FEEDBACK REPORT FOR PARTICIPANTS

Examining the Potential Pitfalls of Using ChatGPT in Situational Judgment Tests

SMU REB # 23-101

Research Project conducted by:

Harley Harwood

Dr. Nicolas Roulin

Dr. Muhammad Zafar Iqbal

Department of Psychology

Saint Mary's University, 923 Robie Street, Halifax, NS B3H 3C3

In collaboration with Acuity Insights

Email: nicolas.roulin@smu.ca

The Issue:

Has ChatGPT signaled the end of hiring (for instance selection, assessment, or testing) as we know it? Many may wonder and struggle to conceptualize how assessment will work with the growth, popularity, and availability of ChatGPT and other generative artificial intelligence (AI) tools and large language models (LLM). Indeed, recent research has shown that Generative AI such as ChatGPT, when prompted effectively, is able to perform well on a variety of assessments (for example, personality tests or knowledge tests).

In the present project, we explored the transformative impact of ChatGPT on applicants' responses and performance in situational judgement tests (SJTs). SJTs generally present applicants with a series of contextualized scenarios, either written or via a short video, and ask them what they would (or in some cases "should") do in that situation. Applicants can then be asked to choose the "best" response from multiple options, rank the options from best to worst, type their own open-ended response, or (more recently) video-record their response.

The present research examined applicants' use of ChatGPT, whether using such tools influence the content of their responses, the impact of such behaviors on selection outcomes (e.g., performance ratings), and the potential solutions to detect such behaviors in the context of SJTs (i.e., faking-prevention mechanisms, especially AI-detection tools like GPTZero),

The Research:

We conducted two complementary studies:

- Study 1 examined how the availability of ChatGPT influenced response content and performance of real applicants (N = 107,805). They represent the entire population of US and Canadian candidates who completed the Casper SJT as part of their application for various health sciences programs (e.g., medicine, dentistry, nursing) between June 2022 and May 2023. We then compared the content of the responses and performance scores of applicants who completed the test before vs. after the release of the technology. We also tested detecting AI use with GPT Zero on a sub-sample of those responses (who completed the test either before or after the release of ChatGPT).

- In Study 2, we used an experimental approach with 138 participants recruited on the Prolific online platform. Participants were randomly assigned to one of three conditions: They completed a mock SJT while being instructed to either (1) use ChatGPT when responding (2) use online resources or (3) use no external resources. Like in Study 1, we then compared the content of the responses and performance scores of participants in those three conditions. We also tested detecting AI use with GPT Zero.

The Findings:

In the first study, we found only small differences in content (e.g., slightly less "authentic" words used) and performance (slight score improvements when controlling for response length, no differences otherwise) of applicants who completed the Casper SJT before or after GPT Zero was released – and thus AI was available to help them.

In the second study, we found only slightly higher SJT scores for the ChatGPT users, but no difference in response content.

Additionally, GPTZero (i.e., a popular AI detection tool) struggled to detect ChatGPT content, and generated many false positives, in both studies.

The Implications:

This research advances our understanding of how the release and popularization of ChatGPT can influence applicant behaviors. For instance, test providers and selection professionals in hiring organizations or education programs are looking for guidance on how such behaviors can impact their selection process, as well as potential remedies. This research thus contributes to better understanding if, when, or how applicants can (and do) use generative AI, while examining how such behaviors can impact test scores, and exploring AI detection tools.

Our findings also highlight the importance of designing assessments to prevent or limit faking. For instance, the Casper SJT uses some video-based scenarios, it provides limited time for applications to respond to the questions, and it prevents copy and pasting content – all of which are likely making faking or cheating using tools like ChatGPT much more difficult. Yet, the ever-evolving nature of AI calls for continuous research on the topic.

If you would like to read the full manuscript, click here.