Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Leveraging AI for Automatic Item Generation for Psychological Scales

Xijuan Zhang,* Kentaro Suzuki, and Kai Wen Zhou

York University, Psychology Department, Toronto, Canada
*Corresponding author. Email: xijuan@yorku.ca

Abstract

This study examined how large language models (LLMs) can be used for automated item generation (AIG) in psychological scale development, focusing on how prompting strategies and LLM brands and versions influence item quality for both an established and a novel construct. Using three prompting conditions across eight LLMs from four major providers, we generated AI items and evaluated them through expert review, AI-psychometric analyses, and lexical diversity metrics. The findings highlight both the potential and limitations of AIG. Many AI-generated items were clear, concise, and theoretically meaningful, and for the novel construct, some items captured nuanced facets beyond the researcher's initial conceptualization. At the same time, other items were overly general or tapped unintended dimensions, underscoring the need for careful human screening. Prompt engineering was the strongest determinant of item quality, requiring deliberate design to avoid reproducing existing scale items and to achieve the desired balance of specificity and conciseness. Overall, AIG appears to be a useful tool for early-stage scale development when paired with thoughtful prompt design and human oversight.

Keywords: automatic item generation, artificial intelligence, large language model, psychometrics

Developing psychological scales has traditionally been labour-intensive and time-consuming. Researchers generate large pools of potential items guided by recommended item-writing principles (e.g., simplicity, clarity, avoiding double-barrelled wording), followed by expert review and extensive validation. Recent advances in Artificial Intelligence (AI), especially Large Language Models (LLMs), offer a faster, scalable, and cost-effective alternative. Using tools like ChatGPT, researchers can automatically generate psychological items—a process known as automatic item generation (AIG).

In recent years, several studies have explored the use of AIG in developing psychological scales (e.g., Götz et al., 2023; Hommel et al., 2022; Lee et al., 2023, 2025; Russell-Lasalandra et al., 2024). However, these studies have several limitations. First, most research to date has focused on generating items for well-established psychological constructs, such as the Big Five personality traits (e.g., Hommel et al., 2022; Lee et al., 2023; Russell-Lasalandra et al., 2024). It remains unclear whether AIG is equally effective for developing items that capture novel psychological constructs. Second, many earlier studies relied on older versions of GPT-based LLMs that lacked modern chat-based interfaces (e.g., Götz et al., 2023; Russell-Lasalandra et al., 2024). More recently, commercially available chat-based platforms, powered not only by GPT but also by other LLMs such as DeepSeek, Claude, and Geminil, have become increasingly popular. Researchers typically interact with these LLMs through chat interfaces rather than direct API access. This distinction is crucial because chatboxes use predefined system prompts that shape the model's behaviour, tone, and output, meaning results from direct API access may not generalize. Finally, although some studies presented the prompts used to generate items (e.g., Lee et al., 2023; Russell-Lasalandra et al., 2024), none have systematically

compared different prompting strategies. Yet, prompt engineering has become a critical component of effectively using modern LLM chatboxes.

The present research addresses these limitations. We examine AIG for both a well-established psychological construct (Global Self-Esteem) and a new construct under development in a social cognition laboratory. We include three prompting strategies that vary in the level of detail provided to the LLM chatboxes and use several of the most recent and popular LLM chatboxes. In this article, we use AI-powered psychometric analytical techniques to compare AI-generated items across different prompting and LLM conditions, as well as against human-generated items. This work is complemented by ongoing data collection, which will allow for a future evaluation of the AI items' key psychometric properties, including factor structure, reliability, and validity.

1. Methods

1.1 Psychological Constructs

As noted earlier, we examined AIG for both a well-established and a novel psychological construct. The well-established construct in this study is global self-esteem, defined as a person's overall sense of self-worth. The most widely used measure of global self-esteem is the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1989), consisting of five positively worded and five negatively worded items. In this study, we used the RSES as the benchmark for evaluating the AI-generated items.

The novel construct is related to an existing construct known as cognitive fluency or processing fluency, defined as the experienced ease of mental processing, characterized by fast, effortless, and accurate cognitive operations (e.g., Alter & Oppenheimer, 2009; Winkielman et al., 2003). One of our co-authors is currently developing a facet of processing fluency that captures an individual's preference for immediate, tangible results and discomfort with delays or unresolved situations. They refer to this facet as *resolution*, as it pertains to how quickly an event resolves itself in the human mind so that it no longer occupies cognitive space. Their lab created an initial item pool through a focus group, collected data, and conducted an exploratory factor analysis (EFA), ultimately retaining three items to measure resolution (see the example items in Table 1). Due in part to the small item count, the scale showed moderate reliability ($\alpha = 0.69$). By leveraging AIG, we aim to lengthen the scale, using the three existing items as the benchmark for comparison.

1.2 Prompt Engineering

For each construct, we designed three prompt conditions—zero-shot, few-shot-basic, and few-shot-advanced—which progressively increased the level of detail provided to the LLM (see Table 1 for complete prompts). The zero-shot prompts offered only minimal instructions: for the global self-esteem construct, a single sentence requesting a balanced scale; for the resolution construct, the same plus a brief definition of the construct. The few-shot-basic condition introduced an LLM persona and provided example items. For global self-esteem, we included examples from the RSES but instructed the model not to replicate them; for resolution, we provided the three human-generated items mentioned earlier. The few-shot-advanced condition built on this by incorporating widely recommended item-writing guidelines for writing, commonly discussed in research methods textbooks and articles (e.g., Clark & Watson, 2016; Cozby et al., 2024; Kyriazos & Stalikas, 2018). For the resolution construct, the few-shot-advanced prompt also included an expanded explanation of the construct.

To ensure comparability across conditions, prompts for the global self-esteem construct requested five positively- and five negatively-worded items, mirroring the RSES. For the resolution construct, prompts requested three items, matching the number of human-generated items.

1.3 Large Language Models (LLMs)

We used LLMs from four major brands: ChatGPT, Gemini, Claude, and DeepSeek. For each brand, we selected two models: one optimized for speed (referred to as *Fast*) and a slower counterpart designed for enhanced reasoning (referred to as *Thinking*). Table 2 provides the specific version for each LLM. For each LLM and prompt condition, we conducted three independent replications of AIG to assess consistency across iterations. To prevent cross-iteration influence, we disabled memory settings in ChatGPT, Gemini, and Claude (this option was unavailable for DeepSeek).

1.4 Analytical Methods

After obtaining the AI-generated items, each co-author reviewed them for quality and clarity. We then conducted three types of statistical analyses to compare the items. The first two analyses rely on text embeddings in LLM processing, which represent text data as numerical vectors that capture semantic relationships between words or sentences. In our study, we used OpenAI's text-embedding-3-large model to convert each psychological item into a high-dimensional vector. Applying embeddings to evaluate psychological scales falls within AI-powered psychometrics, which is an emerging area that integrates LLM-based methods into psychometric analysis (Guenole, 2025).

