Automating Personality-Based Employment Interviews: Development and Validation of an Artificial Intelligence Chatbot

Short title: *AI Chatbot Personality-Based Employment Interview*

Ashley Sylvara[1*], Pengda Wang[2], Tianjun Sun[2], Anna L. Heimann[3], Pia V. Ingold[4]

[1]*Kansas State University*

[2]*Rice University*

[3]*Universität Zürich*

[4]*University of Copenhagen*

[*This is a working paper and has not yet been peer-reviewed.*]

*Correspondence concerning this article should be addressed to Ashley Sylvara, Department of Psychological Sciences, Kansas State University, 1114 Mid-Campus Dr North, Manhattan, KS 66506-5302, United States, email: asylvara@ksu.edu

Author Contributions: Ashley Sylvara: Conceptualization; Data curation; Formal analysis; Investigation; Funding acquisition; Methodology; Writing – Original Draft. Pengda Wang: Data curation; Writing – Review & Editing; Software; Writing – Review & Editing. Tianjun Sun: Conceptualization; Methodology; Supervision; Writing – Review & Editing. Anna L. Heimann: Data curation; Writing – Review & Editing. Pia V. Ingold: Data curation; Writing – Review & Editing.

Abstract

This study examines the use of artificial intelligence (AI) chatbots and natural language processing methods for administering and scoring personality-based employment interviews. We adapted a behavioral description interview to a chatbot interview format and evaluated the construct and criterion-related validity of machine-derived personality scores. Using archival data as a baseline, the study incorporated natural language processing (NLP) methods, including word embeddings extracted with transformers and zero-shot prompt-based scoring using a large language model (LLM). Three key findings emerged. First, chatbot interviews generated significantly lower interviewee word counts than human interviews, limiting trait-relevant cues for raters and machine-based methods. Second, construct validity results demonstrated moderate convergence between machine-derived and human rater scores, with LLM-based scores performing comparably to human ratings. However, limited discriminant validity suggests that method effects outweigh trait-specific variance. Third, machine-derived scores demonstrated incremental validity in predicting organizational citizenship behaviors (OCB) beyond self-reported personality scores, underscoring their potential utility in selection contexts. These findings emphasize the need for refinements in chatbot design to elicit richer responses and improve scoring accuracy, offering promising implications for scalable and efficient personality assessments in organizational settings.

Keywords: Structured interviews, personality assessment, artificial intelligence, natural language processing

Data Availability Statement:

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Practitioner Points:

1. This study developed an AI chatbot to administer an interview designed to assess personality from behavioral description interview questions.
2. To ensure reliable personality assessments, chatbot interviews must elicit detailed, trait-relevant responses. Our findings highlight that low word counts and lack of contextual detail in responses can limit scoring accuracy.
3. Machine-derived scores, particularly those from large language models (LLMs) using zero-shot prompting, showed some support for incremental validity in predicting work-related outcomes beyond traditional self-reported personality scores.

Automating Personality-Based Employment Interviews: Development and Validation of an Artificial Intelligence Chatbot

Personality traits conceptualize an individual's pattern of cognitions, emotions, and behaviors (Goldberg, 1990), making them critical factors in understanding work behavior. Personality inventories are widely used in selection assessments since they are easy to administer, demonstrate low adverse impact (Hough et al., 2001), and have been linked to a wide range of work outcomes, including job satisfaction (Heller et al., 2009), task performance (Kamdar & Van Dyne, 2007), and contextual performance (i.e., organizational citizenship behaviors [OCBs]; Chiaburu et al., 2011). Despite the important links between personality and work outcomes, personality inventories have been critiqued for having relatively low criterion-related validity concerning performance (Morgeson et al., 2007) and heavy reliance on self-reports to accurately report their personality (De Cuyper et al., 2017). Moreover, when used as self-reports, they may be susceptible to socially desirable responding, as candidates may intentionally inflate their scores to appear more favorable, which can negatively impact selection decisions (Mueller-Hanson et al., 2003; Tett & Simonet, 2021). As a complementary approach to self-report personality inventories, open-ended formats, such as structured interviews, may provide additional benefits for assessing personality through job-relevant behaviors. Personality influences many aspects of an interviewee, including qualifications, academic performance, and workplace behavior (Gonzalez-Mulé et al., 2014; Poropat, 2009; Salgado & Moscoso, 2002). Supporting this notion, meta-analyses have found that personality traits, particularly facets of conscientiousness, can be effectively assessed through interview questions (Cortina et al., 2000; Huffcutt et al., 2001). In this context, structured interviews offer an effective method to assess personality traits (Heimann et al., 2021; Van Iddekinge et al., 2005).

Employment interviews are one of the most widely used tools in selection contexts (Macan, 2009). They are positively perceived by candidates (Hausknecht et al., 2004; Anderson et al., 2010), consistently linked to job performance (Sackett et al., 2022), and demonstrate low adverse impact (Levashina et al., 2014). The goal of an employment interview is to predict future job performance and suitability based on a candidate's responses to open-ended questions (McDaniel et al., 1994). These interviews can measure various constructs by asking respondents to describe their behavior in work-related situations (Huffcutt, 2011). Behavior description interviews, a type of structured interview, use predetermined scenario-based questions to which candidates respond, detailing how they behaved in a job-relevant situation in the past (Janz, 1982). These interviews demonstrate criterion-related validity in predicting performance (Weyhrauch & Huffcutt, 2017). However, inconsistencies in interview administration can present challenges to maintaining standardization (Levashina et al., 2014). Additionally, behavior description interviews are resource-intensive, requiring extensive interviewer training, multiple raters to ensure reliability, and significant time to administer and evaluate candidate responses. To address these challenges, researchers and practitioners would benefit from exploring methods that provide a consistent experience for candidates and automate both the interview and scoring processes to reduce administrative demands (Wang et al., 2024).

The advancement of technology offers promising solutions to the challenges of traditional structured interviews. Compared to the processes and data associated with human ratings, technological data collection methods and their resulting data have the potential to be more transparent and guided by expert input, thereby enhancing fairness and accuracy (Woo et al., 2024). For example, automated scoring systems based on artificial intelligence (AI) can be used to analyze text data collected from personnel selection assessments (Campion et al., 2016) and narrative comments of performance appraisals (Speer, 2018; Speer, 2021). It has also been demonstrated that scores derived from such technological assessments can predict academic and job-related success beyond traditional selection predictors (i.e., cognitive ability assessments) while simultaneously reducing subgroup differences (Campion et al., 2024). Similarly, AI chatbots (i.e., computer programs designed to simulate an interaction with a human user) are an innovative method to assess personality in an interview-like format and can streamline the interview process. Compared to a human interview, a chatbot can conduct thousands of interviews per day and provide an objective assessment in less time than it would take a human to analyze textual data (Zhou et al., 2019). Additionally, chatbots can simulate a more natural conversation which mimics that of a human conversation by providing empathetic feedback or handling follow-up questions, creating a conversational experience that mimics human interaction (Zhou et al., 2019; Xiao et al., 2020), which is more engaging than personality inventories.

Previous chatbot interviews have been designed to infer personality through life narrative, where a person details important points in their life as chapters (see Fan et al., 2023; Sun et al., 2024). Narrative interviews focus on one’s narrative identity–the story of an individual’s life experiences (McAdams, 2013). Similar to behavioral description interviews, narrative interviews prompt interviewees to discuss their thoughts, experiences, and behavior across a range of situations. Differently, where behavioral description interviews use pre-determined scenarios tailored to work-related behaviors, narrative interviews are more open-ended, allowing the interviewee to narrate chapters of their lives, including high points, low points, and significant memories (McAdams, 2013). Narrative identity has demonstrated incremental validity in predicting well-being beyond self-reported personality (Adler et al., 2016). In an adapted chatbot format, it has also been found to predict academic and social outcomes beyond self-reported personality (Fan et al., 2024). However, narrative-based interviews are not suitable for selection contexts. Narrative interviews typically ask participants to describe significant moments across their entire life and aspirations on topics including childhood memories, spiritual ideology, and political or social values (see the Life Story Interview II; McAdams, 2008). While this approach has demonstrated utility in inferring personality, in selection contexts questions should be job-relevant rather than asking about important moments and future aspirations.

The current study addresses an important gap in the literature by developing and validating a chatbot-based personality assessment using job-relevant behavior description interview questions. As AI methods are becoming increasingly popular in organizational settings (Budhwar et al., 2022; Lee et al., 2023; Siocon, 2013), rigorous development and validation of selection tools with different design characteristics and scoring approaches are essential (Lievens & Sackett, 2017). Building upon established findings from personality-based employment interviews, we adapt these methods to an automated format using an AI chatbot. We use methodology and archival data from Heimann et al. (2021) to guide our methodology and serve as a baseline for the chatbot development and scoring procedures and collect additional data using an AI chatbot to administer the interview. Our scoring algorithm uses word embeddings extracted using NLP transformers and zero-shot prompt engineering using a large language model (LLM). We then assess the construct validity (convergent and discriminant) and criterion-related validity (incremental validity) of machine-derived scores compared to and beyond self-reported personality scores.

The findings of this study will expand the boundaries of personality measurement in organizations in multiple ways. First, conceptually, this research advances our theoretical understanding of technology-enabled modular factors in selection systems (Lievens & Sackett, 2017). By modularizing key components of traditional behavioral description interviews, specifically the interview modality and response evaluation process, this study demonstrates how an AI chatbot and NLP techniques can be leveraged to enhance efficiency and standardization. Second, methodologically, this study uses traditional NLP approaches (i.e., embedding-based methods) and more modern techniques (i.e., zero-shot learning using LLMs) to score textual data. Third, technologically, this study contributes to efforts to automate the interview process by demonstrating the use of an AI chatbot to administer a behavioral description interview and leveraging NLP techniques to automatically score textual responses.

Theoretical Background and Hypothesis Development

Personality influences many aspects of an interviewee, including their qualifications, academic performance, and workplace behavior, making interviews a valuable method for assessing personality. Trait Activation Theory (TAT) posits that behavior emerges from the interaction between an individual and their environment, where situational cues activate relevant traits, which are expressed in behavior when the context allows (Tett & Burnett, 2003; Tett & Guterman, 2000). In a behavioral description interview, interviewees are prompted to describe their actions in specific work-related scenarios. If candidates are asked to discuss their behavior in interactions with colleagues, the situation may for instance elicit traits related to agreeableness. Therefore, if the goal is to predict workplace behaviors across various scenarios, interview measures should be designed to align with these behaviors. Guided by TAT, behavioral description interviews can be structured to elicit trait-relevant responses, thereby allowing for evaluation of behavior in trait-relevant situations, making them useful tools for predicting workplace behavior.

