Maximizing Impact: Leveraging Assessment and Accountability to Drive Student Learning

National Conference on Student Assessment

Council of Chief State School Officers
NCSA™
National Conference on Student Assessment

Finding a Place for AI in Inclusive Language Assessments

Monday, June 23, 2025 2:45-3:30 PM

SPEAKERS

Dr. Janna Schaeffer, Senior Researcher, ITSC-Group (presenting)

Dr. Kym Taylor, Chief Director, ITSC-Group (co-author)

Increasing Role of LLMs in Assessment

Rapid integration → Efficiency gains → Personalization potential → Widespread adoption → Emerging concerns

Increasing Role of LLMs in Assessment

Rapid integration

LLMs such as ChatGPT and GPT-4 are increasingly used in automated test development and scoring systems.

Increasing Role of LLMs in Assessment

Efficiency gains

Generative AI accelerates test item creation, reducing time and cost compared to traditional methods.

Increasing Role of LLMs in Assessment

Personalization potential

AI enables adaptive assessment through dynamic item generation that can be tailored to learner proficiency.

Increasing Role of LLMs in Assessment

Major companies (e.g., Duolingo, ETS, Pearson) are actively exploring or already deploying AI for creation of assessment tasks.

Widespread adoption

Increasing Role of LLMs in Assessment

Despite their benefits, AI systems raise validity, bias, and transparency issues in high-stakes educational contexts.

Emerging concerns

Council of Chief State School Officers NCSA National Conference on Student Assessment

AI's Potential to Complement Human-Created Assessments

- LLMs can produce large volumes of items across levels, domains, and question types (Malik et. al, 2024).
- AI test items can be refined to enhance quality and promote fairness (Belzak, Naismith, & Burstein, 2023).
- Generative models may suggest novel linguistic and cultural contexts (Li et al., 2024).
- AI can help identify patterns in human-authored materials through large-scale analysis (Ferrara, 2024).

BACKGROUND

PROJECT OVERVIEW

RESEARCH CONTEXT

GAP IN RESEARCH

BACKGROUND

What is the study about?

- Explores bias in assessments
- Focuses on listening comprehension test items
- Compares AI vs. human-created content from G-TELP Level 2 listening tests
- Identifies perceived cultural, linguistic, gender, and socio-economic biases

BACKGROUND

Why investigate bias in AI-generated testing?

- AI used to create test items across test types (Aryadoust et al., 2022).
- AI materials reflect bias from training data (Brown et al., 2020; Buolamwini & Gebru, 2018).
- Subtle bias undermines assessment fairness (Kim and Zabelina, 2015; Kunnan, 2000; Shohamy, 2001).

BACKGROUND

What is missing from our current research understanding?

- Bias manifestation in testing content (Elder, 2012; Kim & Zabelina, 2015)
- Overemphasis on gender bias (Baiqiang, 2007; Karami, 2011)
- Neglection of intersectionality (Brand et al., 2022)
- Limited comparison with human-created assessments (Durak et al., 2024; Herbold et al., 2023)

Questions that guided our research:

- Are there measurable differences in perceived bias between AI-generated and human-created listening comprehension question sets?
- If bias is perceived, which specific dimensions (cultural, linguistic, gender-based, or socio-economic) are most commonly identified in test items?

PARTICIPANTS

25 English language teaching professionals in the US

Graduate degrees in TESOL / Applied Linguistics

Majority (76%) with 15 years or more experience

Each randomly assigned to one of five surveys

DESIGN

COMPARATIVE, MIXED-METHODS APPROACH

ONLINE SURVEY CREATED USING LIMESURVEY 6.5

QUANTITATIVE ANALYSIS AND FINDINGS

Dimension	*p* value	Significance
Cultural bias	$p = .180$	none
Language bias	$p = .166$	none
Gender bias	$p = .763$	none
Socio-economic bias	$p = .593$	none

QUANTITATIVE ANALYSIS AND FINDINGS: AI SETS

	culture	language	socio-econ	gender
culture	—	τ = .377* p = .0497	τ = .484* p = .0106	τ = 0.163 p = ns
language		—	τ = .550* p = .0044	τ = 0.113 p = ns
gender			—	τ = 0.071 p = ns
socio-econ				—