Cosine Similarity Using the text-embeddings, we first examined semantic similarity among human- and AI-generated items by computing cosine similarity between item vectors. Cosine similarity ranges from 0 (no semantic relation) to 1 (identical meaning) and serves as a correlation-like index for comparing textual content.

Principal Component Analyses (PCA) The second embedding-based analysis involved conducting principal component analysis (PCA) on the cosine similarity matrix, which, according to Milano et al. (2025), allows us to examine semantic clusters among items, similar to factor analysis on the covariance matrix between variables. [1] We applied PCA only to the global self-esteem construct because the resolution construct contained only three items. The RSES often yields a two-factor structure with the exploratory factor analysis, with positively worded and negatively worded items loading on separate factors (e.g., Hensley & Roberts, 1976; Zhang & Savalei, 2016). Accordingly, we hypothesized that PCA on the cosine similarity matrix of the self-esteem items would similarly reveal two clusters corresponding to item polarity.

Regression Analyses for Lexical Diversity Finally, we examined lexical diversity using word count and Corrected Type-Token Ratio (CTTR), which is the number of unique words divided by the total number of words, corrected for text length (Carroll, 1964). Higher word counts and CTTR values may reflect unnecessarily complex wording that violates recommended item-writing guidelines. Using these measures, we ran regression models predicting the word count and CTTR of AI-generated items using the prompt condition, LLM brand, model type, and iteration number as predictors, and used pairwise Tukey comparisons to analyze significant predictors and compare AI- and human-generated items.

2. Results

All AI- and human-generated items, analytic code, and supplementary results materials are publicly available in our Open Science Framework (OSF) project folder:https://osf.io/7ek2u/overview?view_only=e4b259fa44e94b17a5feb6577595a21c.

2.1 Expert Review of AI-Generated Items

As noted earlier, each co-author reviewed the AI-generated items, and several patterns emerged. First, item quality varied widely. Within each prompt and LLM condition, there were high-quality items characterized by clear writing and strong face validity (see Table 3 for examples). For the novel construct of resolution, the co-author leading its development noted that some AI-generated items captured some nuanced aspects of the construct, such as the unsettled state of mind while awaiting outcomes, that are worthy of further exploration. At the same time, there were problematic items across conditions (see Table 4 for examples). Minor problems included being too lengthy or using unusual symbols, such as emoticons. More serious problems included using extreme wording that respondents would struggle to endorse, or tapping into unintended constructs. Because resolution is a novel construct with a small item pool, we might expect LLMs to struggle generating items that tap into the intended construct. However, the emergence of unintended constructs in the well-established global self-esteem items indicates a broader trend (see Table 4): regardless of the construct's maturity, LLMs tend to generate content that is semantically related but conceptually adjacent to the intended target.

Second, the prompt condition had a substantial effect on the items. For instance, because the few-shot-advanced prompt instructed the LLM to avoid complex phrasing, the items were considerably shorter than those produced under the few-shot-basic and zero-shot prompts. Furthermore, prompt design also affected whether the LLM reproduced the existing items. For the global self-esteem, the zero-shot prompt did not warn against copying existing items, while the few-shot-basic and few-shot-advanced did (see Table 1). As a result, the zero-shot prompt frequently yielded verbatim RSES items, particularly when using the fast LLM variants, while items generated under the other prompt conditions consistently differed from the originals. With DeepSeek, nearly all zero-shot global self-esteem items were identical to the original RSES items across both the fast and thinking models. Although this results pattern is understandable, it is nonetheless striking that, in the absence of explicit guidance, LLMs reproduced many existing items verbatim.

2.2 Cosine Similarity

Figures 1 and 2 show the visualization for selected cosine similarity matrices for the global self-esteem and resolution constructs, respectively (see Supplementary Materials for complete results). Several noticeable patterns of results emerged from the cosine similarity matrices.

First, consistent with the expert review, for the global self-esteem construct, under the zero-shot prompt, many AI-generated items show a cosine similarity of one with the original RSES items, indicating they were identical (see Figure 1). ChatGPT's thinking model performed better compared to those from other brands. Particularly, for ChatGPT, under the zero-shot prompt, although the fast model generated many identical RSES items, the thinking model did not generate any identical RSES items. On the other hand, for DeepSeek, Claude, and Gemini, both the fast and thinking models generated many verbatim RSES items. Notably, DeepSeek's fast model generated the RSES items in the same order across all three iterations; its thinking model also generated the RSES items verbatim, but with different orders across iterations.

These aforementioned patterns of results for the global self-esteem construct stand in contrast to those for the resolution construct. Because resolution is a novel construct and the human-generated items have not been published online, the zero-shot prompt did not lead LLMs to reproduce the human-generated items (see Figure 2).

Another set of noticeable patterns of results concerns the differences between the items generated using the few-shot-basic versus the few-shot-advanced prompts. For global self-esteem, the few-shot-advanced prompt generated items that were more similar to the original RSES items and more consistent with one another than those generated by the few-shot-basic prompt. As shown in Figure 1, the cosine similarity matrices for the few-shot-advanced conditions appear lighter in colour

compared to those for the few-shot-basic conditions, indicating higher cosine similarities under the few-shot-advanced prompt and lower similarities under the few-shot-basic prompt. This pattern of results can be attributed to the prompt design: the few-shot-advanced prompt includes more detailed instructions about item-writing style, which constrains variation both within and across iterations.

For the resolution construct, the few-shot-advanced prompt similarly yielded items that were more consistent with one another than those generated using the few-shot-basic prompt. However, unlike the global self-esteem results, the resolution items generated under the few-shot-advanced prompt were less similar to the human-generated items compared to those under the few-shot-basic prompt (see Figure 2). One potential explanation is that the human-generated resolution items are relatively long, which is inconsistent with the item-writing guidelines emphasized in the few-shot-advanced prompt.

Finally, another noteworthy pattern is that, for the global self-esteem construct, items with the same wording polarity were more similar to one another than to items with the opposite polarity. This produced a distinctive checkerboard colouring pattern in Figure 1 because items within each iteration were ordered such that the first five were positively-worded and the last five were negatively-worded, resulting in five-by-five blocks of darker red for within-polarity similarities followed by lighter-red blocks for between-polarity similarities.

2.3 PCA

Figure 3 presents the PCA biplot for the original RSES items and AI-generated global self-esteem items from selected conditions (see Supplementary Materials for complete results). Overall, PCA results were very consistent between human- and AI-generated items. For both human and AI-generated items, the first component accounted for 30-40% of the variance, whereas the second component accounted for around 15-25% of the variance. The first component clearly differentiated positively worded items from negatively worded items, with the former loading positively and the latter loading negatively. This result is consistent with the pattern observed in the cosine similarities, where items with the same polarity had higher cosine similarities than those with opposite polarities.