While crafting trait-relevant questions is a necessary first step towards achieving validity, it is essential to additionally ensure that the necessary trait-relevant cues are indeed elicited and observable. According to Funder's (1995) Realistic Accuracy Model (RAM), personality traits manifest through behaviors, and the relevance of these behaviors depends on their availability, detectability, and utilization. For traits to be accurately observed and scored by human raters or NLP methods, questions must prompt interviewees to provide sufficient information on trait-relevant behaviors at work. While personality traits, such as those in the Big Five, are not directly observable, they are typically measured indirectly through self-reports or other-reports, with the latter offering greater predictive validity (Connelly & Ones, 2010). However, the accuracy of these measures depends on the availability and detectability of trait-relevant cues. Previous studies have reported low to moderate convergence between self-reported and interviewer ratings from behavioral description interviews (see Heimann et al., 2021; with correlations ranging from r = .18 for agreeableness to r = .45 for extraversion), automated video interviews (see Hickman et al., 2022; with correlations rating from r = .07 for conscientiousness to r = .29 for emotional stability) and structured interviews (see Van Iddekinge et al., 2005; with correlations ranging from r = .20 for vulnerability [a facet of emotional stability] and r = .43 for altruism [a facet of agreeableness]). Based on the theoretical underpinning of TAT, we posit that convergence between self-reports and chatbot-derived ratings may result from shared trait activation processes, as both methods prompt individuals to reflect on and express trait-relevant behaviors. Additionally, in line with previous research, we hypothesize the following:

Hypothesis 1: Questionnaire-based self-reports of personality will correlate positively with human-based personality ratings derived from the chatbot-based behavior description interview.

AI-based personality assessments, particularly those using chatbot technologies, have increasingly been used to extract trait-relevant features from text data, with machine learning algorithms being trained on self-reported personality scores as the so-called "ground truth" (Fan et al., 2023). When compared to self-reported personality assessments, machine-derived personality scores show good reliability and convergent validity, though they tend to have lower discriminant validity (Azucar et al., 2018; Fan et al., 2023; Hickman et al., 2019; Hickman et al., 2022; Sun, 2024; Tay et al., 2020). Accordingly, based on previous research we hypothesize the following:

Hypothesis 2: Machine-derived personality scores from the chatbot-based behavior description interview will demonstrate construct-related validity on par with previous validity evidence from existing AI-based personality assessments.

Beyond construct validity, machine-derived personality scores may explain unique variance beyond self-reported measures in predicting a range of work-related outcomes. Machine-derived personality trait scores from an AI chatbot interview have demonstrated incremental validity in predicting academic performance above and beyond ACT test (a standardized test used for college admission in the US) scores and self-reported personality (Fan et al., 2023). Heimann et al. (2021) demonstrated that interviewer-rated personality from behavioral description interviews accounted for significant incremental variance beyond verbal cognitive ability and self-reported personality. Beyond task and contextual performance, personality derived from traditional self-report inventories have been significantly related to a wide range of measures commonly used in organizational contexts, including counterproductive work behaviors (CWBs; Chiaburu et al., 2011), job satisfaction (Judge et al., 2002), subjective well-being (SWB; Gutiérrez et al., 2005; Steel et al., 2008), and stress (Luo et al., 2023). Accordingly, we hypothesize the following:

> *Hypothesis 3: Machine-derived personality scores from the chatbot-based behavior description interview will demonstrate incremental criterion-related validity over and above questionnaire-based personality self-reports.*

Materials and Method

Participants and Procedures

Data for the present study were collected from three samples: (1) an archival sample of working adults from Heimann et al. (2021), (2) an undergraduate student sample, and (3) a working adult sample. The archival data informed the development of the chatbot-based interview, which was administered to the undergraduate and working adult samples in the U.S. While the archival data were used for model training with the chatbot data, they were not included in the formal analyses of construct or criterion-related validity.

Sample 1: Archival Data from Human Interviews

The archival data consisted of 203 working adults (M_{age} = 30.56, *SD* = 7.51; 60% male) who completed a job interview in a simulated selection setting at a university in Switzerland. To participate, they needed to provide their supervisor's contact information (for collection of supervisor ratings). The simulation was designed to help participants prepare for future job applications, encouraging them to behave as if it were an actual selection interview. During the simulation, participants completed a 30-minute personality-based interview and a contextualized personality self-report measure. The order of these two were randomized, where half completed the interview first and the other half completed the self-report measure first. Each interview was videotaped and conducted by a panel of two trained interviewers, who also served as raters. Interviewers took notes and independently rated each participant's responses to interview questions using 5-point rating scales with behavioral anchors. Interviewers were instructed to follow a highly structured format and were limited from paraphrasing questions, providing explanation of questions, or probing. Interviewees were instructed to keep their responses brief to ensure interviews could be completed within 30 minutes. Following the simulation, interviewers discussed their ratings and resolved any discrepancies greater than one point. After ratings were complete, participants received detailed feedback on their performance. Full details on the data collection procedures for this sample are available in Heimann et al. (2021).

To compare responses with the chatbot interviews and to use the archival data for model training, we extracted the audio files from the videotaped interviews to transcribe the participants' responses to the interview questions and then translated them from Swiss German to English. We used OpenAI's Speech-to-Text Whisper API (OpenAI, 2023) with the large Whisper model to perform both transcription and translation. This approach makes interview transcripts suitable for model training and to ensure consistency in training alongside the US sample (as most NLP transformers are extensively trained in English). Both the transcriptions and translations were reviewed for accuracy by bilingual speakers fluent in Swiss German and English.

Samples 2 and 3: Data from Chatbot Interviews

Our second sample included 130 undergraduate students (M_{age} = 18.85, *SD* = 1.71; 78% female; 80% White) who were recruited from a subject pool at a U.S. Midwestern University. Data from 185 participants were initially collected, but subsets of participants were removed for failing attention checks (*n* = 33), providing an incorrect participant ID during the chatbot interview (*n* = 14), or submitting irrelevant responses to multiple chatbot interview questions (*n* = 8).

Our third sample included 88 working adults (M_{age} = 35.12, *SD* = 10.27; 50% male; 58% White) who were recruited from Connect, an online crowdsourcing platform by CloudResearch with a high-quality participant pool. To participate, they needed to have at least one year of work experience and be currently employed at least part-time. Data from 102 participants were initially collected, but some participants were removed for failing attention checks (*n* = 3), submitting irrelevant or low-effort responses to multiple interview questions (*n* = 11).

Data collection for Samples 2 and 3 consisted of three sections. Prior to beginning the study, participants provided informed consent and were briefed on the general purpose of the study (i.e., "to validate a personality selection assessment"). Participants were informed that their data would be used for research purposes only and that all responses would remain confidential. In the first section, participants completed surveys regarding basic demographics and work experience. In the second section, participants completed a personality self-report measure and the AI chatbot interview; the order of these two assessments was randomized. The chatbot interview was conducted using Juji Inc.'s AI chatbot platform (https://juji.io). Supplemental Material A includes a screenshot of what the AI chatbot looked like. Participants typed their responses to each interview question asked by the chatbot (i.e., they provided textual input data). In the third section, participants completed outcome measures, including surveys on organizational citizenship behaviors, perceived job performance, and subjective well-being. At the end of the study, participants received feedback on their personality, but detailed feedback on their interview performance was not provided.

Personality Measures

Self-Reported Personality

For the archival data (Sample 1), self-reported personality was assessed using a contextualized version of the 50-item International Personality Item Pool (IPIP-50; Goldberg, 1992). This instrument measures extraversion, agreeableness, conscientiousness, emotional stability, and intellect/openness with ten items each. A combination of instructional contextualization (i.e., participants were explicitly instructed to complete the inventory thinking about their typical cognitions, emotions, and behaviors at work) and tagged contextualization was used (i.e., the tag "at work" was added to each item; see also Lievens et al., 2008). Heimann et al. (2001) reported that internal consistencies ranged from Cronbach's $\alpha = .75$ (for conscientiousness) to $\alpha = .85$ (for emotional stability).

For the student sample (Sample 2), we used the Big Five Inventory (BFI-2; Soto & John, 2017). This instrument measures extraversion, agreeableness, conscientiousness, emotional stability, and intellect/openness with 12 items each. We used the generalized (i.e., not contextualized) version because student participants were not expected to be currently employed, and contextualization could have made the questions feel less relevant to their experiences. Items were ranged on an agreement scale ranging from 1 (strongly disagree) to 5 (strongly agree). Example extraversion item: "*I am someone who is outgoing, sociable.*" Internal consistency for extraversion was Cronbach's α = .86. Example agreeableness item: "*I am someone who is respectful, treats others with respect.*" Internal consistency for agreeableness was Cronbach's α = .76. Example conscientiousness item: "*I am someone who is dependable, steady*" Internal consistency for conscientiousness was Cronbach's α = .81. Example emotional stability item: "*I am someone who is relaxed, handles stress well.*" Internal consistency for emotional stability was Cronbach’s α = .91. Example openness item: “*I am someone who is original, comes up with new ideas*” Internal consistency for openness was Cronbach’s α = .84.

For the working adult sample (Sample 3), we used a contextualized version of the BFI-2 (Soto & John, 2017). To contextualize this measure, we incorporated a combination of instructional contextualization, tagged contextualization, and complete textualization (i.e., redesigning the statement to fit within a work context), as this approach has been shown to improve the criterion-related validity of forced-choice personality measures (Li et al., 2024). An example of this approach was changing “*Has an assertive personality*” to “*Has an assertive personality when engaging in workplace discussions.*” Items were ranged on an agreement scale ranging from 1 (strongly disagree) to 5 (strongly agree). Example extraversion item: “*I am someone who is outgoing, sociable with colleagues.*” Internal consistency for extraversion was Cronbach’s α = .89. Example agreeableness item: “*I am someone who is respectful, treats others professionally at work.*” Internal consistency for agreeableness was Cronbach’s α = .88. Example conscientiousness item: “*I am someone who is dependable, steady in fulfilling work responsibilities.*” Internal consistency for conscientiousness was Cronbach’s α = .89. Example emotional stability item: “*I am someone who is relaxed, handles stress well at work.*” Internal consistency for emotional stability was Cronbach’s α = .89. Example openness item: “*I am someone who is original, comes up with new ideas for work projects.*” Internal consistency for openness was Cronbach’s α = .91.

Personality-Based Employment Interview

The interview questions were adapted from Heimann et al. (2021), which includes 15 behavior description interview questions designed to assess specific work behaviors as indicators of the Big Five personality traits, with three questions per trait. An example interview question and rating scale for each trait is shown in Supplemental Material B. To ensure consistency across the approach used by the interviewers in Heimann et al. (2021) and the chatbot, we designed the chatbot to equally follow a highly structured interview. Thus, similar to the interviewers in Heimann et al. (2021), if participants asked for clarification, the chatbot would repeat the question without providing additional explanation or rephrasing. Although this approach may have limited the depth of participants' responses, it ensured standardization and maintained consistency with the structured format of the human interviews.

Human-Based Personality Ratings in the Employment Interview

For the archival sample (Sample 1), two interviewers took notes on interviewees' responses to each interview question and individually rated responses on a 5-point behaviorally-anchored rating scale. Interrater reliability was calculated using one-way random effects intraclass correlation coefficient (ICC) for each interview question. Across the 15 interview questions, the ICC was .78, reflecting the reliability of the average rating between two interviewers.

