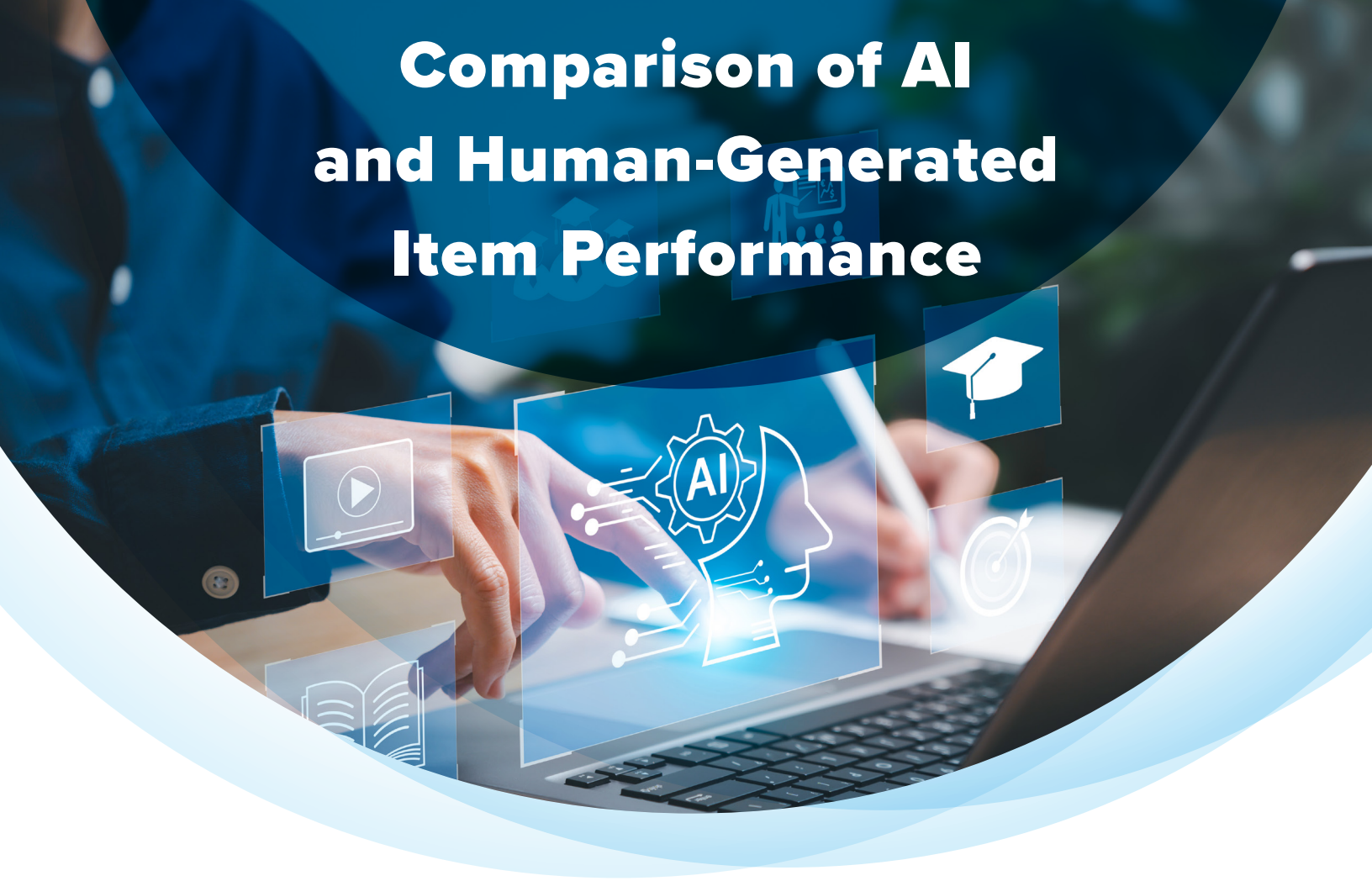
Comparison of AI and Human-Generated Item Performance



Background

Stories about rapid changes in technology are peppering the news and resulting in correspondingly rapid shifts in the way we conduct business, but there is little evidence about whether the use of artificial intelligence (AI) is fundamentally different in terms of the outcomes that are achieved through its use as compared to human-generated work products. Furthermore, little is known about the equitable and responsible use of AI, especially as it relates to people-practices in organizations.

