"Anything You Can Do, I Can Do": Examining the Use of ChatGPT in Situational Judgment Tests for Professional Program Admission

Harley Harwood & Nicolas Roulin

Saint Mary's University

Muhammad Zafar Iqbal

Acuity Insights

Article "in press" in *Journal of Vocational Behavior*

Authors' note: Correspondence should be sent to Harley Harwood, Saint Mary's University, 923 Robie Street, B3H 3C3, Halifax, Canada, Email: harley.harwood@smu.ca

CRediT roles: Harwood: Conceptualization, Data curation, Methodology, Project administration, Formal analysis, Writing - original draft. Roulin: Conceptualization, Data curation, Methodology, Funding acquisition, Writing - review & editing. Iqbal: Funding acquisition, Data curation, Writing - review & editing.

The authors thank Adam Kabbeke for creating GPTZero API code, and Amanda Lapointe and Reegan Prete for their help with the Study 2 scoring; The data (for Study 2) and supplementary materials (related to both studies) are openly available in OSF: https://osf.io/4ncxt/?view_only=a0349e2901d841e7adcad3739c65a2a6; This research was supported by a Mitacs Accelerate Grant (Grant IT36461).

Abstract

We explored the transformative impact of ChatGPT on applicants' responses and performance in situational judgement tests (SJTs), as well as the role played by faking-prevention mechanisms, in two complementary studies. Study 1 examined how the availability of ChatGPT influenced response content and performance of real applicants (*N* = 107,805), who completed an SJT for admission before vs. after the release of the technology. 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). In Study 2, we used an experimental approach with (*N* = 138) Prolific participants completing a mock SJT, while being instructed to use ChatGPT when responding (vs. use online resources or no resources). 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. This research advances our understanding of how the release and popularization of ChatGPT can influence applicant behaviors. Given the "arms race" nature of applicant selection, they also highlight the importance of designing assessments to prevent or limit faking. Yet, the ever-evolving nature of AI calls for continuous research on the topic.

Keywords: Generative AI; ChatGPT; Situational Judgment Tests; Faking

"Anything You Can Do, I Can Do": Examining the Use of ChatGPT in Situational Judgment Tests for Professional Program Admission

"Has ChatGPT signaled the end of assessment as we know it?" (Kolade, n.d.). 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). While Rudolph et al. (2023) claim the need for adaptation, change, and enhancement of our current assessment practices in higher education, those in the world of personnel selection or professional program admission are likely left at best scratching their heads, and at worst panicking. Indeed, Generative AI such as ChatGPT, when prompted effectively, is able to perform well on a variety of assessments, including knowledge tests (Geerling et al., 2023), personality tests (Phillips & Robie, 2024), or situational judgement tests (SJTs; Borchert et al., 2023). Generative AI could similarly help applicants to provide better or more correct answers when completing such assessments. This in turn could potentially damage their validity and usefulness as selection instruments. As the technology under discussion is new, there is scant empirical research on the effects of generative AI on applicants' behaviors or performance in selection or admission tests.

The present research examines applicants' use of ChatGPT, 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. It thus explores an issue that is both conceptually and practically important. Theoretically, applicants using generative AI to improve their responses might represent a novel form of impression management (or even faking). Practically, 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. Thus, advancing the literature on and provides resources for practitioners involved in personnel selection and admissions testing.

SJTs and Applicant Faking

SJTs can take a variety of forms, but they 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 (Corstjens et al., 2017). Applicants can 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. SJTs have demonstrated both reliability and validity to assess competencies beyond traditional knowledge, skills, abilities, and other characteristics (KSAOs) in selection and admission contexts. For instance, SJTs are conventionally used to assess interpersonal skills for medical or dental school admission (Lievens, 2013; Lievens et al., 2005, 2016) or other non-academic skills for pharmacy programs (Patterson et al., 2019). SJTs are also relevant to evaluate professionalism or a lack thereof (Goss et al., 2017; Sahota & Taggar, 2020; Webster et al., 2020) or potential misconduct (Tiffin et al., 2022) in professional programs.

Similar to personality tests, SJTs generally offer incremental validity compared to those measures that assess different KSAOs (e.g., GPA, cognitive ability) but still require some element of self-reporting of behavior (Sackett et al., 2022). Compared to more cognitively-loaded selection methods, SJTs have less of an effect on important equity, diversity, and inclusion variables (e.g., small differences between members of differing socio-economic background), thus offering one of the best validity-diversity trade-offs amongst selection tools (Lievens et al., 2016; Ployhart & Holtz, 2008).

Applicant faking has been a vexing issue for selection researchers and practitioners alike (Morgeson et al., 2007). While SJTs are not immune to applicant faking (Nguyen et al., 2005), they are less susceptible compared to personality tests (Kasten et al., 2020). Fan et al., (2012) define faking as “the tendency to deliberately present oneself in a more positive manner than is accurate in order to meet the perceived demands of the testing situation” (p. 867). Faking is an important concern for hiring managers, organizations, or committees in charge of the admission to education programs for two main reasons. First, applicants who fake tend to be less qualified, less experienced, and have less desirable personality profiles (e.g., Melchers et al., 2020). Second, applicant faking can negatively impact the validity of selection tools (e.g., Wood et al., 2022) and can cause a mean shift and, therefore, increase the likelihood that applicants who fake get selected (Burns & Christiansen, 2011).

Many theoretical models (e.g., Levashina & Campion, 2006; McFarland & Ryan, 2006; Roulin et al., 2016) concur that faking depends on the combination of three elements: (1) applicants’ motivation or willingness to fake, (2) their ability or capacity to fake, and (3) situational or organizational factors that increase the opportunity (and/or decrease the risks) to fake. For instance, applicants who are able to identify the job, or program-relevant requirements (e.g., what personality traits are important to be a good doctor), have a higher capacity to fake, and do engage in more faking (Griffin & Wilson, 2012; Wood et al., 2022). Organizations can also try to design selection processes to reduce opportunities to fake. For instance, by introducing time constraints (e.g., Komar et al., 2010), warnings (e.g., Fan et al., 2012), or other countermeasures (e.g., Bill & Melchers, 2023). However, with the introduction of generative AI tools like ChatGPT, applicants' capacity to fake on a variety of selection methods might be enhanced.

Generative AI and Applicant Faking

The relationship between applicants and hiring organizations or test providers has been described as an "arms race", with the former using new technologies to help them fake and the latter developing countermeasures to prevent or detect faking attempts (Bangerter et al., 2012). Hiring organizations and test providers have an incentive to design assessments to limit applicants' use of generative AI, and thus reduce their opportunity to fake (Roulin et al., 2016). For instance, SJTs with open-ended questions could be vulnerable to applicants using generative AI to fake. Generative AI or LLM can be described as using machine learning models to "generate new content, including text, audio, video, images, software code and simulations" (Budhwar et al., 2023, p. 4). Perhaps the most popular tool to date is ChatGPT, an AI-powered "chatbot" developed by Open AI that can interact with the user and generate specific content or responses based on the questions, prompts, or documents provided. Version 3.5 of ChatGPT was originally released to the public in November 2022, with an updated and more efficient version (ChatGPT-4) released in March 2023.

Preliminary research has already demonstrated that ChatGPT and other generative AI tools can be prompted to effectively answer questions in a variety of tests and assessments. For instance, Phillips and Robie (2024) asked four generative AI tools (ChatGPT-3.5, ChatGPT-4, Jasper, and Google Bard) to complete a multiple-choice personality inventory after being provided with a job description. Not only were the AI tools able to effectively fake good, but ChatGPT-4 even outperformed business students attempting to fake their response for the same job. Borchert et al. (2023) found that ChatGPT was able to pass the UK's national SJT for medical school, which is a multiple-choice and ranking options-based test, with an overall score of 76%. Similar findings have been reported for other standardized tests (e.g., Geerling et al., 2023). These findings highlight that generative AI tools (e.g., ChatGPT) can generate relevant responses to a variety of assessments, and thus have the *potential* to increase applicants' capacity and/or opportunity to fake. Theoretically, applicants could use such tools to intentionally present themselves as more desirable professionals when completing assessments for selection. For instance, they could produce "better" answers for SJTs with either multiple choice or open-ended responses. In other words, Generative AI could increase applicants' capacity to fake (e.g., Roulin et al., 2016).

Nevertheless, there are two crucial limitations to the findings of the aforementioned studies. First, a methodological limitation is that researchers asked generative AI tools to answer each question using the same standard instructions (and the *right* prompts). This ideal scenario may not reflect the way real applicants use ChatGPT when completing a conventional SJT. For example, they might not always use it to generate complete answers (i.e., blatant, or extreme forms of faking), but rather to help them improve or "clean-up" their initial answer (i.e., a more subtle form of faking or exaggeration). Alternatively, applicants might use ChatGPT only for questions they are struggling with, and they might filter or edit suggested responses. They might also struggle to use the right prompts because they lack familiarity with the tool, or simply take a less structured approach to using ChatGPT to aid their answers. Second, tests and assessments can include several protections. For instance, SJTs can be designed to prevent applicants from copy/pasting content, to limit response time, and to include a proportion of video-based (vs. only text-based) scenarios, which are precautions likely to reduce applicants' opportunity to use ChatGPT to fake. Indeed, creating relevant prompts or instructions for ChatGPT based on SJT scenarios (especially video-based ones) and then integrating the proposed solution into one’s typed response might be challenging when time is restricted (i.e., similar to speeded assessments, see Komar et al, 2010). Overall, it remains unclear whether applicants utilize ChatGPT to try to fake an SJT, and whether such attempts are successful and leading to increased scores. This leads to our first research question:

RQ1: How does faking using generative AI help (or hinder) applicant SJT performance?

In addition, the extent to which AI can help with SJT performance might depend on applicants’ familiarity with such tools. A recent review highlighted that what is commonly referred to as “prompt engineering” (i.e., the ability to maximize input[prompt] to get the *right* output[answer]; Chen et al., 2023) is extremely important to obtain the best result when using generative AI. This review suggested that the most effective ways to prompt LLMs include several different components, such as providing clear, concise, and detailed instructions. Alternatively, it can be beneficial to prompt the LLM to take a specific role and behave a certain way (e.g., “I want you to answer the following questions as if you were a highly qualified applicant applying for a residency at [insert] hospital in [insert department]), or to try prompting the LLM several times. While some of these best practices are relatively surface-level, applicants could identify some more effective and less surface-level strategies through trial-and-error prompting (Dang et al., 2022). Additionally, Chen et al. (2023) described more advanced ways to prompt engineer that require either technical skills or more specific knowledge about LLMs in general (e.g., using plug-ins, creating, and using APIs, or using triple quotes to separate portions of prompt). Overall, exposure to (and experience with) using LLMs like ChatGPT may make applicants better able to use effective prompting strategies. This leads to our second research question:

RQ2: Does familiarity with generative AI moderate the effect of using AI on SJT performance?

