

Automating personality-based employment interviews: A systematic review, development and validation of an AI chatbot, and practical considerations

by

Ashley Sylvara

B.S., Missouri Western State University, 2017

M.S., University of Nottingham, 2019

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Abstract

Advancements in Artificial Intelligence (AI) and natural language processing (NLP) have enabled the development of innovative methods to assess personality using textual responses elicited from structured behavior-based interviews, chatbot-based open-ended interviews, and automated video interviews. This dissertation aims to advance the field by conceptually integrating literature on automated, personality, and employment interviews and developing an AI chatbot that administers and assesses personality through situational personality-based employment interviews, providing practical utility for selection purposes. This dissertation comprises three main studies. In Part I, I conducted a systematic literature review of research on employment, personality, and automated interviews to synthesize existing findings, identify key themes and gaps, and explore methodological advancements. Results from structural topic modeling uncovered five latent themes and highlighted overlap and underrepresented themes areas within the domains of internet research. In Part II, I developed and validated an AI chatbot to administer a personality-based employment interview, leveraging word embeddings and zero-shot prompt engineering to assess personality from participants' textual responses. Results demonstrated moderate convergence between chatbot-derived and self-reported personality scores, though response quality (e.g., brevity in chatbot interviews) influenced scoring accuracy. In Part III, I examined the practical considerations of implementing AI-based selection tools, including user perceptions, subgroup differences, and resistance to response distortion. Results revealed mixed support for chatbot-based assessments in selection contexts. While participants found the chatbot engaging, concerns remained regarding usability, fairness, and faking resistance.

Automating personality-based employment interviews: A systematic review, development and validation of an AI chatbot, and practical considerations

by

Ashley Sylvara

B.S., Missouri Western State University, 2017
M.S., University of Nottingham, 2019

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Approved by:

Co-Major Professor
Tianjun Sun, Ph.D.

Approved by:

Co-Major Professor
Jin Lee, Ph.D.

Copyright

© Ashley Sylvara 2025.

Abstract

Advancements in Artificial Intelligence (AI) and natural language processing (NLP) have enabled the development of innovative methods to assess personality using textual responses elicited from structured behavior-based interviews, chatbot-based open-ended interviews, and automated video interviews. This dissertation aims to advance the field by conceptually integrating literature on automated, personality, and employment interviews and developing an AI chatbot that administers and assesses personality through situational personality-based employment interviews, providing practical utility for selection purposes. This dissertation comprises three main studies. In Part I, I conducted a systematic literature review of research on employment, personality, and automated interviews to synthesize existing findings, identify key themes and gaps, and explore methodological advancements. Results from structural topic modeling uncovered five latent themes and highlighted overlap and underrepresented themes areas within the domains of internet research. In Part II, I developed and validated an AI chatbot to administer a personality-based employment interview, leveraging word embeddings and zero-shot prompt engineering to assess personality from participants' textual responses. Results demonstrated moderate convergence between chatbot-derived and self-reported personality scores, though response quality (e.g., brevity in chatbot interviews) influenced scoring accuracy. In Part III, I examined the practical considerations of implementing AI-based selection tools, including user perceptions, subgroup differences, and resistance to response distortion. Results revealed mixed support for chatbot-based assessments in selection contexts. While participants found the chatbot engaging, concerns remained regarding usability, fairness, and faking resistance.

Table of Contents

List of Figures	viii
List of Tables	ix
Acknowledgements.....	xi
Chapter 1 - Introduction.....	1
Chapter 2 - Part I: Systematic Literature Review of Interviews in Personality and Organizational Research.....	14
Method.....	16
Results.....	20
Discussion.....	26
Chapter 3 - Part II: Development and Validation an AI Chatbot and Scoring Algorithm.....	35
Methods	40
Results.....	51
Discussion.....	58
Chapter 4 - Part III: Practical Considerations for Selection Contexts	65
Methods	71
Results.....	80
User Perceptions	81
Subgroup Differences	82
Faking Resistance	83
Discussion.....	84
Chapter 5 - General Discussion	88
References.....	91
Appendix A - Supplementary Materials for Part I.....	160
Example Web Scraper for Bibliographic Database	160
PRISMA 2020 Checklist	163
Appendix B - Supplementary Materials for Part II.....	167
AI Chatbot Interview Example	167
Example Behavior Description Interview Questions and Scoring Guide from Heimann et al. (2021	169

Embedding-based Model Selection Performance	172
Example Zero-shot Prompt	176
Appendix C - Supplementary Materials for Part III	177
Selected Examples from Part II for Few Shot Learning	178
Example Few-shot Prompt.....	181
Condition Assignment Scripts	182
Appendix D - Survey Measures	185
Contextualized Personality	185
User Perceptions and Engagement	188
Outcome Variables	191

List of Figures

Figure 1. Visualization of SLR Article Filtering	156
Figure 2. Visualization of Topics and Top Words.....	157
Figure 3. User Engagement Comparisons Between the AI Chatbot and the Personality Measure	158
Figure 4. Boxplots of Personality Trait Scores by Condition and Assessment Method.....	159
Appendix Figure B.2. Visualization of model performance for each question across Personality Traits	175

