INNOVATIVE OR IMITATIVE?
EXAMINING THE CREATIVE CAPABILITIES OF AI

G-TELP · GENERAL TESTS OF ENGLISH LANGUAGE PROFICIENCY · G-TELP

G-TELP

International Testing Services Center · International Testing Services Center ·

ITSC

International Testing Services Center

Dr. Janna Schaeffer Dr. Kym Taylor

KOTESOL International
May 2025

BACKGROUND

PROJECT OVERVIEW

RATIONALE

RESEARCH CONTEXT

BACKGROUND

- **Project overview:** Exploring creativity in assessments
  - Investigates creativity in language assessments
  - Focuses on listening comprehension test items
  - Compares AI vs. human-created content from G-TELP Level 2 listening tests
  - Examines creative language use and engagement

BACKGROUND

- **Project overview:** Exploration of creativity in assessments
- **Rationale:** Creativity improves learner engagement and relevance
  - Boosts engagement and critical thinking (Lucas, Claxton, & Spencer, 2013)
  - Adds real-world relevance to tasks (Herrington, Reeves, & Oliver, 2010)
  - Supports evolving communication-focused testing (Kaufman & Reiter-Palmon, 2023)

BACKGROUND

- **Project overview:** Exploration of creativity in assessments
- **Rationale:** Creativity improves learner engagement and relevance
- **Research context:** AI's evolving role in test creation (Goksel & Bozkurt, 2023; Shermis & Hamner, 2013)

GAP IN RESEARCH

- Limited research on AI in language assessment design
- Creativity understudied in test item design (Al-Imamah & Halim, 2023; Laverghetta Jr. et al., 2024)
- Need to explore how creativity is perceived in AI-generated assessments

RESEARCH QUESTIONS

Research Questions:

Are AI-generated listening assessment questions perceived as more or less creative than human-created questions?

Do evaluations of certain factors of creativity differ for AI-generated vs. human-created listening questions?

STUDY OVERVIEW

Purpose of the study:

- To compare perceived creativity in AI- vs. human-created English listening comprehension items

What it explores:

- How experienced ELT professionals evaluate creativity in listening test questions
- Whether AI-generated questions are seen as more or less creative than human-created ones
- Which dimensions (e.g., vocabulary, engagement) influence perceptions of creativity

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

MATERIALS AND PROCEDURE

- Part of a larger data collection
  - Background questionnaire
  - AI and human-created listening sets presented in isolation and side by side (blind)
  - Multiple choice and free text questions
    - Complexity (syntactic, morphological, pragmatic)
    - Bias (cultural, linguistic, gender, socioeconomic)
    - Alignment with script content
    - Accessibility
    - Quality
    - Creativity

QUANTITATIVE ANALYSIS AND FINDINGS

Dimension	*p* value	Significance
General creativity	0.887724827	none
Vocabulary variety	0.118920453	none
Initial engagement	0.118920453	none
Maintaining interest	0.671811034	none
Balance btw. creativity and functionality	0.479887662	none