Another area that may be affected by the use of ChatGPT (and other LLMs) is the content of applicants' responses and the way they are structured. LLMs are trained on content originally generated by humans and therefore may replicate observed differences in human-generated deceitful vs. honest language. For instance, studies have used LIWC (a linguistic software that creates scores of word categories) to distinguish honest from deceptive speech (e.g., Hirschberg et al., 2005). Recently, Van Der Zee et al., (2022) used this approach and found differences in proportion of positive and negative emotion words (e.g., related to love, happiness, or hope vs. anxiety, anger, or sadness) in factually correct vs. incorrect tweets. Moreover, AI might generate unique deceptive content. Zhou et al. (2023) compared AI-generated to human generated misinformation content (i.e., news and social media posts) and found several LIWC word categories where humans and AI differed. More specifically, AI-generated news included fewer words classified as analytical (i.e., signaling logical and formal thinking) and authentic (i.e., signaling honesty and genuineness), but more positive and negative emotion words. Overall, it is possible that AI-generated content will differ from authentically human-generated content, however, the specifics are not entirely clear. This led to our third research question:

RQ3: How do applicants' responses differ when using generative AI versus not?

Generative AI Detection

In addition to deterring applicant faking, researchers have explored a variety of tools and techniques to detect when applicants fake (e.g., Burns & Christiansen, 2011; Melchers et al., 2020). Conceptually, this represents an organization's attempt to increase the risks associated with faking (Roulin et al., 2016). In the context of generative AI, there may be ways to detect unique language or structure patterns in applicants' responses to identify attempts to fake using ChatGPT. Indeed, in parallel to the development of generative AI tools, AI detection tools like GPTZero or ZeroGPT have proliferated. Yet, initial research suggests that, while it is possible to detect some types of AI-generated content, there are important practical limitations. For example, such tools generate many "false positives" when used to detect AI-generated content in academic articles (Dalalah & Dalalah, 2023). They are also likely to misidentify responses from non-native English writers as AI-generated content (Liang et al., 2023). That said, more research is needed to further explore the efficacy of these AI detection tools in selection settings. We thus propose to examine the following research question:

RQ4: Are AI-detection tools effective to identify applicants' use of generative AI?

Context and Overview of the Studies

We examine these research questions in two complementary studies relying on the Casper, an open-ended SJT that assesses social intelligence and professionalism (e.g., Saxena et al., 2021). Casper is widely used in the admission process of health sciences, engineering, business education, teacher education, and social science programs in the US, Canada, UK and Australia (*Casper Technical Manual*, n.d.). Applicants are presented with 14 unique social dilemmas (scenarios). Although the test conventionally comprises both typed and video responses, we focused on typed responses only (i.e., nine SJT scenarios per applicant), as they are more likely to be prone to faking using Generative AI. Within the typed responses, Casper includes two types of scenarios: (a) video-based scenarios, where applicants watch relatively short videos from a first-person view and observe interactions between coworkers, fellow students, doctors, friends, etc.; and (b) text-based scenarios, where applicants read a brief statement about a situation. Applicants then respond to three questions specifically designed for each scenario. Typically, two questions focus on one main competency that the scenario is designed to evaluate, while one question will focus on a secondary competency (see Online Supplement for examples).

Study 1 uses a historical dataset with responses from a large sample of real applicants and performance scores from professional raters of Acuity Insights (i.e., the Casper provider). All applicants completed the Casper as a part of their admission applications for health sciences programs between June 2022 and May 2023. This allowed us to explore potential differences in response content, SJT scores, and AI detection before vs. after the release of ChatGPT, thus representing an initial examination of RQs 1, 3, and 4. Despite its benefits in terms of external validity, Study 1 only let us compare applicants who had (vs. did not have) the opportunity to use ChatGPT, but not whether they actually used such tools or not. In complement, Study 2 uses an experimental design with participants instructed to either use no external resources (i.e., control condition), use general online resources (excluding generative AI), or use ChatGPT exclusively. They then completed a shortened six-scenario Casper SJT and answered questions about their behaviors, thus providing more direct (and internally valid) ways to examine all four RQs.[1]

Study 1

Methods

Sample and Procedure

This study uses a large historical dataset with 107,805 real applicants, each with typed responses to nine scenarios (i.e., 969,242 datapoints). They represent the entire population of US and Canadian candidates who completed the Casper as part of their application for various health sciences programs (e.g., medicine, dentistry, nursing) between June 2022 and May 2023. Mean

[1] The two studies have been pre-registered (https://aspredicted.org/blind.php?x=38Z_653). We also provide detailed information in our Online Supplement (https://osf.io/4ncxt/?view_only=a0349e2901d841e7adcad3739c65a2a6).

age was 23.57 (*SD* = 7.01). The sample included 53.2% women and 24.3% men (22.5% N/A), 37.2% White, 9.5% East or South Asian, 6.2% Latino, 1.9% Black (44.6% N/A). The typed answers were content-coded using LIWC-22 to evaluate the type of language used (e.g., authentic words). Using this data allowed us to examine potential differences in applicants' responses or performance before vs. after the release of ChatGPT. Additionally, to test the effectiveness of AI detection tools, we randomly selected a subsample of 1,000 applicants and evaluated whether their responses included potentially AI-generated content using GPTZero (with a Python code developed using GPTZero's API).[2] Half of the randomly selected applicants completed the Casper prior to ChatGPT's release, while the other half did it after the release. This allowed us to examine both the potential effectiveness of using such a tool, as well as false positives (i.e., flagging content as AI-generated for applicants who completed the SJT prior to the release of ChatGPT).

Measures

ChatGPT Release. The effect of ChatGPT release/availability was captured in two ways: first, we coded the release of ChatGPT on Nov 30, 2022 (0 = before the release, 1 = after); second, we computed the number of months since the release (number of days since the release divided by 30, with all participants completing the SJT before the release coded as 0).

SJT Performance. Performance for each scenario was assessed by trained raters from Acuity Insights. Raters are instructed to read applicants' responses to all three questions attached to a scenario and assign a global score using a nine-point scale (i.e., poor to excellent – with explanations for each level). Ratings were also norm-referenced.

[2] The Python code was written by a professional software engineer and the code can be found on pastebin.com, a place to share code (https://pastebin.com/fY3t47xZ)

Response Content. Applicant answers were scored using LIWC-22 (Boyd et al., 2022). LIWC counts and categorizes words based on keywords from textual input, which in this case were answers to each SJT scenario. LIWC then generates proportion scores that represent the percentage of a given category in the text, with a higher scores indicating that the category is more prevalent, while also considering overall length of the textual input (Boyd et al., 2022). For instance, in the sentence "it felt really bad", "bad" is counted as a negative emotion word and represents 1/5 of the words in the sentence, so negative emotion would be scored 20%. We only included the three LIWC word categories that are most relevant in the context of faking, namely: authentic, positive emotion, and negative emotion words (see Zhou et al., 2023 for more detail). These categories were chosen given their connection to AI-generated as well as deceptive content.

Applicant and SJT Characteristics. To isolate the effect of ChatGPT on applicant responses or scores, we included various characteristics of the applicants and the SJT scenarios in analyses. At the applicant level, we included the type of academic program they applied (and completed Casper) for. This includes Canadian undergraduate health programs, Canadian medical schools, US undergraduate health sciences programs, US graduate health sciences programs, US medical residency programs, and US medical schools only (for more detail see supplemental material 1). At the SJT scenario level, we included the type of scenario used (text- vs. video-based), and the core competency assessed in each scenario from the nine-competency model targeted in all Casper SJTs (i.e., empathy, self-awareness, resilience, problem solving, motivation, ethics, equity, communication, and collaboration).

AI Detection. We used the two most relevant scores provided in GPTZero’s API in this context: First, a measure of “completely AI generated probability”, which is defined by GPTZero

(2023) as the overall probability that a document (i.e., an applicant's full response to all three questions related to a Casper scenario) was generated by AI. Second, GPTZero (2023) provides a "document classification", thus classifying SJT responses as either AI only, Human only, or Mixed (i.e., "either a certain section of this document has a strong signature of being AI-generated, or the overall document has a weak signature of being AI-generated"; GPTZero 2023).

Analyses

The structure of the data follows what is equivalent to a repeated measures design, with SJT scenario data (level 1; 969,242 observations) nested within applicants (level 2; n = 107,805). As such, multi-level modeling was conducted to examine the relationship between the release of ChatGPT and our outcome variables: SJT performance and response content using LIWC word categories (negative and positive emotion, and authentic words). The models also included the control variables described above: competency assessed in the scenario, scenario type, and academic program. Five comparative models were created following the general process suggested by Hox et al. (2018): starting with a null/empty model (M0 - with random intercepts), and subsequently entering control variables as fixed effects in blocks (M1 with programs, M2 with competencies assessed, M3 with video vs. text scenario), and adding months since ChatGPT's release in M4.[3] Additionally, M5 was tested for SJT scores, where the word count for applicant response was accounted for. This was only done for models predicting SJT scores, because LIWC scores take word count into account, as they are measured as a percentage (i.e., number of words from a category relative to the total number of words). To examine the effectiveness of GPTZero, similar models were tested to determine the effects of the months

[3] We also tested the same models using a dichotomous variable (i.e., pre- vs. post-ChatGPT release). The results were largely identical. We thus only present the findings using the "months since" variable.

since ChatGPT's release while controlling for applicant and SJT characteristics. Only fixed effects (not random slopes) were tested. We used the lme4 and lmerTest packages in R (see Bates et al., 2015; Kuznetsova et al., 2017) to obtain *p*-values to indicate statistical significance using the Satterthwaite's method to approximate degrees of freedom.

Results

Means, standard deviations, and correlations for study variables can be found in Table 1. Some correlations of note include the associations between word count and SJT score ($r = .62$, $p < .01$), word count and months since ChatGPT's release ($r = -.19$, $p < .01$), months since ChatGPT's release and complete AI generated probability ($r = -.12$, $p < .01$). Generally, these correlations show that longer responses are associated with higher SJT scores, response length seems to decrease after the release of ChatGPT, and the probability of AI-generated responses (per GPTZero) seems to decrease after the ChatGPT's release.

The intraclass corrections (ICC; variance explained at the applicant level) for each model were calculated using the null model for each outcome: $\mathrm{ICC}_{\mathrm{Score}} = .40$, $\mathrm{ICC}_{\mathrm{Authentic}} = .01$, $\mathrm{ICC}_{\mathrm{Negative\ Emotion}} = .02$, $\mathrm{ICC}_{\mathrm{Positive\ Emotion}} = .05$, and $\mathrm{ICC}_{\mathrm{Completey\ Generated\ by\ AI}} = .41$. This indicates that the percentage of variance explained at the applicant level was 40%, 1%, 2%, 5%, and 41% for each of the outcomes, respectively. We compared fit indices across models for all outcomes (see Table 2). The more comprehensive models were systematically the best fitting models (i.e., significantly different from/better than the previous models according to the Chi-square difference tests for log likelihood): Model 5 for SJT score ($\chi^2 = 236609.20$, $p < .001$) and completely generated by AI ($\chi^2 = 16.32$, $p < .001$); Model 4 (i.e., with all variables except word count) for authentic ($\chi^2 = 66.74$, $p < .001$), negative emotion ($\chi^2 = 12.93$, $p < .001$), and positive emotion ($\chi^2 = 80.27$, $p < .001$).