2.4 Regression Analyses for Lexical Diversity

Figure 4 presents the word count and CTTR for the human- and AI-generated items for both constructs across all conditions. Table 5 presents the results from the regression analyses examining how prompt condition, iteration number, fast versus thinking model versions, and LLM brand predict the word count and CTTR of AI-generated items. Table 6 summarizes the pairwise comparisons among AI-generated items and between human- and AI-generated items.

2.4.1 Comparing Among AI-Generated Items

The regression results indicate that for AI-generated items, the prompt condition was the strongest predictor of lexical diversity, as measured by word count and CTTR (see Table 5). For both the global self-esteem and resolution constructs, significant differences in word count and CTTR emerged between the zero-shot and few-shot-advanced conditions and between the few-shot-basic and few-shot-advanced conditions (see Table 6). The items generated using the few-shot-advanced prompt had significantly lower word counts and lower CTTR values than those generated using either the zero-shot or few-shot-basic prompts. More specifically, controlling for all other variables, for the global self-esteem construct, the few-shot-advanced items were on average 2.57 words shorter than the few-shot-basic items; for the resolution construct, they were 6.56 items shorter (see the first panel in Table 6). This pattern aligns with the expert review findings described above and, as explained above, likely reflects the additional instructions in the few-shot-advanced prompt that emphasized concise language, thereby reducing lexical diversity and better following the recommended item-writing guidelines.

Table 1. Prompt Conditions For Each Construct

	Global Self-Esteem Construct	Resolution Construct
Zero-Shot Prompt	Generate 10 self-report items, half positively-worded and half negatively-worded, for assessing the global self-esteem construct using a 5-point Likert scale ranging from 1= Strongly Disagree, 5=Strongly Agree.	Generate 3 self-report items for assessing the resolution facet of the cognitive fluency construct using a 7-point Likert scale ranging from 1=Strongly Disagree to 7=Strongly Agree. The resolution facet captures an individual's preference for immediate, tangible results and discomfort with delays or unresolved scenarios.
Few-Shot-Basic Prompt	You are an expert psychometrician who is developing a psychological scale for assessing the global self-esteem construct. Below are four example items. Each item is designed using a 5-point Likert scale ranging from 1=Strongly Disagree to 5=Strongly Agree. Do not duplicate or closely paraphrase any items from the Rosenberg Self-Esteem Scale (RSES). After the examples, generate 10 original self-report items, half positively-worded and half negatively-worded, for assessing the global self-esteem construct. Example Items: -I feel that I have a number of good qualities. (positively-worded) -I feel inferior to most people around me. (negatively-worded) -I feel I am deserving of respect from others. (positively-worded) -All in all, I am inclined to think that I am a failure. (negatively-worded)	You are an expert psychometrician who is developing a psychological scale for assessing a facet of the cognitive fluency construct, called "resolution". The resolution facet captures an individual's preference for immediate, tangible results and discomfort with delays or unresolved scenarios. Below are three example items for assessing the resolution facet. Each item is designed using a 7-point Likert scale ranging from 1=Strongly Disagree to 7=Strongly Disagree. Do not duplicate or closely paraphrase any existing resolution or decision-making scale items. After the examples, generate 3 new, original items for assessing the resolution facet. Example Items: -I prefer to see my actions lead to immediate responses instead of being delayed. -I enjoy tasks that provide immediate, tangible results. -I feel more in control when I see the direct outcomes of my choices.
Few-Shot-Advanced Prompt	You are an expert psychometrician who is developing a high-quality, self-report psychological scale in the Likert format for assessing the global self-esteem construct. Follow all instructions below carefully to ensure clarity, conceptual precision, and suitability for psychometric analysis. ## Conceptual Focus: -Global self-esteem refers to an individual's overall evaluation of their own worth and value as a person. -Items should tap into broad, trait-level feelings of self-worth, self-acceptance, and self-respect. -Avoid referencing specific abilities, behaviors, or situations; instead, focus on general feelings and beliefs about oneself. ##Writing Style: - Write each item as a first-person, present-tense statement (e.g., "I feel…," "I believe…"). - Use simple, concise, everyday language; avoid technical jargon or multiclause sentences. - Avoid complex phrasing and double negatives. - Avoid double-barreled items; that is, avoid asking two or more distinct issues within a single item. - Avoid words or phrases that could be interpreted differently across cultures (e.g., "having it all together," "feeling like a rock star"). - Keep each item concise (no more than 15 words). ##Response Format: -Items will be rated on a 5-point Likert scale: 1 = Strongly Disagree 2 = Disagree 3 = Neither Disagree nor Agree 4 = Agree 5 = Strongly Agree ##Avoiding Redundancy and Copying: - Do not duplicate or closely paraphrase any items from existing scales (e.g., Rosenberg Self-Esteem Scale). - Ensure each item covers a slightly different nuance of global self-esteem (e.g., general self-acceptance, confidence in intrinsic worth, absence of self-criticism). ##Examples Items: -I feel that I have a number of good qualities. (positively-worded) -I feel inferior to most people around me. (negatively-worded) -I feel I am deserving of respect from others. (positively-worded) -All in all, I am inclined to think that I am a failure. (negatively-worded) ##Generation Task: - After reviewing the examples, generate 10 new, original items that adhere to the guidelines above. - Ensure exactly half of the items are positively-worded and half are negatively-worded. - Verify that no two items are semantically redundant.	You are an expert psychometrician who is developing a high-quality, self-report psychological scale in the Likert format for assessing a facet of the cognitive fluency construct, called the "resolution" facet, which pertains to how immediately an event resolves itself in the human mind so it does not occupy cognitive spaces. Follow all instructions below carefully to ensure clarity, conceptual precision, and suitability for psychometric analysis. ##Conceptual Focus: -The resolution facet captures an individual's preference for immediate, tangible results and discomfort with delays or unresolved scenarios. - The resolution items should tap into how much someone feels discomfort with delays or unresolved situations and how strongly they seek clarity and closure. - Individuals high in the resolution facet value clarity and closure. They prefer immediate feedback or responses to the tasks or actions they conducted. They also feel uneasy when tasks remain open-ended. - Avoid referencing specific tasks (e.g., "I finish my work tasks quickly"); instead, focus on the general cognitive-style preference for swift resolution and mental closure. ##Writing Style: - Write each item as a first-person, present-tense statement (e.g., "I feel…," "I believe…"). - Use simple, concise, everyday language; avoid technical jargon or multiclause sentences. - Avoid complex phrasing and double negatives. - Avoid double-barreled items; that is, avoid asking two or more distinct issues within a single item. - Avoid words or phrases that could be interpreted differently across cultures (e.g., "having it all together," "feeling like a rock star"). - Keep each item concise (no more than 15 words). ##Response Format: - Items will be rated on a 7-point Likert scale: 1 = Strongly Disagree 2 = Disagree 3 = Slightly Disagree 4 = Neither Agree nor Disagree 5 = Slightly Agree 6 = Agree 7 = Strongly Agree ##Avoiding Redundancy and Copying: - Do not duplicate or closely paraphrase any existing Cognitive Fluency or related scale items. - Ensure each item covers a different nuance of Resolution (e.g., emotional discomfort vs. cognitive impatience vs. preference for closure). ##Example Items: - I prefer to see my actions lead to immediate responses instead of being delayed. - I enjoy tasks that provide immediate, tangible results. - I feel more in control when I see the direct outcomes of my choices. ##Generation Task: - After reviewing the example items, generate 3 new, original items that adhere to the guidelines above. - Verify that no two items are semantically redundant.

