ITSC International Testing Services Center G-TELP General Tests of English Language Proficiency

Bias in AI-Generated Test Items

2025

ITSC Research Series
Report ITSC-2025-02

Bias in AI-Generated Test Items

ITSC Research Series

Report ITSC-2025-02

Prepared by Janna Schaeffer, PhD

Abstract

This study examines the use of artificial intelligence (AI), particularly large language models (LLMs) and ChatGPT, in generating multiple-choice items for the G-TELP Level 2 listening assessment. More specifically, the study investigates whether AI-generated multiple-choice questions are perceived to contain more cultural, language, gender, or socio-economic bias than those developed by human experts. The analysis employs a blind evaluation method where 25 experienced English Language Teaching (ELT) professionals rate the presence of bias in the listening comprehension question sets created by humans and AI. Results indicate that AI tools can efficiently produce assessment items, yet they will not be free of subtle forms of bias. Findings suggest that biases are not only present in the AI-generated context but are also multidimensional, intertwining cultural, linguistic, and socio-economic aspects across testing items. This research underscores the need for future investigation into strategies that will effectively shield test takers from bias embedded in content created with the assistance of AI tools.

Introduction

As one of the most popular artificial intelligence (AI) language tools, ChatGPT has become widely utilized in educational practices such as teaching, learning, and the creation of assessments (Busker et al., 2023; Kasneci et al., 2023; TESOL International Association, 2023). It has proven useful for teaching grammar and lexis, designing educational materials, and developing assessments for language learners across all levels of language proficiency (Aryadoust & Luo, 2023; Settles et al., 2020). Recent research indicates that chatbots such as ChatGPT can generate language test items that rival human-created test items in terms of question complexity, creativity, and quality (e.g., Aryadoust, Yu, & Goh, 2022; Aryadoust et al., 2024).

Notably, despite extensive investigation into AI's capabilities, there is limited empirical research investigating whether AI-generated test content exhibits more bias than human-created content. In language testing, issues of fairness and bias are at the forefront, as embedded assumptions about culture, language, gender, and socio-economic status can unfairly disadvantage a diverse population of test takers (Elder, 2017; Kunnan, 2000; Shohamy, 2001).

While studies have identified bias in human-created assessments (McNamara, 2006; Taylor, 2006) and highlighted cultural and gender-related biases in other AI output (Bolukbasi et al., 2016; Buolamwini & Gebru, 2018), few studies offer a direct comparison of bias profiles in AI-generated and human-created language assessments. Even fewer studies examine educators' evaluations of such biases in AI-generated language test items. This study seeks to address that gap by examining whether AI-generated multiple-choice questions are perceived to contain more cultural, language, gender, or socio-economic bias than those developed by human experts. Through a survey of experienced English language educators, this research aims to assess the perception of bias in both AI- and human-created content and explore the implications of using AI in test development.

Background

In recent years, the rise of artificial intelligence technology, including natural language processing and machine learning (ML), has been undeniable (Hagras, 2018; Munoko et al., 2020; Osoba & Welser, 2017). Although initially conceptualized for use in healthcare, commerce, and law, conversational AI tools such as large language models (LLMs) like OpenAI's ChatGPT have become widely incorporated into educational practices (Bellamy et al., 2018; Busker et al., 2023; Feine et al., 2020). As can be seen in the field, the wide availability and affordability of chatbots has notably affected both teaching and learning. Educators across the globe have begun integrating tools such as ChatGPT into their lesson planning, material design, and development; meanwhile, students rely on these tools for help with grammar correction, vocabulary improvement, and even personalized tutoring (Kasneci et al., 2023; Rudolph et al., 2023; TESOL International Association, 2023). The accessibility and popularity of AI have also influenced its use in assessment practices, including test ideation, item design and development, item generation, and automated scoring (Aryadoust & Luo, 2023; Settles et al., 2020).

AI tools such as ChatGPT, which originally served as peripheral support, have now become central in test design, delivery, and scoring, a shift that points to AI's growing role in standardized assessment. Recent research demonstrates the expanding use of chatbots such as ChatGPT in language assessment, specifically, through the generation of listening scripts, speaking prompts, and other test items across the four major skills at all proficiency levels (Aryadoust et al., 2022). Despite concerns expressed by some researchers about AI's semantic redundancy and inadequate differentiation between proficiency levels (McNamara et al., 2014), tools such as Text Inspector and Coh-Metrix confirm that AI-produced test content contains appropriate vocabulary levels as well as variation in lexical and syntactical complexity, affirming its comparability to human-created materials in at least some regards.