For the student sample (Sample 2) and the working adults sample (Sample 3), three raters scored participants' responses to each interview question on the same 5-point behaviorally-anchored rating scale as the archival sample (Sample 1). Before scoring, the raters were familiarized with the context of the interviews (i.e., that they were designed to assess an individual's personality), provided definitions of the Big Five personality traits, and received training on how to score responses using the behavioral anchors. Consistent with Heimann et al. (2021), raters had no access to self-reported scores. This rater training process was consistent to interviewer training for the archival sample, except they did not undergo a formal one-day frame-of-reference training on how to administer an interview, and these raters had more familiarity with the general purpose of the study but were not informed of the specific hypotheses being tested. After completing their initial ratings, the raters met to discuss discrepancies where ratings differed by two or more points. Following the same procedures for the archival sample (Sample 1), they were not required to agree on the same final rating, but they were able to adjust their rating based on discussion. Across the 15 interview questions, the ICC for the average ratings across the three raters was .85 for the student sample (Sample 2) and .84 for the working adult sample (Sample 3), indicating good agreement. The final score for each participant was calculated by averaging the scores from the three raters.

Machine-Derived Personality Scores in the Employment Interview

We adopted two natural language processing (NLP) approaches to score the textual data from the employment interview (i.e., the transcribed and translated interview responses from Sample 1 and the textual chatbot input from Samples 2 and 3): (a) word embeddings extracted using the DistilBERT transformer model (Sanh et al., 2019) and (b) zero-shot prompting with the Llama 3.1 model (Grattafiori et al., 2024; Meta AI, 2024). These approaches represent two widely used, yet distinct methods for analyzing text.

Word embeddings are numeric representations of text, and when generated using transformers, these embeddings are dense vector representations that capture the contextual relationships between words. Transformers, like DistilBERT, have been interpreted as useful tools for text analytics since their introduction because they encode the nuanced context of language (Vaswani et al., 2017). This embedding-based approach has demonstrated utility for evaluating responses in assessment center exercises to assess job-relevant competencies (Thompson et al., 2023) and automated video interviews to assess cognitive ability (Hickman et al., 2024). In this study, text from each interview question (including both the question and the participant’s response) was used to train models. Including the question text provides additional context for the participant’s response, ultimately allowing for a deeper understanding of the data. Consistent with earlier studies (e.g., Hickman et al., 2024; Speer et al., 2023), the embeddings were trained using *k*-fold cross-validation, a method that splits the data into sections (folds). Specifically, we used 10-fold, which means in each iteration, 90% of the data is used for training, and the remaining 10% is used for testing. This process is repeated across all 10 folds, reducing overfitting and generating cross-validated scores for each participant's response to each interview question.

Large language models (LLMs) are advanced machine learning models, typically built on transformer-based architectures, that are trained on massive amounts of text to process and mimic human-like language. Differently from embeddings, these models more closely represent human approaches to scoring. To extract scores for this study, we used zero-shot prompting with the Llama 3.1 model. Zero-shot prompting involves providing the model with a task without labeled examples and LLMs have been found to perform well when task instructions are provided (Sanh et al., 2021). This approach has been shown to perform similarly to human raters and, when using LLMs such as GPT-3.5 and GPT-4, to outperform BERT transformer models for certain personality traits, such as extraversion, though not for conscientiousness (Zhang et al., 2024). In this study, we used a single zero-shot prompt where the model was given the following elements: a task and role assignment (e.g., “You will play the role of a scoring expert and assess the answer based on the given behavioral anchors”), the question text, the participant’s response, and explicit instructions on how to use the behavioral anchors to score the response. This approach was first piloted on the archival data (Sample 1) to ensure ratings converged well with human-rated scores. From our pilot, we obtained an average correlation of $\bar{r}$ =.48, ranging from *r*

= .42 for conscientiousness to *r* = .56 for extraversion. An example of this prompt can be found in Supplemental Material C. By structuring the prompt in this way, the LLM was guided to generate both a numerical score and an explanation of how the behavioral cues in the response aligned with the given anchors. This approach leverages the LLM's ability to process instructions and perform reasoning tasks in a flexible, human-like manner, enabling it to provide nuanced and context-sensitive evaluations without requiring prior task-specific training (Kojima et al., 2022).

Outcome Variable Measures

To assess the incremental validity of machine-derived personality scores from the chatbot interview over and above the questionnaire-based personality self-reports, we collected the following outcome variables in Sample 2 and 3 as self-report.

Organizational Citizenship Behavior (OCB) was measured using ten items from Spector et al. (2010) to assess extra-role behaviors. Items were rated on a frequency scale ranging from 1 (*never*) to 5 (*every* day). Example item: "*In the past year, how often have you helped new employees get oriented to the job*?". Internal consistency was Cronbach's α = .83.

Counterproductive Work Behavior (CWB) was measured using ten items from Spector et al. (2010), designed to assess harmful workplace behaviors. Items were rated on a frequency scale ranging from 1 (never) to 5 (every day). Example item: "*In the past year, how often have you ignored someone at work*?". Internal consistency was Cronbach's α = .86.

Task Performance was measured using three self-developed items designed to assess an individual's perceived competence in their tasks and responsibilities. Items were rated on an agreement scale ranging from 1 (strongly disagree) to 5 (strongly agree). Example item: "*I am very competent at what I do*.". Internal consistency was Cronbach's α = 84.

Job Satisfaction was measured using eight items from Russell et al. (2004) assessing individuals' overall attitudes toward their work. Participants were asked to think about their current school or professional work and rate items (e.g., "*Good*", "*Enjoyable*", and "*Poor*") on a 3-point scale: 1 = Yes, it describes my work; 2 = Cannot decide; 3 = No, it does not describe my work. Internal consistency was Cronbach's α = .81.

Stress was measured using ten items from Cohen et al. (1983), assessing stress levels over the past month. Items were rated on a 1 to 5 frequency scale ranging from 1 (*never*) to 5 (*very often*). Example item: "*In the past month, how often have you been upset because of something that happened unexpectedly*?". Internal consistency was Cronbach's α = .88.

Table 1 reports the correlations for self-reported personality, human-based personality ratings in the interview, machine-derived personality ratings in the interview, and outcome variables.

Results

Response Quality in the Chatbot Interview

Hypothesis 1 stated that questionnaire-based personality self-reports would correlate positively with human-based personality ratings from the chatbot interview. Table 2 presents the correlations between personality scores derived from self-reports and human raters in the human interview (Sample 1) and the chatbot interview (Samples 2 and 3). As can be seen, self-reported personality and human ratings from the chatbot interview were significantly correlated for extraversion in the student sample (Sample 1; r = .19, p = .026) and working adult sample (Sample 2; r = .28, p = .007), as well as agreeableness in the student sample (Sample 1; r = .29, p < .001) and the working adult sample (Sample 2; r = .36, p < .001).

In contrast, for conscientiousness and emotional stability, only the student sample (*r* = .35, *p* < .001, and *r* = .18, *p* = .038, respectively) but not the working adult sample (*r* <.01, *p* > .05, and *r* = .07, *p* = .491, respectively) showed significant correlations between self-reports and human ratings in the chatbot interview. For openness/intellect, neither the student sample (*r* = .14, *p* = .114) nor the working adult sample (*r* = .14, *p* = .206) showed significant correlations between self-reports and human ratings in the chatbot interview. This is likely due to construct differences between openness assessed from the BFI-2, whereas the interview was designed to assess intellect/openness (as defined by Goldberg, 1990; 1992). Taken together, Hypothesis 1 found full support only for agreeableness and extraversion, and partial support for conscientiousness and emotional stability.

Exploratory analysis

To further explore participants' response quality in the chatbot interview, we screened the responses in all three samples. We observed that participants in the chatbot interview (Samples 2 and 3) provided less explanation and context in their responses to interview questions. To quantify this observation, we examined the word count for each interview question in each sample. Table 3 shows the average word count for each question across the three samples, with participants in the human interviews providing an average of 166 words per question compared to only 48 words per question in the chatbot interviews.

Construct-Related Validity of Machine-Derived Scores from the Chatbot Interview

Hypothesis 2 predicted that machine-derived personality scores from the chatbot interview would demonstrate construct-related validity. To examine this, we used the multitrait-multimethod (MTMM) analytical framework (Campbell & Fiske, 1959) and adopted generalizability theory (G-theory) based variance decomposition to model MTMM data (Woehr et al., 2012). This approach allowed us to examine and explain variances at the person, trait, method, and their interaction levels. Tables 4 and 5 present the convergent and discriminant mean correlations between the self-report inventory, the human-based ratings from the chatbot interview, and the machine-derived ratings from the chatbot interview (i.e., word embeddings and zero-shot prompting) averaged across all Big Five traits for Sample 2 (see Table 4) and Sample 3 (see Table 5).

Convergent validity was assessed using the convergence index (C1), which represents the coverage of correlations for the same traits assessed across different methods (i.e., monotrait-heteromethod [MTHM] correlations). A large C1 indicates that trait scores converge well across methods. C1 across the four methods was .404 for Sample 2 and .427 for Sample 3, suggesting that 40.4% and 42.7% of the observed variance can be attributed to person-level main effects and trait-specific variance, respectively. Notably, the embedding-based scores from the chatbot interview demonstrated high monotrait-heteromethod (MTHM; i.e., same traits assessed by different methods) with human-based ratings from the chatbot interview, with average MTHM correlations of .553 for Sample 2 and .661 for Sample 3. This result is expected given that the embedding-based scores were trained on the human-based ratings. Similarly, the zero-shot prompt-based scores showed high MTHM correlations with human-based ratings, with average MTHM correlations of .680 for Sample 2 and .725 for Sample 3, suggesting that the LLM followed the behavioral anchors in a manner similar to human raters. Discriminant validity was assessed using the discriminant indices D1 and D2. The first discriminant index (D1) is calculated by subtracting the average correlations for different traits assessed across different methods (i.e., heterotrait-monomethod [HTMM] correlations) from C1 and the second discriminant index (D2) is calculated by subtracting the average of correlations for different traits assessed by the same method (i.e., heterotrait-heteromethod [HTHM] correlations) from C1. Large D1 and D2 indices indicate items discriminate well across the three methods. While D1 represents the extent to which different traits using the same method are distinct, D2 represents the extent to which the method effect is stronger than the trait effect. D1 for was .196 for Sample 2 and .110 for Sample 3, indicating that, respectively, 19.6% and 11.0% of the observed variance can be attributed to trait-specific variance after accounting for overlap among different traits measured across different methods. D2 was -.074, indicating that the method effect was stronger than the trait effect.

Lastly, method variance (MV) is calculated by subtracting the average of HTMM correlations from the average of HTHM correlations, which represents the proportion of variance attributable to methods. The method variance was .142 for Sample 2 and .245 for Sample 3, indicating that, respectively, 14.2% and 24.5% of the observed variance can be attributed to differences in the methods used to assess traits (e.g., questionnaire-based self-report, human-based interview ratings, word embeddings in the interview, or zero-shot prompting in the interview). This reflects the influence of method-specific factors, such as differences in the scoring mechanisms and how trait-relevant information is extracted from limited text in the chatbot interviews. The observed method variance is likely amplified by the significant differences in format between self-reported questionnaires and machine-derived scores obtained from a chatbot interview.

