-scale study, we selected learning rates of 1e-6 for BERT layers and 2e-6 for classification heads (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.

BERT models nearly achieved parity with human raters for items 1 and 2, and outperformed human raters for item 3. See Appendix C for details regarding the performance of each of the six BERT models in terms of accuracy, correlation, and quadratic weighted kappa (QWK).

3 Results

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.

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 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.

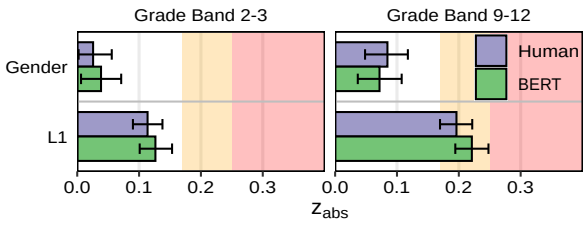

Figure 1: Estimates of overall DIF. Error bars indicate 95% confidence intervals. Yellow shaded regions correspond to moderate DIF, and red shaded regions correspond to strong DIF.

Human vs. BERT overall DIF Overall DIF of automated scores was highly similar to human scores. As seen in Figure 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.

Human vs. BERT individual item DIF In addition to overall DIF, we examined DIF of each individual item. Figure 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 tended to be higher across all items; however, only item 3 reached statistical significance ($\Delta z_{abs} = .032, CI_{95\%} = [.010, .055], p = 3.3 \cdot 10^{-3}$). Again, an effect size of .032 standard deviations is very small.

3.2 DIF increases with item length

Based on human rater scores, longer speaking items tended to exhibit more DIF than shorter speaking items. Automated scores did not exacerbate this trend.

By design, item 3 was longer than item 2, which in turn was longer than item 1. Figure 2 shows that, in general, item 3 had more DIF than item 2, which in turn had more DIF than item 1. Table 3 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

[5] We set aside a larger percentage of data for testing (50% as opposed to the conventional 20%) because we required a more robust calculation of DIF in the testing set for a related study on debiasing (Kwako, 2023).

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

Table 3: Differences in DIF between longer and shorter items, within each grade band, based on human ratings. "*" indicates that an estimate is statistically significant using B-H adjusted p-values. 95% confidence intervals are presented in square brackets.

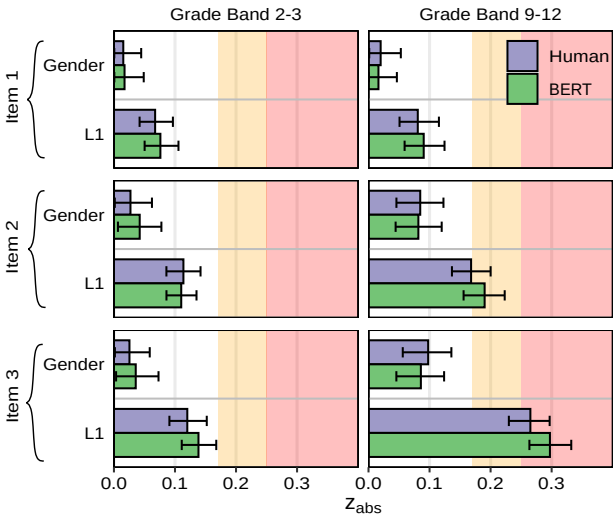

Figure 2: Estimates of DIF for each of the 3 speaking items. Error bars indicate 95% confidence intervals. Yellow shaded regions correspond to moderate DIF, and red shaded regions correspond to strong DIF.

scores (i.e., the gray bars in Figure 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 3, many (but not all) between-item $\Delta z_{abs,ij}$ were statistically significant.

Although longer items tend 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.

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). Automated scores, however, did not exacerbate this trend.

There was significantly more DIF in grade band 9–12, compared to grade band 2–3, in terms of both gender and L1. This trend can be seen clearly in Figure 1. Based on bootstrapped estimates for gender, $\Delta z_{abs} = .059$ ($CI_{95\%} = [.011, .100], p = 4.9 \cdot 10^{-3}$); and for 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 2. The Δz_{abs}, based on human ratings, are presented in Table 4. For example, in item 1, the difference between DIF observed in grade band 2-3 versus grade band 9-12 is $\Delta z_{abs} = .013$, 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 ($\Delta z_{abs} = .054$ and $\Delta z_{abs} = .145$, respectively).

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.

Factor	Item 1	Item 2	Item 3
Gender	.005 [-.033, .042]	.058 * [.011, .105]	.072 * [.019, .118]
L1	.013 [-.029, .057]	.054 * [.012, .098]	.145 * [.098, .193]

Table 4: Differences in DIF between grand bands, based on human ratings, for each of the three speaking items. "*" indicates that an estimate is statistically significant using B-H adjusted p-values. 95% confidence intervals are provided in square brackets.

