Ensuring Fairness of Human- and AI-Generated Test Items

William C. M. Belzak[1[0000-0001-6594-1651]], Ben Naismith[1[0000-0001-8347-3142]], and Jill Burstein[1[0000-0001-7725-7574]]

[1] Duolingo, Pittsburgh PA 15209, USA
wbelzak@duolingo.com

Abstract. Large language models (LLMs) have been a catalyst to the increased use of AI for automatic item generation on high-stakes assessments. Standard human review processes applied to human-generated content are also important for AI-generated content because AI-generated content can reflect human biases. However, human reviewers have implicit biases and gaps in cultural knowledge which may emerge where the test population is diverse. Quantitative analyses of item responses via differential item functioning (DIF) can help to identify these unknown biases. In this paper, we present DIF results based on item responses from a high-stakes English language assessment (Duolingo English Test - DET). We find that human- and AI-generated content, both of which were reviewed for fairness and bias by humans, show similar amounts of DIF overall but varying amounts by certain test-taker groups. This suggests that humans are unable to identify all biases beforehand, regardless of how item content is generated. To mitigate this problem, we recommend that assessment developers employ human reviewers which represent the diversity of the test-taking population. This may lead to more equitable use of AI in high-stakes educational assessment.

Keywords: Assessment, Fairness and Bias, Differential Item Functioning

1 Introduction

Automated item generation (AIG) has been used in the past decade as an efficient way to generate item content for high-stakes assessments (e.g., question prompts, response options) [1]. Advances in large language models (LLMs), such as Generative Pre-trained Transformer models (e.g., GPT-4) have accelerated the capabilities of AIG, as these methods can produce wider varieties of item content with less human input. On one hand, this has drastically increased the efficiency of content development, allowing for cheaper, more accessible assessments. On the other hand, LLMs are not immune to generating the same types of biased content that humans may produce (e.g., gender and ethnic stereotypes [2]). Thus, humans remain integral to reviewing items according to fairness guidelines [e.g., 3, 4].

Theoretical fairness guidelines such as those by Zieky [3] provide assessment developers with a sensible approach for evaluating whether item content is offensive, inappropriate, or may give unfair advantages to certain groups of test takers. For instance, Zieky [3] recommends avoiding topics that may elicit strong, negative

emotions in test takers such as religion, death, and slavery, unless they are required for valid measurement. However, there is little empirical research demonstrating that human reviewers who use these guidelines can sufficiently filter out all such "construct-irrelevant" content, especially given that humans are prone to implicit biases [4]. For assessments with a diverse population of test takers, this problem may be exacerbated if reviewers are not representative of the test-taking population.

The first aim of this paper is to evaluate whether human reviewers can sufficiently identify all construct-irrelevant content in human- and AI-generated item content. We use statistical analyses (i.e., differential item functioning, DIF) on a high-stakes assessment of English proficiency, the Duolingo English Test (henceforth, the DET), to answer the following research question: Do human- and AI-generated items exhibit DIF despite being human-reviewed? The second aim is to compare human- and AI-generated item content with respect to bias; in other words: do AI-generated items exhibit more or less DIF compared to human-generated items?

2 Background

Currently, GPT-4 is the state-of-the-art LLM driving the most widely used text generation systems (such as ChatGPT). AI researchers openly acknowledge that while there are many benefits of LLMs, LLMs can contribute to fairness and bias (FAB) issues [2, 5]. Although there has been a considerable amount of research on fairness in machine learning and LLMs, the scope of this paper is on FAB reviews with regard to AI-generated content for high-stakes assessment. We focus on this below.

LLMs are trained on human-created texts. Human biases, e.g., stereotypes and prejudices, may be embedded in these texts and then reflected in automatically generated text. Findings from FAB word co-occurrence experiments in Brown et al. [6] suggested stereotypes related to gender, race, and religion. For instance, words describing women were more often related to their appearance than those describing men. From an assessment perspective, language reflecting stereotypes may introduce construct-irrelevant factors which do not assess the intended skill and could unfairly impact test-taker performance, such as distracting the test taker from the task at hand.