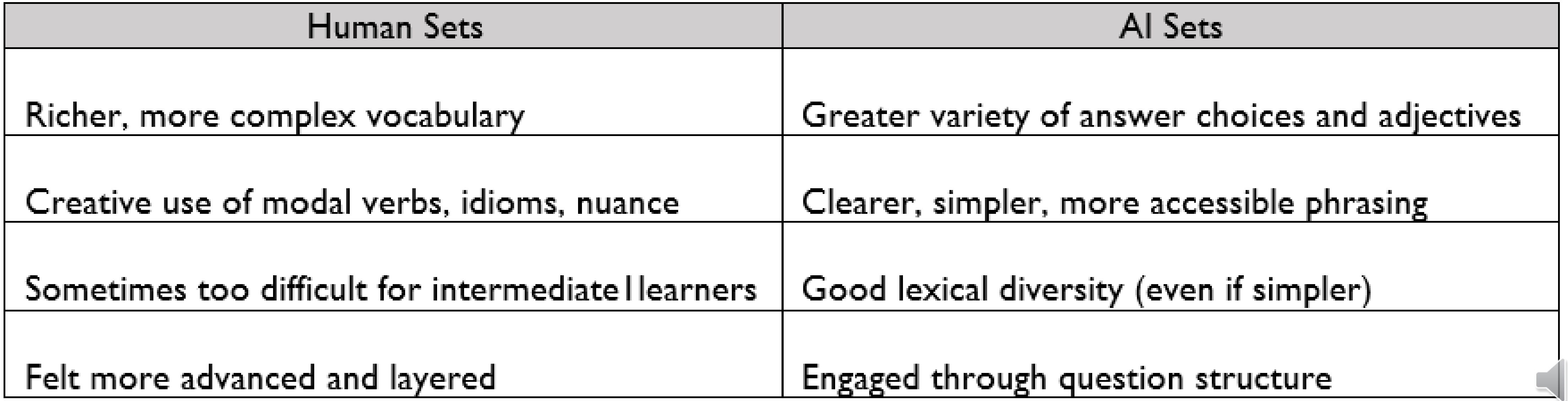
QUALITATIVE ANALYSIS

- Qualitative data were analyzed through Reflexive Thematic Analysis (Braun & Clarke, 2006)
- Of the five categories for *quantitative* analysis, three emerged as notable categories for discussion in the *qualitative* findings:

  1. Creativity in vocabulary
  2. Perceived engagement
  3. Balance between creativity and functionality

QUALITATIVE FINDINGS: Creativity in Vocabulary

Which set of questions uses a wider variety of vocabulary? How does this vocabulary variety affect interest or engagement with the text?



Human Sets	AI Sets
Richer, more complex vocabulary	Greater variety of answer choices and adjectives
Creative use of modal verbs, idioms, nuance	Clearer, simpler, more accessible phrasing
Sometimes too difficult for intermediate l learners	Good lexical diversity (even if simpler)
Felt more advanced and layered	Engaged through question structure

QUALITATIVE FINDINGS: Creativity in Vocabulary

“A less frequent vocabulary (e.g., hesitant, confident, concerned) might require a higher level of engagement.”

Re: human set

“For a native speaker, the slightly more expressive and idiomatic language seems more interesting / engaging.”

Re: human set

"Set A being shorter and more direct, less reading time is better. It holds more attention and wouldn't make the listener think as much".

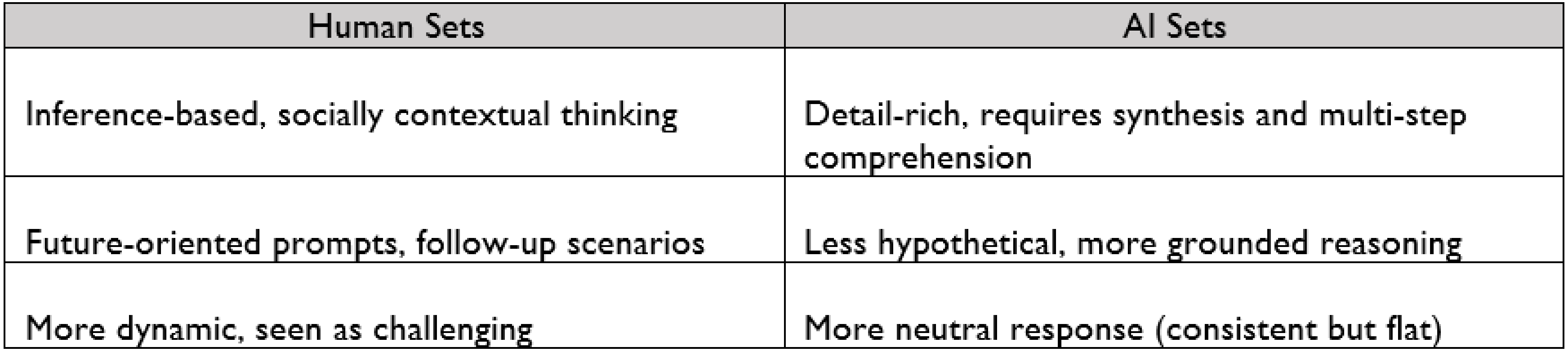
Re: AI set

QUALITATIVE FINDINGS: Creativity in Vocabulary

Finding 1: Human sets had richer lexis and used language more creatively yet seemed more challenging for B1–C1 learners; AI sets had simpler structures and were more accessible to more levels of proficiency.