Issues such as proficiency level mismatches, cognitive load imbalances, and semantic redundancy are not the only challenges present in AI-generated language content. Like human-created materials, AI-generated content also exhibits elements of bias. Statistical methods have long been used to detect and eliminate biased content in human-created materials that may disadvantage test takers from certain socio-economic or linguistic backgrounds (Shohamy, 2001; Taylor, 2006). These validation practices remain a central objective for content developers aiming to promote culturally responsive and equitable assessment design—and must now be applied similarly to identify and mitigate bias in AI-generated content.

As observed by Xue et al. (2023), biases in chatbot systems can arise from multiple interconnected sources and manifest in various forms. For example, they can stem from the cultural and linguistic backgrounds, personal life experiences, and educational or societal norms of those who design and train the systems. Since AI and machine learning models are designed by humans and trained on human-generated data, human-like biases are inevitably learned and embedded into the content they generate. More specifically, biases are inherent in the training data and underlying algorithmic frameworks of AI and are particularly prevalent in LLMs, which are typically trained on vast web-based corpora containing skewed representations of social groups and embedded societal prejudices (Brown et al., 2020; Buolamwini & Gebru, 2018). Once deployed, AI systems may further reinforce such biases through user interaction and feedback, compounding preprogrammed biases over time. The resulting output can be harmful in the long term, as it perpetuates stereotypes and contributes to unequal access for certain cultural, linguistic, and social groups (Aryadoust & Goh, 2020; Campbell et al., 1997; Smith & Rustagi, 2020).

Current theoretical understandings of potential biases in LLMs have raised growing concern about the practical implications of ChatGPT's algorithmic design and training data, especially given its popularity and expanding use in educational settings. These biases, whether embedded in AI-generated or human-created content, can subtly shape instructional materials, assessment item design, and evaluation processes. Because biased assumptions are often interwoven with neutral cultural or disciplinary content, they can be difficult to identify and isolate (Kunnan, 2000; McNamara, 2006). This challenge is particularly troubling in high-stakes contexts, such as language instruction and assessment.

Although AI has been widely used in educational and, more specifically, language assessment practices, empirical research into the presence of bias in language proficiency assessment materials remains skeletal. More specifically, there are few studies focused on investigating the presence and nature of bias in AI-generated versus human-created multiple-choice items on English proficiency tests, and research into the use of chatbots in listening assessments is particularly sparse. In order for existing biases to be reduced, possible limitations of ML technology ought to be identified and recognized (Malik, 2020; Stine & Kavak, 2023). To do so and to help build this body of research, this study examines the use of AI in the generation of multiple-choice items for the General Test of English Language Proficiency (G-TELP) Level 2 listening test, focusing on the presence of bias. This study closely examines the bias profiles of AI-generated versus human-created listening items and proposes strategies to support more equitable and valid language assessment design.

The Current Study

The current study examines the use of AI, particularly Chat-GPT, in generating multiple-choice items for the G-TELP Level 2 listening assessment. The study's objective is to determine whether human-created or AI-generated content is perceived as having bias (i.e., cultural, linguistic, gender, or socio-economic bias) by English language teaching (ELT) professionals. In this study, both the human-created and AI-generated sets were presented to a group of 25 experienced ELT professionals. These educators, randomly and evenly divided into five groups, conducted a blind evaluation of the two question sets. Through a comparison of participants' responses regarding bias across question sets and subsequent detailed thematic analysis, this study aims to inform more equitable assessment practices, with the additional goal of proposing actionable strategies to reduce bias and improve the fairness, validity, and inclusivity of language assessments. This study explores the following research questions:

1. Are there measurable differences in perceived bias between AI-generated and human-created listening comprehension question sets?

2. If bias is perceived, which specific dimensions—cultural, linguistic, gender-based, or socio-economic—are most commonly identified in AI-generated versus human-created test items?

Method

Participants

Participants included 25 ELT professionals based in North America. All survey respondents had extensive experience in English language instruction, including roles at community colleges, four-year colleges, and universities. All held graduate degrees in TESOL or applied linguistics. Most participants (76%) reported having 15 or more years of teaching experience, and 92% reported having experience designing multiple-choice assessments for English language learners.

