KORN FERRY
BE MORE THAN

MODERN ASSESSMENT IN THE AGE OF GENERATIVE AI

Thought Leadership

Introduction

The release of applications such as ChatGPT, Copilot, and Gemini have made large language models readily accessible to the public. With just a click, people can use generative artificial intelligence (GenAI) to power how they approach everyday tasks at home and work, from simple document summarization to more complex activities such as generating code and scraping information from social media. Recent studies have found that the productivity of knowledge workers, depending on the role, may surge by as much as 30% to 60% in the coming years. These large language model (LLM) tools are primed to revolutionize the workplace, prompting a need for careful consideration of their impacts across all enterprise functions.

Within talent management, recruiting and hiring employees are critical work processes. Assessment is often a key part of these processes, with tools ranging from psychometric-based self-assessments to interviews to simulations. The broad availability and generalized use cases of LLMs are generating new concerns about an issue at the heart of assessment use: To what extent do assessment results accurately predict how someone will perform in a given role? Specifically, will the use of tools like ChatGPT on unproctored assessments give applicants an advantage over other candidates, thus undermining the accuracy of assessment scores? This apprehension has been fueled by OpenAI's sharing how GPT-4 performed on standardized assessments used for educational admissions and professional licensing. In many instances, GPT-4 performed at or above the 90th percentile of human test takers (Hickman, Dunlop, & Wolf, 2024).

With this in mind, the Korn Ferry Institute systematically investigated the extent to which the use of readily available GPT models affects results on Korn Ferry's assessments. Our basic approach was to prompt an LLM to assume the role of a job applicant. Items were then fed into the LLM with a request for it to generate appropriate responses. The responses were used to generate assessment scores. To capture a diversity of conditions, our prompt engineers used two different LLMs (GPT-3.5 Turbo and GPT-4) and varied the approach to prompts (basic and expert). We also studied different assessment types, including Korn Ferry's self-assessments of competencies, traits, and drivers (CTD); cognitive assessments; and a situational judgment test. For the self-assessments of CTD, we also had the LLMs take on the role of job applicants for 12 different roles covering a variety of management levels. Here are five key takeaways from our investigation.

Key Point #1: GenAI tools do not effectively take role-relevant information into account when completing Korn Ferry self-assessments.

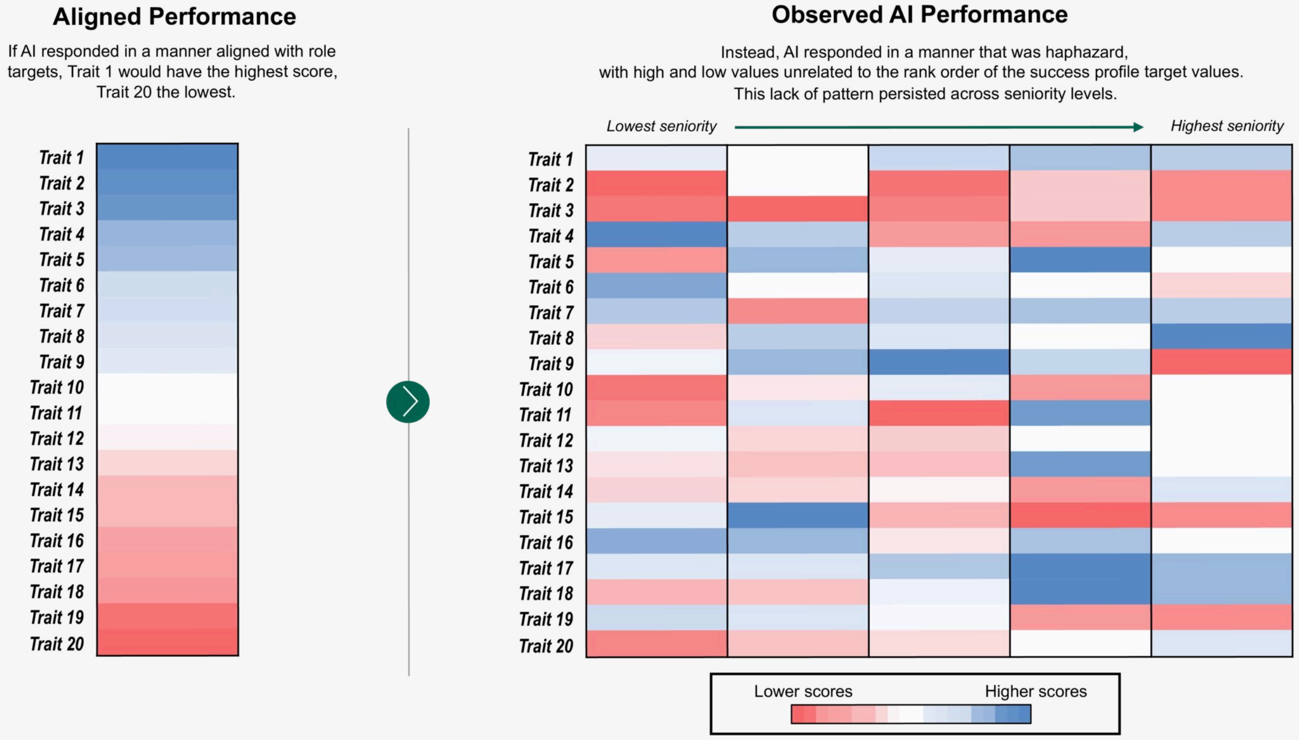
Making decisions about people at work—whether for role selection or differential developmental activities—is consequential. Therefore, the reporting from psychometric-based assessment results used for these decisions strongly benefits from being precise and tailored to the context (Deege, Hezlett, & Severinsen, 2021). This typically involves comparing a person to what "good" looks like, or what is required for success through norms and benchmarking to enhance interpretation of individual assessment results. Norms reflect the average score of a specific population, and a high or low score is defined by the distance from the average or "norm" for that population. In addition, it is important to take into consideration the specific demands of the job. This can be reflected through benchmarks, or a profile that defines job success.