SJT Performance

As an initial attempt to explore RQ1, we examined the effect of ChatGPT release on SJT scores using both Model 4 and Model 5. We also briefly comment on the effect of SJT scenario type. Results are presented in Table 3, which also provides detailed information about the effect of programs or competencies on SJT scores. When examining Model 4 (i.e., without word count), the number of months since ChatGPT's release was slightly negatively associated with applicant SJT performance (b = -0.02, SE = 0.00, p < .001). In other words, with each passing month since ChatGPT became available, applicant scores were reduced by 0.02 points. Scores were also slightly higher for text-based than video-based scenarios (b = 0.04, SE = 0.00, p < .001).

However, given the protections in place in Casper (e.g., no copy/pasting, limited response time, virtual proctoring), applicants attempting to use ChatGPT and integrate the recommendations obtained in their responses might face difficulties. For instance, they might not have enough time to effectively do so. As such, we also explored whether response length changed once ChatGPT became available. We found that applicants provided shortened answers (b = -11.10, SE = 0.16, p < .001), that is, about 11 fewer words with each month since ChatGPT's release. This led us to re-examine RQ1 using Model 5, which includes word count as an additional variable. Word count was an impactful predictor of SJT performance (b = .019, SE = .00, p < .001), suggesting that each extra 100 words in applicants' responses led to 1.9 points increase in scores (recall that the SJT is scored 1-9). Importantly, the effect of the number of months since ChatGPT's release on SJT scores became stronger and turned positive (b = 0.14, SE = 0.02, p < .001), when controlling for word count.

SJT Responses Content

To explore RQ3, we examined the effect of ChatGPT availability on applicants' response content using the LIWC categories (see Table 4). The number of months since ChatGPT's release was associated with a lower percentage of authentic words in applicants' responses (*b* = -0.33, *SE* = 0.04, *p* < .001). Applicants also used more authentic words in text-based than video-based scenarios (*b* = 33.97, *SE* = 0.09, *p* < .001). In addition, the number of months since ChatGPT's release was associated with a slightly lower percentage of negative emotion words (*b* = -0.003, *SE* = 0.00, *p* < .001), and a slightly higher percentage of positive emotion words (*b* = 0.01, *SE* = 0.00, *p* < .001). Text-based SJT scenarios lead to a higher percentage of negative emotion words (*b* = 0.09, *SE* = 0.00, *p* < .001), and a higher percentage of positive emotion words (*b* = 0.07, *SE* = 0.00, *p* < .001) than video-based SJT scenarios.

AI Detection

When exploring RQ4 about the potential effectiveness of GPTZero as a tool to detect AI-generated responses in Casper, we first examined frequencies (see Table 5). Given that we have no objective data about which response included AI-generated content, we focused on false positive cases. GPTZero inaccurately flagged only 12 responses (out of 4500 responses; less than 1%) as being *completely AI generated* prior to ChatGPT's release. However, more problematically, it flagged 1748 responses as *mixed* before the release (around 38.84%). Responses that are classified as *mixed* have probabilities that the text was completely AI generated, ranging from 10 to 88%. It is evident that this may be problematic, as GPTZero flags many potential responses as containing some form of AI when Generative AI and LLMs were not yet in the spotlight, nor commonly in use. In addition, the number of responses flagged as *completely AI generated* (*n* = 10, < 1%) or *mixed* (*n* = 1291, about 29%) after the ChatGPT release were slightly lower than pre-release.

Furthermore, we examined the relationship between the number of months since ChatGPT's release and the GPTZero probability that responses were completely generated by AI in Model 4 (without word count) and Model 5 (with word count). Results are reported in Table 6. In both models, the probability slightly, but significantly, decreases as time passed following ChatGPT's release (e.g., b = -0.01, SE = 0.00, p < .001 in M5). Such a negative relationship would suggest *less* use of AI (according to GPTZero) as Generative AI tools became more available and popular. Interestingly, GPTZero indicated a higher probability of responses being generated by AI when applicants responded to text-based compared to video-based scenarios (e.g., b = 0.05, SE = 0.00, p < .001 in M5).

Discussion

Availability of Generative AI and Applicant Performance

The results of Study 1 based on a large sample of real applicants suggests that since the release of ChatGPT, applicants obtained no better (or even slightly lower) SJT scores, used a smaller percentage of authentic or negative emotion words, while using a higher percentage of positive emotion words. Ultimately, despite the ease of access to a mainstream Generative AI tools, such as ChatGPT, we observed a small decline in applicant performance. In contrast to more alarmist comments raised by companies and researchers alike (e.g., Arctic Shores, 2023; Borchert et al., 2023; Rudolph et al., 2023), the availability of ChatGPT did not help (and did even slightly hurt) applicants' performance on Casper. However, we also observed that applicants' response length (i.e., word count) decreased after ChatGPT's release by an average of 11 words every month, likely due to the protections embedded in the SJT (e.g., no copy/pasting, time restrictions). Response length was also positively associated with SJT scores. This finding suggests that applicants providing longer responses are better able to explain their reasoning (and/or provide more information about their competencies) and therefore receive a higher score. Further, when controlling for word count, the availability of ChatGPT slightly helped applicants: about 1.5% increase in performance scores per month.

Taken together, our findings suggest that the availability of generative AI tools like ChatGPT might have triggered two competing mechanisms: On the one hand, (some) applicants might use Generative AI to produce better-quality responses that helped them achieve higher performance scores. On the other hand, raters seem to reward longer and more detailed responses. Moreover, the SJT incorporated several elements preventing the effective use of AI (i.e., reducing the opportunity to fake; Roulin et al., 2016), such as making it impossible to copy and paste content and imposing strict time limits (five minutes to answer a set of three questions around each scenario). In addition, several scenarios are video-based, which might be more complex or time-consuming for applicants to translate into effective prompts for ChatGPT (or similar Generative AI chatbots). These factors might explain why applicants' responses were substantially shorter post ChatGPT's release, and why scores did not improve overall.

There are, of course, other possible explanations. For instance, only a small portion of applicants might have attempted to use ChatGPT, and the effect of this behavior on SJT scores might have been largely hidden by a majority of applicants not using this technology. In addition, ethical or moral reasons might have limited applicants' use of Generative AI to improve their responses (e.g., considering it "cheating" and thus unethical). Many applicants may have also lacked familiarity with the technology and struggle to use it. Alternatively, professional raters might have become stricter in their ratings as a reaction to the anticipated use of AI by applicants (although the company did not provide any instructions to do so). Study 2 was thus designed to better understand how applicants use ChatGPT to complete SJTs and what makes their strategy effective (vs. not).

Availability of Generative AI and Applicant Response Content

Study 1 findings suggest that the content of applicants' responses or the type of language used has changed only slightly post-ChatGPT release. They contained less authentic language and more positive emotion words, which is consistent with previous findings from Zhou et al. (2023), who showed that human text contained significantly more authentic words, while AI-generated text had more positive emotion words. However, contrary to Zhou et al. (2023) who found more negative emotion words in AI-generated content, Casper applicants used fewer negative words post ChatGPT release. Noteworthily, Study 1 did not allow us to directly connect changes in response content to Generative AI use. This was further examined in Study 2.

GPTZero False Positives

While we are unable to objectively identify if applicants actually used ChatGPT or other Generative AI tools (or who did) in the data, we used a popular AI detection tool (GPTZero) to explore its potential effectiveness. Overall, Study 1 findings suggest that GPTZero did not fare well. For instance, it classified 1,760 (out of 4,500) responses from applicants who completed the SJT *prior* to the release of ChatGPT as either being partially or completely generated by AI. Additionally, in multilevel models, the probability of a response being AI-generated *decreased* every month since ChatGPT's release, whereas we should logically expect that probability to increase as generative AI tools became more available and popular.

Study 2

Methods

Participants

We recruited 159 US or Canadian participants from Prolific with an undergraduate degree, to mimic real Casper applicants. Participants were excluded if they failed attention checks ("If you were to arrange the list of movies below in alphabetical order, which movie title would come first?" and "Please select Agree.") or seriousness check ("I answered the survey questions seriously."), leading to a final sample of 138. Mean age was 39.62 (SD = 14.16), with a majority of White participants (59.90%, with 15.30% Black, 3.60% Hispanic, 17.50% South or East Asian), who identified as 47.10% male, 52.20% female and 0.70% as non-binary.

Procedure

Participants were asked to imagine they were applying for a competitive medical school which they really want to be accepted into, and they were invited to complete an SJT as part of the admission process. They received detailed information about the SJT and a practice scenario with three questions (see Supplementary Materials 2 for more detail). They were then randomly assigned to one of three experimental conditions: (1) participants in the "control" condition were instructed to spend five minutes to prepare for the test by reflecting on their life (work, school, etc.) experiences and then complete the SJT honestly, without relying on any external resources; (2) those in the "online resources" condition were instructed to prepare using online resources available for Casper (we provided links to four example resources, including YouTube videos and Reddit posts), but excluding AI tools, and use only these online resources when completing the SJT; (3) those in the "ChatGPT" condition were instructed to practice prompting ChatGPT to help them provide better answers to the SJT (we provided links to ChatGPT tutorials, instructions to split screen, etc.) and then use it while completing the SJT.

Participants then completed a mock Casper SJT that included six scenarios (three video-based scenarios and three text-based scenarios), for which they were asked to type responses to three questions. We used “retired” scenarios/questions provided by Acuity Insights. The SJT was designed to mimic the actual Casper test experience. Participants were given 30 seconds to reflect after watching video scenarios (but no reflection time for text-based scenarios) and could not replay the scenarios; copying/pasting any text was blocked. They were given a maximum of five minutes to respond to the three questions for each scenario, after which they would automatically move to the next scenario (but had to spend a minimum of two minutes before they were allowed to move forward). After completing the SJT, participants answered questions about how they used their respective instructions to complete the SJT, their willingness and capacity to fake using ChatGPT, experience with ChatGPT, attention, seriousness, and manipulation checks, and some demographic questions[4].

Like in Study 1, SJT responses were content-coded using LIWC-22 and GPTZero to further address RQs 3 and 4. Responses were rated for performance by trained research assistants to examine RQs 1 and 2 – taking experience with ChatGPT into account.

Measures

SJT Performance. Participant responses were scored by two research assistants who were blind to the experimental conditions. Raters were trained using training resources provided by Acuity Insights, and responses to the three questions for each scenario were scored together on a 1-9 scale (*Casper Technical Manual*, n.d.). To mimic how Casper raters evaluate real applicants, each rater was assigned three of the six scenarios to rate, and rated all responses for one scenario before rating the next one. Inter-rater consistency was checked twice: Following the training, each rater scored all six scenarios for 10 randomly selected participants (Mean ICC (2,1) = .84). After all the rating was completed, raters scored another 10 participants for each

[4] See Online Supplement for all materials (e.g., instructions, experimental manipulation, items).

scenario they had not scored to confirm inter-rater consistency (Mean ICC (2,1) = .73; see Supplemental Table 6 for scenario-level ICCs).