Table 2. Large Language Models (LLMs) Used in the Study

Brand	Fast Version	Thinking Version
ChatGPT	5.0 (Instant Mode)	5.0 (Thinking Mode)
Gemini	2.5 (Flash)	2.5 (Pro)
Claude	Haiku 4.5 (fastest)	Sonnet 4.5 (smartest)
DeepSeek	V3 without DeepThink mode	V3 with DeepThink mode

Table 3. Examples of High-Quality AI-Generated Items

Condition	Item
Construct: Global Self-Esteem	
Prompt: Few-Shot-Advanced Machine: Claude (Fast)	I tend to be critical and disapproving of myself. (Negatively-Worded)
Prompt: Zero-Shot Machine: Gemini (Fast)	I am able to do things as well as most other people. (Positively-Worded)
Prompt: Few-Shot-Advanced Machine: DeepSeek (Thinking)	I tend to focus on my flaws. (Negatively-Worded)
Prompt: Few-Shot-Basic Machine: ChatGPT (Thinking)	I believe I deserve to be treated well. (Positively-Worded)
Construct: Resolution	
Prompt: Few-Shot-Advanced Machine: Claude (Thinking)	Unresolved matters occupy my thoughts until they are settled.
Prompt: Few-Shot-Basic Machine: Claude (Thinking)	If a result is pending, it's hard for me to concentrate until there's something concrete to show.
Prompt: Zero-Shot Machine: ChatGPT (Fast)	I prefer situations where I can see immediate results from my actions.
Prompt: Few-Shot-Advanced Machine: Gemini (Thinking)	My mind feels settled once an outcome is clear.

Table 4. Examples of Problematic AI-Generated Items

Condition	Item	Problems
Construct: Global Self-Esteem		
Prompt: Few-Shot-Basic Machine: ChatGPT (Thinking)	I regard myself as someone who merits kindness and care.	- Taps into other constructs, such as self-compassion.
Prompt: Zero-Shot Machine: Gemini (Thinking)	On the whole, I am satisfied with myself. 😊	- Use of emoticon.
Prompt: Few-Shot-Advanced Machine: DeepSeek (Thinking)	My life lacks purpose and meaning.	- Too general. - Taps into other constructs related to the meaning of life.
Construct: Resolution		
Prompt: Few-Shot-Basic Machine: Claude (Fast)	I strongly prefer situations where I can quickly determine whether my efforts have succeeded or failed, rather than waiting for delayed feedback.	- Too long. - Taps into other constructs related to the worthiness of effort.
Prompt: Zero-Shot Machine: Gemini (Thinking)	I would rather complete several small projects with quick turnarounds than one large project that takes a long time to finish.	- adds additional context which researchers may or may not want.
Prompt: Zero-Shot Machine: Gemini (Fast)	I strongly dislike not knowing what will happen next.	- extreme wording, making it very difficult to endorse.
Prompt: Zero-Shot Machine: ChatGPT (Thinking)	I'm comfortable leaving issues open-ended for a while if resolution takes time.	- negatively-worded even though the prompt did not request negatively-worded items.

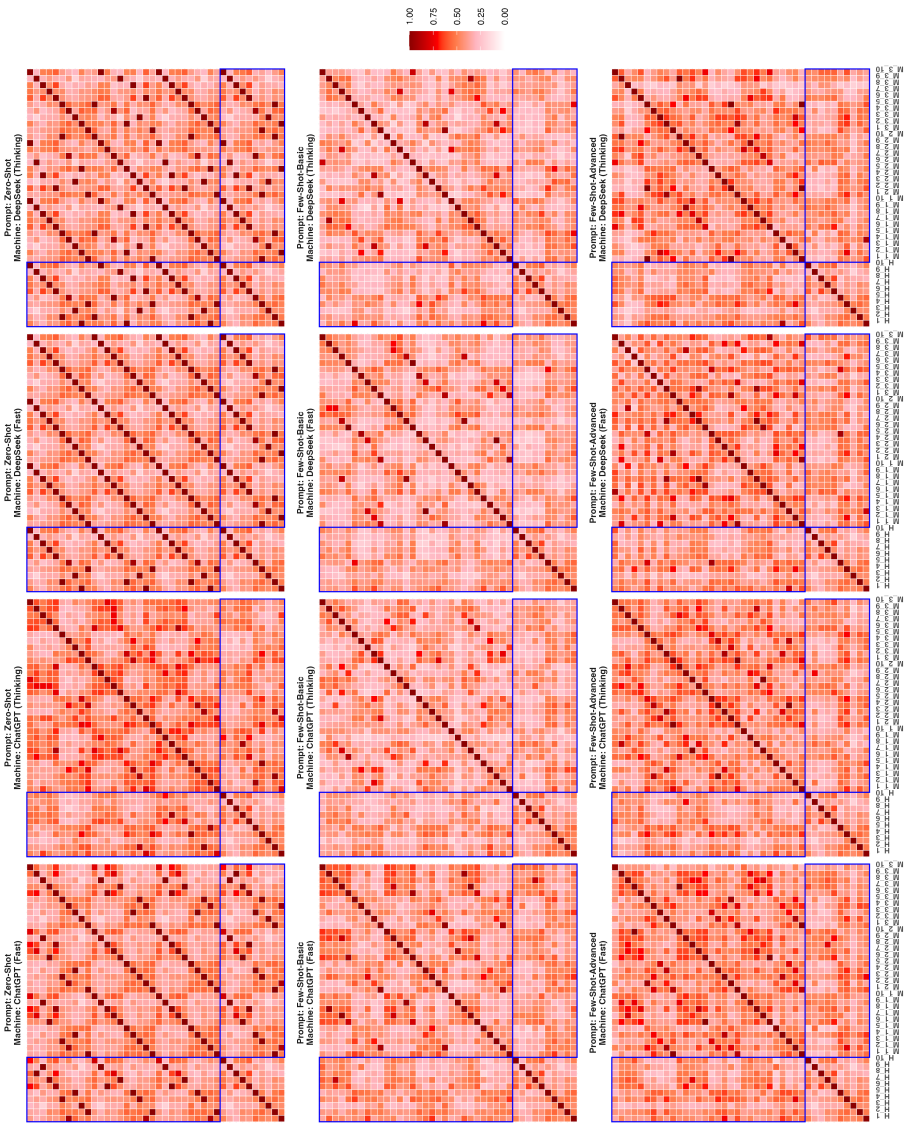

Figure 1. Cosine similarity matrices for the global self-esteem items generated by ChatGPT and DeepSeek LLMs. H stands for human-generated items (i.e., RSES items). H_1 stands for item 1 from the RSES. M stands for machine- or AI-generated items. M_1_1 refers to AI-generated item 1 from iteration 1, where the first number indicates the iteration number and the second number indicates the item number. Items 1–5 are positively-worded items; items 6–10 are negatively-worded items. The blue box highlights the cosine similarities between AI- and human-generated items in each condition.