Korn Ferry's extensive expertise in work measurement and people is embedded in our 10,000+ "Success Profiles." These Success Profiles include target values for 30 work-related competency scores, 20 specific personality traits, grouped into five super factors analogous with the "Big Five" model of personality, and six motivational drivers. These attributes have a significant bearing on getting the job done effectively. The job-specific targets for each component in our Success Profiles are based on multiple quantitative data points describing the characteristics of the job, the culture of the organization, and the level of seniority of the profile, all while accounting for regional differences. By comparing assessment results to Success Profile target scores, we are able to make conclusions around role fit that are contextualized by these inputs. For example, our research suggests that, in psychometric terms, people with a strong fit to a role are typically 7.5 to 8 times more engaged than those with a poor fit, and typically 2.7 times more engaged compared to those with an average fit (Lewis, Lams, Hezlett, & Firat, 2023). Additionally, those who fit well are more than five times more likely to be high performers compared to those who fit poorly (Lewis, Starke, Foster, & Hezlett, 2024).

In our study, we prompted the LLMs to act as an applicant to a diverse range of job levels and functions. Traits, drivers, and self-efficacy for behavioral competencies were measured using Korn Ferry's world-class psychometric-based self-assessments. We examined the degree to which the AI responses to the assessment items varied across roles. Scores did vary across roles; the GenAI tools did not complete our competency, trait, and driver self-assessments the same way each time. However, the differences were not aligned with the target values outlined in the respective Success Profiles. Typically, the AI's scores fell below many target values, including those notably important for the role. In other words, even though the AI was prompted to apply for a specific role, it was not effective at tailoring responses to present as a strong candidate for those roles.

The illustration below summarizes trait results across the level-based profiles we investigated in our study. We've masked the actual traits and roles, but the visual shows how the GenAI responses fall short. Fit to role matters.

Figure 1. Aligned Performance vs. Observed AI Performance on Korn Ferry Traits Assessment*.



*Traits 1-20 will vary in importance by seniority level, but specific traits are masked here. In other words, Trait 1 in the first column may not be the same as Trait 1 in the second column, but the values in both cases represent the traits with the highest target score for that seniority level.

Key Point #2: Korn Ferry's forced-choice item response theory approach mitigates "faking good" for both humans and AI.