Among the advancements in technology, generative artificial intelligence (GenAI)—a type of AI capable of generating text, images, or other media in response to prompts—has emerged as a powerful tool. In recent years, GenAI has revolutionized the field of talent assessment, enabling the development of assessment items at an unprecedented speed and volume. For example, companies are now considering using GenAI to create selection tests that help select suitable candidates for specific roles.

> In this research study, we qualitatively and psychometrically compared personality assessment items developed by ChatGPT-4 and human experts.

AUTHORS

Bharati B. Belwalkar, PhD

Christina Curnow, PhD

Chelsi Campbell, MA, ABD

AIR®
Advancing Evidence.
Improving Lives.

This burgeoning interest in GenAI's capabilities prompted us to conduct a comparative study to ***evaluate the performance of assessment items generated by ChatGPT-4 against those crafted by human experts.*** Personality assessments are commonly used by organizations for assessing job candidates. Examples of common personality characteristics include extraversion and conscientiousness. These types of assessments generally present a person with a series of statements and ask them to make a rating (usually on a 5-point scale) about how well the statement describes them. The format of these assessments lends itself to the use of AI. Our study uncovered some interesting insights into the content and psychometric properties of both types of items—GenAI- and human-generated items—which are presented in this report in an infographic style.

To generate items for comparison, we identified three personality-related constructs. A personality construct is a psychological concept that represents aspects of an individual's traits, behaviors, and patterns of thought. We selected ▼**Learning Agility**, ▼**Proactivity**, and ▼**Self-Regulation** for this comparative study. Exhibit 1 shows their definitions.

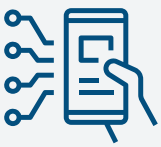
With the help of ChatGPT4, we generated items based on these constructs and compared them to the items developed by human item-writers.

Exhibit 1. Three Personality Constructs and Their Definitions

Learning Agility (LA)	Proactivity (PRO)	Self-Regulation (SR)
A tendency to develop oneself by acquiring new skills, mastering new situations, and improving one's competence; a set of complex skills that enables an individual to learn something new in one place and then apply what is learned elsewhere, in a different context.	A self-initiating tendency to act in advance of a future situation in order to either take control of one's situation or cause a desired change in others' and/or one's environment.	An ability to perceive, use, understand, manage, and handle one's emotions and one's capacity to blend thinking and feeling to make optimal decisions.

These personality constructs were selected based on specific criteria to ensure that adequate information was available for each selected construct to develop theoretically and psychometrically sound assessment items. For human-generated items, we selected psychometrically valid and reliable items from the existing literature that were developed by human item-writers. After a thorough evaluation of the constructs based on these criteria (Exhibit 2), our GenAI item-development process commenced.

Exhibit 2. Criteria for Selecting Personality Constructs for This Study



Theoretically broad constructs
offer more room for diverse item generation, allowing GenAI models to showcase flexibility.

Clearly defined constructs
ensure GenAI creates items that reflect their core meaning without diverging.

Well-researched constructs
provide a standard for comparison, making it easier to assess GenAI-item quality against human-generated items.

Results

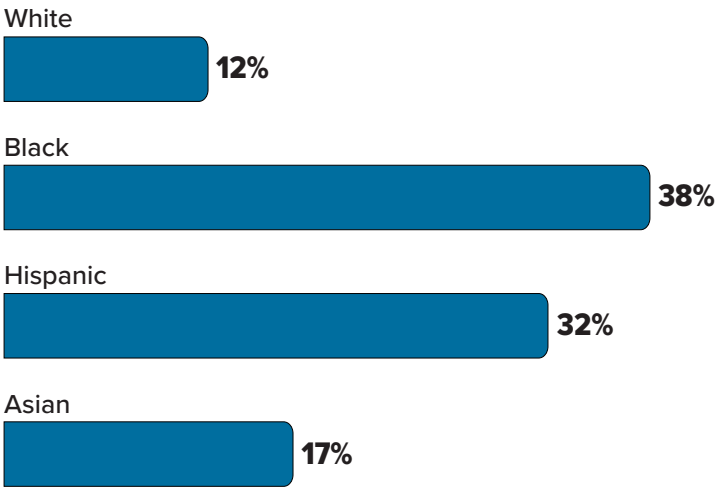
Our study sample included a total of 686 individuals who took an online survey questionnaire including a mix of 250+ human- and GenAI-generated personality items, along with a short demographic questionnaire. Their task was to read a given personality-assessment item (statement) and rate their level of agreement on a 5-point rating scale. See example in Exhibit 3.