3.4 Severity of DIF depends on L1 and grade band

The direction and magnitude of DIF varied by L1 background, and patterns were generally not consistent across grand bands. Figure 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 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 D).

4 Discussion

4.1 Main findings

Analysis of differential item functioning (DIF) revealed several patterns of biases in L2 English speaking assessment based on human rater scores, some of which biases were exacerbated by BERT-based automated scores. With respect to human scores, we found that 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' native language (L1) backgrounds. Automated scores generated by off-the-shelf BERT models 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 this bias, however, was very small.

4.2 Causes of DIF

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

Transcription (in)accuracy Prior research showed that there were discrepancies in transcription accuracy based on speakers' L1 background B. Specifically, automated transcription struggled with speakers of Vietnamese L1 backgrounds in grade band 9-12. 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.

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 the text-based BERT model. In other words, removal of acoustic input did not reduce bias. From this, we 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 still affected human raters' judgment.

4.3 Limitations

Our analyses were based around one metric of uniform DIF, z. The benefits of z are 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).

Consistent with other analyses of DIF, our study struggles to identify sources of DIF (Zumbo, 2007).

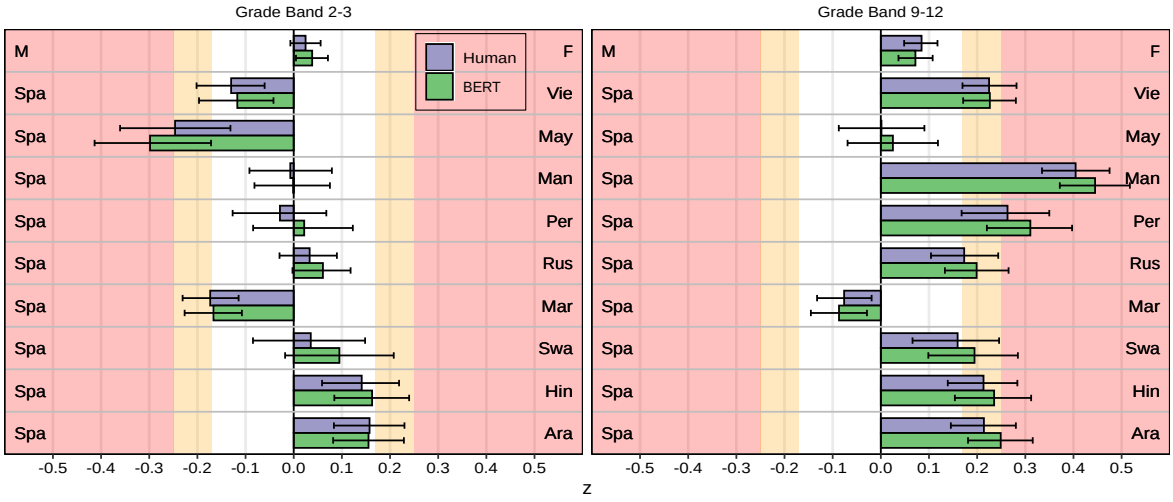

Figure 3: Estimates of direction and magnitude of overall DIF. Error bars indicate 95% confidence intervals. Yellow shaded regions correspond to moderate DIF, and red shaded regions correspond to strong DIF. Reference groups are listed on the left of each chart (M = Male, Spa = Spanish); focal groups 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.

Although it is outside the scope of this study, a fine-grained analysis of examinees' language, especially based on L1, could provide insight. Additionally, it could be beneficial to explore the possibility of modifying BERT using debiasing techniques (Sun et al., 2019). These techniques could potentially reveal sources of DIF and reduce DIF. Follow-up analyses along these lines of inquiry may be found in Kwako (2023).

References

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

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

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

Yoav Benjamini and Yosef Hochberg. 1995. 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.

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

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

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

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

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

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

Katrijn Denies, Liesbet Heyvaert, Jonas Dockx, and Rianne Janssen. 2022. Mapping and explaining the gender gap in students' second language proficiency across skills, countries and languages. *Learning and Instruction*, 80:101618.

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

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

Neil J Dorans and Edward Kulick. 1986. 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.

Neil J Dorans and Karin Zeller. 2004. 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.

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

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

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

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

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

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

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

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

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

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

Alexander Kwako. 2023. *Mitigating Gender and Racial Bias in Automated English Speaking Assessment*. University of California, Los Angeles.

Alexander Kwako, Yixin Wan, Jieyu Zhao, Kai-Wei Chang, Li Cai, and Mark Hansen. 2022. 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.