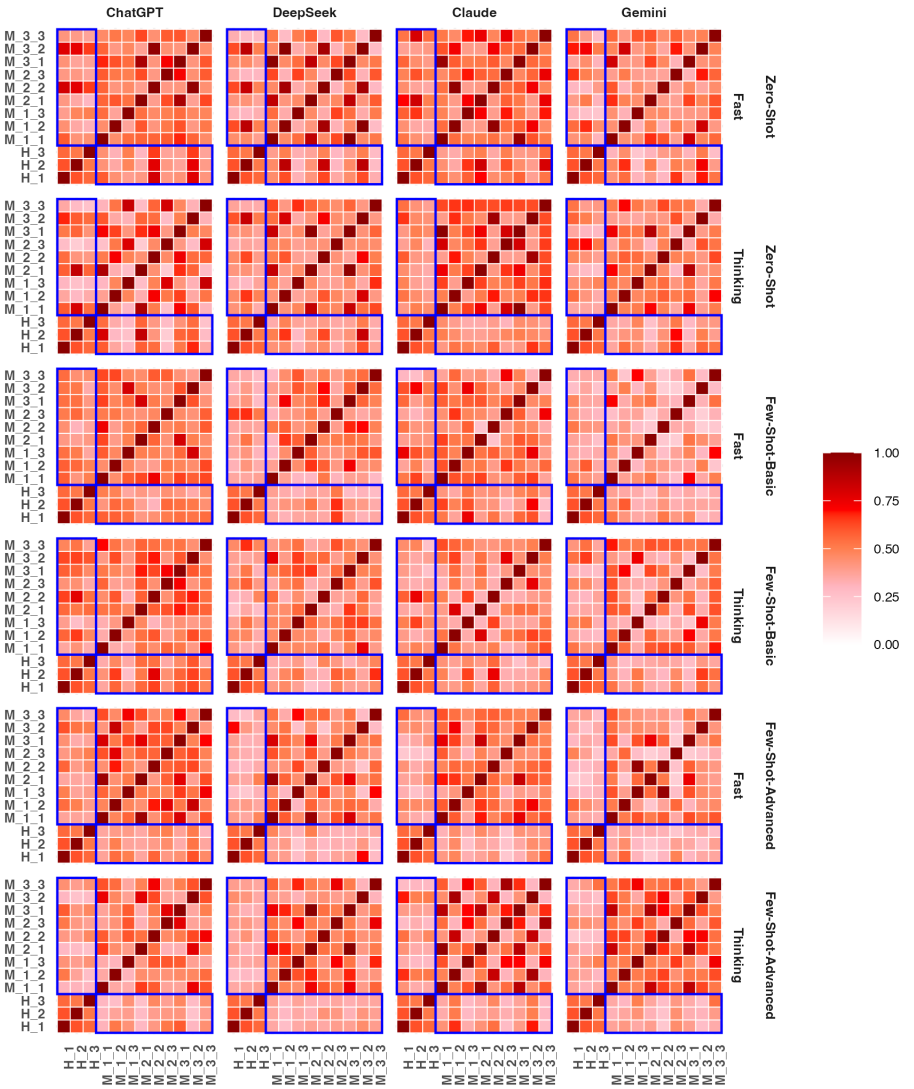

Figure 2. Cosine similarity matrices for the resolution construct items. H stands for human-generated items. H_1 stands for item 1 of the human-generated items. M stands for machine- or AI-generated items. M_1_1 refers to AI-generated item 1 from iteration 1, where the first number indicates the iteration number and the second number indicates the item number. The blue box highlights the cosine similarities between AI- and human-generated items in each condition.

However, the discrepancy between the few-shot-basic and zero-shot conditions varied by construct. For the global self-esteem construct, controlling for all other variables, the few-shot-basic items had significantly higher word counts and CTTR values than the zero-shot items (see the first panel in Table 6). In contrast, for the resolution construct, there were no significant differences in