Exhibit 3. Example Survey Item

	Strongly Disagree	Disagree	Neither Disagree Nor Agree	Agree	Strongly Agree
I take charge when the situation needs a solution.	○	○	○	○	○

The demographic makeup of our study sample is summarized in Exhibit 4. In addition, Exhibit 5 shows descriptive and psychometric statistics of human and GenAI items in terms of their counts, average ratings, and reliability indices of final assessment scales.

Exhibit 4. Demographic Characteristics of Study Sample



Note: 1% of the participants did not respond.

Average Age = 24 years
(standard deviation = 3.16)

78.25%
reported annual income below $75,000

93.35%
reported English as a first language

Exhibit 5. Human and GenAI Performance Comparison

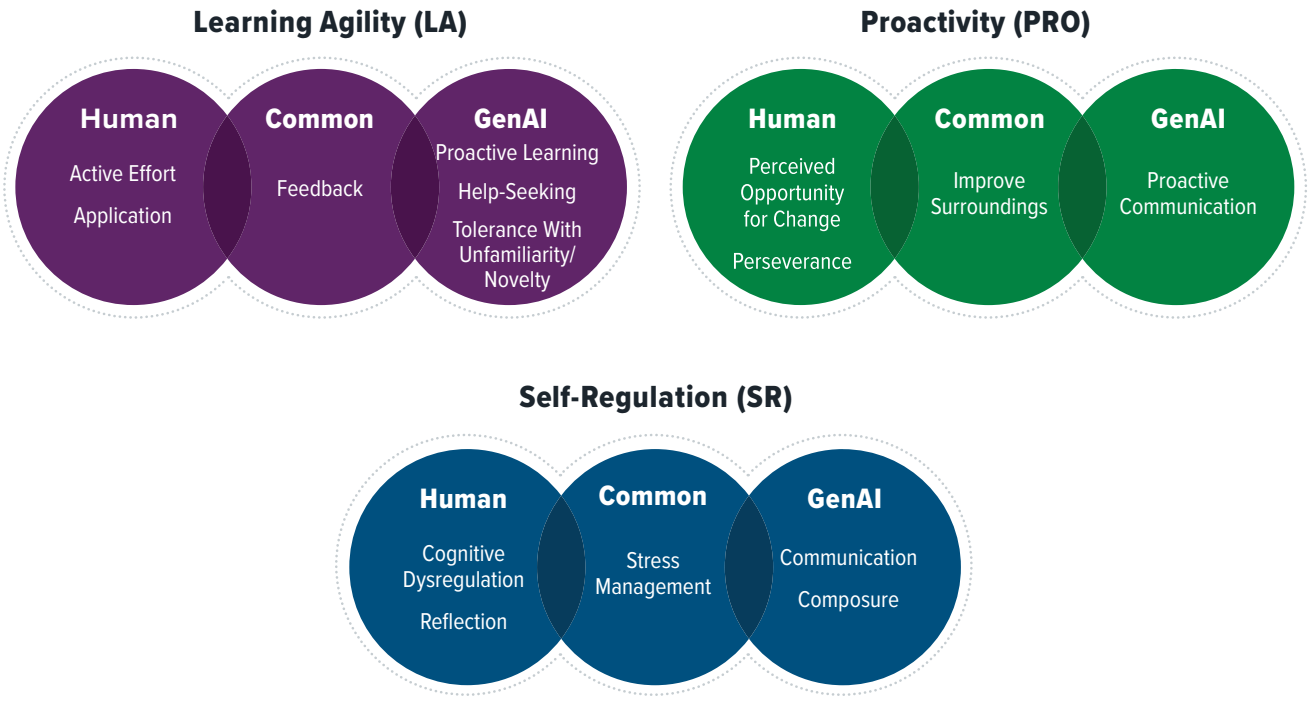
Assessment Development Steps and Statistics	Learning Agility (LA)		Proactivity (PRO)		Self-Regulation (SR)	
	Human	GenAI	Human	GenAI	Human	GenAI
Total Number of Items Developed	43 Items	45 Items	40 Items	45 Items	43 Items	45 Items
Number of items in Final Assessment	13 Items	15 Items	9 Items	7 Items	12 Items	9 Items
Final Assessment Statistics	Average = 3.82 SD = .26 Reliability = .89	Average = 3.83 SD = .17 Reliability = .92	Average = 3.62 SD = .41 Reliability = .88	Average = 3.49 SD = .26 Reliability = .83	Average = 3.27 SD = .39 Reliability = .92	Average = 3.56 SD = .14 Reliability = .87