Materials

For the purpose of this study, the G-TELP in-house writing team created ten sets of listening scripts with four script genres per set: two conversations and two monologues. Those scripts were provided to the company's trained question writers, who designed a set of four questions per script. Each question set was based on the G-TELP content development guidelines and consisted of three question types per set (what, why, how) and an inferential question of any of the aforementioned types. In keeping with all G-TELP Level 2 listening tests, test items were designed to assess intermediate to advanced (B1–C2) listening skills. ChatGPT-4 was then used to generate additional question sets utilizing the same listening scripts and guidelines as those used by the human question creators.

Procedure

Data were collected using a customized survey developed on the LimeSurvey 6.2 platform. The survey was pilot tested, and feedback from the pilot study participants informed the revision and refinement of the item presentation and overall survey structure to improve usability, functionality, and clarity. Once revised, the survey was opened to 25 study participants. The participants were invited to identify any bias present in the listening question sets and, if found, to identify which of four specific bias types (cultural, language, gender, socio-economic) were perceived within each question set. To do so, participants rated each bias dimension on a scale of 1–3 (no bias, some bias, significant bias). They were then prompted to elaborate on their responses, discussing issues they had noticed in their review of each question set. While each group evaluated both an AI-generated and a human-created set presented sequentially, the specific content presented to participants varied, with different scripts and question sets assigned to each of the five survey groups.

Analysis

A mixed-methods approach to data collection yielded output for both quantitative and qualitative analysis. Quantitative data analysis was conducted using non-parametric statistical methods to evaluate whether there were significant differences in median bias ratings between the two groups. To compare the perceived levels of bias between human-created and AI-generated questions across multiple metrics, a Wilcoxon signed-rank test was conducted across four bias dimensions: cultural, language, gender, and socio-economic. In addition, Kendall's Tau correlation coefficients were calculated to examine the direction and strength of relationships between different bias dimensions within AI-generated and human-created question sets. This correlational analysis was undertaken to provide insight into the interrelatedness of bias types within each set of questions.

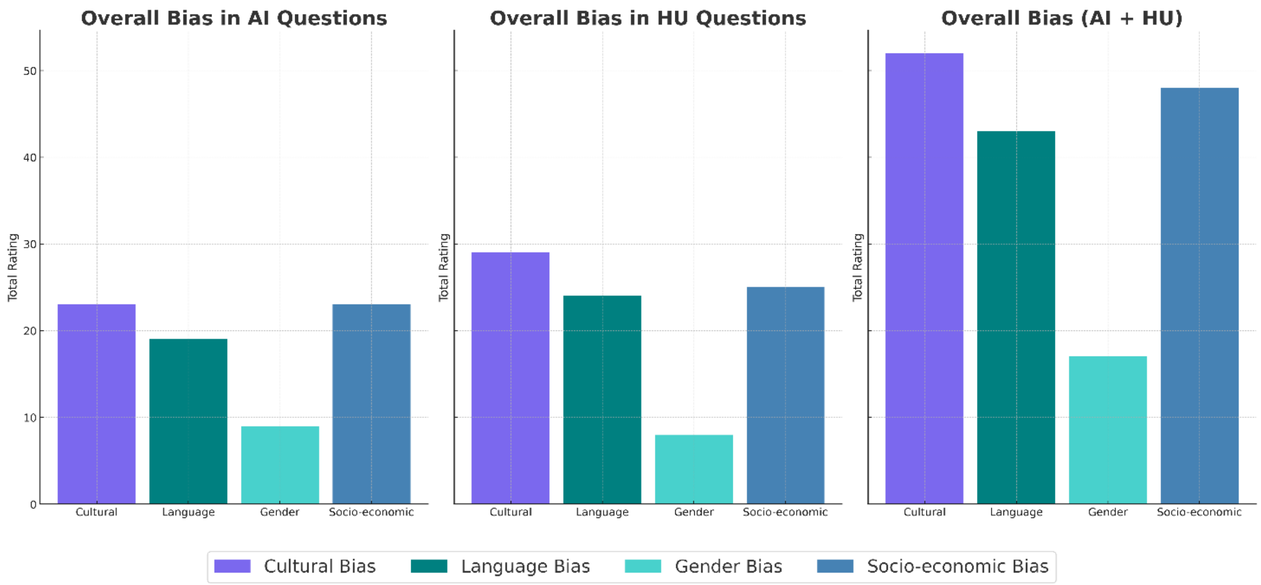
To evaluate the qualitative data, a coding analysis of participant responses was conducted to identify patterns and gather insight, following established qualitative research methodologies (Creswell & Poth, 2018; Flick, 2009). The data were manually coded by multiple coders. A reflexive thematic analysis was applied to uncover patterns and underlying themes related to the four bias categories, and a hybrid approach combining deductive and inductive coding was used. Although four bias categories were predefined, additional topics and patterns emerged during the coding process.