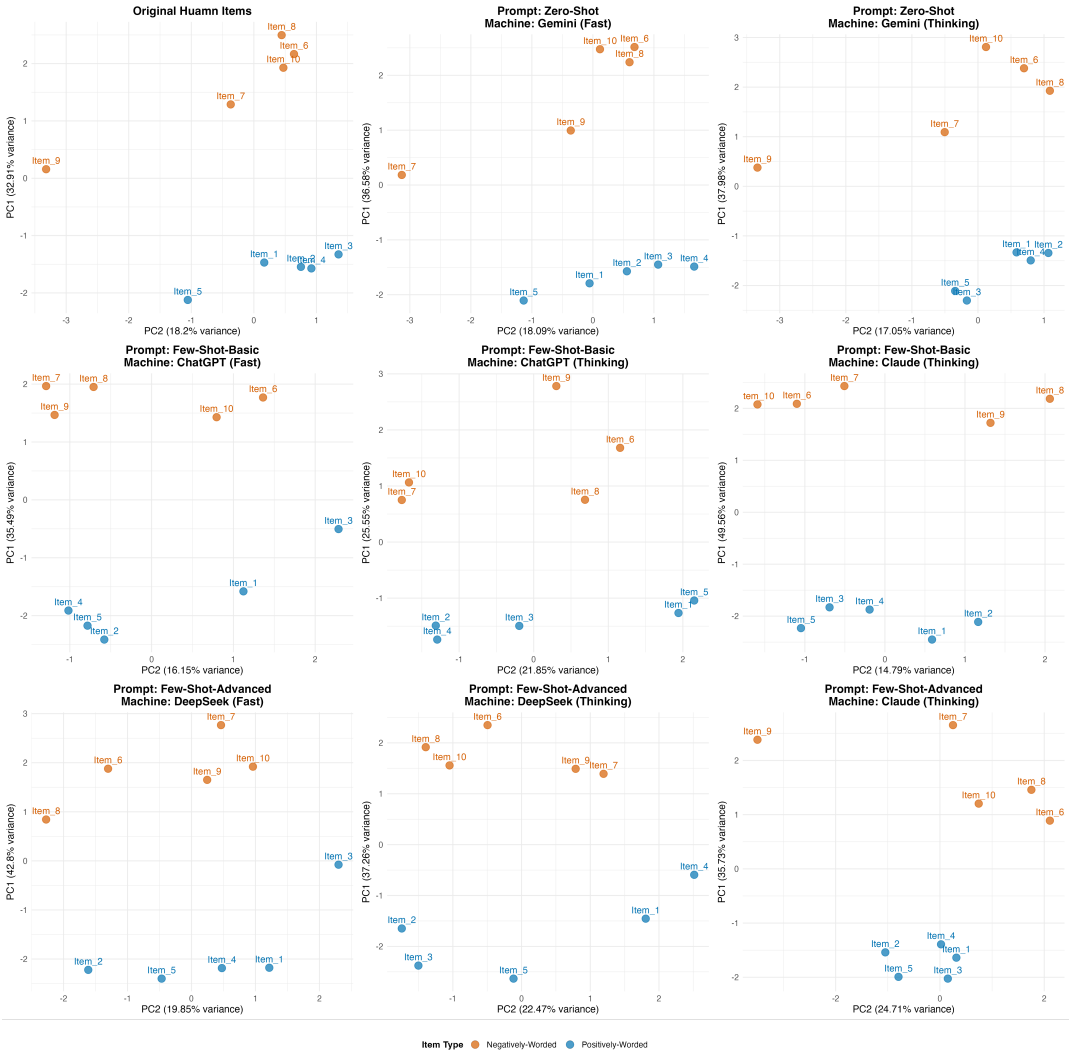

Figure 3. PCA Biplots for the Global Self-Esteem Construct. The first biplot is for the Original RSES items. The other biplots are for AI-generated items from selected conditions.

word count or CTTR between the few-shot-basic and zero-shot items. One possible explanation for this discrepancy is that, for the global self-esteem construct, the zero-shot prompt often produced items identical to the existing RSES items, which are relatively simple and concise, whereas for the resolution construct, no such existing items were available online, preventing the LLMs from copying existing items.

Somewhat unexpectedly, variables other than the prompt condition, such as LLM brand and model version, had relatively small effects on lexical diversity. For the global self-esteem construct, controlling for other variables, the LLM version had a small but significant effect, with the thinking version producing slightly lower word counts and CTTR values, suggesting that the thinking mode may improve item conciseness (see Table 5). Additionally, controlling for other variables, DeepSeek generated self-esteem items with slightly but significantly higher word counts and CTTR values than ChatGPT and Gemini, indicating that DeepSeek tended to produce somewhat wordier items

(see the third panel in Table 6).

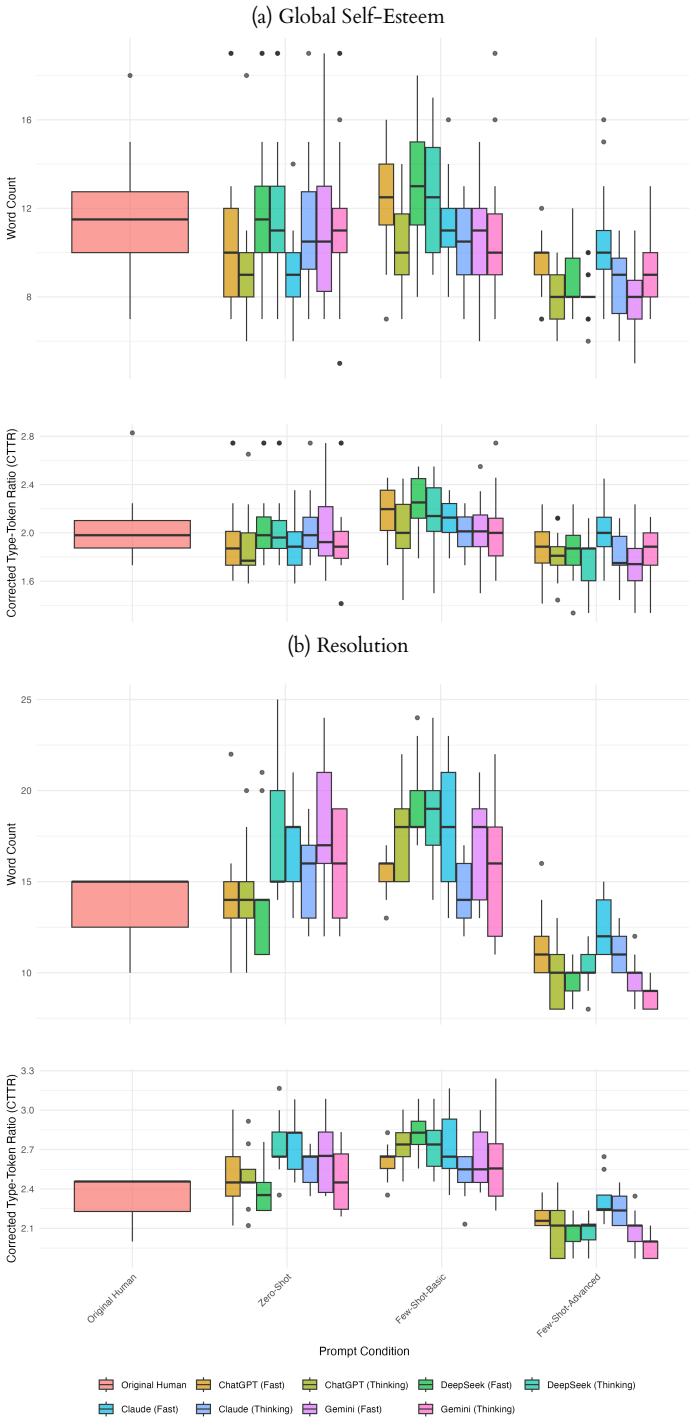

Figure 4. Word Count and CTTR For the Global Self-Esteem and Resolution Items Across Conditions

Table 5. Results for Regression Analyses Predicting Word Count or CTTR

Predictors	Model 1	Model 2	Model 3	Model 4
	Construct: Global Self-Esteem		Construct: Resolution	
	Predicting Word Count	Predicting CTTR	Predicting Word Count	Predicting CTTR
(Intercept)	10.69***	1.98***	15.64***	2.58****
PromptConditionFew-Shot-Basic	0.77***	0.12***	1.21*	0.10**
PromptConditionFew-Shot-Advanced	-1.80***	-0.13***	-5.35***	-0.43***
IterationNumber2	-0.07	0.01	-0.15	-0.03
IterationNumber3	-0.31	-0.03	0.10	0.00
VersionThinking	-0.41*	-0.06**	-0.68	-0.04
LLMBrandClaude	0.20	0.03	0.08	0.08
LLMBrandDeepSeek	0.87***	0.08**	1.02	0.05
LLMBrandGemini	0.08	-0.01	0.19	-0.03
Adjusted R^2	0.18	0.17	0.50	0.52
F-stats	20.45***	19.98***	27.42***	30.65***
(df_1, df_2)	(8,711)	(8,711)	(8,207)	(8, 207)