Response Content. Similar to Study 1, responses were coded using LIWC-22 to capture the proportion of authentic, as well as negative and positive emotions words.

GPTZero. SJT responses were scored using GPTZero, similar to Study 1, focusing on two indicators: “document classification” and “completely AI generated probability”.

Willingness and Capacity to Fake. We measured participants’ willingness (2 items; α = .86; e.g., “I would be willing to use ChatGPT to cheat on a test like this”) and capacity (3 items; α = .83; e.g., “I could have provided inaccurate information from ChatGPT without anyone knowing”) to use ChatGPT to fake in the context of a test with items adapted from Law et al., (2016).

ChatGPT Questions. Experience with ChatGPT was measured by asking approximately how many times they have used ChatGPT.

Manipulation Checks. At the end of the study, participants answered a multiple-choice question about how they completed the SJT (i.e., responding honestly, using online resources, using ChatGPT). They also reported (using open textboxes) how they used their pre-SJT reflection/preparation time, their strategies when completing the SJT (e.g., how they used online resources or ChatGPT for those conditions), and any difficulties they experienced. We analyzed the data in two ways: first, we reviewed participants’ response to the open-ended questions and, when appropriate, reallocated participants into other conditions (e.g., some participants assigned to the ChatGPT condition refused to use it for ethical reasons), leading to a sample of $N = 138$. Second, we used a conservative approach by retaining only those participants who passed the multiple-choice manipulation check question (i.e., answered it according to the condition they were assigned to), leading to a sample of *N* = 101. The results were similar for both approaches. We thus report the findings for the "reallocation" approach here, and the more conservative one in our Online Supplement (see Supplemental Tables 3-5).

Analysis

We used a similar multi-level modeling approach as in Study 1, with SJT scenario data (level 1; 826 observations) nested within participants (level 2; *n* = 138). This was done to examine the relationship between instructions to use ChatGPT and our outcome variables (i.e., SJT performance, response content, completely AI generated probability). Like in Study 1, we created comparative models (see Table 7), starting with a null/empty model (M0 - with random intercepts), entering control variables as fixed effects (M1 with age, gender, ethnicity; M2 with video vs. text scenario). Experimental conditions were added in M3, and word count in M4 (only for SJT scores and completely AI generated probability, as LIWC outcomes already account for word count). For SJT scores, M5 included participants' experience with ChatGPT and the interaction between the ChatGPT condition and experience (to examine RQ2). Only fixed effects (not random slopes) were tested, again using the lme4 and lmerTest packages in R.

Results

The ICCs (variance explained at the participant level) for each model were calculated using the null model for each outcome: $\text{ICC}_{\text{Score}} = .37$, $\text{ICC}_{\text{Completely AI Generated}} = .62$, $\text{ICC}_{\text{Authentic}} = .07$, $\text{ICC}_{\text{Negative}} = .00$, $\text{ICC}_{\text{Positive}} = .02$. This indicates a large percentage of between participant variance for the first two outcomes. We first compared fit indices across models for all outcomes. Models 4 were the systematically best fitting models for SJT score ($\chi^2 = 237.43$, $p < .001$) and completely AI generated probability ($\chi^2 = 67.34$, $p < .001$). For SJT scores, Model 5 (including the ChatGPT experience moderation) was non-significantly superior to Model 4 ($\chi^2 = 1.79$, $p =$

.409). Additionally, Models 3 were *not* significantly better for any of the LIWC outcomes than Model 2. This suggests that the experimental conditions did not significantly contribute to response content, and we thus only report these findings in our Online Supplement (see Supplemental Table 2).

Research Questions

ChatGPT Use and Performance. Like in Study 1, we examined SJT scores both with and without word count in our models to explore RQ1. We also briefly comment on the effect of SJT scenario type. Results are presented in Table 8. In Model 3, participants in the ChatGPT condition scored half a point higher than those in the control (b = 0.53, SE = 0.20, p < .001). A similar pattern was found in Model 4 when controlling for word count, however the effect was halved (b = 0.27, SE = 0.11, p < .05). We also modeled conditions and scenario type to predict word count. Participants' responses did not include significantly more words in the online resource condition or the ChatGPT condition compared to the control condition. Similar to Study 1, word count was positively associated with SJT scores: 100 extra words increased scores by one full point (b = 0.01, SE = 0.00, p < .001). Regarding RQ2, including ChatGPT experience as a moderator did not improve model fit, and did not influence SJT scores.

Detecting ChatGPT. We first explored the frequencies and percentages of AI detection across our three experimental conditions using GPTZero to further examine RQ4 (see Table 9). GPTZero identified 1.99% of the responses from participants instructed to use ChatGPT as AI-generated and 55.78% as a mix of human and AI text. In addition, it labelled responses from those in the control condition as 1.17% AI-generated and 41.06% mixed (0.85% and 55.98% respectively, for the online resource condition). Moreover, Model 4 (in Table 8) showed that being in the ChatGPT condition was associated with a significantly higher completely AI generated probability (*b* = .07, *SE* = 0.03, *p* < .05). Although this finding suggests that GPTZero may be effective, the practical significance is only an increase of 7% compared to the control condition.

Exploratory Analyses

In addition to our main analyses, we also explored if participants experienced any difficulties using ChatGPT to complete the assessment. All participants in the ChatGPT condition (*n* = 42) were asked two open-ended questions relating to the use and difficulty of ChatGPT. Using the three main protections from Casper as a guide (i.e., use of video scenarios, disabling copy/pasting, and the time limit), we content-coded responses. Overall, one participant (2.40%) mentioned difficultly of translating the video into ChatGPT to prompt it to assist in answering the question, six participants (14.30%) mentioned difficultly surrounding copy/pasting from ChatGPT, and 16 participants (38.10%) mentioned difficulties related to time (i.e., not having enough time to go between and answer the questions using ChatGPT). Overall, this provides evidence that the protections that are designed in the Casper *do* indeed hamper the use of ChatGPT.

While examining participants' willingness and capacity to use ChatGPT to cheat on future SJTs like Casper, participants reported low levels for both ($M_{\text{Willingness}}$ = 1.84, *SD* = 0.96, $M_{\textit{Capacity}}$ = 2.74, *SD* = 1.09)[5]. So, while there may be an increased opportunity to cheat with the availability of Generative AI tools, it seems that people are largely unwilling (particularly) and feel limited capability of doing so.

[5] A between groups ANOVA revealed no significant difference between experimental conditions for capacity, however there was a significant difference for willingness, *F*(2,135) = 3.98, *p* = .021. This is likely due to some participants using ChatGPT directly prior to answering ($M_{\textit{ChatGPT}}$ = 2.15, $M_{\textit{Online Resources}}$ = 1.83, $M_{\textit{Control}}$ = 1.61).

In addition to testing the effectiveness of GPTZero on participants' SJT responses instructed (vs. not) to use ChatGPT, we tested it on responses directly provided by ChatGPT to the six scenario used in Study 2 (similar to past research like Borchert et al., 2023). For the video-based scenarios, we transcribed the content of the interactions between co-workers. We prompted with "ChatGPT, can you please read the discussion below [between two co-workers], and then tell me how the coworkers should address the situation and answer the three questions at the bottom?" We then pasted the transcript and the three questions. For text-based questions, we used the same prompt but with "[…] read the short scenario below, and then tell me how I should address the situation […]." We obtained responses from both ChatGPT 3.5 and ChatGPT-4 (using Bing in MS Edge). We then ran these (fully-AI-generated) responses in GPTZero, results are presented in Table 10. All six responses from ChatGPT 3.5 were classified as "likely to be a mix of human and AI text", with probability of the text being AI-generated ranging from 50 to 61%. Detection was slightly better for ChatGPT-4, with two responses to video-based scenarios as "likely to be written by AI" (91 and 92% AI probabilities), one as "moderately likely to be written AI" (67% AI probability), and all three responses to text-based scenarios as "likely to be a mix of human and AI text" (56-72% AI probabilities).

Discussion

Instructions to Use ChatGPT, SJT Performance, and Response Content

Results of Study 2 replicated some important findings from Study 1. For instance, we found that participants instructed to use ChatGPT only performed slightly better than those instructed to use no online resources. They also performed similarly to participants using online resources but excluding generative AI. SJT scores were about 0.5 points higher in the ChatGPT condition than the control (on a 1-9 scale). Interestingly, this positive effect was reduced when controlling for word count (contrary to Study 1), and participants in the ChatGPT condition neither provided shorter nor longer responses. This can be explained by differences in samples and context: Overall SJT scores were lower in Study 2 than in Study 1 (M_{Study1} = 5.05; M_{Study2} = 3.82) and responses were shorter (M_{Study1} = 196.96 words; M_{Study2} = 124.70 words). This was expected given that Study 2 included Prolific respondents who were unprepared to complete the Casper. Moreover, these respondents likely had a limited motivation to perform as compared to real applicants completing the test to be admitted into a health science program, which is an important next step in their education and career. Importantly, Study 2 participants did put some effort into their responses: they spent around four minutes answering questions for each scenario. Yet, most did not use the full five minutes allocated to them. Therefore, the protections embedded into Casper might have had a lesser effect to limit the opportunity of respondents using ChatGPT to benefit from it, as compared to the highly prepared and motivated applicants in Study 1. For instance, time restrictions are less effective to prevent the use of ChatGPT to produce better responses when respondents do not use all the time allocated anyway. That said, participants in the ChatGPT conditions still mentioned these protections as limiting factors. More specifically, 38.10% of participants mentioned difficulties surrounding the time limit, and 14.30% noted that not being able to copy/paste limited their effectiveness when using ChatGPT.

In terms of response content, contrary to Study 1 which found less authentic and negative emotions words (but more positive emotion words) post ChatGPT release, we found no differences in the proportion of those three types of words across conditions in Study 2. These findings could be explained by the same differences in motivation (and stakes) between the two samples described above. The overall lack of motivation may contribute to less personal and reflective content and may be reflected by the overall lower performance.

Detection Using GPTZero

Similar to Study 1, the findings of Study 2 provide limited support for the use of AI detection tools like GPTZero in the context of SJTs. Indeed, the proportion of responses identified as AI-only, or a mix of human and AI was largely similar for participants instructed to use ChatGPT vs. not. And, while the "completely generated AI probability" was significantly higher for responses in the ChatGPT condition, the difference was practically small (i.e., only 7%). In addition, GPTZero performed poorly on responses fully generated by ChatGPT, labelling most of them as a mix of human and AI (although it did slightly better for ChatGPT-4 responses than 3.5).

We also asked our raters to score those AI-generated responses (without knowing they were), alongside the participants' responses, and their scores were substantially higher than the average participants' scores (i.e., *M* = 7.50 for ChatGPT 3.5 and 7.83 for ChatGPT-4). This suggests that if applicants could use ChatGPT to its full potential (i.e., without any copy/pasting or time restrictions), the benefits in terms of SJT performance could be superior, thus further highlighting the importance of efforts to reduce opportunity to fake (e.g., Roulin et al., 2016).