Results

The quantitative analysis revealed a similar pattern in the distribution of perceived bias types across AI-generated, human-generated (HU), and combined (AI + HU) question sets, as illustrated in Figure 1 below. In both sets, cultural bias was the form of bias most frequently noted by respondents. There was also a high perceived presence of socio-economic bias, which consistently appeared as the second most prevalent bias type. Perceptions of language bias were noted at moderate levels across the question sets, suggesting that while present, it was not as prominent as cultural and socio-economic concerns. Gender bias was identified as the least frequently occurring bias. These patterns suggest that manifestations of gender bias may be less immediately recognizable, whereas issues related to cultural familiarity and socio-economic assumptions are more overt and frequently observed.

Figure 1

Overall Ratings of Bias by Dimension



Results of the Wilcoxon signed-rank test indicate that none of the differences between AI-generated and human-created question sets for any dimension were statistically significant. Specifically, cultural bias ratings did not differ significantly between AI and human items (w = 51, p = .180), nor did ratings for language bias (w = 15, p = .166), gender bias (w = 16, p = .763), or socio-economic bias (w = 27.5, p = .593), indicating that survey respondents did not perceive AI-generated or human-created content to contain significantly more bias than the other.

Results of the Kendall's Tau correlation analysis revealed significant relationships, defined as those with p-values < .05, but only in certain dimensions. These associations varied between AI-generated and human-created question sets, but the strongest interconnectedness between bias types was found in the AI-generated sets. Here, several positive correlations emerged, indicating that survey respondents tended to rate multiple biases together when they perceived bias in the questions. More specifically, the test showed a positive correlation (τ = .377, p = .0497) between cultural and language bias, implying that when cultural bias was perceived, language bias was also identified. An even stronger correlation was observed between cultural and socio-economic bias (τ = .484, p = .0106), suggesting a co-occurrence of cultural framing with assumptions of socio-economic privilege. Lastly, a significant correlation was observed between language and socio-economic bias (τ = 0.550, *p* = 0.0044). This outcome indicates that the items respondents considered linguistically challenging may have also contained class-based assumptions. Notably, gender bias, the bias least identified across all question sets, did not significantly correlate with any other bias types.

The correlation analysis revealed only one significant correlation in the human-created question sets, which was between cultural and language bias (τ = 0.671, *p* = 0.0004), and the relationship was relatively strong. All other correlations between cultural, socio-economic, and gender bias were nonsignificant. Specifically, a negative correlation between cultural and gender bias (τ = -0.195, *p* = 0.33) was observed that was not statistically significant.

The patterns observed in the quantitative analysis were also reflected in the qualitative findings. In their responses, survey participants commented extensively on the frequent co-occurrence of cultural, linguistic, and socio-economic bias, which supports the statistically significant correlations identified in the quantitative data. Similarly, the infrequent occurrence of gender bias in the quantitative findings was echoed in the qualitative responses. Overall, the qualitative findings reinforce the complex interrelation of bias types and provide a deeper understanding of how cultural, linguistic, socio-economic, and gender biases are interpreted by test takers across both AI-generated and human-created content. These patterns are explored in more detail in the section below.

Discussion

Cultural Bias

In the qualitative analysis for this study, survey responses identified instances of Western-centric bias in AI-generated question sets. Respondents noted that these questions often focused on leisure activities such as yoga or watching horror movies, which they described as cultural experiences typical of the middle class in the United States. These observations, which highlighted bias rooted in Western lifestyles and pop-cultural familiarity, align with claims that chatbots often rely on cultural references more familiar to those with culture-specific capital (Busker et al., 2023). Such references to everyday Western practices, as flagged by survey respondents, may make questions more accessible to some test takers with a deeper understanding or experience of dominant cultural norms while disadvantaging others—a pattern also noted by Akcan and Kabasakal (2019). Although some of these notions were mentioned in the scenarios, respondents emphasized that the framing of the AI-generated question sets required test takers to draw upon a lived schema, prompting them to activate their cultural background knowledge of and lived experiences with Western cultural practices. Those unable to reliably contextualize such references due to a lack of personal experience with these cultures could find themselves at a disadvantage (Mousa & Ali, 2022).

The presence of culture-specific vocabulary and embedded social assumptions identified in the AI sets by the respondents, such as hiring cleaners or participating in graduation rituals (common in the US) can also skew an assessment in favor of culturally aligned test takers. It could be argued that such instances of bias are a direct result of training language models on English-language corpora sourced primarily from Western- or US-centric data, which tends to produce output that reflects and reinforces Western cultural norms (Buolamwini & Gebru, 2018).