The predictors are PromptCondition with the zero-shot condition being the reference, IterationNumber with the first iteration being the reference, Version with fast being the reference, and LLM brand with ChatGPT being the reference. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

2.4.2 Comparing Between Human- and AI-Generated Items

Given the small number of human-generated items (ten for global self-esteem and three for resolution), the comparisons between human- and AI-generated items should be interpreted descriptively rather than inferentially.[2] Nonetheless, a few interesting patterns of results emerged. First, across both constructs, human-generated items tended to have higher word counts and CTTR values than AI items generated using the few-shot-advanced prompt (see the second panel in Table 6). Compared with the few-shot-advanced AI items, human items were on average 2.75 words longer for the global self-esteem items and 2.89 words longer for the resolution items. This suggests that the AI-generated items in the few-shot-advanced condition better followed the recommended item-writing guidelines than the human-generated items, although the conciseness instructions may have been too restrictive if the goal is for AI items to more closely match the human-generated items.

Some divergences between human- and AI-generated items varied by construct. For the global self-esteem construct, human items were comparable in word count and CTTR to the AI items under the zero-shot and few-shot-basic prompts; in contrast, for the resolution construct, human items had lower word counts and CTTR values than AI items in those conditions (see the second panel in Table 6). Furthermore, when disaggregated by LLM brand, human items generally had slightly higher word counts and CTTR values than AI items for the global self-esteem construct, but slightly lower values for the resolution construct (see the fourth panel in Table 6).

3. Discussion

This study examined the effectiveness of LLMs for AIG in psychological scale development, focusing on comparing how prompting strategies and LLM brands and versions affect AIG for both an established construct (global self-esteem) and a novel construct (resolution). Several key findings emerged from our study.

3.1 Varied Quality of AI-Generated Items

Expert review showed substantial variability in the quality of AI-generated items. Some items were clear, construct-relevant, and comparable to existing human-generated items, while other items had notable issues, such as tapping into unintended constructs or being overly general. Such nuances cannot be detected through AI-powered psychometric analyses that rely on embedding matrices,

Table 6. Pairwise Comparisons For Word Count and CTTR

Comparisons	Construct: Global Self-Esteem		Construct: Resolution	
	Word Count	CTTR	Word Count	CTTR
Comparing Among AI-Generated Items Across Prompt Conditions				
Few-Shot-Basic – Zero-Shot	0.77**[0.25,1.29]	0.12***[0.07 , 0.18]	1.21[0.08, 2.34]	0.10[0.01, 0.19]
Few-Shot-Advanced – Zero-Shot	-1.80***[-2.32, -1.28]	-0.13***[-0.18, -0.08]	-5.35***[-6.48, -4.21]	-0.43***[-0.52, -0.34]
Few-Shot-Advanced – Few-Shot-Basic	-2.57***[-3.09, -2.05]	-0.25***[-0.31, -0.20]	-6.56***[-7.69, -5.42]	-0.53***[-0.61, -0.44]
Comparing Between Human- and AI-Generated Items Across Prompt Conditions				
Human – Zero-Shot	0.95[-1.09, 3.00]	0.08[-0.13, 0.28]	-2.46[-6.87, 1.95]	-0.26 [-0.60, 0.08]
Human – Few-Shot-Basic	0.18[-1.86, 2.23]	-0.05[-0.25, 0.16]	-3.67[-8.08, 0.75]	-0.36* [-0.70, -0.02]
Human – Few-Shot-Advanced	2.75**[0.71 , 4.80]	0.21[0.01, 0.41]	2.89[-1.52 ,7.30]	0.17 [-0.18, 0.51]
Comparing Among AI-Generated Items Across LLM conditions				
Claude – ChatGPT	0.20[-0.49, 0.89]	0.03[-0.04 , 0.10]	0.82[-0.62, 2.25]	0.08[-0.03, 0.19]
DeepSeek – ChatGPT	0.88**[0.21, 1.53]	0.08*[0.01 , 0.14]	1.02[-0.42, 2.45]	0.05[-0.06, 0.16]
Gemini – ChatGPT	0.09[-0.56, 0.75]	-0.01[-0.08 , 0.05]	0.19[-1.25, 1.62]	-0.03[-0.14, 0.08]
Claude – DeepSeek	-0.67[-1.33, -0.01]	-0.05[-0.11 , 0.02]	-0.20[-1.64, 1.23]	0.03[-0.08, 0.14]
Claude – Gemini	0.11[-0.55, 0.77]	0.04[-0.02 , 0.11]	0.63[-0.81, 2.07]	0.11[-0.00, 0.22]
DeepSeek – Gemini	0.78*[0.13, 1.44]	0.09**[0.02 , 0.15]	0.83[-0.60, 2.27]	0.07[-0.04 0.18]
Comparing Between Human- and AI-Generated Items Across LLM Conditions				
Human – Claude	1.39[-0.98, 3.76]	0.07 [-0.16, 0.31]	-1.39[-8.00 , 5.22]	-0.21 [-0.73, 0.31]
Human – DeepSeek	0.72 [-1.65, 3.08]	0.03 [-0.21, 0.26]	-1.59[-8.20 , 5.02]	-0.18 [-0.69, 0.34]
Human – Gemini	1.50 [-0.87, 3.87]	0.11 [-0.12,0.35]	-0.77[-7.37 , 5.85]	-0.10 [-0.62, 0.42]
Human – ChatGPT	1.59 [-0.78 , 3.96]	0.10 [-0.13, 0.34]	-0.57[-7.19 , 6.04]	-0.13 [-0.65, 0.39]

Note. The pairwise comparisons were done with Tukey Honest Significance Tests to control for family-wide Type I error. The numbers in the square brackets are 95% confidence intervals. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

underscoring the importance of human screening after the initial AIG. Lee et al. (2025) examined a multi-agent framework where separate agents were designed for item generation and item screening. This approach shows promise as a potential alternative to human review, although further research is needed to evaluate its generalizability across psychological constructs.

3.2 Effectiveness of AIG for Established vs. Novel Constructs

Another major finding from our study concerns the difference between AIG for established versus novel constructs. For the global self-esteem construct, the zero-shot prompt often produced the existing RSES items verbatim for all LLMs except ChatGPT thinking. These results highlight a risk when using LLMs for well-known constructs: without explicit anti-copying instructions, LLMs tend to reproduce existing items rather than generate new item content.