General Discussion

Main Findings and Theoretical Implications

The emergence of Generative AI and LLMs represents a technological revolution that has the potential to meaningfully impact personnel selection and admission in higher education (Rudolph et al., 2023). Indeed, tools like ChatGPT can be prompted to provide correct or job-relevant responses to a variety of assessments (e.g., Borchert et al., 2023; Phillips & Robie, 2024). Yet, research examining how the availability of this technology influences applicants' behaviors or performance, and potential solutions for organizations, remains limited. The present paper thus aimed to contribute to the emerging literature on the role played by Generative AI and LLMs in selection/admission, and to help ensure that research is not (too) outpaced by practice.

We first examined how ChatGPT can impact applicant performance in SJTs, both when considering the availability of the technology (Study 1) and instructions to use it (Study 2). Across both studies, we observed a significant but small positive relationship between ChatGPT availability or use and SJT scores, especially when accounting for response length. Our findings can be interpreted in light of theories of applicant faking (e.g., Levashina & Campion., 2006; Roulin et al., 2016), which emphasize the importance of applicants' capacity, willingness and opportunity to fake. Findings from Study 2 illustrate how much ChatGPT *can* help applicant performance, showing only small effects (i.e., 0.53 score improvement on a 1-9 scale; half as small when controlling for word count). This provides evidence that ChatGPT has the potential to increase applicants' *capacity* to fake. This is also consistent with some preliminary evidence from assessment companies (e.g., Arctic Shores, 2023). Furthermore, participants in Study 2 reported an average capacity to fake using Generative AI in future assessments (*M* = 2.74 out of 5), with slightly higher capacity for participants who used ChatGPT in the study. Generative AI and LLMs therefore could challenge the benefits of SJTs as generally less susceptible to faking compared to similar assessment tools (e.g., Kasten et al., 2020). Additionally, in Study 2 we found that most participants used ChatGPT 3.5, which has demonstrated weaker performance on personality tests than ChatGPT 4.0 (Phillips & Robie, 2023). Applicants' capacity to fake could be further bolstered by using better LLMs (e.g., ChatGPT 4.0) or future models specifically developed for the purpose of answering a particular assessment. Interestingly, we did not find any evidence that experience moderated the relationship between the use of ChatGPT and performance. However, we did not directly examine the prompting strategies participants used, or whether some strategies were more effective than others. That said, it is possible that prompt engineering may matter less when several protections (i.e., limited time, lack of copy/paste functionality) are incorporated into the assessment itself.

What is perhaps more reassuring is the overall low *willingness* to use ChatGPT to fake answers in future assessments observed in Study 2 (*M* = 1.84). Limited willingness might explain the findings from Study 1 as well. Since that study was based on a large sample of real applicants, it might also represent a more realistic estimate of how much ChatGPT *does* help applicant performance. The effect of ChatGPT availability was small, with almost no effect overall (0.02 score reduction per month), and a positive but small effect (0.14 score improvement) when accounting for word count. The differences between models including and excluding word count in Study 1 also shed some light on the *opportunity* to fake using ChatGPT. Indeed, the availability of generative AI was associated with shorter answers to SJT questions and shorter answers led to lower performance ratings. When combined, the availability of ChatGPT was associated with slightly higher performance when accounting for word count. Consistent with predictions in Roulin et al.'s (2016) framework, these findings suggest that the investments made by test developers to make faking more difficult or risky (i.e., no copy/pasting, time restrictions, video-based scenario, online proctoring) can be beneficial. Indeed, it might prevent/ deter applicants from effectively using ChatGPT, because they are unable (or do not have enough time) to write relevant prompts and integrate suggested content when typing their responses. This explanation was also supported by examining the strategies used and difficulties experienced by participants in the ChatGPT condition in Study 2, many of whom noted the time limit or copy/pasting prevention.

We then explored whether (and how) ChatGPT could influence the content of applicants' responses in SJTs. We only found differences in content when examining responses from real applicants before vs. after the release of ChatGPT (in Study 1). Applicant responses included significantly less authentic words and negative emotion words, but more positive emotion words after the release of ChatGPT (though the effects and thus practical value were very small for the last two). However, no differences in content were found when participants were instructed to use ChatGPT (vs. not - in Study 2). The more limited use of authentic words in Study 1 was promising in this context since it should reflect lower honesty and genuineness according to the LIWC dictionary, and was consistent with recent work (Zhou et al., 2023). Yet, the overall inconsistent findings and small effects suggest that ChatGPT might not generate a unique type or style of content that meaningfully differs from human-generated responses. In addition, Tu et al., (2023) demonstrated that ChatGPT's responses change over time, and therefore trying to determine the response content that ChatGPT provides may be a futile exercise.

Finally, we examined the potential value of AI detection tools, relying on GPTZero, which is one of the most popular ones on the market. Both studies provided similar evidence and conclusions about the limited effectiveness of GPTZero. In Study 2, it was only slightly better than chance level at identifying responses using ChatGPT as being AI or human-AI mix (57.77% combined). This makes GPTZero largely equivalent to humans in their (in)ability to detect deception in everyday life (e.g., Bond & DePaulo, 2006) or applicant faking in job interviews (e.g., Roulin et al., 2015), despite the advantage one may suspect when thinking about AI detection tools (e.g., the ability to "reverse engineer" AI's content generation rules). Importantly, GPTZero led to many "false positives" across both studies, for example labelling more responses as a human-AI mix prior than after ChatGPT's release in Study 1. It also struggled to identify responses completely generated by ChatGPT.

Overall, these findings are largely concerning from a theoretical perspective, as this type of faking is likely not indicative of an applicant's job-relevant qualifications. For example, Marcus (2009) argues that faking involves two different skills that may be job relevant: (1) knowing what to say, and (2) knowing how to say it. Yet, the rapid advancement of LLMs and generative AI will likely lead to a general decline in the traditionally "skilled" faking. Moreover, this technological (r)evolution illustrates a key principle of signaling theory, namely an 'arms race' between applicants and organizations (Bangerter et al., 2012). If assessment and selection scholars and practitioners (or test developers) do not find better ways to deter or detect the use of AI to fake, escalation will likely continue. As a result, organizations' (and applicants') trust in vulnerable and unprotected assessments might diminish, as these assessments will no longer represent reliable signals of applicants' abilities. In some cases, it might result in organizations abandoning some assessments. Our findings suggest that detection does not currently seem to be an option, and we instead suggest that efforts be put into deterrence (i.e., making faking using ChatGPT a more difficult and/or risky task).

We also note a few theoretically and practically relevant findings related to the differences in SJT performance and response content between video- and text-based SJT scenarios. First, in line with previous research showing no performance differences between video- and text-based scenarios (Lievens & Sackett, 2006; Webster et al., 2020), we found very small and inconsistent differences in performance across the two studies (i.e., .04 points higher scores for text-based scenarios in Study 1, but .32 points lower scores in Study 2). Second, there were relatively large differences in the proportion of authentic words used by applicants between the two types of scenarios (i.e., 33.97% more authentic words in responses to text-based compared to video-based scenarios in Study 1, and 11.86% in Study 2). This could be because video-based scenarios require applicants to refer to a specific situation and characters (i.e., actors playing the role of colleagues, supervisors, etc.), leading to less genuine language content. However, differences in content were relatively small for negative and positive emotion words, with less than one word difference across studies. Taken together, these findings suggest that while applicants respond somewhat differently to video-based scenarios as compared to text-based scenarios, these differences in content have little to no impact on their SJT scores.

Practical Implications

Assessment providers, test developers, hiring organizations, and admission programs could be cautiously optimistic when considering our findings. Although ChatGPT certainly has the potential to help applicants and can thus increase their capacity to fake, we found low levels of willingness to use it in future assessments (in Study 2). We also noted limited improvements in performance when test-takers are instructed to use ChatGPT (in Study 2) and especially when examining the availability of such tools with real applicants (in Study 1). Additionally, findings from Study 2 demonstrated that having experience with ChatGPT does not seem to matter. Overall, organizations could explore three ways to deal with applicants' use of AI/LLMs in selection and assessment: (1) detect; (2) deter; or (3) incorporate.

The first option would be to develop effective methods to detect AI use. Although this option might look appealing to practitioners, our results suggest that the use of publicly available AI detection tools such as GPTZero is not a worthwhile endeavor. Not only is their detection accuracy limited, but they also generate an alarming number of "false positives". Additionally, AI detection tools bolster errors when classifying non-native English writers (i.e., falsely identifying non-native English writers as AI; Liang et al., 2023). These findings might not be surprising to generative AI or LLM experts. In fact, OpenAI (i.e., the organization responsible for ChatGPT) relinquished pursuit of their AI detector after roughly half a year (and until further notice), due to the limited accuracy of their detector (Dreibelbis, 2023). Given the current state of AI detection tools, we recommend that organizations avoid the use/implementation of such tools to identify test-takers as potential generative AI users (i.e., fakers). In addition, although there might be more potential for the development of assessment-specific detection tools, it might not take long for applicants to find ways to bypass such detection systems. As an example, applicants can easily modify randomness (or "temperature") settings in LLMs, which may reduce the effectiveness of detectors that rely on measures of "randomness" in the text.

The second option involves deterring, or at least limiting, the use of AI by applicants. This might be a promising approach, at least in the short term. For instance, our findings showing the limited impact of ChatGPT on performance were observed for a SJT that included many protections (e.g., time limits, limiting copy and paste functionality, video scenarios). Consistent with theoretical models (e.g., Roulin et al., 2016), we thus encourage organizations to implement similar preventative elements when designing assessments, to make the use of generative AI more difficult for applicants and thus limit their opportunity to fake. That said, rapidly evolving AI technology might also hinder the development and implementation of effective deterrents in the long term. As an example, organizations might be tempted to stick to exclusively recording applicants' responses via a webcam and microphone. However, technologies either exist already to help combat these (e.g., NVIDIA Broadcast can "fix" where a person's eyes are set to make "eye contact") or will eventually exist (e.g., live AI avatars that look like applicants and respond to questions using an LLM to live generate scripts).

A third and final option might entail incorporating AI tools as part of the selection process. One argument is that LLMs will likely act as ‘co-pilots’ or ‘virtual assistants’ in many jobs, and employees will be expected to use them in their daily work activities (e.g., summarizing data, writing email drafts, proof-reading documents). For such jobs, LLM usage can conceptually become part of their job performance (i.e., the criteria space), and thus excluding such tools from assessments (i.e., the prediction space) might be ill-advised. That said, this might require developing and incorporating local or task-specific LLMs, rather than giving access to ChatGPT or the like. And, of course, more research is needed to demonstrate the benefits and drawbacks of such an approach before it could be safely implemented in practice.