Overall, Hypothesis 2 found partial support. What speaks for the construct-related validity of machine-derived personality scores from the chatbot interview is that the variance attributable to traits (C1) was considerably higher (at 40.4% and 42.7%) than the variance attributable to methods (MV; at 14.2% and 24.5%). What speaks against the construct-related validity of machine-derived personality scores are the low discriminant validity indices (D1 and D2) suggesting a limited ability to distinguish between traits within methods. These results highlight challenges in achieving accurate ratings of chatbot interview responses, particularly when little trait-relevant text input is available.

Criterion-Related Validity of Machine-Derived Scores from the Chatbot Interview

Hypothesis 3 stated that machine-derived personality scores from the chatbot interview would demonstrate incremental criterion-related validity over and above questionnaire-based personality self-reports. To examine the criterion-related validity of machine-derived scores, we conducted 30 hierarchical regression analyses (five personality domains × six relevant outcome variables) for Samples 2 and 3 separately. The outcome variables (i.e., self-reported OCB, CWB, job performance, job satisfaction, well-being, and perceived stress) served as dependent variables. Each outcome was regressed on the questionnaire-based self-report personality scores (Step 1), and the machine-derived scores from the chatbot interview (i.e., embedding-based and zero-shot scores) were added separately in Step 2. Tables 5 through 10 report the results of the regressions.

For OCB (see Table 6), results demonstrate that embedding-based scores from the chatbot interview accounted for significant incremental variance beyond self-reported scores for Sample 3 for agreeableness ($\Delta R^2 = .10$, $p = .003$), conscientiousness ($\Delta R^2 = .08$, $p = .008$), and openness ($\Delta R^2 = .08$, $p = .006$). They did not account for significant incremental variance for Sample 2. The zero-shot prompt-based scores from the chatbot interview accounted for significant incremental variance beyond self-reported scores for agreeableness in both Sample 2 ($\Delta R^2 = .04$, $p = .024$) and Sample 3 ($\Delta R^2 = .08$, $p = .006$), conscientiousness only for Sample 2 ($\Delta R^2 = .04$, $p = .013$), and openness only for Sample 3 ($\Delta R^2 = .08$, $p = .006$).

For CWB (see Table 7), the embedding-based scores accounted for significant incremental variance beyond self-reported scores for agreeableness just for Sample 3 ($\Delta R^2 = .05$, $p = .007$), but not for any of the other Big Five traits. The zero-shot prompt scores accounted for significant incremental variance beyond self-reported scores just for openness for Sample 3 ($\Delta R^2 = .04$, $p = .044$).

For job performance (see Table 8), neither embedding-based nor zero-shot scores demonstrated significant incremental variance beyond self-reported personality scores. It is worth noting that personality has been found to be more relevant for predicting OCB and CWB than job performance (Gonzalez-Mulé, 2014).

For job satisfaction (see Table 9), embedding-based scores accounted for significant incremental variance beyond self-reported personality scores for Sample 3 for extraversion ($\Delta R^2 = .03$, $p = .026$) and agreeableness ($\Delta R^2 = .03$, $p = .036$). Zero-shot prompt-based scores accounted for significant incremental variance for Sample 3 for agreeableness ($\Delta R^2 = .05$, $p = .011$) and openness ($\Delta R^2 = .05$, $p = .049$). For Sample 2, neither embedding-based nor zero-shot scores demonstrated significant incremental variance beyond self-reported personality scores in predicting job satisfaction.

For subjective well-being (see Table 10), embedding-based scores accounted for significant incremental variance beyond self-reported scores for conscientiousness ($\Delta R^2 = .03$, $p = .035$) and openness ($\Delta R^2 = .02$, $p = .012$). Zero-shot prompt-based scores did not account for significant incremental variance.

Finally, for stress (see Table 11), embedding-based scores accounted for significant incremental variance for conscientiousness for both Sample 2 ($\Delta R^2 = .03$, $p = .046$) and Sample 3 ($\Delta R^2 = .05$, $p = .012$) and for openness for Sample 3 ($\Delta R^2 = .04$, $p = .038$). Zero-shot prompt-based scores accounted for significant incremental variance for Sample 3 for agreeableness ($\Delta R^2 = .05$, $p = .030$) and conscientiousness ($\Delta R^2 = .04$, $p = .040$).

Taken together, providing partial support for Hypothesis 3, results indicate that embedding-based scores from the chatbot interview provide small but meaningful incremental validity for predicting a range of outcome variables, particularly for traits such as extraversion, agreeableness, conscientiousness, and openness. This was most apparent for OCB, job satisfaction, subjective well-being, and stress. By contrast, zero-shot prompt-based scores demonstrated more limited incremental validity, with significant contributions observed primarily for agreeableness in OCB and job satisfaction. The key difference between embedding-based scores and zero-shot prompt-based scores is their approach to processing text. Embeddings are generated using NLP transformers and encode contextual relationships between words into dense vector representations. Zero-shot prompt-based scores using LLMs evaluate responses more similar to humans, using explicit instructions to generate ratings. Because all outcome measures were self-reported, it is somewhat expected that self-reported personality scores aligned more closely with these outcomes compared to embedding-based scores trained on rater scores and zero-shot prompt-scores that mimic human ratings.

Discussion

As AI becomes increasingly prevalent in organizational settings, the rigorous development and validation of AI-based selection tools with varying design characteristics and scoring approaches is critical (Lievens & Sackett, 2017). The present study had two primary aims: (1) to apply natural language processing (NLP) techniques to score text from personality-based employment interviews administered through an AI chatbot and (2) to evaluate the psychometric validity of machine-derived scores from the chatbot interview. Through the initial testing of an AI chatbot to administer a personality-based employment interview, our study yielded three key findings that offer important insights into the development and application of AI-based selection tools. Additionally, embedding-based methods may require more robust input data to reduce reliance on method variance and improve discriminant validity.

First, our results underscore the challenges of adapting interviews designed for human administration to an automated, chatbot-based format. One notable finding was the substantially lower word counts in chatbot interviews compared to human interviews. Participants in chatbot interviews provided, on average, only 48 words per question, compared to 166 words in human interviews, which may substantially limit the availability of trait-relevant cues for accurate rater scoring. According to Funder's (1995) realistic accuracy model (RAM), accurate trait assessment requires the availability of relevant cues that raters or algorithms can detect and utilize. While word counts for human interviews (Sample 1) typically include filler words, the significantly lower word count in chatbot interviews (Samples 2 and 3) suggests that the lack of convergence with self-reported scores may be due to insufficient trait-relevant cues for raters to evaluate, most notably for the working adult sample (Sample 3). The written response format of chatbot interviews, which requires participants to type their answers, may contribute to the reduced word count by being more cumbersome and less interactive than speaking in a human-based interview, aligning with Lievens and Sackett's (2017) framework emphasizing the importance of interactivity and motivation in modular assessment formats.

Second, these findings highlight the importance of ensuring that chatbot interviews are designed to elicit richer, more detailed responses. Though both the student sample (Sample 2) and working adult sample (Sample 3) had motivation to complete the chatbot interview effortfully: students needed to finish the chatbot interview to earn research credit for course requirement and working adults received payment only upon effortful completion of the interview. These motivations may still fall short of the motivation in the archival data (Sample 1), where participants treated the interviews as developmental exercises and received feedback on their performance. For the present study, the aim was to closely mimic the structure of the interviews in the archival data (Sample 1), one potential solution is incorporating follow-up prompts when participants fail to address all parts of a question or provide overly brief answers. While this reduces the structure of the interview, it also ensures there is enough context for raters to accurately score responses. Such enhancements could increase the availability of trait-relevant cues, improving both human and machine-based scoring accuracy.

Third, our findings provide modest support for the construct validity of machine-derived scores. These results align with prior research (e.g., Azucar et al., 2018; Fan et al., 2023; Hickman et al., 2019, 2022) demonstrating that machine-derived personality scores exhibit good convergent validity but less than optimal discriminant validity. However, low discriminant validity highlights challenges in distinguishing between traits, a common issue in situational assessment methods like structured interviews and assessment centers. These results indicate that method effects, driven by differences in response format and scoring mechanisms, remain a significant factor influencing scores. Overall, our results demonstrate the utility of NLP for personality assessment. Zero-shot prompt-based scores exhibited good alignment with human raters in the chatbot interview, particularly when behavioral anchors were clearly defined in the prompts, which suggests that such LLMs can “understand” and apply scoring criteria in a way that mimics human judgment. Notably, while embedding-based methods were more sensitive to the brevity of responses, both approaches were still able to provide decent scoring accuracy even when trait-relevant cues were limited. These findings highlight the potential of embedding-based methods and LLMs to score personality responses, but they also emphasize the need for further refinement in chatbot interview design and scoring procedures to enhance construct validity.

Lastly, a particularly promising finding is the evidence for the incremental criterion-related validity of machine-derived scores. Both embedding-based and zero-shot prompt-based scores demonstrated significant incremental variance in predicting OCB beyond traditional self-reported personality scores. For instance, embedding-based scores accounted for significant incremental variance for extraversion, agreeableness, conscientiousness, and openness, while zero-shot scores accounted for incremental variance for agreeableness, conscientiousness, and openness. These results suggest that machine-derived scores provide unique and valuable information that complement self-reported personality measures, even when text responses are relatively brief and limited in detail. This incremental validity highlights the potential utility of machine-derived scores in real-world organizational contexts, where they could serve as a complementary tool to traditional selection methods. Additionally, the fact that these scores demonstrated predictive utility despite limitations in the interview format (e.g., lack of follow-up probing) suggests that further refinements in the design of chatbot interviews could enhance their predictive power even further.

Limitations

The present study had several limitations, which present opportunities and important considerations for future research. First, participants were not actual job candidates, which may have influenced their responses during the chatbot interviews. Without the motivation of a real-world, high-stakes selection context, participants may not have engaged with the interview questions as seriously, resulting in shorter and less informative responses to interview questions or strategically. This lack of incentive might have contributed to some validity challenges, particularly the limited text provided to the chatbot by participants. However, participants who did not effortfully complete the interview were excluded from analyses, the written response format likely also constrained the length and depth of responses compared to verbal interactions in traditional interviews. Second, our study was designed with a highly structured chatbot interview format to be consistent with procedures from the archival data. Our primary goal was to maintain consistency with the human condition, which was highly structured. However, this design choice meant that participants who provided minimal or overly brief responses were not prompted to elaborate further. Consequently, these short responses might have lacked the richness necessary for both human raters and machine-learning models to generate more accurate personality scores. Future research could examine the degree to which chatbot systems with dynamic follow-up prompts encourage participants to elaborate and to what degree this affects the quality of responses. Third, outcome variables for both chatbot samples (Samples 2 and 3) were all measured through self-report. We would encourage future research to gather other, preferably supervisor reports. Though both of these were honest reporting conditions and there was no motivation to inflate responses.