Theoretical FAB review guidelines have been used to review human content creation for high-stakes assessments [3, 4]. These guidelines were developed to mitigate the generation of construct-irrelevant test item content which may unfairly impact test-taker performance. Different factors could contribute to test fairness, such as content knowledge. This may be accentuated when an assessment is intended for diverse populations, where test takers represent varied linguistic and cultural contexts. We examine a high-stakes test of English language proficiency to investigate the sufficiency of human reviewers and FAB guidelines in this context, and to compare human- and AI-generated content with respect to the presence and magnitude of bias.

3 Duolingo English Test

The Duolingo English Test (DET) is an online English proficiency test that measures a test-taker's ability to communicate in English-medium settings (e.g., universities).

DET test items are developed by experts in language testing, primarily by leveraging automatic item generation based on authentic sources of English language content [7]. From 'seed' items, a large item bank is generated with items appropriate for test takers of all proficiency levels from A1-C2 on the Common European Framework of Reference [8]; e.g., reading comprehension passages and questions were created using GPT-3 [9]. Items are then reviewed by humans using established FAB guidelines [3].

The DET is taken by a highly diverse testing population, making it an ideal case for investigating a wide range of potential human biases. During 2022 and 2023, there were test takers from 218 countries / dependent territories (most commonly India and China), with an approximately even distribution of self-reported male and female identities (52.4% vs. 47.5%), a median age of 22 (80.9% between 16-30), and a larger proportion of Windows operating system users compared to Mac (74.3% vs. 25.2%).[1]

4 Methods and Data

We analyze the DET for Differential Item Functioning (DIF), a standard statistical analysis used in assessment research. DIF is defined as systematic test-taker differences in item correctness, controlling for true differences in English language proficiency.[2] Test-taker differences refer to differences in background characteristics, including gender, age, computer operating system, and nationality. All items on the DET are reviewed by humans using theoretical FAB guidelines [3]. Our first goal is to determine whether statistical analyses identify biases in item content that human reviewers are unaware of, and which the FAB guidelines do not account for. Our second goal is to compare DIF results between human- and AI-generated items.

We focus on a single item type from the DET – namely, a C-test task of reading comprehension – because this task contains more content per item compared to other item types, and thus may be more liable to exhibiting DIF due to content familiarity. A C-test task is a passage that contains a number of 'damaged' words (the last part of the word is missing) which the test taker must complete (e.g., see p. 5). Partial credit is given for correctly completing the missing text of a single word.[3] DIF analysis is done at the C-test task (or passage) level, rather than at the damaged word level.

From May 2022 to May 2023, we collected DET response data from 1,967 C-test tasks. Humans generated 62% of the tasks, and GPT-3 generated the rest. After excluding C-test tasks with fewer than 250 responses, the mean number of responses per human-generated task was N = 566 and per GPT-3-generated task was N = 540.

We evaluate 23 background variables for DIF in each C-test task (see Table 1). All background variables except age are dummy coded (e.g., gender is 1 if female, 0 if male/other). Age is coded in years (e.g., 20 years old). We use linear regression and Wald tests of statistical significance [11] to test for DIF, such that item correctness is predicted by the overall DET score and each background variable separately. The

[1] Test-taker operating system is a proxy for socioeconomic status (SES), with Mac correlating to higher SES.

[2] This definition describes "intercept DIF" but not "slope DIF". We focus on evaluating intercept DIF here to simplify the presentation of results. See Millsap [10] for a more general definition of DIF.

[3] A score of .5 means that half of the damaged words in the passage were completed correctly.

DET score controls for true differences in English language proficiency, whereas the effect of each background variable on item correctness provides a direct test of DIF.

Although there are a multitude of DIF methods, we use the regression approach because it is particularly flexible and powerful [11]. It can handle large amounts of missing data, and it has high sensitivity in detecting DIF (low Type II error). This is critical in cases where the failure to identify bias is highly detrimental. A downside of this method is that it has lower specificity (high Type 1 error) in larger samples. To adjust for this lack of specificity, we use effect sizes to determine the severity of DIF.