Responses to assessments can be collected and scored in a variety of ways. Assessments with conventional Likert-type scales give multiple response options that are ordered on a continuum. For example, people are given a statement and asked to choose one of five response options, ranging from "Strongly Disagree" to "Strongly Agree." People often prefer to have multiple choices, but they vary in their tendencies to use these choices. For example, people differ in the extent to which they use the full range of responses or how often they use the most extreme responses. In addition, Likert-type scales are susceptible to response distortions, where people intentionally or unintentionally adjust their responses to manage impressions of themselves.

The forced-choice response test format helps to mitigate these issues, asking participants to rank a set of words or statements. For example (Brown & Maydeu-Olivares, 2011):

> *Rank the items in the list below from most like you to least like you:*
>
> 1. I am careful over details.
>
> 2. I manage to relax easily.
>
> 3. I set high personal standards.
>
> 4. I enjoy working with others.

Items that force a participant to choose can be scored in different ways. Forced-choice item response theory (FC-IRT) methods offer multiple advantages (Deege, Hezlett, & Severinsen, 2021). First, this approach to scoring avoids the ipsativity that stems from more basic approaches of scoring forced-choice responses. Ipsativity makes assessment responses unsuitable for comparisons among individuals and unusable for many analytics. Korn Ferry's competency, trait, and driver self-assessments use FC-IRT. Second, research has found that FC-IRT assessments are less susceptible to human impression management or "faking good" than Likert-type tools (Cao & Drasgow, 2019; Wetzel, Frick, & Brown, 2021; Lewis, Starke, Foster, & Hezlett, 2024).

To date, there is limited research on whether this holds for AI. Focusing on extraversion and conscientiousness, Phillips and Robie (2024) examined how several LLMs performed when prompted to respond to personality inventories as if they were applying for a sales job. Response format mattered. When Likert-type response scales were used, the LLMs tended to perform at or above the median of students who were pretending to respond as ideal applicants. When a forced-choice response format was used, GenAI tended to perform worse, scoring either below or at the median. An exception was GPT-4, which scored over the 85th percentile of the student distribution in all cases.

However, GPT-4 did not perform nearly as well on Korn Ferry's self-assessments. Although further evaluation is needed, we believe this may be due to several differences between our FC-IRT tools and those examined by Phillips and Robie (2024). Korn Ferry's psychometric-based self-assessments use blocks of statements, rather than adjectives. We also use multiple comparisons per response block, rather than pairs. Finally, Korn Ferry tools use IRT-based scoring, rather than applying a simple scoring key based on the social desirability of responses. These distinctions in tandem likely make the assessments more difficult for the AI to fake, as the "best" responses may be more difficult to determine. This difficulty is compounded by the observed weakness in tailoring responses to the required role (see Key Point #1).

To gain further insight, our prompt engineers requested the chat tools provide "rationale" for the item rankings. The text from the AI transcripts was characterized by a tendency to simply repeat the nature of a given item using different terms. For example, the rationale for the placement of an item tied to a creativity-oriented trait merely mentioned that creative problem-solving was important. The rationale did not include why that item was ranked lower or higher relative to other items in the block, perhaps illustrating the LLM's challenge with the FC-IRT assessment.

Key Point #3: Prompt engineering makes a difference with assessment output, but less than expected.

LLMs are trained to evaluate the context of a given query using the words provided, and the relationships between them, and then predict subsequent words ("likely answers") (Meaden, Sturdivant, & Theys, 2024). And, while GenAI models are designed to respond to natural language, natural language is often imprecise or inconsistent, as the same combination of words can have different meanings from one sentence to another (Smith, 2023). Adding to the complexity, LLMs have a limited "attention span" that performs best when prompts are short and more focused as opposed to long and broad or meandering. Furthermore, a GenAI's prior outputs in a given chat conversation also influence future outputs in that same chat session, so it is important to know when to end a conversation and start a new one. The practice of focusing and streamlining queries in these chat tools is called prompt engineering.

Effective prompt engineering is a focus of many web posts, webinars, and courses. An instantly viral quote from Robin Li, co-founder and CEO of the Chinese AI company Baidu, frequently pops up across these diverse sources. Li declared, "In 10 years, half of the world's jobs will be in prompt engineering. And those who cannot write prompts will be obsolete" (Smith, 2023). While this may be an extreme view, there is no doubt that prompt engineering will be an important skill, and indeed even the entire focus of some roles in the future. The message is clear: knowing what and how to communicate to an AI model via prompt engineering is essential for ensuring GenAI outputs are as accurate and usable as possible.