Practical Implications

The findings of this study have significant implications for both researchers and practitioners seeking to develop automated, scalable personality assessment tools. Our results demonstrate the feasibility of using AI-based personality assessments to reduce administrative burdens during the hiring process while maintaining predictive validity. Machine-derived scores, particularly those generated via zero-shot prompting, provide a scalable, objective, and cost-effective alternative to traditional human-rated interviews. These tools can streamline selection processes, allowing organizations to evaluate larger candidate pools with greater efficiency and consistency. However, to maximize the utility of chatbot-administered interviews, it is essential to design systems that elicit richer and more detailed responses. This can be achieved by incorporating follow-up prompts and refining the wording of questions to ensure they are engaging and relevant for diverse candidate populations, such as students, early-career professionals, and experienced workers.

Future Research

This study contributes to the growing literature on AI-driven personality assessment by offering evidence of construct and criterion-related validity for machine-derived scores. These findings highlight the potential for AI chatbots to serve as a viable alternative to traditional methods of assessing personality traits. However, challenges remain in improving the discriminant validity of machine-derived scores. Future research should explore strategies to enhance the availability and detectability of trait-relevant cues in textual responses, such as developing question formats that encourage candidates to provide more specific and behaviorally rich examples. Additionally, researchers should continue to investigate ways to improve embedding-based and LLM-based scoring approaches. For example, leveraging techniques such as few-shot learning and advanced feature extraction could enhance the interpretability and robustness of machine-inferred personality scores. These advancements could aid in refining the predictive accuracy of AI-based systems, particularly in contexts requiring nuanced personality assessments. Finally, exploring the practical applications of these tools in high-stakes, real-world selection scenarios will be critical for validating their scalability and fairness across diverse organizational settings.

Conclusion

This study provides an important step forward in exploring the feasibility and potential of AI chatbots for administering personality-based employment interviews. Despite some challenges, including limited response elaboration and mixed evidence for construct validity, the findings demonstrate that machine-derived personality scores can provide incremental validity in predicting key workplace outcomes. While the results are not without limitations, they highlight promising directions for improving AI-driven assessments. Enhancements in chatbot design, such as incorporating dynamic follow-up prompts and refining question formats, have the potential to elicit richer responses and improve the accuracy of both human and machine-based evaluations. Similarly, continued advancements in LLMs and embedding-based approaches could address issues of discriminant validity, creating more robust and interpretable scoring systems. Overall, this study lays the groundwork for future research and practical applications of AI in administering personality-based interviews.

References

Adler, J. M., Lodi-Smith, J., Philippe, F. L., & Houle, I. (2016). The incremental validity of narrative identity in predicting well-being: A review of the field and recommendations for the future. *Personality and Social Psychology Review*, *20*(2), 142-175. https://doi.org/10.1177/1088868315585068

Anderson, N., Salgado, J. F., & Hülsheger, U. R. (2010). Applicant reactions in selection: Comprehensive meta-analysis into reaction generalization versus situational specificity. *International Journal of Selection and Assessment*, *18*(3), 291-304. https://doi.org/10.1111/j.1468-2389.2010.00512.x

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150-159. https://doi.org/10.1016/j.paid.2017.12.018

Budhwar, P., Malik, A., De Silva, M. T. T., & Thevisuthan, P. (2022). Artificial intelligence – challenges and opportunities for international HRM: a review and research agenda. The *International Journal of Human Resource Management, 33*(6), 1065–1097. https://doi.org/10.1080/09585192.2022.2035161

Campion, E. D., Campion, M. A., Johnson, J., Carretta, T. R., Romay, S., Dirr, B., Deregla, A., & Mouton, A. (2024). Using natural language processing to increase prediction and reduce subgroup differences in personnel selection decisions. *Journal of Applied Psychology, 109*(3), 307–338. https://doi.org/10.1037/apl0001144

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology, 101*(7), 958–975. https://doi.org/10.1037/apl0000108

Chiaburu, D. S., Oh, I.-S., Berry, C. M., Li, N., & Gardner, R. G. (2011). The five-factor model of personality traits and organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology*, *96*(6), 1140–1166. https://doi.org/10.1037/a0024004

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A Global Measure of Perceived Stress. *Journal of Health and Social Behavior*, *24*(4), 385–396. https://doi.org/10.2307/2136404

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*(6), 1092–1122. https://doi.org/10.1037/a0021212

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology*, *53*(2), 325-351. https://doi.org/10.1111/j.1744-6570.2000.tb00204.x

De Cuyper, K., De Houwer, J., Vansteelandt, K., Perugini, M., Pieters, G., Claes, L., & Hermans, D. (2017). Using indirect measurement tasks to assess the self–concept of personality: A systematic review and meta–analyses. *European Journal of Personality, 31*(1), 8-41. https://doi.org/10.1002/per.2092

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., Glorioso, M., & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. Journal of Applied Psychology, 108(8), 1277–1299. https://doi.org/10.1037/apl0001082

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Goldberg, L. R. (1990). An alternative “Description of personality”: The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*, 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Gonzalez-Mulé, E., Mount, M. K., & Oh, I.-S. (2014). A meta-analysis of the relationship between general mental ability and nontask performance. *Journal of Applied Psychology, 99*(6), 1222–1243. https://doi.org/10.1037/a0037547

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., ... Synnaeve, G. (2024). The Llama 3 herd of models. https://doi.org/10.48550/arXiv.2407.21783

Gutiérrez, J. L. G., Jiménez, B. M., Hernández, E. G., & Pcn, C. (2005). Personality and subjective well-being: Big five correlates and demographic variables. *Personality and Individual Differences, 38*(7), 1561-1569. https://doi.org/10.1016/j.paid.2004.09.015

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel psychology*, *57*(3), 639-683. https://doi.org/10.1111/j.1744-6570.2004.00003.x

Heimann, A. L., Ingold, P. V., Debus, M. E., & Kleinmann, M. (2021). Who will go the extra mile? Selecting organizational citizens with a personality-based structured job interview. *Journal of Business and Psychology*, *36*(6), 985-1007. https://doi.org/10.1007/s10869-020-09716-1

Heller, D., Ferris, D. L., Brown, D., & Watson, D. (2009). The influence of work personality on job satisfaction: Incremental validity and mediation effects. *Journal of Personality*, *77*(4), 1051-1084. https://doi.org/10.1111/j.1467-6494.2009.00574.x

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, *107*(8), 1323–1351. https://doi.org/10.1037/apl0000695

Hickman, L., Tay, L., & Woo, S. E. (2019). Validity evidence for off-the-shelf language-based personality assessment using video interviews: Convergent and discriminant relationships with self and observer ratings. *Personnel Assessment and Decisions*, *5*(3), 3. https://doi.org/10.25035/pad.2019.03.003

Hickman, L., Tay, L., & Woo, S. E. (2024). Are automated video interviews smart enough? Behavioral modes, reliability, validity, and bias of machine learning cognitive ability assessments. *Journal of Applied Psychology.* Advance online publication. https://doi.org/10.1037/apl0001236

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*(1-2), 152-194. https://doi.org/10.1111/1468-2389.00171

Huffcutt, A. I. (2011). An empirical review of the employment interview construct literature. *International Journal of Selection and Assessment*, *19*(1), 62–81. https://doi.org/10.1111/j.1468-2389.2010.00535.x

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86*(5), 897–913. https://doi.org/10.1037/0021-9010.86.5.897

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology*, *67*(5), 577–580. https://doi.org/10.1037/0021-9010.67.5.577

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87*(3), 530–541. https://doi.org/10.1037/0021-9010.87.3.530

Kamdar, D., & Van Dyne, L. (2007). The joint effects of personality and workplace social exchange relationships in predicting task performance and citizenship performance. *Journal of Applied Psychology, 92*(5), 1286. https://doi.org/10.1037/0021-9010.92.5.1286

Kamdar, D., & Van Dyne, L. (2007). The joint effects of personality and workplace social exchange relationships in predicting task performance and citizenship performance. *Journal of Applied Psychology, 92*(5), 1286–1298. https://doi.org/10.1037/0021-9010.92.5.1286

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). Large language models are zero-shot reasoners. https://arxiv.org/abs/2205.11916

Lee, M. C., Scheepers, H., Lui, A. K., & Ngai, E. W. (2023). The implementation of artificial intelligence in organizations: A systematic literature review. *Information & Management, 60*(5), 103816. https://doi.org/10.1016/j.im.2023.103816

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, *67*(1), 241-293. https://doi.org/10.1111/peps.12052

Li, L., Zhang, B., Sun, T., & Drasgow, F. (2024). The more contextualized, the more valid: Effects of contextualization strategies on forced-choice measurement. *Journal of Business and Psychology*, 1-19. https://doi.org/10.1007/s10869-024-09983-2

Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology, 102*(1), 43–66. https://doi.org/10.1037/apl0000160

Lievens, F., De Corte, W., & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology, 93*(2), 268–279. https://doi.org/10.1037/0021-9010.93.2.268

Luo, J., Zhang, B., Cao, M., & Roberts, B. W. (2023). The stressful personality: A meta-analytical review of the relation between personality and stress. *Personality and Social Psychology Review*, *27*(2), 128-194. https://doi.org/10.1177/10888683221104002

Macan, T. (2009). The employment interview: A review of current studies and directions for future research. *Human Resource Management Review*, *19*(3), 203-218. https://doi.org/10.1016/j.hrmr.2009.03.006

McAdams, D. P. (2008). The Life Story Interview. https://www.sesp.northwestern.edu/foley/instruments/interview/

McAdams, D. P. (2013). The Psychological Self as Actor, Agent, and Author. *Perspectives on Psychological Science, 8*(3), 272-295. https://doi.org/10.1177/1745691612464657

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, *79*(4), 599–616. https://doi.org/10.1037/0021-9010.79.4.599

Meta AI (2024). *Meta Llama 3.1*. *https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1*

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

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. III. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology*, *88*(2), 348–355. https://doi.org/10.1037/0021-9010.88.2.348

OpenAI. (2023). Whisper API. https://openai.com/api/whisper

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*(2), 322–338. https://doi.org/10.1037/a0014996

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 28492–28518). PMLR.