QUALITATIVE FINDINGS: Perceived Engagement 1

Which aspects of the set help capture attention?



Human Sets	AI Sets
Inference-based, socially contextual thinking	Detail-rich, requires synthesis and multi-step comprehension
Future-oriented prompts, follow-up scenarios	Less hypothetical, more grounded reasoning
More dynamic, seen as challenging	More neutral response (consistent but flat)

QUALITATIVE FINDINGS: Perceived Engagement 1

“The answers are deeper in that they allude to more information. They would inspire more follow-up questions, maintaining interest in the further story.”

Re: human set

“For me, the well-crafted complex sentences in Set A have a rhythm and flow that feel poetic or sophisticated, making the language more enjoyable to read.”

Re: human set

QUALITATIVE FINDINGS: Perceived Engagement 1

"Set A grabs my attention at first because the questions are "easier" - more straightforward."

Re: AI set

"Set B holds my attention because I have to work harder to understand and answer them."

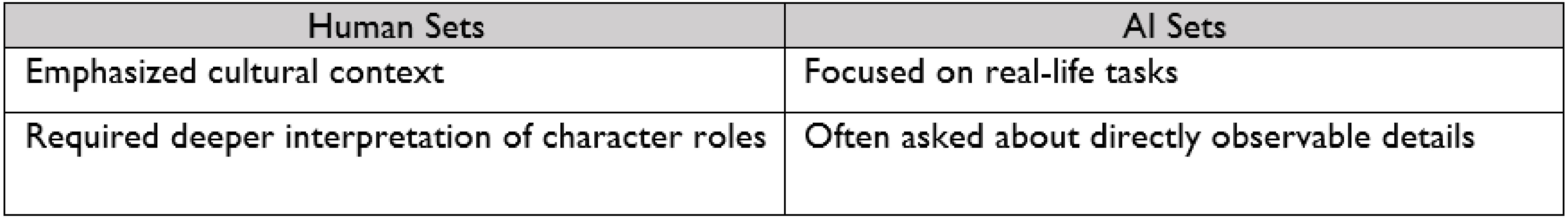
Re: human set

QUALITATIVE FINDINGS: Perceived Engagement 1

Finding 2a: Human sets encouraged imaginative, contextual thinking and felt more dynamic, while AI sets offered clear, structured tasks that required detailed comprehension but felt more predictable.

QUALITATIVE FINDINGS: Perceived Engagement 2

Which aspects of the set help hold attention?



Human Sets	AI Sets
Emphasized cultural context	Focused on real-life tasks
Required deeper interpretation of character roles	Often asked about directly observable details

QUALITATIVE FINDINGS: Perceived Engagement 2

"This set of questions provides practical insights on what to do in the event of an accident, a scenario that mirrors real-life situations. The questions encourage problem-solving and prompt the listener to think more critically and analyze the situation in a deeper way."

Re:AI set

QUALITATIVE FINDINGS: Perceived Engagement 2

"Set B would require one to pay closer attention - to avoid the trap of just listening to hear a specific phrase being said vs. listening for the CORRECT phrase being said."

Re: human set

"The second dialogue contains a more complex interplay of language, idioms, cultural references and - for an advanced student - this is what you look for when trying to keep improving your proficiency."