QUANTITATIVE ANALYSIS AND FINDINGS: AI SETS

	culture	language	socio-econon	gender
culture	—	τ = .377* p = .0497	τ = .484* p = .0106	τ = 0.163 p = ns
language		—	τ = .550* p = .0044	τ = 0.113 p = ns
gender			—	τ = 0.071 p = ns
socio-econ				—

significant

QUANTITATIVE ANALYSIS AND FINDINGS: AI SETS

	culture	language	socio-econ	gender
culture	—	$\tau = .377^*$ $p = .0497$	$\tau = .484^*$ $p = .0106$	$\tau = 0.163$ p = ns
language		—	$\tau = .550^*$ $p = .0044$	$\tau = 0.113$ p = ns
gender			—	$\tau = 0.071$ p = ns
socio-econ				—

significant

QUANTITATIVE ANALYSIS AND FINDINGS: AI SETS

	culture	language	socio-econ	gender
culture	—	$\tau = .377^*$ $p = .0497$	$\tau = .484^*$ $p = .0106$	$\tau = 0.163$ $p = ns$
language		—	$\tau = .550^*$ $p = .0044$	$\tau = 0.113$ $p = ns$
gender			—	$\tau = 0.071$ $p = ns$
socio-econ				—

significant

KEY TAKEAWAYS

- Cultural and language bias co-occurred in the same question content.
- Cultural and socio-economic bias were closely linked.
- Language and socio-economic bias showed the strongest association.

These intersections of bias raise concerns about compound disadvantage in test performance for linguistically or socio-economically marginalized learners.

QUANTITATIVE ANALYSIS AND FINDINGS: HUMAN SETS

	culture	language	socio-econon	gender
culture	—	τ = .671* p = .0004	τ = 0.201 p = ns	τ = -0.195 p = 0.33
language		—	τ = 0.116 p = ns	τ = -0.091 p = ns
gender			—	τ = -0.056 p = ns
socio-econ				—

QUANTITATIVE ANALYSIS AND FINDINGS: HUMAN SETS

	culture	language	socio-econon	gender
culture	—	$\tau = .671^*$ $p = .0004$	$\tau = 0.201$ $p =$ ns	$\tau = -0.195$ $p = 0.33$
language		—	$\tau = 0.116$ $p =$ ns	$\tau = -0.091$ $p =$ ns
gender			—	$\tau = -0.056$ $p =$ ns
socio-econ				—

significant

KEY TAKEAWAYS

- Language and culture are tightly interwoven.
- Cultural references often included linguistically challenging elements.

Other bias pairings (e.g., gender, socio-economic) showed no significant correlations in the human set.

QUALITATIVE ANALYSIS

- Qualitative data were analyzed through Reflexive Thematic Analysis (Braun & Clarke, 2006)
  - Manual coding conducted by multiple coders
  - Hybrid approach: deductive (predefined) + inductive (emerging) coding
  - Focused on four bias categories; new patterns also emerged

QUALITATIVE ANALYSIS: AI SETS

Cultural bias

- Comprehension relied on familiarity with Western cultural norms
- Reflection of implicit cultural assumptions from Western corpora

QUALITATIVE ANALYSIS: AI SETS

"… (in the questions) They are planning to find a convertible couch since there is no guest room. So there is definite culture bias in this segment. It really resonates with white, middle-class American millennials more than anyone, I think..."

"In the US it's normal to take a full day off from scholarly instruction as a reward at the end of the school year. But is this the case in other parts of the world? I think some western cultures might have a significant advantage for this set of questions."

QUALITATIVE ANALYSIS: HUMAN SETS

Cultural bias

relied more on institutional/workplace references and US-centric vocabulary like "attic," "security deposit," and "field day".

QUALITATIVE ANALYSIS: HUMAN SETS

"Discussing part-time jobs, loss of funds (instead of has no money), attic, inherited, treat in the testing items -- these are all words that may be unfamiliar because of the ways employment works in other countries, interior and architecture of other cultures..."

" 'Field days' that are discussed in the questions are likely different in different cultures and even in different socio-economic areas in the US. People who attended a US elementary school may be more likely to have experienced such an event."