Roulin, N., Bourdage, J. S., & Wingate, T. G. (2019). Who is conducting "better" employment interviews? Antecedents of structured interview components use. *Personnel Assessment and Decisions*, *5*(1), 2. https://doi.org/10.25035/pad.2019.01.002

Russell, S. S., Spitzmüller, C., Lin, L. F., Stanton, J. M., Smith, P. C., & Ironson, G. H. (2004). Shorter can also be better: The abridged job in general scale. *Educational and Psychological Measurement, 64*(5), 878-893. https://doi.org/10.1177/0013164404264841

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

Salgado, J. F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology*, *11*(3), 299–324. https://doi.org/10.1080/13594320244000184

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://doi.org/10.48550/arXiv.1910.01108

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Le Scao, T., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., … Rush, A. M. (2021). *Multitask prompted training enables zero-shot task generalization* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2110.08207

Siocon, G. 2023. Ways AI is changing HR departments. Business News Daily. https://www.businessnewsdaily.com/how-ai-is-changing-hr

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Spector, P. E., & Fox, S. (2010). Counterproductive work behavior and organisational citizenship behavior: Are they opposite forms of active behavior? *Applied Psychology*, *59*(1), 21-39. https://doi.org/10.1111/j.1464-0597.2009.00414.x

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, *71*(3), 299-333. https://doi.org/10.1111/peps.12263

Speer, A. B. (2021). Scoring dimension-level job performance from narrative comments: Validity and generalizability when using natural language processing. *Organizational Research Methods*, *24*(3), 572-594. https://doi.org/10.1177/1094428120930815

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin, 134*(1), 138–161. https://doi.org/10.1037/0033-2909.134.1.138

Sun, T., Roberts, B., Drasgow, F., & Zhou, M. X. (2024). Development and validation of an artificial intelligence chatbot to assess personality. https://doi.org/10.31234/osf.io/ahtr9

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality*, *34*(5), 826-844. https://doi.org/10.1002/per.2290

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*(3), 500–517. https://doi.org/10.1037/0021-9010.88.3.500

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, *34*(4), 397-423. https://doi.org/10.1006/jrpe.2000.2292

Tett, R. P., & Simonet, D. V. (2021). Applicant faking on personality tests: Good or bad and why should we care?. *Personnel Assessment and Decisions*, *7*(1), 2. https://doi.org/10.25035/pad.2021.01.002

Thompson, I., Koenig, N., Mracek, D. L., & Tonidandel, S. (2023). Deep learning in employee selection: Evaluation of algorithms to automate the scoring of open-ended assessments. *Journal of Business and Psychology*, *38*(3), 509-527. https://doi.org/10.1007/s10869-023-09874-y

Van Iddekinge, C. H., Raymark, P. H., & Roth, P. L. (2005). Assessing personality with a structured employment interview: construct-related validity and susceptibility to response inflation. *Journal of Applied Psychology*, *90*(3), 536–552. https://doi.org/10.1037/0021-9010.90.3.536

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1706.03762

Wang, P., Myeong, H., & Oswald, F. L. (2024). On putting the horse (raters and criteria) before the cart (variance components in ratings). *Industrial and Organizational Psychology*, *17*(3), 309–313. https://doi.org/10.1017/iop.2024.16

Weyhrauch, W. S., & Huffcutt, A. I. (2017). A tale of two formats: Direct comparison of matching situational and behavior description interview questions. *Human Resource Management Review*, *27*, 167-177. https://doi.org/10.1016/j.hrmr.2016.09.009

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of G-theory methods for modeling multitrait–multimethod data: Clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods*, *15*(1), 134-161. https://doi.org/10.1177/10944281114086

Woo, S. E., Tay, L., & Oswald, F. (2024). Artificial intelligence, machine learning, and big data: Improvements to the science of people at work and applications to practice. *Personnel Psychology*, *77*(4), 1387-1402. https://doi.org/10.1111/peps.12643

Xiao, Z., Zhou, M. X., Liao, Q. V., Mark, G., Chi, C., Chen, W., & Yang, H. (2020). Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI), 27*(3), 1-37. https://doi.org/10.1145/3381804

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105. https://doi.org/10.1037/h0046016

Zhang, T., Koutsoumpis, A., Oostrom, J. K., Holtrop, D., Ghassemi, S., & de Vries, R. E. (2024). Can large language models assess personality from asynchronous video interviews? A comprehensive evaluation of validity, reliability, fairness, and rating patterns. *IEEE Transactions on Affective Computing*, *15*(3), 1769-1785. https://doi.org/10.1109/TAFFC.2024.3374875

Zhou, M. X., Chen, W., Xiao, Z., Yang, H., Chi, T., & Williams, R. (2019, March). Getting virtually personal: chatbots who actively listen to you and infer your personality. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion* (pp. 123-124).

Table 1

Means, Standard Deviations, Reliabilities, and Intercorrelations of Study Variables for Chatbot Interview Samples.

	Students (Sample 2)	Working Adults (Sample 3)										
Variables	*M* (*SD*)	*M* (*SD*)	1	2	3	4	5	6	7	8	9	10
Self-reports												
1. Extraversion	3.27 (0.63)	3.33 (0.74)	-	.44**	.43**	.56**	.53**	.28*	.36**	.01	-.04	.14
2. Agreeableness	3.79 (0.48)	4.01 (0.63)	.19*	-	.82**	.74**	.48**	.18	.36**	.02	.03	.06
3. Conscientiousness	3.66 (0.50)	4.12 (0.63)	.25*	.36**	-	.71**	.47**	.12	.27*	.00	-.04	.04
4. Emotional Stability	3.10 (0.74)	3.65 (0.72)	.32**	.41**	.41**	-	.45**	.25*	.36**	.12	.07	.04
5. Openness	3.60 (0.61)	3.66 (0.74)	.28*	.19*	.18*	.06	-	.16	.23*	.13	-.11	.14
Human ratings												
6. Extraversion	3.13 (0.58)	3.21 (0.67)	.19*	.06	.11	.15	.10	-	.40**	.55**	.47**	.60**
7. Agreeableness	3.23 (0.66)	3.23 (0.77)	.14	.29**	.25*	.13	.13	.00	-	.46**	.24*	.44**
8. Conscientiousness	3.11 (0.63)	3.15 (0.65)	.07	.14	.35**	.34**	.11	.19*	.37**	-	.36**	.55**
9. Emotional Stability	3.06 (0.74)	3.06 (0.82)	-.06	.16	.07	.18*	.08	.28*	.23*	.32**	-	.33*
10. Openness	2.97 (0.65)	3.31 (0.67)	-.08	.16	.22*	.09	.14	.16	.36**	.43**	.36**	-
Embedding												
11. Extraversion	3.14 (0.14)	3.27 (0.22)	.11	.17	.05	.12	.27	.38**	.28**	.33**	.37**	.40**
12. Agreeableness	3.26 (0.17)	3.31 (0.23)	-.01	.19*	.01	.05	.14	.07	.62**	.40**	.29**	.41**
13. Conscientiousness	3.14 (0.19)	3.23 (0.27)	.08	.11	.05	.20*	.27*	.07	.37**	.62**	.30**	.38**
14. Emotional Stability	3.08 (0.18)	3.15 (0.25)	-.05	.16	.03	.05	.30**	.14	.32**	.39**	.60**	.36**
15. Openness	3.08 (0.20)	3.27 (0.33)	.05	.11	.01	.02	.27*	.04	.28*	.39**	.17	.56**
Zero-shot												
16. Extraversion	2.80 (0.71)	3.14 (0.94)	.21*	-.03	.09	.13	.24*	.56**	.14	.28*	.18*	.29**
17. Agreeableness	3.14 (0.64)	3.15 (0.89)	.14	.23*	.13	.03	.21*	.02	.70**	.23*	.23*	.25*
18. Conscientiousness	2.84 (0.75)	2.87 (0.96)	.03	.15	.20*	.17	.21*	.10	.25*	.71**	.26*	.41**
19. Emotional Stability	3.52 (0.77)	3.44 (0.86)	.00	.21*	.14	.17	.12	.18*	.20*	.32**	.72**	.32**
20. Openness	2.44 (0.75)	2.91 (0.95)	-.01	.13	.09	.04	.21*	.11	.25*	.40**	.23*	.71**

Outcome variables												
21. OCB	3.11 (0.59)	2.95 (0.68)	.32**	.15	.26*	.01	.20*	.22*	.16	.18*	.12	.15
22. CWB	1.79 (0.53)	1.63 (0.68)	-.06	-.46**	-.36**	-.28*	-.16	-.11	-.22*	-.21*	-.13	-.28*
23. Task Performance	4.04 (0.54)	4.30 (0.60)	.37**	.22*	.54**	.21*	.08	-.02	.18*	.15	-.06	.10
24. Job Satisfaction	2.69 (0.32)	2.69 (0.45)	.20*	.12	.17	.19*	.03	.12	.02	.11	.05	-.12
25. SWB	3.57 (0.85)	3.48 (0.93)	.37**	.20*	.27*	.40**	-.10	.05	.07	.16	-.13	-.15
26. Stress	2.85 (0.64)	2.61 (0.76)	-.32**	-.34**	-.42**	-.81**	.01	-.07	-.06	-.23*	-.10	-.02

Notes. Correlations for the student sample (Sample 2) are shown below the diagonal and correlations for the working adult sample (Sample 3) are shown above the diagonal. Alphas for the chatbot-based sample are presented on the diagonal. For rater, embedding, and zero-shot scores, reliability was calculated using scores for trait-level questions. Meaning reliability for emotional stability was calculated using scores from the three questions targeted at emotional stability. Notably, this means reliability, especially for rater and zero-shot scores, is limited since it's based on just three questions. $N_{\text{Sample 2}}$ = 130; $N_{\text{Sample 3}}$ = 88. * p <.05 and ** p <.001.

Table 1 continued

	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Self-reports																
1. Extraversion	.09	.15	-.03	-.14	.06	.30*	.34*	.09	-.11	.10	.48**	-.21*	.53**	.61**	.34*	-.49**
2. Agreeableness	.13	.17	.01	.01	.06	.22*	.40**	.16	.11	.07	.09	-.64**	.55**	.42**	.22*	-.55**
3. Conscientiousness	.00	.10	-.05	-.05	-.04	.09	.30*	.12	.09	-.01	.10	-.57**	.60**	.32*	.13	-.53**
4. Emotional Stability	.11	.19	.01	-.01	.03	.20	.41**	.18	.11	.03	.23*	-.40**	.62**	.56**	.46**	-.81**
5. Openness	.14	.09	.01	-.12	.00	.23*	.22*	.08	-.06	.18	.29*	-.30*	.51**	.36**	.22*	-.38**
Human ratings																
6. Extraversion	.61**	.54**	.56**	.53**	.60**	.67**	.56**	.51**	.38**	.50**	.21*	.03	.19	.35**	.20	-.31*
7. Agreeableness	.45**	.64**	.49**	.43**	.52**	.43**	.79**	.53**	.23*	.51**	.35**	-.16	.37**	.38**	.28*	-.39**
8. Conscientiousness	.65**	.64**	.77**	.61**	.73**	.51**	.46**	.74**	.28*	.70**	.35**	.22*	.09	.27*	.19	-.21
9. Emotional Stability	.33**	.38**	.41**	.66**	.44**	.37**	.27*	.36**	.69**	.28*	.00	.04	.01	.05	.00	-.12
10. Openness	.56**	.63**	.62**	.52**	.62**	.48**	.52**	.46**	.33*	.74**	.28*	.16	.14	.21	.14	-.16
Embedding																
11. Extraversion	-	.74**	.77**	.68**	.79**	.74**	.57**	.67**	.33**	.73**	.19	.05	.12	.25	.24*	-.22
12. Agreeableness	.52**	-	.80**	.64**	.77**	.54**	.75**	.71**	.36**	.73**	.33*	.11	.18	.27*	.18	-.26*
13. Conscientiousness	.55**	.61**	-	.71**	.84**	.57**	.53**	.79**	.42**	.81**	.27*	.17	.09	.18	.17	-.20
14. Emotional Stability	.62**	.52**	.58**	-	.74**	.55**	.44**	.59**	.61**	.60**	.09	.16	.06	.05	.06	-.10
15. Openness	.54**	.49**	.63**	.48**	-	.58**	.57**	.69**	.38**	.81**	.28*	.14	.09	.18	.26*	-.21
Zero-shot																
16. Extraversion	.55**	.22*	.36**	.28*	.33**	-	.58**	.55**	.33*	.57**	.24*	-.05	.20	.28*	.13	-.29*
17. Agreeableness	.30**	.56**	.37**	.41**	.29**	.28*	-	.64**	.31*	.58**	.30*	-.16	.29*	.39**	.23*	-.40**
18. Conscientiousness	.41**	.42**	.69**	.42**	.49**	.34**	.22*	-	.44**	.66**	.19	.01	.15	.21	.06	-.25*
19. Emotional Stability	.41**	.30**	.30**	.60**	.21*	.19*	.27*	.34**	-	.27*	-.15	-.15	.08	-.09	-.03	-.15