5 Findings

In Table 1, we show percentages of items that exhibited DIF for each background variable and each method of item generation at varying effect sizes (ES). The ESs of .0025, .01, .05, and .1 can be thought about in terms of the proportion of damaged words completed. For instance, if a DIF effect for gender (female) equals .05, this is equivalent to females correctly completing 5% more or less of the damaged C-test passage compared to males, controlling for gender differences in English proficiency.

Notably, DIF effects were small in absolute magnitude for both human- and GPT-3-generated items despite a large percentage of C-test tasks exhibiting DIF for certain groups of test takers (e.g., China, India). It is important to point out, however, that intersections of test-taker backgrounds, such as female Indian test takers may receive an item that has many (dis-)advantageous effects of DIF, leading to more bias.

We find that human- and GPT-3-generated items exhibited similar proportions of DIF across all items. However, some background variables appeared to be (dis-)advantaged more often by human-generated items compared to GPT-3, and vice versa. For instance, human-generated items tended to show slightly higher rates of DIF for countries in the eastern hemisphere (e.g., China, India), whereas GPT-3 items tended to show slightly higher rates of DIF for countries in the western hemisphere (e.g., Canada, United States). This pattern was not observed perfectly however.

The presence of DIF does not indicate the direction of bias, i.e., whether DIF advantages or disadvantages a particular background variable. This is also shown in Table 1 as “Average Effect Size”. Most average effect sizes of DIF are in the same direction for both GPT-3 and human-generated items, although the absolute magnitude of DIF is larger or smaller for some background variables. This is consistent with varying proportions of items with DIF across background variables.

Table 1. Percentage of Items with DIF and Average Effect Size of DIF (GPT-3 vs. Human)

	GPT-3					Human				
	Percentage of Items[1] with Effect Size >=				Average Effect Size[2]	Percentage of Items[1] with Effect Size >=				Average Effect Size[2]
Background variable	.0025	.01	.05	.1		.0025	.01	.05	.1	
Gender	19	19	0	0	0.002	**21**	**21**	**1**	0	**0.021**
Age	**9**	0	0	0	0.000	7	0	0	0	-0.001
Operating System	**19**	**19**	1	0	**0.025**	14	14	**1**	0	0.006
China	41	41	12	0	0.024	**46**	**46**	**21**	**1**	**0.034**
India	37	37	13	0	0.012	**43**	**43**	**24**	**3**	**-0.041**

Canada	**9**	**9**	**7**	**2**	**-0.073**	7	7	6	1	-0.051
Brazil	**11**	**11**	**10**	2	**-0.037**	10	10	10	**2**	-0.017
South Korea	**10**	**10**	**9**	1	**-0.050**	8	8	7	**1**	-0.049
Indonesia	**11**	**11**	**8**	0	**-0.016**	6	6	5	**0**	-0.009
United States	**9**	**9**	**8**	1	**-0.075**	5	5	5	**1**	-0.054
Iran	**9**	**9**	8	1	0.023	9	9	**8**	**2**	**0.041**
Mexico	**9**	**9**	**8**	**3**	**-0.049**	7	7	7	2	0.002
Pakistan	**7**	**7**	**7**	1	-0.037	6	6	6	**2**	**-0.080**
Bangladesh	3	3	3	**1**	-0.029	**4**	**4**	**4**	1	**-0.048**
Columbia	3	3	3	1	-0.001	**7**	**7**	**7**	**2**	**0.035**
France	4	4	4	1	0.020	**6**	**6**	**6**	**2**	**0.030**
Ukraine	0	0	0	0	**0.086**	**5**	**5**	**5**	**2**	0.071
Japan	**4**	**4**	**4**	1	0.024	4	4	4	**1**	**0.058**
Saudi Arabia	3	3	3	2	**-0.120**	**5**	**5**	**5**	**3**	-0.096
Nigeria	3	3	3	1	0.015	**4**	**4**	**3**	**1**	**0.029**
Turkey	**2**	**2**	**2**	0	**-0.078**	2	2	2	**1**	-0.061
Vietnam	1	1	1	0	0.046	**2**	**2**	**2**	**1**	**0.067**
Russia	2	2	2	1	**0.081**	**2**	**2**	**2**	**1**	0.075

1. The values indicate the percentage of items where the absolute DIF effect (for each background variable) is greater than or equal (>=) to .0025, .01, .05, and .1. Bold values indicate the effect is larger for either GPT-3 or human generated items.
2. The average effect size is computed using all non-zero DIF effect sizes for each background variable. Bold values indicate the average effect size is larger *in absolute value* for either GPT-3 or human generated items.