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

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

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

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

Christopher Ormerod, Susan Lottridge, Amy E Harris, Milan Patel, Paul van Wamelen, Balaji Kodeswaran, Sharon Woolf, and Mackenzie Young. 2022. 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.

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

Python Software Foundation. 2022. The python language reference.

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

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

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

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

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

Valerie SL Williams, Lyle V Jones, and John W Tukey. 1999. 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.

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

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, and Sam Shleifer. 2020. 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.

Carol M Woods, Li Cai, and Mian Wang. 2013. 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.

Klaus Zechner. 2009. 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*.

Bruno D Zumbo. 2007. 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.

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

A 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 (in order to remain consistent with a prior study). 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 5. 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 limitation, in the context of English speaking assessment, we believe L1 is a more relevant construct than, say, conventional racial categories (e.g. White, Asian, Black).

B BERT Performance Metrics

We conducted an analysis of the accuracy and bias of Amazon's automated transcription service. The methodology and results of this study are reported in detail in Kwako (2023); however, pertinent aspects of the study are also presented here. Briefly, we evaluated transcription accuracy by computing word error rate (WER), a common metric that represents the number of transcription errors (i.e. insertions, deletions, and substitutions) as a percentage of words in a given utterance. Transcripts generated by Amazon were compared to a set of manually-generated ("ground truth") transcripts.

Figure 4 presents the WER of automated transcription for grade bands 2-3 and 9-12. Overall, examinees in grand band 2–3 had a higher WER, on average, than examinees in grand band 9–12 (20.5% versus 16.5%, respectively). Note that this level of accuracy is on par with human-human levels of (dis)agreement for L2 English speech, which typically ranges from 15-20% (Zechner, 2009).

There were no statistically significant differences in either grade band with respect to gender. There were also no statistically significant differences in grade band 2-3 with respect to examinees' L1. Yet in grade band 9-12, examinees' whose L1 was Arabic had a lower WER (9.1%), on average, compared to other L1 groups. In contrast, examinees whose L1 was Vietnamese had a higher WER (26.3%) than other L1 groups.

As discussed in Section 3.4, there were no statistically significant differences with respect to overall DIF, when comparing human and BERT scores, based on examinees' L1 groups. Given the close

Language	Grade Band 2-3		Grade Band 9-12	
	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

Table 5: Languages of composite L1 groups by grand band.

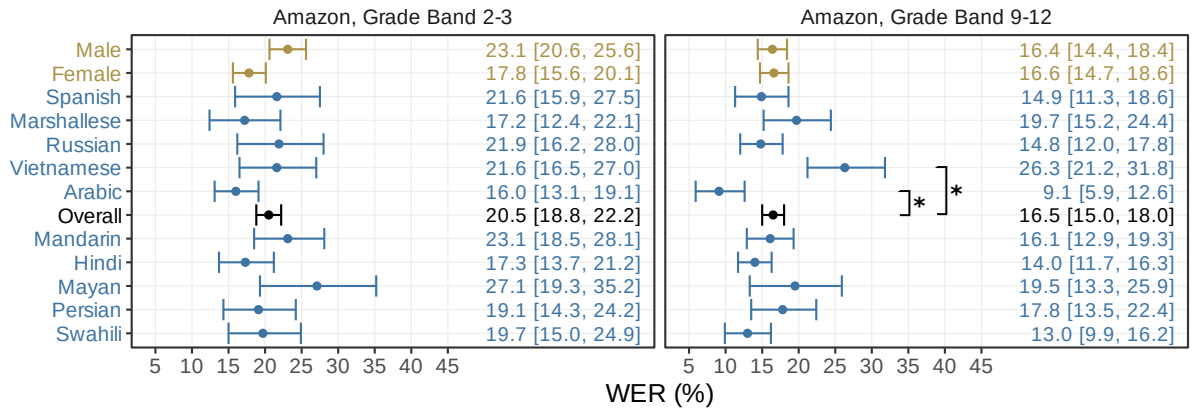

Figure 4: Average word error rate (WER) estimates produced by Amazon's automated transcription service. Overall WER appear in black, and disaggregated WER appear in gold (gender) and blue (L1); whiskers indicate 95% confidence intervals; brackets with asterisks indicate statistically significant pairwise comparisons.

	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

Table 6: Performance of off-the-shelf BERT scoring models for items 1–3, compared to human-human agreement, with respect to accuracy, correlation (r), and quadratic weighted kappa (QWK). "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.

correspondence between human and BERT scores, it is unlikely that transcription inaccuracies engendered lower or higher scores.

