AIR®
Advancing Evidence.
Improving Lives.

Using Generative AI for Item Development

Lessons Learned

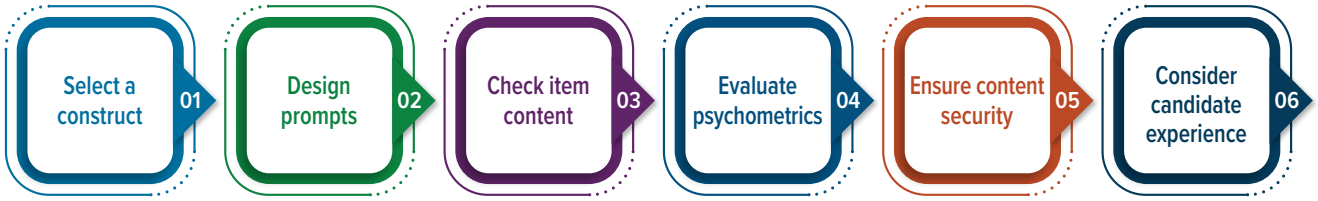
AUTHORS | Yuliya Cheban-Gore, PhD | Bharati B. Belwalkar, PhD | Christina Curnow, PhD

Educational and employment-related testing has increasingly embraced the use of artificial intelligence (AI) to support the assessment development process. Assessment developers are further exploring how AI can enhance efficiency, reduce costs, and maintain high-quality and healthy item banks. The integration of AI into the assessment development lifecycle represents a significant shift toward more innovative and adaptive testing solutions. One approach that has gained momentum in recent years is augmenting item development using generative AI (GenAI).

Recently, researchers and practitioners have been exploring the use of large language models (a form of GenAI) to develop assessment items (e.g., personality scales [Götz et al., 2023]); however, there is limited evidence-based guidance on the practice of using GenAI for item development. Therefore, we explored the use of Generative Pre-trained Transformer 4 (ChatGPT4) to develop assessment items for three constructs. This article details the practical considerations and lessons learned from this process.

Exhibit 1 provides a snapshot of our item generation process, which was used to develop a total of 45 items per construct. The results of our study can be seen in this report.

Exhibit 1. AI Augmented Assessment Development Process



I. Practical Considerations and Lessons Learned

In our exploration of item development via ChatGPT4, in line with the process laid out in Exhibit 1, we identified six areas of the assessment development process that shaped our strategies and operational thinking: (1) selection of constructs, (2) prompt engineering, (3) quality assurance checks, (4) psychometric review, (5) content security, and (6) transparency. Paired with each of these considerations, we share our "lessons learned" to guide future efforts in leveraging GenAI for item development.

Selection of Constructs

PRACTICAL CONSIDERATION

- As with any scale development process, we expected that we needed to ***define our constructs carefully*** because constructs lacking a robust theoretical/research foundation could lead to items that do not accurately measure the intended traits or abilities due to the risks of construct underrepresentation (e.g., Messick, 1995). In other words, having a well-researched construct provides a clearer content space from which to develop assessments.

LESSON LEARNED

- **Invest time in assessing how well-developed the construct in question is.** Without a solid research foundation, the risk of generating invalid or unreliable items increases because without a thorough understanding of the construct and its definition, we may fail to develop a well-rounded scale. Therefore, we recommend ***thorough evaluation and understanding of the constructs before commencing the item generation*** process. This involves reviewing existing literature and taking a close look at the various ways the construct can be defined.
- Our background research led to three well-researched constructs.

Prompt Engineering

PRACTICAL CONSIDERATION

- The quality of AI-generated items depends significantly on the ***quality of prompts***. Effective prompt engineering involves crafting clear, concise, and specific instructions that guide the AI in generating relevant and high-quality items. We determined what information to feed the AI through careful prompt engineering. We reviewed best practices (e.g., Ekin, 2023; Li et al., 2023) and tailored that advice to our assessment development effort.

LESSON LEARNED

- **Create effective prompts** specifying the following details:
  - Clear definition of the construct.
  - Item type (e.g., Likert-type) and its corresponding measurement scale (e.g., 5-point scale
  - Desired difficulty level, one that is relevant to your prospective participants/assessment takers
  - Other specific requirements, such as the number of reverse-coded items
- **AI operates best within clearly defined boundaries**. Therefore, in our prompts, we also specified the purpose of item/scale use (e.g., selection and/or development) and the target population of our assessments.
- Further, consider conducting multiple iterations to fine-tune the prompt based on your qualitative inspection of the items received. We conducted two iterations; however, the number of iterations may also be determined by time or financial constraints.
- While we used the default settings, you may also need to adjust the AI's parameters (e.g., temperature, penalty) to modulate the item content (e.g., giving the platform more or less creative power).