Notably, some cultural bias was also identified in human-created question sets. Respondents observed that, similar to the AI-generated sets, these also contained questions with Western-centric notions carried over from the script scenarios. However, items labeled by respondents as biased were perceived to rely on institutional terminology and workplace norms, emphasizing cultural practices related to education, leisure, and work. These instances of domain-specific, middle-class framing and culturally specific vocabulary prompted respondents to comment on how such framing might benefit some learners while posing difficulties for others. Words such as “attic,” “security deposit,” and “field day” were cited as culture-specific lexis that, in their view, assumes background knowledge and affluence, reflecting experiences not universally shared. However, studies of Korean EFL test takers, the target audience for this test, indicate that they typically possess large English vocabularies, often exceeding the 5,000–7,000-word range associated with B2–C1 proficiency levels (Kim, 2015; Park, 2024). This suggests that, although some items may include culturally specific references, Korean test takers are generally well-equipped to comprehend vocabulary flagged by the respondents.

Language Bias

In their analysis of AI-generated question sets for language bias, most respondents perceived the bias to be relatively subtle. However, when bias was identified, respondents rarely described the "biased" questions as linguistically inaccessible or overtly exclusionary to test takers. These findings echo Busker et al. (2023), who concluded that language bias in AI-generated content is often more contextual and latent, appearing through implicit social assumptions rather than explicitly discriminatory language. Respondents also noted that language bias rarely appeared in isolation; rather, it frequently intersected with cultural and socio-economic (class-based and institutional) bias embedded in the scenarios and question sets. In their responses, they highlighted instances where vocabulary items became more difficult due to culturally or socio-economically laden contexts—such as those referencing financially privileged, academic, or workplace settings. Notably, this co-occurrence of bias types was not exclusive to AI-generated content. Respondents reported that both AI-generated and human-created question sets demonstrated varying degrees of overlap between linguistic and socio-economic assumptions. For example, as mentioned above, they pointed out that seemingly simple phrases such as "RV" or "graduation speaker" required familiarity with privileged class experiences to be fully understood.

Respondents further observed that the main difference between AI-generated and human-created question sets lay in the framing and context in which language bias emerged. AI-generated sets were seen to rely more on casual, lifestyle-oriented language (e.g., references to furniture, shopping, moving), whereas human-created sets tended to include more technical or academic vocabulary. According to respondents, manifestations of language bias in the human-created sets often included features of pragmatics, register, and figures of speech. For instance, they flagged terms like "security deposit" and "loss of funds" as potentially unfamiliar to learners without direct experience of US banking practices. Similarly, phrases such as "negotiating a raise" or "calling the police after a minor incident" were cited as examples of how language can intersect with socio-economic privilege. Respondents argued that although these lexical items may be relatively common in English, they can still act as barriers for test takers who lack economic stability or cultural familiarity.

In terms of lexical familiarity, respondents reported more low-frequency vocabulary in the AI-generated question sets, whereas human-created items appeared more likely to contain collocations and features related to pragmatic function. These observations align with findings by Feine et al. (2020), who emphasized the limitations of chatbot systems in capturing dominant communicative norms and pragmatic variation. Their study concluded that chatbot-generated assessments often fail to reflect the full range of socio-linguistic diversity present in human-authored content. Finally, respondents noted that AI-generated questions tended to be framed in more conversational and "everyday" language, while human-created question sets were more heavily grounded in workplace or institutional discourse. Overall, both sets were perceived to contain some elements of language bias, even if tied to different types of cultural or socio-economic experience.

Gender Bias

In their discussion of perceived gender bias, survey respondents identified subtle manifestations of gender bias in the AI-generated set. When compared to all other types of bias examined in this study, gender bias was noted and flagged the least across both AI-generated and human-created question sets. This could be explained by recent efforts undertaken by tech companies to mitigate manifestations of gender bias in LLMs (Gupta et al., 2022; Liu et al., 2020; Thakur et al., 2023). Gender bias is also seen by some as more readily identifiable and even more easily addressed compared to more nuanced linguistic and cultural biases, as it tends to be more overt and measurable (Tremewan, 2024; West et al., 2019).