Limitations and Future research Directions

This research has several limitations, which might represent valuable avenues for future research. First, given the historical nature of Study 1 data, we were unable to identify which applicants used ChatGPT (and how or to what extent they used it), and can only speak about the effect of its availability. In addition, differences observed before and after the release of ChatGPT (e.g., in terms of SJT performance) could be affected by other/external factors, such as seasonal performance fluctuations or other resources available to applicants. It is also important to consider some of the significant but small effects in Study 1 in light of the large sample sizes and thus extensive statistical power.

Second, the sample used in Study 2 represents a limitation in several ways. Prolific participants were older than the typical applicants who complete the Casper for health science programs (*M* = 39.62, *SD* = 14.16). They might also have been less familiar with generative AI tools, like ChatGPT. That said, this sample might be more representative of a general applicant pool, therefore making our results more generalizable to a variety of selection contexts. These participants were also less prepared and motivated to perform on the SJT than real applicants. This explains the shorter responses and performance discrepancy with the applicant sample from Study 1. The lower overall performance across conditions in Study 2 might also partly explain the larger effects observed for ChatGPT (vs. Study 1), given that participants had more room for improvement. Alternatively, applicants might be more motivated to use all the time and resources available to improve their scores rather than Prolific participants who complete a study for money. Although there would certainly be logistical difficulties, future research should examine the use of ChatGPT by actual applicants in a high-stakes context.

A third and final limitation is related to the ever-changing nature of Generative AI and LLMs (Tu et al., 2023) and the nature of selection as a whole. Consistent with the “arms race” prediction by Bangerter et al. (2012), Generative AI/LLMs, applicants, and organizations or test providers are expected to learn and adapt over time. New generations of LLMs will likely become better at answering assessment questions, especially as more job applicants use them to practice or when completing various assessments. For instance, recent versions of LLM like ChatGPT-4 are now able to analyze images or listen to audio, which might help them with video-based SJT scenarios. This could further applicants’ capacity to fake and contribute to widening the gap between those who choose to fake using Generative AI and those who do not. In response, test developers might be pressured to implement more advanced protections. New SJT formats (e.g., video-recording answers vs. typing them) or different forms of assessments (e.g., game-based), less prone to the effects of Generative AI, might become more popular in the future. As well, AI-detection tools might become more effective. We thus recommend that researchers and organizations alike continue to monitor progress in generative AI, and for future studies to replicate and expand our findings as new LLMs and assessments emerge.

Conclusions

Findings from two studies (field data from a large sample of real applicants and an experiment) suggest that the availability of ChatGPT can potentially help applicants provide stronger answers and obtain slightly higher scores on SJTs. However, this positive effect can be countered by design elements that make AI use more difficult, transforming the use of ChatGPT into a zero-sum game. Applicants' responses included slightly more authentic words since the release of ChatGPT, but not when instructed to use generative AI. AI detection tools were also limited in their ability to identify ChatGPT use. However, it is important to consider the early landscape this research is positioned in, its exploratory nature, and the need for continuing efforts for research to keep pace with developments in AI.

References

Arctic Shores. (2023, September 13). *ChatGPT vs Situational Judgement Tests: How it performs vs a human*. https://www.arcticshores.com/insights/chatgpt-vs-situational-judgement-tests-how-it-performs-vs-a-human

Bangerter, A., Roulin, N., & König, C. J. (2012). Personnel selection as a signaling game. *Journal of Applied Psychology*, *97*(4), 719–738. https://doi.org/10.1037/a0026078

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bill, B., & Melchers, K. G. (2023). Thou Shalt not Lie! Exploring and testing countermeasures against faking intentions and faking in selection interviews. *International Journal of Selection and Assessment*, *31*(1), 22–44. https://doi.org/10.1111/ijsa.12402

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of Deception Judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2

Borchert, R. J., Hickman, C. R., Pepys, J., & Sadler, T. J. (2023). Performance of ChatGPT on the Situational Judgement Test—A Professional Dilemmas–Based Examination for Doctors in the United Kingdom. *JMIR Medical Education*, *9*. https://doi.org/10.2196/48978

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22.* University of Texas at Austin. https://www.liwc.app

Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., Boselie, P., Cooke, F. L., Decker, S., DeNisi, A., Dey, P. K., Guest, D., Knoblich, A. J., Malik, A., Paauwe, J., Papagiannidis, S., Patel, C., Pereira, V., Ren, S., … Varma, A. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal*, *33*(3), 1–54. https://doi.org/10.1111/1748-8583.12524

Burns, G. N., & Christiansen, N. D. (2011). Methods of Measuring Faking Behavior. *Human Performance*, *24*(4), 358–372. https://doi.org/10.1080/08959285.2011.597473

Casper Technical Manual. (n.d.). Acuity Insights. Retrieved August 22, 2023, from https://view-su2.highspot.com/viewer/64c0083d47c548a5b18d7d89

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). *Unleashing the potential of prompt engineering in Large Language Models: A comprehensive review* (arXiv:2310.14735). arXiv. http://arxiv.org/abs/2310.14735

Corstjens, J., Lievens, F., & Krumm, S. (2017). Situational Judgement Tests for Selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection and Employee Retention* (1st ed., pp. 226–246). Wiley. https://doi.org/10.1002/9781118972472.ch11

Dalalah, D., & Dalalah, O. M. A. (2023). The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. *The International Journal of Management Education*, *21*(2), 100822. https://doi.org/10.1016/j.ijme.2023.100822

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). *How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models* (arXiv:2209.01390). arXiv. http://arxiv.org/abs/2209.01390

Dreibelbis, E. (2023, July 25). *OpenAI Quietly Shuts Down AI Text-Detection Tool Over Inaccuracies*. https://www.pcmag.com/news/openai-quietly-shuts-down-ai-text-detection-tool-over-inaccuracies

Fan, J., Gao, D., Carroll, S. A., Lopez, F. J., Tian, T. S., & Meng, H. (2012). Testing the efficacy of a new procedure for reducing faking on personality tests within selection contexts. *Journal of Applied Psychology*, *97*(4), 866–880. https://doi.org/10.1037/a0026655

Geerling, W., Mateer, G. D., Wooten, J., & Damodaran, N. (2023). ChatGPT has Aced the Test of Understanding in College Economics: Now What? *The American Economist*, *68*(2), 233–245. https://doi.org/10.1177/05694345231169654

Goss, B. D., Ryan, A. T., Waring, J., Judd, T., Chiavaroli, N. G., O'Brien, R. C., Trumble, S. C., & McColl, G. J. (2017). Beyond Selection: The Use of Situational Judgement Tests in the Teaching and Assessment of Professionalism. *Academic Medicine*, *92*(6), 780–784. https://doi.org/10.1097/ACM.0000000000001591

GPTZero. (2023). *AI-detection on an array of files*. https://gptzero.stoplight.io/docs/gptzero-api/0a8e7efa751a6-ai-detection-on-an-array-of-files

Griffin, B., & Wilson, I. G. (2012). Faking good: Self-enhancement in medical school applicants: Self-enhancement in medical students. *Medical Education*, *46*(5), 485–490. https://doi.org/10.1111/j.1365-2923.2011.04208.x

Hirschberg, J., Benus, S., Brenier, J. M., Enos, F., Friedman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L., Pellom, B. L., Shriberg, E., & Stolcke, A. (2005). Distinguishing deceptive from non-deceptive speech. *Interspeech 2005*, 1833–1836. https://doi.org/10.21437/Interspeech.2005-580

Hox, J. J., Moerbeek, M., & Van De Schoot, R. (2018). The Basic Two-Level Regression Model. In J. J. Hox, M. Moerbeek, & R. Van De Schoot, *Multilevel Analysis* (3rd ed., pp. 8–26). Routledge. https://doi.org/10.4324/9781315650982-2

Kasten, N., Freund, P. A., & Staufenbiel, T. (2020). "Sweet Little Lies": An In-Depth Analysis of Faking Behavior on Situational Judgment Tests Compared to Personality Questionnaires. *European Journal of Psychological Assessment*, *36*(1), 136–148. https://doi.org/10.1027/1015-5759/a000479

Kolade, S. (n.d.). *Has ChatGPT Signalled the End of Assessment as We Know It?* Retrieved August 21, 2023, from https://charteredabs.org/has-chatgpt-signalled-the-end-of-assessment-as-we-know-it/

Komar, S., Komar, J. A., Robie, C., & Taggar, S. (2010). Speeding Personality Measures to Reduce Faking: A Self-Regulatory Model. *Journal of Personnel Psychology*, *9*(3), 126–137. https://doi.org/10.1027/1866-5888/a000016

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Law, S. J., Bourdage, J., & O'Neill, T. A. (2016). To Fake or Not to Fake: Antecedents to Interview Faking, Warning Instructions, and Its Impact on Applicant Reactions. *Frontiers in Psychology*, *7*, 1771. https://doi.org/10.3389/fpsyg.2016.01771

Levashina, J., & Campion, M. A. (2006). A Model of Faking Likelihood in the Employment Interview. *International Journal of Selection and Assessment*, *14*(4), 299–316. https://doi.org/10.1111/j.1468-2389.2006.00353.x

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). *GPT detectors are biased against non-native English writers* (arXiv:2304.02819). arXiv. http://arxiv.org/abs/2304.02819

Lievens, F. (2013). Adjusting medical school admission: Assessing interpersonal skills using situational judgement tests: Interpersonal skills and medical school admission. *Medical Education*, *47*(2), 182–189. https://doi.org/10.1111/medu.12089

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The Operational Validity of a Video-Based Situational Judgment Test for Medical College Admissions: Illustrating the Importance of Matching Predictor and Criterion Construct Domains. *Journal of Applied Psychology*, *90*(3), 442–452. https://doi.org/10.1037/0021-9010.90.3.442

Lievens, F., Patterson, F., Corstjens, J., Martin, S., & Nicholson, S. (2016). Widening access in selection using situational judgement tests: Evidence from the UKCAT. *Medical Education*, *50*(6), 624–636. https://doi.org/10.1111/medu.13060

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology*, *91*(5), 1181–1188. https://doi.org/10.1037/0021-9010.91.5.1181

McFarland, L. A., & Ryan, A. M. (2006). Toward an Integrated Model of Applicant Faking Behavior. *Journal of Applied Social Psychology*, *36*(4), 979–1016. https://doi.org/10.1111/j.0021-9029.2006.00052.x

Melchers, K. G., Roulin, N., & Buehl, A. (2020). A review of applicant faking in selection interviews. *International Journal of Selection and Assessment*, *28*(2), 123–142. https://doi.org/10.1111/ijsa.12280

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering The Use of Personality Tests in Personnel Selection Contexts. *Personnel Psychology*, *60*(3), 683–729. https://doi.org/10.1111/j.1744-6570.2007.00089.x

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of Response Instructions on Faking a Situational Judgment Test. *International Journal of Selection and Assessment*, *13*(4), 250–260. https://doi.org/10.1111/j.1468-2389.2005.00322.x

Patterson, F., Galbraith, K., Flaxman, C., & Kirkpatrick, C. M. J. (2019). Evaluation of a Situational Judgement Test to Develop Non-Academic Skills in Pharmacy Students. *American Journal of Pharmaceutical Education*, *83*(10), 2092–2101. https://doi.org/10.5688/ajpe7074