20. Openness	.47**	.37**	.56**	.35**	.70**	.38**	.27*	.53**	.26*	-	.32*	.15	.14	.26*	.24*	-.16
Outcome variables																
21. OCB	.12	.05	.12	.14	.06	.19*	.22*	.26*	.14	.04	-	.26*	.24*	.52**	.38**	-.25*
22. CWB	.01	-.09	.01	-.07	-.04	-.01	-.11	-.09	-.12	-.12	-.05	-	-.35**	-.20	-.12	.41**
23. Task Performance	.06	.04	.04	.00	.03	.08	.14	.13	.12	.09	.23*	-.22*	-	.41**	.41**	-.55**
24. Job Satisfaction	.13	.09	.02	.00	-.04	.03	.03	.10	.18*	-.04	.10	-.06	.22*	-	.47**	-.55**
25. SWB	.09	.06	.13	-.04	.02	.04	.09	.10	-.04	-.01	.08	-.08	.43**	.33**	-	-.62**
26. Stress	-.12	-.02	-.18*	-.07	-.02	-.06	.00	-.08	-.16	-.02	.04	.20*	-.30**	-.28*	-.47**	-

Notes. Correlations for the student sample (Sample 2) are shown below the diagonal and correlations for the working adult sample (Sample 3) are shown above the diagonal. Alphas for the chatbot-based sample are presented on the diagonal. For rater, embedding, and zero-shot scores, reliability was calculated using scores for trait-level questions. Meaning reliability for emotional stability was calculated using scores from the three questions targeted at emotional stability. Notably, this means reliability, especially for rater and zero-shot scores, is limited since it's based on just three questions. $N_{\text{Sample 2}}$ = 130; $N_{\text{Sample 3}}$ = 88. * p <.05 and ** p <.001.

Table 2

Correlations between Self and Rater Personality Scores

	Human interview			**Chatbot interview**					
	Sample 1 (working adults) *N* = 203			Sample 2 (students) *N* = 130			Sample 3 (working adults) *N* = 88		
	Self-report	Human interview rating		Self- report	Human interview rating		Self- report	Human interview rating	
	M (SD)	*M* (*SD*)	*r*	*M* (*SD*)	*M* (*SD*)	*r*	*M* (*SD*)	*M* (*SD*)	*r*
Extraversion	3.68 (0.50)	3.81 (0.58)	.43**	3.27 (0.63)	3.13 (0.58)	.19*	3.33 (0.74)	3.21 (0.67)	.28*
Agreeableness	3.87 (0.41)	3.80 (0.55)	.39**	3.79 (0.48)	3.23 (0.66)	.29**	4.01 (0.63)	3.23 (0.77)	.36**
Conscientiousness	4.14 (0.43)	3.98 (0.48)	.27**	3.66 (0.50)	3.11 (0.63)	.35**	4.12 (0.63)	3.15 (0.65)	.00
Emotional Stability	3.90 (0.52)	3.86 (0.52)	.20*	3.10 (0.74)	3.06 (0.74)	.18*	3.65 (0.72)	3.06 (0.82)	.07
Openness	3.89 (0.47)	3.93 (0.60)	.40**	3.60 (0.61)	2.97 (0.65)	.14	3.66 (0.74)	3.31 (0.67)	.14

Notes. *M* = Mean; *SD* = Standard deviation; *r* = Pearson’s correlation coefficient. * *p* <.05 and ** *p* <.001.

Table 3

Means and Standard Deviations of Word Count by Interview Question

	Human interview	Chatbot interview	
	Sample 1 (working adults)	Sample 2 (students)	Sample 3 (working adults)
Extraversion			
Question 1	203.02 (108.38)	55.22 (34.79)	54.22 (37.81)
Question 2	206.86 (108.93)	45.26 (24.84)	53.61 (32.65)
Question 3	133.09 (61.42)	41.40 (24.22)	44.42 (27.88)
Agreeableness			
Question 1	162.10 (88.53)	52.63 (35.68)	50.18 (32.74)
Question 2	164.00 (99.79)	43.45 (25.97)	42.68 (29.50)
Question 3	169.21 (97.91)	39.38 (22.93)	46.47 (35.67)
Conscientiousness			
Question 1	148.19 (62.83)	50.52 (32.52)	55.00 (32.66)
Question 2	157.51 (73.55)	42.32 (28.71)	45.16 (27.26)
Question 3	162.00 (87.39)	39.01 (18.99)	42.51 (28.10)
Emotional Stability			
Question 1	217.69 (105.33)	65.49 (42.64)	60.20 (36.17)
Question 2	156.95 (67.48)	43.95 (23.23)	50.61 (33.88)
Question 3	157.60 (89.14)	43.07 (24.29)	46.19 (29.10)
Openness			
Question 1	149.49 (67.62)	47.55 (34.08)	55.13 (34.24)
Question 2	187.22 (90.71)	47.45 (30.90)	49.92 (29.72)
Question 3	120.63 (57.45)	32.15 (18.66)	42.23 (34.70)
Combined	2495.55 (779.46)	688.84 (336.87)	738.60 (380.83)

Notes. Values represent means and standard deviations (in the parentheses) for word count for each question across the three samples. Combined includes combined response text from all interview questions. $N_{\text{Sample 1}}$ = 203; $N_{\text{Sample 2}}$ = 130; $N_{\text{Sample 3}}$ = 88.

Table 4

Multitrait-Multimethod Statistics for Machine-Derived Personality Trait Scores for Sample 2 (students)

	Self-report	Human rating (chatbot interview)	Embedding (chatbot interview)	Zero-shot (chatbot interview)			Average *r*
Heterotrait-monomethod (HTMM)	.264	.270	.554	.308			.349
	Self-report to human rating	Self-report to embedding	Human rating to embedding	Zero-shot to self-report	Zero-shot to human rating	Zero-shot to embedding	
Heterotrait-heteromethod (HTHM)	.132	.111	.292	.116	.233	.361	.207
Monotrait-heteromethod (MTHM)	.232	.136	.553	.203	.680	.619	.404
	C1	D1	D2	MV			
Variance Partitioning	.404	.196	.055	.142			

Note. Convergence Index (C1) = average of monotrait-heteromethod correlations. Discrimination Index 1 (D1) = C1 – average of heterotrait-heteromethod correlations. Discrimination Index 2 (D2) = C1 – average of heterotrait-monomethod correlations. Method variance (MV) = average of hetero-monomethod correlations – average of heterotrait-heteromethod correlations.

Table 5

Multitrait-Multimethod Statistics for Machine-Derived Personality Trait Scores for Sample 3 (working adults)

	Self-report	Human rating (chatbot interview)	Embedding (chatbot interview)	Zero-shot (chatbot interview)			Average *r*
Heterotrait-monomethod (HTMM)	.562	.441	.748	.494			.561
	Self-report to human rating	Self-report to embedding	Human rating to embedding	Zero-shot to self-report	Zero-shot to human rating	Zero-shot to embedding	
Heterotrait-heteromethod (HTHM)	.136	.075	.531	.154	.433	.568	.316
Monotrait-heteromethod (MTHM)	.171	.044	.661	.221	.725	.739	.427
	C1	D1	D2	MV			
Variance Partitioning	.427	.110	-.134	.245			

Note. Convergence Index (C1) = average of monotrait-heteromethod correlations. Discrimination Index 1 (D1) = C1 – average of heterotrait-heteromethod correlations. Discrimination Index 2 (D2) = C1 – average of heterotrait-monomethod correlations. Method variance (MV) = average of hetero-monomethod correlations – average of heterotrait-heteromethod correlations.

Table 6

Regression of Organizational Citizenship Behavior Outcome on Personality Self-Reports and Machine-Derived Personality Scores in the Chatbot Interview

	Sample 2 (student)						Sample 3 (working adult)					
	Step 1		Step 2		Step 2		Step 1		Step 2		Step 2	
			Embedding		Zero-shot				Embedding		Zero-shot	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Extraversion (S)	.29**	.08			.27**	.08	.44**	.09			.42**	.09
Extraversion (M)					.11	.07					.08	.07
R^2	.10**				.12**		.23**				.24**	
ΔR^2					.02						.01	
Agreeableness (S)	.18	.11	.17	.11	.12	.11	.09	.12	.04	.11	-.04	.12
Agreeableness (M)			.08	.31	.18*	.08			.95*	.31	.24*	.09
R^2	.02		.02		.06*		.01		.11*		.09*	
ΔR^2			.00		.04*				.10*		.08*	
Conscientiousness (S)	.30*	.10	.30*	.10	.25*	.10	.11	.12	.13	.11	.08	.12
Conscientiousness (M)			.31	.26	.17*	.07			.70*	.26*	.13	.08
R^2	.07*		.08*		.11**		.01		.09*		.04	
ΔR^2			.01		.04*				.08*		.03	
Emotional Stability (S)	<.01	.07	<.01	.07	-.01	.07	.22*	.10	.22*	.10	.24*	.10
Emotional Stability (M)			.45	.29	.11	.07			.26	.29	-.14	.08
R^2	<.01		.02		.02		.05*		.06		.08	
ΔR^2			.02		.02				.01			
Openness (S)	.19*	.08	.19*	.09	.20*	.08	.26*	.10	.26*	.09	.22*	.09
Openness (M)			.02	.26	-.01	.07			.57*	.20	.20*	.07
R^2	.04*		.04		.04		.08*		.16**		.16**	
ΔR^2			.00		.00				.08*		.08**	

Notes. S = self-reported score; M = machine-derived score. **p* < .05. ***p* < .001. $N_{\text{Sample 2}}$ = 130; $N_{\text{Sample 3}}$ = 88.