Whenever cases of bias were identified, respondents noted that gender bias was subtly communicated through character roles and interaction dynamics. They observed that women were more often portrayed using passive and emotionally coded language, while men were depicted as competent and assertive. For example, respondents pointed to an AI-generated question set where a female character was depicted as vulnerable in an exchange with her supervisor (a man) when asking for a raise, which they perceived as reinforcing an emotionally coded and passive image of a woman. These patterns align with findings by Brown et al. (2020), who demonstrated LLMs' propensity for role-based differentiation, where women are described more sentimentally, while men are associated with rational, abstract concepts. Additionally, in the AI-generated sets, respondents noted instances where men were depicted as capable and knowledgeable, whereas women were less so. Such patterns align with findings by Troske et al. (2022), who found that AI systems may utilize and universalize culturally specific (Western/US) gender expectations unless prompted to generate content that accounts for other regional variations and norms. Respondents also observed that the AI-generated questions portrayed a more distinct and intentional power imbalance between the two genders; this bias is more commonly embedded in who resolves a problem, who leads a conversation, or who is portrayed as seeking help. These findings are consistent with patterns noted by Bolukbasi et al. (2016), which indicate that AI-generated content tends to link women with nurturing, domestic activities but portray men in leadership or technical roles.

Respondents also perceived subtle gender bias manifestations in the human-created sets. They acknowledged, however, that these sets contained more nuanced cultural expectations in terms of gender. In their view, male or female characters in human-created question sets were not explicitly depicted in stereotypical gender roles. Instead, the framing in human-created sets suggested that certain activities potentially align with a specific gender, but those activities may vary depending on cultural context. For example, in some cultures, going to yoga or sharing recipes could be perceived as ‘feminine,’ while in other cultures, these activities may not be marked; they are seen as gender-neutral or even masculine. Respondents also reflected that the wording used to frame activities and characters in human-created sets struck a more reflexive tone regarding gender expectations, suggesting a higher degree of interpretive variance.

The respondents also observed that gender bias never appeared in isolation; instead, it was commonly intertwined with cultural and socio-economic assumptions. For example, one respondent pointed to an instance of traditional gender roles in the workplace, where the boss or supervisor was male with demonstrated authority, whereas the employee requesting a raise was female. This observation echoes research by Caliskan et al. (2017), which describes such intersectionality of bias as a type of co-occurrence bias, often encoded alongside other cultural norms and social practices and further promoting stereotypical narratives. In our case, respondents flagged such co-occurring bias in questions where gender roles were embedded within cultural and socio-economic assumptions of Western lifestyles.

Socio-Economic Bias

In terms of socio-economic bias, respondents flagged instances in both AI-generated and human-created question sets. They noted, however, that the manifestation of bias in the test items differed between the two types of question sets. In AI-generated content, respondents identified cases where assumptions about consumer-based lifestyle habits were prominent. AI-generated questions often referred to experiences such as purchasing high-value items and planning vacations. Examples labeled by survey respondents as biased assumed a certain level of wealth, convenience-oriented consumption, and access to disposable income, which may not reflect the lived experiences of all test takers. In the respondents' view, the human-created questions did not explicitly focus on privileged lifestyle activities, but they did assume some understanding of financial and institutional systems, making the test items more accessible to test takers from affluent educational and/or occupational backgrounds.

The findings of the study demonstrate that both AI-generated and human-created assessment items may contain subtle representations of cultural, linguistic, gender, and socio-economic bias. However, the nature, framing, and breadth of these biases often differ. The human-created sets were seen as containing some institutional language and presupposing test takers' basic cultural familiarity and financial stability; meanwhile, cultural and socio-economic assumptions reflecting Western, middle-class lifestyles appeared more prominently in the AI-generated sets. Gender assumptions were also more frequently observed in the AI-generated sets, even though gender bias was the least frequently identified overall. Notably, bias was rarely cited as occurring in isolation; rather, it intersected across multiple categories. These findings highlight the importance of examining all assessment content critically to ensure fairness and accessibility for diverse groups of test takers.

Limitations, Implications, and Future Directions

A few study limitations are worth noting. First, many of the issues respondents raised when assessing the presence of bias in the question sets pertained more specifically to the listening scripts (scenarios) themselves rather than the questions designed to test understanding of these scripts. Future investigations would benefit from explicit directions as to which pieces of content, specifically, are meant to be evaluated. Second, It must also be noted that many instances of socio-economic bias flagged by respondents seemed to be the same issues raised in their comments regarding cultural or linguistic bias, making it difficult to isolate bias types. This further attests to the clear co-occurrence between the bias dimensions discussed in earlier sections.