6 Discussion

This paper evaluated whether human review of item content (according to established FAB guidelines) is sufficient for identifying biased test content. In particular, we wanted to know the following: (1) Do human- and AI-generated items exhibit DIF despite being human-reviewed? (2) Do AI-generated items exhibit more or less DIF compared to human-generated items?

First, our DIF analyses suggest that human-reviewed items exhibit DIF despite being human-reviewed, as nearly 40% of the items analyzed showed some degree of DIF (ES >= .01) regardless of the method of item generation. Although the effects for each background variable were often small in isolation, test takers with particular combinations may be affected more severely due to a multitude of DIF effects. For instance, the following is an example of an item with larger DIF effects favoring younger, female, French, Mac-using test takers (in total, ES ≅ .25):

> Since my company was experiencing success, I wanted to fine-tune my processes by finding a more efficient way to promote our services. After researching what others have done, I decided to hold a charity event supporting local children. I contacted a local children's hospital and worked out a deal with them to host our annual fundraising dinner. The event was a huge success, and I hope to make it even bigger next year. I'm sure that the exposure will help us reach new clients and get more jobs.
>
> *Note: The underlined parts of the passage indicate damaged words.*

Fig. 1. Example of AI-generated C-test passage demonstrating DIF effects

This item demonstrates that seemingly innocuous content (e.g., “charity event”) may (dis-)advantage certain test takers (e.g., French vs. non-French). Additionally, there are no clear guidelines from Zieky [4] about “charity events” or similar kinds of content which would lead human reviewers to reject this item. Other variables may also explain the source of DIF, e.g., test takers’ socioeconomic status (as suggested by

DIF in favor of Mac operating system users). Nevertheless, the fact that no cultural explanation is evident highlights the potentially 'invisible' nature of bias. This task was retired from the DET operational item bank based on this analysis.

Second, we did not find strong evidence that GPT-3-generated items exhibited more or less DIF compared to human-generated items. This suggests that human reviewers evaluate item content in similar ways, regardless of the item generation method.

Of course, statistical analyses of DIF are used frequently in high-stakes assessments. Zieky [4] recommends that developers use DIF analyses to investigate whether test items show unfair disadvantages. With the rise of LLMs, however, reviewing content rather than creating it has become the primary job of humans. The findings here thus emphasize previous recommendations of using multi-pronged FAB reviews which integrate both theoretically-motivated human reviews and empirically-motivated quantitative analyses [4].

Based on our findings, we recommend that human reviewers represent the diversity of the test-taking population. This ensures that more diverse perspectives are considered during the human review process, which may reduce the number of items exhibiting DIF due to unknown reviewer biases.

References

1. Gierl, M. J., & Haladyna, T. M. (Eds.). (2012). *Automatic item generation: Theory and practice*. Routledge.
2. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54*(6), 1–35.
3. Zieky, M. J. (2016). Fairness in test design and development. *In: Dorans, N. J., Cook, L. L. (eds.) Fairness in educational assessment and measurement*, pp. 9–31. Routledge.
4. Zieky, M. J. (2015). Developing fair tests. *In: Downing, S. M., Haladyna, T. M. (eds.) Handbook of test development*, pp. 97–115. Routledge.
5. Sherman, J. E. (2010). Multiple levels of cultural bias in TESOL course books. *RELC Journal 41*(3), 267–281.
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems. 33*, 1877–1901.
7. Cardwell, R., LaFlair, G., Naismith, B., & Settles, B. (2022). *Duolingo English Test: Technical Manual.* Duolingo. https://go.duolingo.com/dettechnicalmanual
8. Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Press Syndicate of the University of Cambridge, UK.
9. Attali, Y., Runge, A., LaFlair, G., Yancey, K., Goodwin, S., Park, Y. and von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence, 5*.
10. Millsap, R. E. (2011). Statistical approaches to measurement invariance. Routledge.
11. Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.