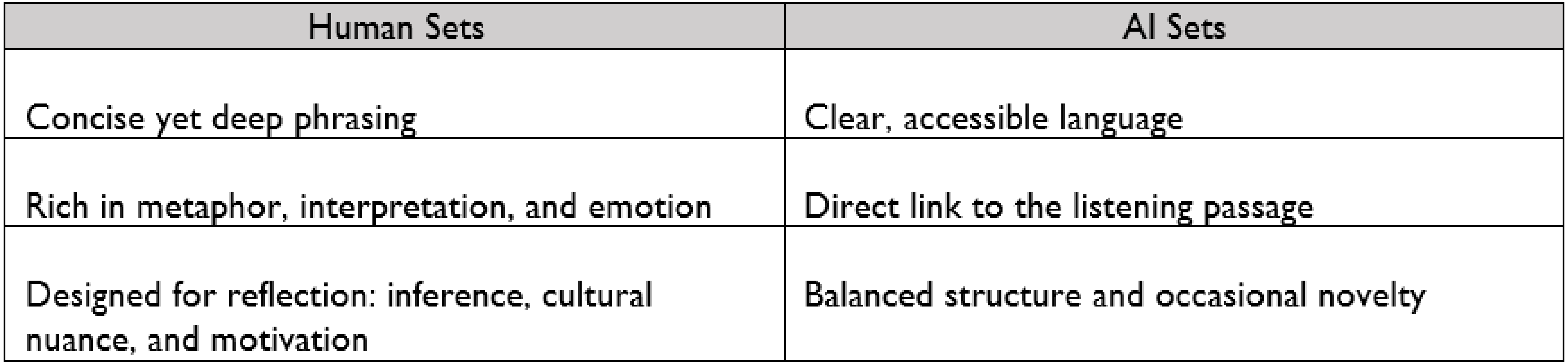
Re: human set

QUALITATIVE FINDINGS: Perceived Engagement 2

Finding 2b: AI-generated questions were seen as more direct, while human-created items offered deeper cultural and personal engagement.

QUALITATIVE FINDINGS: Creativity and Functionality

Which set of questions better balances creativity and functionality?



Human Sets	AI Sets
Concise yet deep phrasing	Clear, accessible language
Rich in metaphor, interpretation, and emotion	Direct link to the listening passage
Designed for reflection: inference, cultural nuance, and motivation	Balanced structure and occasional novelty

QUALITATIVE FINDINGS: Creativity and Functionality

“Set B feels more formal, like a formal test or assessment. Language is more dry / formal / academic.”

Re: AI set

“Set A feels more conversational and interesting, like I'm talking about it with a friend.”

Re: human set

QUALITATIVE FINDINGS: Creativity and Functionality

"Set A questions require the listener to think a little more about the content of the listening exercise whereas..."

Re: human set

"Set B questions use language taken from the exercise."

Re: AI set

QUALITATIVE FINDINGS: Creativity and Functionality

Finding 3: AI sets prioritized clarity, structure, and a direct link to the listening text, while human sets were more reflective and interpretive, using concise yet emotionally rich language.

DISCUSSION AND CONCLUSION

Summary of key findings:

- Human-created sets were perceived as having richer, more creative language but were thought to be potentially more challenging for mid-level learners.
- Participants perceived human-created content to encourage imagination in terms of contextual and cultural engagement.
- AI sets were perceived to be clearer and more structured, making them more accessible and predictable.
- AI items were evaluated as having greater focus on direct comprehension, while human items invited deeper reflection.

DISCUSSION AND CONCLUSION, continued

- For the listening question sets in this study, AI and human content was indistinguishable by quantitative measures (aligns with Köbis & Mossink, 2021).
- Participants trended toward slightly favoring human-generated content in their qualitative judgments (aligns with McCormack, & d'Inverno, 2014).
- We can see that creativity matters (aligns with Thomas, 2021).

IMPLICATIONS

- Emphasis on balanced item design
- ‘Human in the loop’ as essential in test design
- Awareness of strengths / weaknesses for research
- Need for more creativity-focused studies
- Exploration of hybrid models of assessment

LIMITATIONS AND FUTURE DIRECTIONS

Limitations

- Number of participants
- Number of question set examples

Future directions

- Comparison of additional AI vs. human content creation
- Comparison of AI vs. human scoring

THANK YOU!