A third limitation of the study is the fixed sequence in which the item sets were presented to survey respondents. All participants reviewed the AI-generated questions first, followed by the human-created set. This ordering may have primed them to examine the second set with greater scrutiny for bias. Evidence of this potential priming effect appeared in several comments, where respondents explicitly stated that they were looking for bias because they had been prompted to do so. Some also reported becoming hyper-aware of more subtle biases after initially encountering clearer examples in the first set they reviewed. Such priming may have influenced the types of bias respondents identified in the second set, as prior exposure to a particular stimulus can shape how subsequent information is perceived (Cargile & Giles, 1997; Blackwell et al., 2023). This presentation may, in fact, explain why participants perceived bias to be slightly more prominent in the human-created question sets. Regardless, to better control for order effects and isolate perceptions of bias, future studies should consider counterbalancing the presentation of question sets.

The findings of this study have several important implications for the design of language assessments. When using AI tools to generate assessment items, greater attention must be paid to the content produced to ensure that the items measure what they are intended to measure. Specifically, it is crucial to avoid test items that assess test takers' cultural familiarity rather than their vocabulary knowledge or overall language proficiency, and the study revealed that learners who lack experience with Western lifestyles or relevant lived schemata may be unfairly disadvantaged. Therefore, these test takers may underperform due to issues beyond lower language ability that encompass cultural and socio-economic assumptions embedded in the content. This possibility further underscores the need to ensure that assessments are accessible and fair to all test takers, particularly those from linguistically and culturally diverse backgrounds.

Furthermore, results of this research highlight the importance of continued human oversight in test development, particularly when AI is used to generate assessment items. Human item review (i.e., a "human-in-the-loop" approach) remains essential for detecting nuanced or intersectional biases that LLMs may not yet be able to fully recognize. Furthermore, the current study's findings demonstrate the need to develop more systematic and scalable approaches to bias detection in automated test generation. Finally, greater scrutiny should be applied to the training data used to improve chatbot performance, ensuring that the underlying web-based corpora are free from cultural, linguistic, gender, and socio-economic bias, as well as from Western or US-centric norms that may not be universally applicable.

Conclusion

This study, which examined the presence of four dimensions of bias in AI-generated and human-created question sets, highlights the multidimensional nature of bias in English language assessment—extending beyond simple cultural or linguistic insensitivity. Findings reveal that when bias is present, cultural, linguistic, and socio-economic dimensions often intertwine rather than occur independently. At the same time, the study underscores the need for future research into strategies for shielding test materials from bias embedded in content that has been generated with the assistance of LLMs.

The study also emphasizes the importance of adopting a more holistic approach to the design and development of language assessment materials. The focus of bias detection should expand beyond identifying discrete and overt racial, ethnic, or gender stereotypes to include subtler assumptions about lifestyle, economic status, and cultural familiarity embedded in test items. As the role of AI in assessment continues to grow, so too does the need for more robust bias detection and oversight mechanisms. A commitment to inclusive assessment design practices that accurately measure the language proficiency of diverse learner populations must remain a central objective to ensure fairness and validity in language assessment.

References

Akcan, S., & Kabasakal, E. (2019). An investigation of item bias of English test: The case of 2016 year undergraduate placement exam in Turkey. *International Journal of Assessment Tools in Education, 6*(1), 48–62. https://doi.org/10.21449/ijate.508581

Aryadoust, V., & Goh, C. C. M. (2020). Exploring listening strategy use and its relationship to listening test performance with the use of a cognitive diagnostic model. *Language Assessment Quarterly, 17*(1), 17–37. https://doi.org/10.1080/15434303.2019.1692342

Aryadoust, V., & Luo, L. (2023). Automated writing evaluation systems: Advances, limitations, and future directions. *Language Testing, 40*(1), 55–78. https://doi.org/10.1177/02655322221103624

Aryadoust, V., Yu, C. H., & Goh, C. C. M. (2022). Artificial intelligence in language assessment: A scoping review. *Language Testing, 39*(2), 163–188. https://doi.org/10.1177/02655322211062652

Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence, 6*, Article 100204. https://doi.org/10.1016/j.caeai.2024.100204

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv*. https://arxiv.org/abs/1810.01943

Blackwell, M., Brown, J. R., Hill, S., Imai, K., & Yamamoto, T. (2023). Priming bias versus post-treatment bias in experimental designs. *Political Analysis, 31*(2), 251–268. https://doi.org/10.1017/pan.2023.14

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems, 29*, 4349–4357.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT)*, 77–91.