Quality Assurance Checks

PRACTICAL CONSIDERATION	LESSON LEARNED
• Quality checks are necessary to ensure that the items accurately measure the intended constructs and meet the required standards and best practices. Therefore, we focused on maintaining consistency and accuracy of AI-generated items by ***verifying their content, as well as ensuring their adherence to specified item length and format.***	• **Build in human intervention in the process** to ensure high-quality items: ▪ Include a thorough review of the items for content accuracy, clarity, uniform interpretation, and alignment with the intended construct. ▪ However we developed an extensive prompt, a qualitative review of the items revealed several issues, such as redundant items. ▪ Provide item-writing best practice guidance in the AI prompt such as avoiding double negatives and idiosyncratic language. ▪ Supplement such guidance with examples of high-quality items from a trusted source.

Psychometric Review

PRACTICAL CONSIDERATION	LESSON LEARNED
• As with any assessment development process, we knew we would need to ***conduct psychometric analysis*** of our AI-generated items. Psychometric analyses help identify items that do not perform well and ensure that the overall assessment meets the required standards of measurement accuracy. Thus, we conducted a large-scale pilot (n = ~700) to assess the psychometric properties of generated items. To see the results of this study, refer to the Comparison of AI and Human-Generated Item Performance report.	• **Perform a thorough psychometric evaluation of AI-generated items.** • Despite prompting AI about the target population and item writing best practices, as already noted, issues such as statistical flags still emerged still emerged, highlighting the need for continuous review and testing.

The last two practical considerations and lessons learned (i.e., content security and transparency) are especially important when working with high-stakes assessment content. The current landscape of AI-based test development, from a best practices and legal standpoint, is limited in providing guidance. Therefore, our team had numerous conversations around these topic areas and conceptualized the following two lessons learned. Although we did not encounter any issues (e.g., security leaks or negative applicant reactions) during our pilot testing process, we include them here as a list of proactive measures one should take when using AI for item development.

Content Security

PRACTICAL CONSIDERATION

- The use of AI, especially large language models like ChatGPT4 in this case, raised questions about ***how secure our generated content is*** within the AI model.

LESSON LEARNED

- To mitigate the concerns related to content security, we recommend the following:
- **Using secure platforms (e.g., Microsoft Azure) that limit access** to the generated content to your organization.
- **Establishing clear policies and protocols regarding the use of AI-generated items** within your test development team.

Transparency

PRACTICAL CONSIDERATION

- The ethical implications of using AI to develop assessment content must be considered such that candidates and test-takers are required to be notified if AI was involved in creating the assessments, particularly in light of regulations like the automated employment decision tools (AEDT) law (The New York City Council, 2021), which mandates transparency in employment evaluations.

LESSON LEARNED

- Transparency can significantly impact the candidate experience, therefore, we recommend the following*:
  - **Informing candidates about the use of AI in item generation** to enhance their understanding of the assessment process and alleviate concerns about bias and fairness.
  - **Developing clear communication strategies** that explain the role of AI in the assessment process, how items are generated and validated, and the measures taken to ensure fairness and accuracy.

*Note. This transparency, when paired with a justification for AI's use, can help increase candidates' perceived fairness (Langer et al., 2021) of assessments.

II. Discussion and Outstanding Questions

Our work with GenAI for item development raised practical and theoretical questions. Further research, experimentation, and policy development are needed to find their answers.

Memory/Training Effects

- What mechanisms do GenAI systems use to retain and utilize information from previous prompts and interactions, and how do they affect the generation of new items?

Appropriateness of Assessment Type

- How effective is GenAI in developing different types of assessments, such as personality assessments and cognitive ability tests?
- What prompt engineering strategies are most effective for optimizing the generation of items for different assessment types?

Parameters

- Given the customizability of parameters (e.g., temperature, penalty) available in GenAI platforms, how qualitatively different are the items that are produced through parameter adjustments?

Logistics

- As automatic item development becomes more popular, how can testing programs accommodate such large-scale item development into pilot and/or field testing?

III. Conclusion

Our experience with item generation using ChatGPT4 underscores the ***indispensable role of human intervention*** in the development process. Despite the advanced capabilities of GenAI, human expertise in prompt engineering, quality assurance, and psychometric evaluation remains crucial for ensuring the validity, reliability, and relevance of generated items. Our findings highlight the importance of a ***hybrid approach in which GenAI augments human capabilities rather than replaces them.*** Moving forward, future research should focus on understanding memory effects, studying the impact on test takers, and tailoring strategies for different assessment types. This will pave the way for more efficient, accurate, and fair assessment solutions, ultimately enhancing the utility of GenAI in educational and employment testing.

References

Ekin, S. (2023). *Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices.* TechRxiv. https://doi.org/10.36227/techrxiv.22683919.v2

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, 29*(3), 494–518.

Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D., & Speith, T. (2021). Spare me the details: How the type of information about automated interviews influences applicant reactions. *International Journal of Selection and Assessment, 29*(2), 154–169.

Li, C., Wang, J., Zhu, K., Zhang, Y., Hou, W., Lian, J., & Xie, X. (2023). *Emotion prompt: Leveraging psychology for large language models enhancement via emotional stimulus* [preprint arXiv:2307.11760]. arXiv.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

The New York City Council. (2021). *Automated employment decision tools.* https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7CText%7C&Search=

AIR® Headquarters
1400 Crystal Drive, 10th Floor | Arlington, VA 22202-3289 | +1.202.403.5000 | **AIR.ORG**

2539-v6.4/25