Phillips, J., & Robie, C. (2024). Can a computer outfake a human? *Personality and Individual Differences*, *217*, 112434. https://doi.org/10.1016/j.paid.2023.112434

Ployhart, R. E., & Holtz, B. C. (2008). The Diversity–Validity Dilemma: Strategies for Reducing Racioethnic and Sex Subgroup Differences and Adverse Impact in Selection. *Personnel Psychology*, *61*(1), 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x

Roulin, N., Bangerter, A., & Levashina, J. (2015). Honest and Deceptive Impression Management in the Employment Interview: Can It Be Detected and How Does It Impact Evaluations? *Personnel Psychology*, *68*(2), 395–444. https://doi.org/10.1111/peps.12079

Roulin, N., Krings, F., & Binggeli, S. (2016). A dynamic model of applicant faking. *Organizational Psychology Review*, *6*(2), 145–170. https://doi.org/10.1177/2041386615580875

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching*, *6*(1). https://doi.org/10.37074/jalt.2023.6.1.9

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, *107*(11), 2040–2068. https://doi.org/10.1037/apl0000994

Sahota, G. S., & Taggar, J. S. (2020). The association between Situational Judgement Test (SJT) scores and professionalism concerns in undergraduate medical education. *Medical Teacher*, *42*(8), 937–943. https://doi.org/10.1080/0142159X.2020.1772466

Saxena, A., Desanghere, L., Dore, K., & Reiter, H. (2021). Incorporating situational judgment tests into postgraduate medical education admissions: Examining educational and organizational outcomes. *Academic Medicine*, *96*(11S), S203–S204. https://doi.org/10.1097/ACM.0000000000004280

Stanley, D. J. (2022). *Package "apaTables"* [Computer software]. https://cran.r-project.org/web/packages/apaTables/apaTables.pdf

Tiffin, P. A., Sanger, E., Smith, D. T., Troughton, A., & Paton, L. W. (2022). Situational judgement test performance and subsequent misconduct in medical students. *Medical Education*, *56*(7), 754–763. https://doi.org/10.1111/medu.14801

Tu, S., Li, C., Yu, J., Wang, X., Hou, L., & Li, J. (2023). *ChatLog: Recording and Analyzing ChatGPT Across Time* (arXiv:2304.14106). arXiv. http://arxiv.org/abs/2304.14106

Van Der Zee, S., Poppe, R., Havrileck, A., & Baillon, A. (2022). A Personal Model of Trumpery: Linguistic Deception Detection in a Real-World High-Stakes Setting. *Psychological Science*, *33*(1), 3–17. https://doi.org/10.1177/09567976211015941

Webster, E. S., Paton, L. W., Crampton, P. E. S., & Tiffin, P. A. (2020). Situational judgement test validity for selection: A systematic review and meta-analysis. *Medical Education*, *54*(10), 888–902. https://doi.org/10.1111/medu.14201

Wood, J. K., Anglim, J., & Horwood, S. (2022). Effect of job applicant faking and cognitive ability on self-other agreement and criterion validity of personality assessments. *International Journal of Selection and Assessment*, *30*(3), 378–391. https://doi.org/10.1111/ijsa.12382

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3544548.3581318

Table 1

Means, Standard Deviations, and Correlations for Study 1 Variables.

Variable	***M***	***SD***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1. Self-awareness	0.11	0.31											
2. Resilience	0.11	0.31	-.12**										
3. Problem Solving	0.11	0.31	-.12**	-.12**									
4. Motivation	0.11	0.31	-.12**	-.12**	-.12**								
5. Ethics	0.12	0.32	-.13**	-.13**	-.13**	-.13**							
6. Equity	0.11	0.31	-.12**	-.12**	-.12**	-.12**	-.13**						
7. Communication	0.11	0.31	-.13**	-.12**	-.13**	-.13**	-.13**	-.12**					
8. Collaboration	0.11	0.31	-.12**	-.12**	-.12**	-.12**	-.13**	-.12**	-.13**				
9. Empathy	0.10	0.30	-.12**	-.12**	-.12**	-.12**	-.12**	-.12**	-.12**	-.12**			
10. Scenario Type	0.33	0.47	.30**	.24**	-.08**	-.07**	-.13**	-.03**	-.03**	.02**	-.21**		
11. Canada undergrad.	0.11	0.31	.00	.00*	.00	.00	.01**	-.02**	.00	.00	.01**	.00	
12. Canada medical	0.20	0.40	.00	-.01**	.00	.00	-.01**	.00**	.00	.00	.03**	.00	-.17**
13. US undergrad.	0.01	0.12	.00	.00	.00	.00	-.00**	.00	.00	.00	.00**	.00	-.04**
14. US graduate	0.26	0.44	.00	.00**	.00	.00	.00	.00**	.00	.00	-.02**	.00	-.20**
15. US residency	0.06	0.24	.00	.00	.00	.00	-.01**	.00	.00	.00	.01**	.00	-.09**
16. Months Since GPT	0.46	1.02	.00	.00*	.00	.00	.01**	-.02**	.00	.00	.02**	.00	.55**
17. Word Count	196.96	58.19	.02**	.02**	-.02**	.01**	-.04**	-.04**	.00	.01**	.03**	.00**	-.20**
18. SJT Score	5.05	1.67	.01**	.01**	-.02**	.01**	-.02**	-.01**	-.01**	.02**	.02**	.01**	-.03**
19. Negative emotions	0.50	0.69	.00**	.08**	-.04**	-.06**	-.06**	-.06**	.05**	-.02**	.10**	.06**	.01**
20. Positive emotions	0.59	0.75	.05**	.10**	-.06**	.09**	-.11**	-.05**	.00**	-.02**	.01**	.08**	.02**
21. Authentic	36.40	31.35	.31**	.20**	-.06**	-.00*	-.12**	.00**	-.11**	-.06**	-.15**	.58**	.03**
22. Completely AI Generated[1]	0.14	0.21	.05**	.02*	-.02*	0.01	-.05**	-0.02	0.01	-0.01	0.02	.11**	NA

Table 1 (continued)

Variable	12	13	14	15	16	17	18	19	20	21
13. US undergrad.	-.06**									
14. US graduate	-.29**	-.07**								
15. US residency	-.13**	-.03**	-.15**							
16. Months Since GPT	.13**	.06**	-.15**	-.11**						
17. Word Count	.11**	-.02**	-.16**	-.02**	-.19**					
18. SJT Score	.08**	.00	-.07**	.00	-.01**	.62**				
19. Negative emotions	-.02**	.00	.01**	.00	.00**	.03**	.01**			
20. Positive emotions	.02**	.00	.01**	-.03**	.03**	-.01**	-.00**	.05**		
21. Authentic	-.06**	.01**	.02**	-.02**	-.00*	.02**	.02**	.00	.02**	
22. Completely AI Generated^{1}	NA	.04**	-.09**	NA	-.12**	.13**	.12**	.02*	.05**	.08**

Note. ^{1} used a subset of the data (*n* = 9000 responses), NA = some correlations not possible due to subset being used. ** *p* <.01, * *p* < .05. Months Since GPT = number of months since ChatGPT release. SJT Scenario type coded as 0 = Video, 1 = Text. This table used the apaTables package in R (Stanley, 2022).

Table 2

Comparison of Multilevel Models for all Outcome Variables.

Outcome	Model	AIC	BIC	Loglikelihood	χ2	*df*
SJT Score	M0	3458445	3458481	-1729220	-	-
	M1	3456428	3456522	-1728206	2027.52***	5.00
	M2	3453530	3453730	-1726748	2916.01***	9.00
	M3	3453392	3453604	-1726678	139.27***	1.00
	M4	3453371	3453594	-1726666	23.78***	1.00
	M5	3216763	3216999	-1608362	236609.20***	1.00
Authentic	M0	9428143	9428178	-4714068	-	-
	M1	9423672	9423766	-4711828	4481.09***	5.00
	M2	9234121	9234321	-4617043	189569.24***	9.00
	M3	8919456	8919680	-4459709	314668.99***	2.00
	M4	8919391	8919627	-4459675	66.74***	1.00
Negative Emotion	M0	2029707	2029742	-1014851	-	-
	M1	2029416	2029510	-1014700	301.06***	5.00
	M2	2000362	2000563	-1000164	29071.69***	9.00
	M3	1997573	1997785	-998768	2791.69***	1.00
	M4	1997562	1997786	-998762	12.93***	1.00
Positive Emotion	M0	2175084	2175119	-1087539	-	-
	M1	2173830	2173924	-1086907	1264.21***	5.00
	M2	2136922	2137122	-1068444	36926.30***	9.00
	M3	2135158	2135370	-1067561	1765.81***	1.00
	M4	2135079	2135303	-1067521	80.27***	1.00
Completely AI Generated[1]	M0	-5552.36	-5531.05	2779.18	-	-
	M1	-5567.04	-5531.51	2788.52	18.68***	2.00
	M2	-5677.34	-5584.97	2851.67	126.30***	8.00
	M3	-5832.87	-5733.40	2930.43	157.53***	1.00
	M4	-5852.69	-5746.12	2941.34	21.82***	1.00
	M5	-5867.01	-5753.33	2949.50	16.32***	1.00

Note. *** $p < .001$, ** $p < .01$. M0 random intercepts models (null); M1 adding education program type; M2 adding competencies assessed, M3 adding video vs. text scenario; M4 adding months since ChatGPT's release M5 adding word count.[1] Depicts models with a sub sample n = 1,000 (9000 responses), otherwise used full sample N = 107,805 applicants (969,242 responses).

Table 3

Fixed Effects Models Predicting SJT Scores and Word Count (Study 1).

	SJT Score (Model 4)				SJT Score (Model 5)				Word Count			
Predictors	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*
Intercept	5.01	0.02	966200	294.85***	1.01	0.02	915500	61.55***	20.02	.18	107800	1148.14***
Months Since GPT	-0.02	0.00	107800	-4.88***	0.14	0.00	107800	66.77***	-11.10	.16	107800	-70.73***
Self-awareness	0.07	0.02	885400	4.39***	-0.04	0.02	937600	-2.71**				
Resilience	0.11	0.02	885500	6.53***	-0.02	0.02	937800	-1.08				
Problem Solving	-0.03	0.02	885800	-1.69	-0.02	0.02	938600	-1.21				
Motivation	0.11	0.02	885800	6.34***	0.03	0.02	938500	1.87				
Ethics	-0.03	0.02	886200	-1.58	0.06	0.02	939800	3.82***				
Equity	-0.00	0.02	885800	-0.08	0.10	0.02	938800	6.25***				
Communication	0.03	0.02	885700	1.75	-0.02	0.02	938500	-1.12				
Collaboration	0.17	0.02	885700	10.18***	0.09	0.02	938400	6.06***				
Empathy	0.14	0.02	888600	8.46***	-0.00	0.02	945500	-0.27				
Scenario Type	0.04	0.00	861600	11.80***	0.06	0.00	861400	19.08***				
Canada undergrad.	-0.13	0.02	107900	-8.53***	0.48	0.01	108400	66.42***				
Canada medical	0.23	0.01	107800	22.97***	0.18	0.01	107300	37.84***				
US undergrad.	-0.01	0.03	107700	-0.35	0.29	0.02	107100	19.56***				
US graduate	-0.22	0.01	107800	-24.73***	0.35	0.00	109200	78.88***				
US residency	-0.04	0.02	107800	-2.85**	0.35	0.01	107500	47.26***				
Word Count					0.02	0.00	185300	661.15***				

Note. *N* = 107,805 applicants (969,242 responses). *** *p* < .001, ** *p* <.01, * *p* < .05. Months Since GPT = number of months since ChatGPT release; All competencies compared to NA (Not applicable). All program types compared to the US Medicine. SJT Scenario type coded as 0 = Video, 1 = Text.