Busker, M., Dousay, T. A., & Sundeen, T. (2023). AI in education: Perceptions, practices, and possibilities. *International Journal of Educational Technology in Higher Education, 20*(1), 1–20. https://doi.org/10.1186/s41239-023-00374-7

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1997). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Cargile, A. C., & Giles, H. (1997). Understanding language attitudes: Exploring listener affect and identity. *Language & Communication, 17*(3), 195–217. https://doi.org/10.1016/S0271-5309(97)00016-5

Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). SAGE Publications.

Elder, C. (2017). Language assessment in higher education. In E. Shohamy, I. Or, & S. May (Eds.), *Language testing and assessment* (pp. 1–9). Springer. https://doi.org/10.1007/978-3-319-02261-1_35

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2020). A taxonomy of social cues for conversational agents. *International Journal of Human–Computer Studies, 132*, 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

Flick, U. (2009). *An introduction to qualitative research* (4th ed.). SAGE Publications.

Gupta, R., Patel, S., & Sharma, A. (2022). Automating updates: Leveraging AI for real-time decision support. *Journal of Information Systems, 15*, 45–58.

Hagras, H. (2018). Toward human-understandable, explainable AI. *Computer, 51*(9), 28–36. https://doi.org/10.1109/MC.2018.3620965

Kasneci, E., Sessler, K., Küchenhoff, H., Bannert, M., Dementieva, D., Fischer, F., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kim, Y. (2015). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *Korean Association of Teachers of English, 69*(4), 45–67.

Kunnan, A. J. (2000). Fairness and validation in language assessment. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–13). Cambridge University Press.

Liu, V. X. (2020). The future of AI in critical care is augmented, not artificial, intelligence. *Critical Care, 24*, Article 673. https://doi.org/10.1186/s13054-020-03494-y

Malik, M. (2020). A hierarchy of limitations in machine learning. *arXiv preprint arXiv:2002.05193*.

McNamara, T. (2006). *Language testing: The social dimension*. Blackwell Publishing.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Munoko, I., Brown-Liburd, H. L., & Vasarhelyi, M. A. (2020). The ethical implications of using artificial intelligence in auditing. *Journal of Business Ethics, 167*(2), 209–234. https://doi.org/10.1007/s10551-019-04407-1

Osoba, O. A., & Welser, W. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. RAND Corporation.

Park, H. I. (2024). Validation of the Korean bilingual version of the vocabulary size test. *English Teaching, 79*(2), 139–162. https://doi.org/10.15858/engtea.79.2.202406.139

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching, 6*(1), 1–13. https://doi.org/10.37074/jalt.2023.6.1.9

Settles, B., Brust, C. A., Gustafson, E., Hagiwara, M., & Madnani, N. (2020). Second language acquisition modeling: An ensemble approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4237–4245).

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Longman.

Smith, G., & Rustagi, K. (2020). Mitigating bias in AI models. *AI & Society, 35*, 803–813. https://doi.org/10.1007/s00146-020-00965-9

Stine, J. M., & Kavak, H. (2023). Measuring model robustness and fairness: A review. *AI and Ethics, 3*(1), 45–58. https://doi.org/10.1007/s43681-022-00143-2

Taylor, L. (2006). The changing landscape of English language assessment. *ELT Journal, 60*(3), 215–223. https://doi.org/10.1093/elt/ccl010

TESOL International Association. (2023). *ChatGPT and TESOL: Opportunities and considerations for English language professionals*. https://www.tesol.org/

Thakur, A., Hinge, P., & Adhegaonkar, V. (2023). Use of artificial intelligence (AI) in recruitment and selection. In *Proceedings of the International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022)* (pp. 632–640). Atlantis Press. https://doi.org/10.2991/978-94-6463-136-4_54

Tremewan, L. (2024, April 23). *Gender bias in AI: Is there a problem with representation?* Finder. https://www.finder.com/uk/stats-facts/gender-bias-in-ai

Troske, S., Garg, N., & Binns, R. (2022). The algorithmic construction of gender in AI systems. *Patterns, 3*(2), Article 100416. https://doi.org/10.1016/j.patter.2022.100416

West, M., Kraut, R., & Ei Chew, H. (2019). How AI bots and voice assistants reinforce gender bias. *Brookings Institution*. https://www.brookings.edu/articles/how-ai-bots-and-voice-assistants-reinforce-gender-bias/

Xue, V. W., Lei, P., & Cho, W. C. (2023). The potential impact of ChatGPT in clinical and translational medicine. *Clinical and Translational Medicine, 13*(2), e1216. https://doi.org/10.1002/ctm2.1216