Let's look at Exhibit 6 including three Venn diagrams for each of the three personality constructs:

- The left circle represents the items created by human item writers, and the other circle to the right represents the items created by ChatGPT-4.
- The part where both circles overlap shows the topics or dimensions that both GenAI and humans covered in their items.
- The parts of the circles that do not overlap show the unique topics or dimensions that were covered only by GenAI or only by human item-writers.

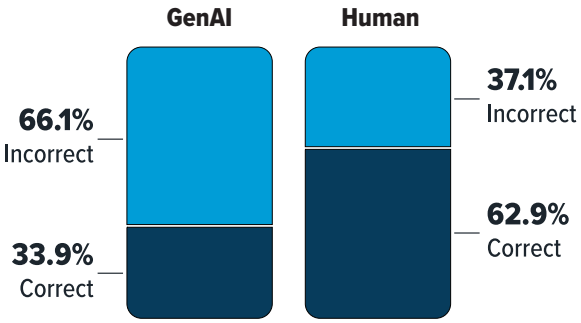
Evidently, there was at least one area/dimension of the construct that both GenAI and humans focused on. And, comparatively, there were a greater number of areas/dimensions that were unique to each.

Exhibit 6. Overlap Between GenAI- and Human-Generated Items



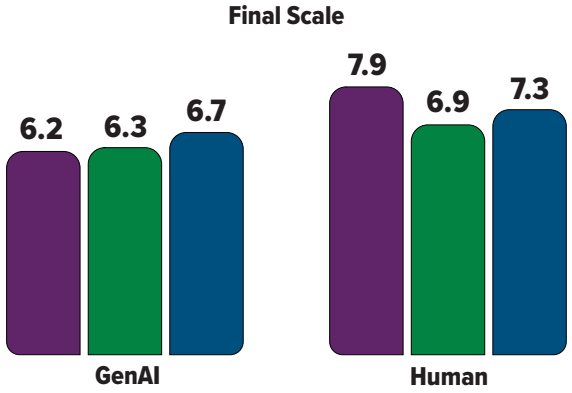
For study participants, the GenAI items were ***indistinguishable*** from human-generated items. Interestingly, participants ***misattributed GenAI items that included "emotions," "feelings"*** to human generation (Exhibit 7).

Exhibit 7. Proportion of Correct Identification GenAI Versus Human-Generated Items



The final GenAI assessments required ***lower grade-level reading*** compared to the human-generated final assessments (Exhibit 8), which could lead to easier reading and likely better understanding for assessment takers.

Exhibit 8. Reading Level of GenAI and Human Items



Insights

Based on careful prompting in ChatGPT-4, we developed the initial item pool. Prompting in ChatGPT-4 is like having a focused conversation with someone (in this case a GenAI model) who can provide detailed answers, so long as we guide the "conversation" with clear and specific questions/instructions. Each item thereafter was reviewed by a team of four industrial and organizational (I/O) researchers, and they offered the following specific observations about these items:

- Many of the GenAI items were very similar to each other, **with only minor differences that did not add much value.**
- About 20% of the items created by GenAI **did not relate to the construct we were trying to assess.**
- **Negatively worded (also known as "reverse-scored") items** from GenAI simply included the word "not" to make them negative, which was **not effective or meaningful.**
- Psychometrically speaking, a much larger proportion of GenAI items were **flagged for demographic differences** than human-generated items.
  - GenAI items showed **differences in response patterns for English- and Non-English-speaking participants.**

Conclusion

GenAI excels in generating diverse items, offering a fresh perspective, and potentially uncovering novel insights. However, it currently falls short in replicating the nuanced understanding that humans bring to the item-writing process. GenAI-generated items require thorough review to ensure accuracy and relevance. To learn more about this study, stay tuned for our next “lessons learned” piece!

Reflections

Our findings highlight that GenAI is a powerful tool to enhance and drive innovation in the field of assessment. However, it should be used to augment, not fully automate, the assessment development process.

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

25389-v07_12/24