In our study, we observed a slight trend for more accurate assessment outputs in the group provided with specific prompting and more prescriptive instructions to follow, when compared to the group for which instructions were loose and participants had more latitude to develop their own prompts. This means a savvy prompt engineer using an LLM to complete assessments might get slightly better results than a novice user, but not spectacularly so. A person with the mix of competencies, traits, drivers, and cognitive capabilities needed for a role will typically stand out on our assessments, and ahead of a GenAI fed with well-engineered prompts.

Key Point #4: Using GenAI to respond to assessments is a cumbersome process that many users dislike.

Applicants' experiences in the hiring process can affect their subsequent behavior and decisions. For example, applicants who have more positive views of an employer's online hiring website (e.g., user-friendliness, navigation, accessibility) or selection procedures have better perceptions of the organization and are more likely to pursue employment (Mooney, 2020). Although favorable reactions to selection procedures do not always translate into an offer acceptance (Van Iddekinge, Lievens, & Sackett, 2023), it may have an impact on the organization's brand. Applicants may share their experiences on social media, which in turn may impact the organization's ability to recruit participants (Woods, Ahmed, Nikolaou, Costa, & Anderson, 2019).

A variety of factors may influence applicant experiences and user reactions throughout the hiring process. Conventional wisdom holds that one of these is how long the process takes. Although research to date finds little support for a relationship between assessment process length and attrition (Hardy, 2017, 2019), it seems probable that many people would like to avoid activities that would prolong an application process or increase its tediousness.

Our test takers found the experience of using AI to complete the Korn Ferry self-assessments both time-consuming and annoying. It generally took longer to complete the assessments using either version of the chat tools as a go-between than it would have if the study participants had simply read the items and responded themselves. Some of the extra time spent was attributable to chat tool "hallucinations" that necessitated restating questions, in some cases multiple times. The most common hallucinations involved prompt misinterpretations, answering assessment questions that had been posed earlier (often several items prior) in the chat session. Additionally, users reported that Bing Chat (now known as Copilot) GPT-4 exhibited a tendency to provide excessive and unsolicited information beyond what was requested by the prompt engineer. Using GenAI to complete Korn Ferry assessments takes more time and effort, without yielding results that might make this process worthwhile.

Key Point #5: GenAI does not excel at all cognitive activities.

As mentioned previously, when GPT-4 was released, its strong performance on some standardized tests was announced to highlight its functionality. This has led many people to wonder if GenAI will be particularly effective at completing cognitive ability tests sometimes used as part of hiring processes (Hickman, Dunlop, & Wolf, 2024). Research that has dug into this issue has yielded interesting results, including that LLMs such as GPT-4 do not consistently perform well across all assessments. AI performance on quantitative tests can be particularly poor, especially for some test formats (Hickman, Dunlop, & Wolf, 2024). Similarly, our investigation of Korn Ferry's cognitive ability tests found GenAI's performance can be suboptimal.

We learned that some widely used GenAI applications struggle to accurately process complicated tables of numerical or image-based data. In other words, GenAI cannot be counted on to perform on all cognitive ability tests. There are a number of tools in assessment providers' toolkits that block GenAI's capacity to respond well. Many job applicants will place themselves at a disadvantage when attempting to use GenAI to complete assessments. Of course, different GenAIs have different capabilities. As GenAIs evolve, there will be an ongoing need to evaluate their performance. Where there are still concerns, in-person testing and online proctoring remain viable options to verify that scores came from humans, rather than being AI-generated.

Conclusion

Generative AI is an exciting new tool that will be increasingly leveraged at work for the foreseeable future. Our current investigation highlights that Korn Ferry's self-assessment approach, utilizing Success Profiles and FC-IRT methodology, remains robust. Consistent with the numerous appropriate applications of GenAI, stronger impacts were observed when ideal prompt engineering and more advanced GenAI models were deployed. However, in our examination, these trends were negligible. Additionally, GenAI does not consistently succeed on cognitive ability assessments.