QUALITATIVE ANALYSIS: AI SETS

Language bias

- more subtle and contextually embedded
- included more casual, lifestyle-oriented
- contained more low-frequency terminology
- less effective in reproducing pragmatic norms

QUALITATIVE ANALYSIS: AI SETS

"Items include phrasal verbs and cultural references that would give students some difficulty."

"Adjectives used in the questions are low-frequency and may require some pre-teaching of vocabulary."

QUALITATIVE ANALYSIS: HUMAN SETS

Language bias

- more formal, technical, and pragmatic (often tied to academic or workplace discourse)
- included more collocations, idiomatic expressions, and pragmatic features
- relied more on functional language and register variation

QUALITATIVE ANALYSIS: HUMAN SETS

"There is some language bias: use of complex vocabulary (phrasal verbs), grammar and syntax..."

"Language bias appears in some items; use of the word 'engaged' in this context; words like 'protest'; 'inspirational' vs 'humorous'; 'packed'; 'notable alumni'; 'going viral'."

QUALITATIVE ANALYSIS: AI SETS

Gender bias

- appeared through role assignments and interactional dynamics
- reflected culturally specific (Western/US) gender expectations

QUALITATIVE ANALYSIS: AI SETS

"The man appeared to have more knowledge than the woman, reinforcing traditional gender biases."

"(In the questions) Voice of guilty party… young girl; victim… older man. Reversal of roles would change language and behavior."

QUALITATIVE ANALYSIS: HUMAN SETS

Gender bias

- included gender roles that are more culturally nuanced
- contained more reflexive wording, allowed for greater interpretive variance in gender representation

QUALITATIVE ANALYSIS: HUMAN SETS

"Gender roles of husband and wife in many conversations seem to be reversed; when one retires and the other continues to work. Also, woman inherits property, not "typical" items for a woman. "

QUALITATIVE ANALYSIS: AI SETS

Socio-economic bias

- reflected in consumer-driven interactions
- wealth and disposable income

QUALITATIVE ANALYSIS: AI SETS

"I don't think socio-economic bias can be completely avoided in the questions. With that said, some of the socio-economic bias was the discussion of chess. This is usually associated with middle income to higher income as well as educated populations."

QUALITATIVE ANALYSIS: HUMAN SETS

Socio-economic bias

- relied on understanding financial systems or institutional norms
- more accessible to educationally or economically privileged test takers

QUALITATIVE ANALYSIS: HUMAN SETS

"Similar to the first conversation, I believe the situations are relatable across cultures. However, due to socio-economic differences, not all students may feel comfortable having this type of discussion with their boss. The questions reflect experiences more typical of individuals from middle to upper-middle-class backgrounds, such as the concepts of working from home or taking a year off."

DISCUSSION OF KEY FINDINGS

- AI- and human-created items contained subtle cultural, linguistic, gender, and socio-economic bias.

DISCUSSION OF KEY FINDINGS

- AI- and human-created items contained subtle cultural, linguistic, gender, and socio-economic bias.

- Bias differed in framing:
  - AI sets emphasized Western, middle-class lifestyles.
  - Human sets reflected institutional language and assumed cultural/financial familiarity.

DISCUSSION OF KEY FINDINGS

Gender assumptions were perceived more in AI items, though gender bias was the least flagged bias type overall.

DISCUSSION OF KEY FINDINGS

- Gender assumptions appeared more in AI items, though gender bias was least flagged overall.
- Biases rarely appeared in isolation—intersections across multiple categories were common.

IMPLICATIONS, LIMITATIONS, AND FUTURE RESEARCH

Implications

- AI items must measure language skills, not cultural familiarity, to ensure fairness.
- Human review is essential to catch subtle and complex biases AI may miss.

Limitations and Future Research

- Examination of additional AI vs. human bias
- Comparison of AI vs. human scoring

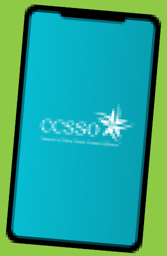
Let’s practice!

SCREENING FOR BIAS

Original prompts

Revised prompts

TAKE-HOME MESSAGES



Don’t forget to log in the mobile app to complete the session survey!

Council of Chief State School Officers

NCSA

National Conference on Student Assessment

THANK YOU

Save the Date - #NCSA2026

Austin, Texas • June 22-24, 2026