C BERT Performance Metrics

Performance metrics of all six BERT models are presented in Table 6. Approximately 10% of all responses were scored by two human raters, independently, which provides the basis for comparisons between human and BERT performance. Off-the-shelf BERT models performed marginally worse for items 1 and 2, but were more consistent than human raters for item 3, across most metrics.

D Human vs. BERT DIF for each item

Figure 5 presents the magnitude and direction of DIF of items 1-3 for grand bands 2-3 and 9-12, based on gender and all nine L1 focal groups separately.

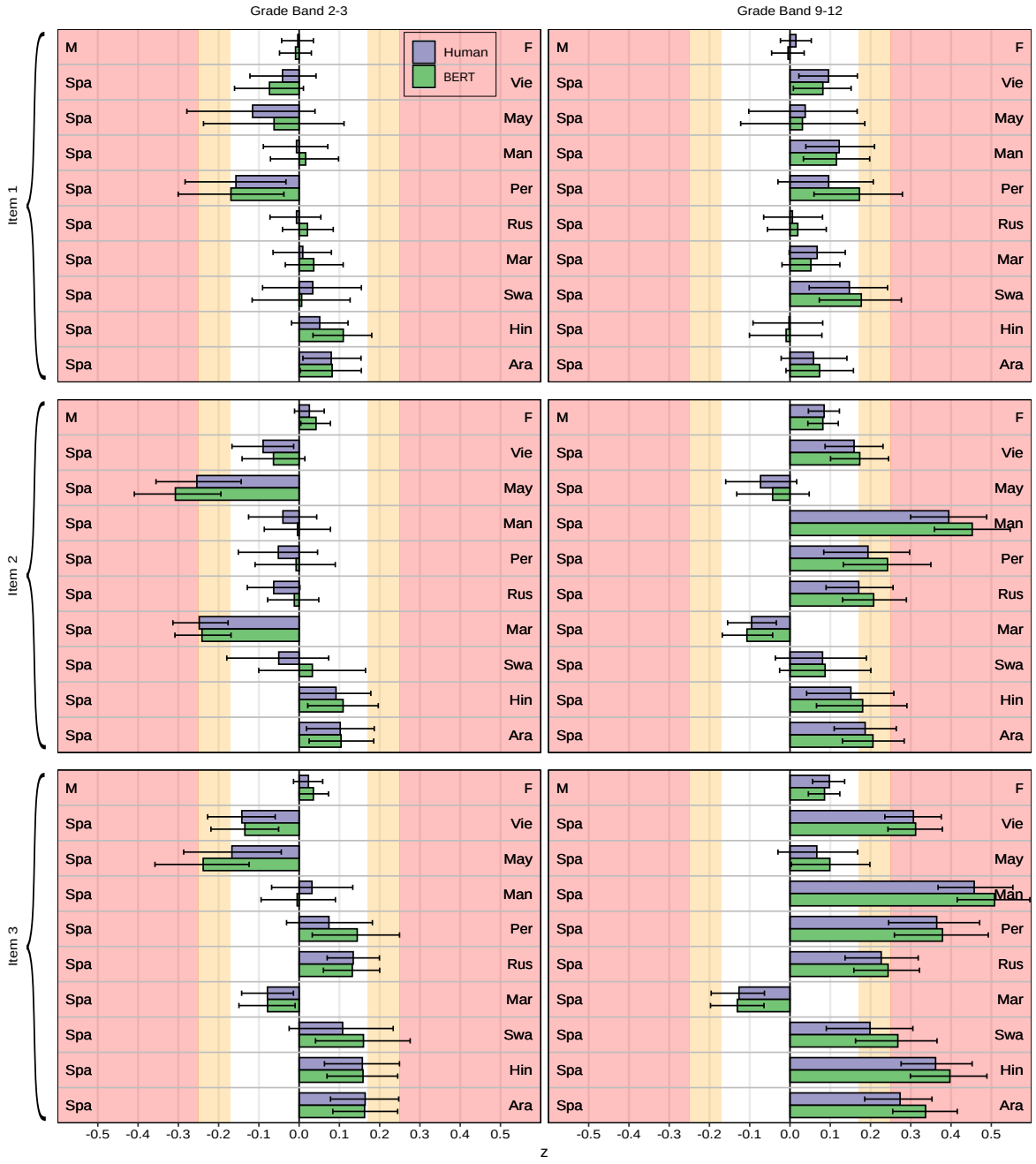

Figure 5: Estimates of direction and magnitude of DIF for each of the three speaking items. Error bars indicate 95% confidence intervals. Yellow shaded regions correspond to moderate DIF, and red shaded regions correspond to strong DIF. Reference groups are listed on the left of each chart (M = Male, 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.