Table 4

Fixed Effect Models for LIWC Variables (Study 1).

	Authentic				Negative emotions				Positive emotions			
Predictors	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*
Intercept	12.11	0.29	964200	41.85***	0.83	0.01	968600	101.87***	0.44	0.01	966600	50.95***
Months Since GPT	-0.35	0.04	107700	-8.71***	-0.00	0.00	107600	3.60***	0.01	0.00	107700	8.96***
Self-awareness	27.33	0.30	933800	91.95***	-0.38	0.01	955900	-45.10***	0.18	0.01	946000	19.75***
Resilience	21.15	0.30	934000	71.30***	-0.23	0.01	956100	-27.24***	0.28	0.01	946200	31.43***
Problem Solving	12.72	0.30	934700	43.18***	-0.42	0.01	956700	-50.06***	-0.03	0.01	946900	-2.90**
Motivation	17.08	0.30	934700	57.97***	-0.47	0.01	956700	-56.35***	0.30	0.01	946900	33.81***
Ethics	9.00	0.29	935700	30.62***	-0.45	0.01	957700	-54.07***	-0.12	0.01	948000	-13.13***
Equity	15.74	0.30	934600	53.39***	-0.46	0.01	956700	-55.28***	0.00	0.01	946900	-0.04
Communication	5.82	0.30	934600	19.75***	-0.25	0.01	956700	-30.01***	0.11	0.01	946800	12.56***
Collaboration	7.90	0.30	934500	26.77***	-0.39	0.01	956600	-46.84***	0.06	0.01	946800	6.90***
Empathy	9.84	0.30	941600	33.19***	-0.12	0.01	962700	-14.82***	0.15	0.01	953800	17.00***
Scenario Type	33.97	0.06	861900	602.64***	0.09	0.00	862000	52.89***	0.07	0.00	861900	42.05***
Canada undergrad.	2.46	0.14	108200	17.45***	0.01	0.00	108300	1.88	0.04	0.00	108200	9.37***
Canada medical	-4.67	0.10	107900	-49.00***	-0.03	0.00	107800	-15.07***	0.05	0.00	107800	17.619***
US undergrad.	1.96	0.29	107700	6.78***	-0.01	0.01	107500	-1.47	0.00	0.01	107600	0.14
US graduate	0.14	0.09	107800	1.61	0.00	0.00	107700	0.14	0.03	0.00	107700	12.47***
US residency	-3.54	0.14	107800	-24.52***	-0.01	0.00	107700	-1.44	-0.07	0.00	107700	-17.66***

Note. N = 107,805 applicants (969,242 responses). *** *p* < .001, ** *p* <.01, * *p* < .05. Months Since GPT = number of months since ChatGPT release; All competencies compared to NA (Not applicable). All program types compared to the US Medicine. SJT Scenario type coded as 0 = Video, 1 = Text.

Table 5

GPTZero Response Classification Pre- vs. Post-ChatGPT Release (Study 1).

	Pre ChatGPT's Release		Post ChatGPT's Release	
	Frequencies	Percentages	Frequencies	Percentages
AI Only	12	0.26	10	0.22
Mixed	1748	38.84	1291	28.69
Human Only	2740	60.90	3199	71.09
Total	4500	100	4500	100

Note. N = 1000 applicants (9000 responses). AI only (Completely AI generated probability > 0.88), Human only (Completely AI generated probability < 0.10).

Table 6

Fixed Effect Models for GPTZero Completely AI Generated Probability (Study 1).

	Model 4				Model 5			
Predictors	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*
Intercept	0.17	0.01	3016	19.72***	0.12	0.02	3594	8.42***
Months Since GPT	-0.01	0.00	1003	-4.70***	-0.01	0.00	1009	-4.35***
Self-awareness	0.00	0.01	8027	0.02	-0.00	0.01	8042	-0.26
Resilience	-0.01	0.01	8028	-0.78	-0.01	0.01	8026	-0.79
Problem Solving	-0.02	0.01	8029	-2.36*	-0.02	0.01	8060	-2.15*
Motivation	-0.02	0.01	8027	-2.98**	-0.02	0.01	8044	-2.83**
Ethics	-0.05	0.01	8077	-6.42***	-0.05	0.01	8086	-6.34***
Equity	-0.02	0.01	8029	-2.25*	-0.02	0.01	8068	-2.01*
Communication	-0.04	0.01	8025	-4.85***	-0.04	0.01	8023	-4.85***
Collaboration	-0.03	0.01	8027	-4.07***	-0.03	0.01	8027	-4.06***
Scenario Type	0.05	0.00	8001	12.60***	0.05	0.00	8021	12.34***
US undergrad.	0.02	0.02	1000	0.62	0.02	0.02	998	0.73
US graduate	-0.02	0.01	1000	-2.09*	-0.02	0.01	1016	-1.56
Word Count					0.00	0.00	4085	4.06***

Note. *N* = 1000 applicants (9000 responses). *** *p* < .001, ** *p* <.01, * *p* < .05. Months Since GPT = number of months since ChatGPT release; All competencies compared to Empathy (i.e., NA was not present in sub sample). Program type (i.e., US undergraduate and graduate health science only present in sub sample) compared to the US Medicine. SJT Scenario type coded as 0 = Video, 1 = Text.

Table 7

Model comparisons for participants' score and completely AI generated probability (Study 2).

Outcome	Model	AIC	BIC	Loglikelihood	χ2	*df*
SJT Score	M0	2832.90	2847.10	-1413.50	-	-
	M1	2815.70	2843.90	-1401.80	23.28***	3
	M2	2795.00	2828.00	-1390.50	22.70***	1
	M3	2791.70	2834.10	-1386.80	7.30*	2
	M4	2556.30	2603.40	-1268.10	237.43***	1
	M5	2558.50	2615.10	-1267.20	1.79	2
Completely AI Generated[1]	M0	-504.69	-490.54	255.35	-	-
	M1	-518.85	-490.55	265.43	20.16***	3
	M2	-533.83	-500.81	273.92	16.98***	1
	M3	-536.38	-493.93	277.19	6.55*	2
	M4	-601.72	-554.55	310.86	67.34***	1

Note. *** $p < .001$, * $p < .05$. M0 random intercepts models (null); M1 adding controls (gender; Ethnicity; Age); M2 adding scenario type; M3 adding ChatGPT and online preparation conditions; M4 adding word count; M5 adding main effect of ChatGPT experience and interaction of ChatGPT condition and ChatGPT experience. [1]N = 138 (826 responses), otherwise N = 138 (825 responses).

Table 8

Fixed effects for models predicting SJT scores, word count, and Completely AI Generated Probability (Study 2).

Outcome	Predictors	*b*	*SE*	*df*	*t*
SJT Score (Model 3)	Intercept	4.10	0.27	144.13	15.17***
	Online Resources	0.28	0.19	137.99	1.46
	ChatGPT	0.53	0.20	138.12	2.73**
	Scenario Type	-0.39	0.08	687.30	-4.80***
	Ethnicity	0.67	0.17	138.07	3.91***
	Gender	0.30	0.16	138.02	1.85
	Age	-0.02	0.01	138.02	-3.84***
SJT Score (Model 4)	Intercept	2.36	0.17	175.60	13.69***
	Online Resources	0.12	0.10	128.50	1.20
	ChatGPT	0.27	0.11	129.70	2.55*
	Scenario Type	-0.32	0.08	679.30	-4.08***
	Ethnicity	0.19	0.10	134.50	1.98*
	Gender	0.07	0.09	129.60	0.82
	Age	-0.01	0.00	133.30	-2.26*
	Word Count	0.01	0.00	321.80	20.20***
Word Count	Intercept	121.52	7.99	146.86	15.21***
	Online Resources	8.11	11.56	137.98	0.70
	ChatGPT	11.04	11.63	138.04	0.95
	Scenario Type	-5.66	2.81	688.07	-2.02*
Completely AI Generated	Intercept	0.07	0.04	150.80	1.55
	Online Resources	0.04	0.03	120.80	1.19
	ChatGPT	0.07	0.03	121.60	2.47*
	Scenario Type	0.05	0.01	670.10	4.76***
	Ethnicity	0.08	0.03	125.30	3.06**
	Gender	0.03	0.03	121.60	1.37
	Age	-0.00	0.00	124.50	-2.92**
	Word Count	0.00	0.00	789.90	9.04***

Note. *** $p < .001$, ** $p < .01$ * $p < .05$. Online resource condition = 1, other conditions (i.e., ChatGPT and control) = 0. ChatGPT condition = 1 other conditions (i.e., online resources and control) = 0. Scenario Type: Video = 0, Text = 1. Gender: 1 = Female, 0 = Male & Non-binary; Ethnicity: 1 = white, 0 = non-white; Age (continuous). *N* = 138 (825 responses). Model 5 not presented as fixed effects were *ns*.

Table 9

Frequencies and Percentages of Response Classification from GTPZero by Experimental Condition (Study 2).

	Control		Online Resources		ChatGPT	
	Frequencies	Percentage	Frequencies	Percentage	Frequencies	Percentage
AI	4	1.17	2	0.85	5	1.99
Mixed	140	41.06	131	55.98	140	55.78
Human	197	55.77	101	43.16	106	42.23
Total	341	100	234	100	251	100

Note. 826 responses, *n* = 138 but two participants did not answer one scenario.

Table 10

GPTZero Classification of Fully-AI-Generated Responses (Study 2).

	ChatGPT 3.5			ChatGPT-4 (in MS Bing)		
		GPTZero Indicators			GPTZero Indicators	
SJT Scenario	Response length (words)	Document classification	Probability of AI generated	Response length (words)	Document classification	Probability of AI generated
Video-based 1	493	Mix human/AI	50%	498	Likely AI	92%
Video-based 2	373	Mix human/AI	51%	422	Moderately Likely AI	67%
Video-based 3	399	Mix human/AI	50%	521	Likely AI	91%
Text-based 1	428	Mix human/AI	61%	553	Mix human/AI	68%
Text-based 2	412	Mix human/AI	53%	539	Mix human/AI	56%
Text-based 3	442	Mix human/AI	51%	602	Mix human/AI	72%