Assessment practices for both employers and applicants will continue to evolve alongside GenAI. Not only will GenAI be increasingly used to create, score, and report on assessments, it will be incorporated into a broad variety of work tasks. With these changes, we anticipate organizations will explore new approaches to assessment design and delivery. Just as the widespread availability of calculators led to their use both on the job and in testing situations, we expect there will be opportunities to integrate GenAI into testing and work.

Authors

Sarah Hezlett
Vice President, Assessment Science
Korn Ferry Institute

Andrea Deege
Senior Director, Assessment Science and Scoring
Korn Ferry Institute

Andrew Hall
Senior Manager, Client Analytics
Korn Ferry Institute

James Lewis
Senior Director and Senior Scientist
Korn Ferry Institute

References

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3). 460-502. ISSN 0013-1644.

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, *104*(11), 1347–1368. https://doi.org/10.1037/apl0000414

Deege, A., Hezlett, S., & Severinsen, K. (2021). 11 truths about personality assessment. Korn Ferry Institute. https://www.kornferry.com/content/dam/kornferry/docs/pdfs/kfi-the-real-truth-about-personality-assessment.pdf

Gerkin, J., & Misiaszek, T. (n.d.) Harnessing Generative AI: A Win-Win for Employers and Employees. Korn Ferry Institute. https://www.kornferry.com/institute/harnessing-generative-ai-a-win-win-for-employers-and-employees

Hardy, J. H., III, Gibson, C., Sloan, M., & Carr, A. (2017). Are applicants more likely to quit longer assessments? Examining the effect of assessment length on applicant attrition behavior. *Journal of Applied Psychology*, *102*(7), 1148–1158. https://doi.org/10.1037/apl0000213

Hardy, J. H., III, Gibson, C., Carr, A., & Dudley, N. (2021). Quitters would not prosper: Examining the relationship between online assessment performance and assessment attrition behavior. *International Journal of Selection and Assessment*, *29* (1), 55-64. https://doi.org/10.1111/ijsa.12313

Hickman, L., Dunlop, P. D., & Wolf, J. L. (2024). The performance of large language models on quantitative and verbal ability tests: Initial evidence and implications for unproctored high-stakes testing. *International Journal of Selection and Assessment*, 1–13. https://doi.org/10.1111/ijsa.12479

Lewis, J., Lams, S., Hezlett, S., & Firat, R. (2023). Fit matters everywhere: A high-engagement recipe with 8-fold better odds. Korn Ferry Institute. https://www.kornferry.com/institute/fit-matters-everywhere

Lewis, J., Starke, M., Foster, J., & Hezlett, S. (2024). Korn Ferry Assess™ – leadership solution: A modern approach to employee assessment. Korn Ferry. https://www.kornferry.com/content/dam/kornferry-v2/pdf/institute/kfi-kfls-modern-assessment.pdf

Meaden, J., Sturdivant, M., & Theys, E. R. (2024, April 20). Harnessing the power of generative AI though effective prompt engineering [Master tutorial]. Society for Industrial and Organizational Psychology Annual Conference, Chicago, IL, United States.

Mooney, D. J., (2020). A meta-analysis of e-recruitment applicant experience, perception, and behavior. *Walden Dissertations and Doctoral Studies.* 8761. https://scholarworks.waldenu.edu/dissertations/8761

Phillips, J., & Robie, C. (2024). Can a computer outfake a human? *Personality and Individual Differences, 217*(1-4). https://doi.org/10.1016/j.paid.2023.112434

Smith, C.S. (2023, April 5). Mom, Dad, I want to be a prompt engineer. *Forbes*. https://www.forbes.com/sites/craigsmith/2023/04/05/mom-dad-i-want-to-be-a-prompt-engineer/

Van Iddekinge, C. H., Lievens, F. & Sackett, P. R. (2023). Personnel selection: A review of ways to maximize validity, diversity, and the applicant experience. *Personnel Psychology, 76*(2): 651–686. https://doi.org/10.1111/peps.12578

Wetzel, E., Frick, S., & Brown, A. (2021). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment, 33*(2), 156–170. https://doi.org/10.1037/pas0000971

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. (2019). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology*, *29*, 64–77. https://doi.org/10.1080/1359432X.2019.1681401