In contrast, for the novel construct of resolution, where no human-generated items exist online, LLMs did not reproduce any existing items under the zero-shot prompt. Instead, they relied exclusively on the construct definition and often produced items that captured meaningful nuances extending beyond the researcher's initial conceptualization. This finding suggests that AIG may be particularly valuable for emerging constructs, where LLMs can serve as creative partners that help researchers probe and elaborate on different facets of an emerging construct.

3.3 Importance of Prompt Engineering for AIG

Across analyses, prompt engineering emerged as the strongest determinant of item characteristics. As just explained, when the prompt did not contain explicit anti-copying instructions, as in the zero-shot prompt for the global self-esteem, LLMs often reproduced the existing items available online. When the prompt contained explicit item-writing guidelines, as in the few-shot-advanced prompts, LLMs

generated items that were more concise and easier to read, as indicated by lower lexical diversity, and more semantically homogeneous, as indicated by higher cosine similarity. Overall, these findings highlight the importance of carefully designing prompts to obtain high-quality items using LLMs.

3.4 Limitations and Future Directions

Several limitations of the present study should be noted. First, although prompting strategies play an important role in AIG, the three prompt conditions in our study were not enough to examine how specific nuances within a prompt influenced AI-generated items. Future research should systematically vary the prompt features to identify which aspects most strongly affect item quality.

Second, the number of human-generated items for the two constructs was very small, which limited our ability to draw reliable comparisons between human- and AI-generated items. Studies with larger sets of human-generated items are needed to more rigorously evaluate these differences.

Third, we used OpenAI's text-embedding-3-large model to compute semantic similarity across all items. Although our results did not show any systematic advantage for ChatGPT-generated items, using an embedding model from a single provider could introduce bias. Future research should replicate these analyses using additional embedding models, including open-source alternatives.

Fourth, the fact that the co-authors of this study served as the experts to review the items for quality and face validity may have introduced some experimenter bias. Although we authors have no incentive to favour AIG or any specific LLM, independent subject matter experts blind to model conditions would strengthen the design. Future research should incorporate such blinded evaluations.

Finally, the biggest limitation is that we have not yet fully evaluated the key psychometric properties of the AI-generated items, including their factor structure, reliability, and predictive and convergent validity. We are currently conducting a follow-up study that collects data from real participants on these items. Preliminary analyses indicate that, across most conditions, the AI-generated items show reliability, factor structures, and validity patterns that are broadly comparable to those of human-created items. We will report the full analysis results in a forthcoming paper.

3.5 Conclusion

In sum, our findings highlight both the promise and the limitations of AIG for psychological scale development. LLMs can efficiently produce large numbers of items, many of which are clear, concise, and theoretically meaningful. For novel constructs in particular, AI-generated items may capture nuanced facets that extend beyond the researcher's initial conceptualization and merit further exploration. At the same time, the results underscore the need for careful human screening, as some items showed problems such as tapping into unintended constructs and being overly general. The study also demonstrates that prompt engineering plays a central role in determining the quality of AI-generated items. Depending on researchers' goals, prompts must be deliberately crafted to prevent the reproduction of existing scale items and to achieve the desired balance of specificity and conciseness. Together, these insights suggest that AIG can be a valuable tool in early-stage scale development when paired with thoughtful prompt design and human oversight.

Funding Statement This research was supported by Insight Grants from the Social Sciences and Humanities Research Council of Canada (SSHRC), given to the first author and the OpenAI Researcher Access Program Fund given to the first author.

Competing Interests None.

Notes

1 Guenole et al. (2025) proposed applying factor analysis to the cosine similarity matrix, a method they referred to as pseudo factor analysis. We did not adopt this approach because factor analysis models observed variables as combinations of latent

constructs plus measurement error—an assumption that does not align with the properties of embedding systems or cosine similarity matrices. Instead, we consider Milano et al. (2025)'s use of PCA more appropriate, because the embedding system is a semantic high-dimensional space and PCA is simply a dimension reduction technique without assumptions about latent constructs.

2 For consistency, we still presented the significance tests comparing human- and AI-generated items in Table 6.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Carroll, J. B. (1964). *Language and thought*. Prentice-Hall.

Clark, L. A., & Watson, D. (2016). Constructing validity: Basic issues in objective scale development. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (4th ed., pp. 187–203). American Psychological Association. https://doi.org/10.1037/14805-012

Cozby, P. C., Mar, R. A., & Bates, S. (2024). *Methods in behavioural research* (4th ed.) [ISBN-10: 1264860994]. McGraw-Hill Education.

Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*. https://doi.org/10.1037/met0000540

Guenole, N. (2025). *Psychometrics.ai: Transforming behavioral science with machine learning* [Online textbook. London, May 2025.]. https://psychometrics.ai

Guenole, N., D'Urso, E. D., Samo, A., Sun, T., & Haslbeck, J. M. B. (2025). Enhancing scale development: Pseudo factor analysis of language embedding similarity matrices [Preprint]. https://doi.org/10.31234/osf.io/vf3se_v2

Hensley, W. E., & Roberts, M. K. (1976). Dimensions of Rosenburg's Self-Esteem Scale. *Psychological Reports*, *38*(2), 583–584. https://doi.org/10.2466/pr0.1976.38.2.583

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*, *87*(2), 749–772. https://doi.org/10.1007/s11336-021-09823-9

Kyriazos, T. A., & Stalikas, A. (2018). Applied psychometrics: The steps of scale development and standardization process. *Psychology*, *9*(11), 2531–2560. https://doi.org/10.4236/psych.2018.911145

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2023). A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. *Journal of Business and Psychology*, *38*(1), 163–190. https://doi.org/10.1007/s10869-022-09864-6

Lee, P., Son, M., & Jia, Z. (2025). Ai-powered automatic item generation for psychological tests: A conceptual framework for an llm-based multi-agent aig system. *Journal of Business and Psychology*, 1–29. https://doi.org/10.1007/s10869-025-10067-y

Milano, N., Luongo, M., Ponticorvo, M., & Marocco, D. (2025). Semantic analysis of test items through large language model embeddings predicts a-priori factorial structure of personality tests. *Current Research in Behavioral Sciences*, *8*, 100168. https://doi.org/10.1016/j.crbeha.2025.100168

Rosenberg, M. (1989). *Society and the adolescent self-image*. Wesleyan University Press.

Russell-Lasalandra, L. L., Christensen, A. P., & Golino, H. (2024). Generative psychometrics via ai-genie: Automatic item generation and validation via network-integrated evaluation. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/fgbj4

Winkielman, P., Schwarz, N., Fazendeiro, T. A., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 189–217). Lawrence Erlbaum Associates Publishers.

Zhang, X., & Savalei, V. (2016). Improving the factor structure of psychological scales: The Expanded format as an alternative to the Likert scale format. *Educational and Psychological Measurement*, *76*(3), 357–386. https://doi.org/10.1177/0013164415596421