Table 7

Regression of Counterproductive Work Behavior Outcome on Personality Self-Reports and Machine-Derived Personality Scores in the Chatbot Interview

	Sample 2 (student)						Sample 3 (working adult)					
	Step 1		Step 2		Step 2		Step 1		Step 2		Step 2	
			Embedding		Zero-shot				Embedding		Zero-shot	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Extraversion (S)	-.04	.07	-.04	.07	-.05	.08	-.20*	.10	-.20*	.10	-.20	.10
Extraversion (M)			.05	.38	<.01	.07			.19	.30	.01	.08
R^2	<.01		<.01		<.01		.05*		.05		.05	
ΔR^2			.00		.00						.00	
Agreeableness (S)	-.51**	.09	-.51**	.09	-.51**	.09	-.69**	.09	-.73**	.09	-.75**	.10
Agreeableness (M)			-.01	.26	<.01	.07			.66*	.24	.09	.07
R^2	.21**		.21**		.21**		.41**		.46**		.43**	
ΔR^2			.00		.00				.05*		.02	
Conscientiousness (S)	-.38**	.09	-.38**	.09	-.38**	.09	-.62**	.10	-.61**	.09	-.63**	.10
Conscientiousness (M)			.08	.23	-.01	.06			.34	.22	.06	.06
R^2	.13**		.13**		.13**		.33**		.35**		.34**	
ΔR^2			.00		.00				.02		.01	
Emotional Stability (S)	-.20*	.06	-.20*	.06	-.19*	.06	.38**	.09	-.38**	.09	-.37**	.09
Emotional Stability (M)			-.17	.25	-.05	.06			.43	.27	-.08	.08
R^2	.08*		.08*		.09*		.16**		.18**		.17**	
ΔR^2			.00		.01				.02		.01	
Openness (S)	-.13	.08	-.13	.08	-.12	.08	-.28*	.09	-.28*	.09	-.31*	.09
Openness (M)			.01	.24	-.06	.06			.29	.21	.15*	.07
R^2	.03		.03		.03		.09*		.11*		.13*	
ΔR^2			.00		.00				.02		.04*	

Notes. S = self-reported score; M = machine-derived score. **p* < .05. ***p* < .001. $N_{\text{Sample 2}} = 130$; $N_{\text{Sample 3}} = 88$.

Table 8

Regression of Job Performance Outcome on Personality Self-Reports and Machine-Derived Personality Scores in the Chatbot Interview

	Sample 2 (student)						Sample 3 (working adult)					
	Step 1		Step 2		Step 2		Step 1		Step 2		Step 2	
			Embedding		Zero-shot				Embedding		Zero-shot	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Extraversion (S)	.31**	.07	.31**	.07	.31**	.07	.43**	.07	.43**	.07	.42**	.08
Extraversion (M)			.06	.36	<.01	.06			.25	.23	.03	.06
R^2	.13**		.13**		.13**		.29**		.30**		.29**	
ΔR^2			.00		.00				.01		.00	
Agreeableness (S)	.25*	.10	.25*	.10	.23*	.10	.52**	.08	.51**	.09	.49**	.09
Agreeableness (M)			<.01	.29	.08	.07			.24	.24	.05	.07
R^2	.05*		.05*		.06*		.31**		.32**		.31**	
ΔR^2			.00		.01				.01		.00	
Conscientiousness (S)	.58**	.08	.58**	.08	.57**	.08	.56**	.08	.57**	.08	.55**	.08
Conscientiousness (M)			.04	.21	.02	.05			.26	.19	.05	.05
R^2	.29**		.29**		.29**		.36**		.37**		.36**	
ΔR^2			.00		.00				.01		.00	
Emotional Stability (S)	.15*	.06	.15*	.06	.14*	.06	.52**	.07	.52**	.07	.52**	.07
Emotional Stability (M)			-.02	.26	.06	.06			.16	.21	<.01	.06
R^2	.04*		.04		.05*		.39**		.39**		.39**	
ΔR^2			.00		.01				.00		.00	
Openness (S)	.07	.08	.07	.08	.06	.08	.41**	.07	.41**	.07	.41**	.08
Openness (M)			.04	.25	.05	.06			.16	.16	.03	.06
R^2	.01		.01		.01		.26**		.27**		.27**	
ΔR^2			.00		.00				.01		.01	

Notes. S = self-reported score; M = machine-derived score. **p* < .05. ***p* < .001. $N_{\text{Sample 2}}$ = 130; $N_{\text{Sample 3}}$ = 88.

Table 9

Regression of Job Satisfaction Outcome on Personality Self-Reports and Machine-Derived Personality Scores in the Chatbot Interview

	Sample 2 (student)						Sample 3 (working adult)					
	Step 1		Step 2		Step 2		Step 1		Step 2		Step 2	
			Embedding		Zero-shot				Embedding		Zero-shot	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Extraversion (S)	.10*	.04	.09*	.04	.10*	.04	.37**	.05	.37**	.05	.35**	.05
Extraversion (M)			.27	.22	-.01	.04			.34*	.16	.05	.04
R^2	.04*		.05*		.04		.38**		.41**		.39**	
ΔR^2			.01		.00				.03*		.01	
Agreeableness (S)	.08	.06	.07	.06	.08	.06	.30**	.07	.27**	.07	.22*	.07
Agreeableness (M)			.14	.17	<.01	.04			.40*	.19	.13*	.05
R^2	.01		.02		.01		.18**		.22**		.24**	
ΔR^2			.01		.00				.03*		.05*	
Conscientiousness (S)	.11	.05	.11	.06	.10	.06	.23*	.07	.24*	.07	.22*	.07
Conscientiousness (M)			.01	.15	.03	.04			.32	.16	.08	.05
R^2	.03		.03		.03		.11*		.14*		.13*	
ΔR^2			.00		.00				.03		.02	
Emotional Stability (S)	.08*	.04	.08*	.04	.07	.04	.35**	.06	.35**	.06	.36**	.06
Emotional Stability (M)			-.01	.15	.06	.04			.10	.16	-.08	.05
R^2	.03*		.03		.06*		.31**		.31**		.33**	
ΔR^2			.00		.03						.02	
Openness (S)	.01	.05	.02	.05	.02	.05	.22**	.06	.22**	.06	.20*	.06
Openness (M)			-.09	.14	-.02	.04			.25	.13	.10*	.05
R^2	<.01		<.01		<.01		.13**		.16**		.17**	
ΔR^2			.00		.00				.03		.05*	

Notes. S = self-reported score; M = machine-derived score. **p* < .05. ***p* < .001. $N_{\text{Sample 2}} = 130$; $N_{\text{Sample 3}} = 88$.

Table 10

Regression of Subjective Well-being Outcome on Personality Self-Reports and Machine-Derived Personality Scores in the Chatbot Interview

	Sample 2 (student)						Sample 3 (working adult)					
	Step 1		Step 2		Step 2		Step 1		Step 2		Step 2	
			Embedding		Zero-shot				Embedding		Zero-shot	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Extraversion (S)	.49**	.11	.48**	.11	.50**	.11	.43*	.13	.42*	.12	.41*	.13
Extraversion (M)			.27	.56	-.04	.10			.78	.39*	.03	.11
R^2	.13**		.14**		.13**		.12*		.16**		.12*	
ΔR^2			.01		.00				.04*		.00	
Agreeableness (S)	.36*	.15	.36*	.16	.35*	.16	.32*	.15	.28	.16	.22	.17
Agreeableness (M)			.11	.45	.05	.12			.61	.43	.17	.12
R^2	.04*		.04		.04		.05*		.07*		.07*	
ΔR^2			.00		.00				.02		.02	
Conscientiousness (S)	.46*	.14	.45*	.14	.45*	.15	.19	.16	.20	.16	.18	.16
Conscientiousness (M)			.53	.38	.05	.10			.61	.36	.04	.10
R^2	.08*		.09*		.08*		.02		.05		.02	
ΔR^2			.01		.00				.03		.00	
Emotional Stability (S)	.45**	.09	.46**	.09	.47**	.09	.59**	.12	.59**	.12	.60**	.12
Emotional Stability (M)			-.28	.39	-.12	.09			.23	.36	-.09	.10
R^2	.16**		.16**		.17**		.21**		.21**		.22**	
ΔR^2			.00		.01				.00		.01	
Openness (S)	-.14	.12	-.16	.13	-.14	.13	.28*	.13	.28*	.13	.24	.13
Openness (M)			.23	.39	.02	.10			.73*	.28	.20	.10
R^2	.01		.01		.01		.05*		.12*		.09*	
ΔR^2			.00		.00				.07*		.04	

Notes. S = self-reported score; M = machine-derived score. *p < .05. **p < .001. $N_{\text{Sample 2}}$ = 130; $N_{\text{Sample 3}}$ = 88.

Table 11

Regression of Stress on Personality Self-Reports and Machine-Derived Personality Scores in the Chatbot Interview

	Sample 2 (student)						Sample 3 (working adult)					
	Step 1		Step 2		Step 2		Step 1		Step 2		Step 2	
			Embedding		Zero-shot				Embedding		Zero-shot	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Extraversion (S)	-.32**	.08	-.31**	.09	-.32**	.09	-.50**	.10	-.50**	.10	-.45**	.10
Extraversion (M)			-.42	.43	<.01	.08			-.50	.30	-.13	.08
R^2	.10**		.11**		.10*		.24**		.26**		.26**	
ΔR^2			.01		.00				.02		.02	
Agreeableness (S)	-.46**	.11	-.47**	.11	-.48**	.11	-.67**	.11	-.64**	.11	-.57**	.12
Agreeableness (M)			.18	.33	.08	.09			-.58	.30	-.18*	.08
R^2	.12**		.12**		.12**		.31**		.34**		.34**	
ΔR^2			.00		.00				.03		.03*	
Conscientiousness (S)	-.54**	.10	-.53**	.10	-.54**	.10	-.65**	.11	-.66**	.11	-.62**	.11
Conscientiousness (M)			-.54*	.27	<.01	.07			-.63*	.25	-.15*	.07
R^2	.18**		.21**		.18**		.29**		.34**		.32**	
ΔR^2			.03*		.00				.05*		.03*	
Emotional Stability (S)	-.70**	.05	-.70**	.05	-.69	.05	-.86**	.07	-.86**	.07	-.85**	.07
Emotional Stability (M)			-.10	.19	-.02	.04			-.33	.20	-.05	.06
R^2	.65**		.65**		.65**		.65**		.66**		.65**	
ΔR^2			.00		.00				.01		.00	
Openness (S)	.01	.09	.01	.09	.01	.09	-.39**	.10	-.39**	.10	-.38**	.10
Openness (M)			-.08	.29	-.02	.08			-.47*	.22	-.08	.08
R^2	<.01		<.01		<.01		.15**		.19**		.15**	
ΔR^2			.00		.00				.04*		.00	

Notes. S = self-reported score; M = machine-derived score. **p* < .05. ***p* < .001. $N_{\text{Sample 2}}$ = 130; $N_{\text{Sample 3}}$ = 88.