List of Tables

Table 1. Model Selection Indices for Structural Topic Modeling	129
Table 2. Overview of Topic Information	130
Table 3. STM Results of Search Term Effect on Topic Prevalence	131
Table 4. Means, Standard Deviations, and Intercorrelations of Part II Study Variables for Chatbot Interview Samples.....	132
Table 5. Correlations between Self and Rater Personality Scores	134
Table 6. Means and Standard Deviations of Word Count by Interview Question	135
Table 7. Multitrait-Multimethod Statistics for Machine-Derived Personality Trait Scores for Sample 2 (students).....	136
Table 8. Multitrait-Multimethod Statistics for Machine-Derived Personality Trait Scores for Sample 3 (working adults)	137
Table 9. Regression of Organizational Citizenship Behavior Outcome on Personality Self- Reports and Machine-Derived Personality Scores in the Chatbot Interview.....	138
Table 10. Regression of Counterproductive Work Behavior Outcome on Personality Self-Reports and Machine-Derived Personality Scores in the Chatbot Interview	139
Table 11. Regression of Job Performance Outcome on Personality Self-Reports and Machine- Derived Personality Scores in the Chatbot Interview	140
Table 12. Regression of Job Satisfaction Outcome on Personality Self-Reports and Machine- Derived Personality Scores in the Chatbot Interview	141
Table 13. Regression of Subjective Well-being Outcome on Personality Self-Reports and Machine-Derived Personality Scores in the Chatbot Interview	142
Table 14. Regression of Stress on Personality Self-Reports and Machine-Derived Personality Scores in the Chatbot Interview	143
Table 15. Correlations for Study Variables for Part III for Combined and Honest condition....	144
Table 16. Correlations for Study Variables for Part III for Induced and Motivated Faking conditions.....	146
Table 17. Means and Standard Deviations of Word Count by Interview Question	148
Table 18. Means and Standard Deviations of Personality Trait for Part III	149

Table 19. Correlations Between Machine-derived and Self-reported Personality Trait Scores for Part III	150
Table 20. Paired <i>t</i> -Tests User Engagement Scores Between the Scale and the AI Chatbot	151
Table 21. Means, Standard Deviations, and Comparisons of User Engagement and Perceptions Across Conditions.....	152
Table 22. Comparisons of Personality Traits by Gender and Age Across Different Measures .	153
Table 23. Comparisons of Personality Traits by Race Across Different Measures for Samples from Part III	154
Table 24. Mixed Design ANOVA Results and Pairwise Comparisons.....	155
Appendix Table A.1. PRISMA 2020 Checklist.....	163
Appendix Table B.2. Domain- and Question-Level Correlations of Embedding-Based Model Predictions.....	174
Appendix Table C.1. Descriptive Statistics for Word Count in Responses with Perfect Agreement Across Human Raters and LLM from Part II.....	177
Appendix Table C.2. Examples for Few Shot Learning	178

Acknowledgements

This academic journey has been quite the rollercoaster, and I could not have done it without the support of my advisors, peers, friends, partner, and of course, my dog. First, I would like to acknowledge my advisors, Dr. Tianjun Sun and Dr. Jin Lee, without whom I would not have become the academician I am today. Dr. Tianjun Sun, thank you for your unwavering support, patience, and mentorship, and especially for always keeping the “1-800-Tianjun hotline” open. Dr. Jin Lee, thank you for your invaluable guidance, encouragement, and mentorship throughout my academic development. I would also like to thank my committee members, Dr. Sarina Maneotis and Dr. Rachael Xiong, for their guidance and feedback on this dissertation, as well as my lab members for their friendship and collaboration over the years. Lastly, I would like to thank my partner, Phillip, and my dog for their constant support and listening to all my practice presentations.

Chapter 1 - Introduction

Personality measures have long played a key role in organizational contexts, demonstrating their significance from academic performance and achievement (McAbee & Oswald, 2013; Noftle & Robins, 2007) to job performance, satisfaction, and organizational citizenship behaviors (Barrick et al., 2001; Chiaburu et al., 2011; Sackett & Walmsley, 2014). These inventories are widely used in selection assessments since they are relatively cost-effective and easy to administer (Evans & Mathur, 2005) and demonstrate low adverse impact compared to other selection methods such as biodata, situational judgement tests, and general mental ability tests (Hough et al., 1990; Sackett et al., 2022). Yet, despite these important links, 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). 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. As personality influences many aspects of an interviewee, including qualifications, academic performance, and workplace behavior (Gonzalez-Mulé et al., 2014; Poropat, 2009; Salgado & Moscoso, 2002), 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). 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 (McLellan et al., 2003). 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, 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) have been developed to infer an individual's Big Five personality traits through an interactive, narrative-based interview (Fan et al., 2023; Sun et al., 2024). While this approach has demonstrated utility in inferring personality, narrative-based interviews ask participants to describe significant moments across their life and are not suitable for selection contexts.

This dissertation addresses an important gap in the literature by conceptually integrating literature on automated, personality, and employment interviews and developing an AI chatbot that administers and assesses personality through situational personality-based employment interviews, providing practical utility for selection purposes. This dissertation includes three parts. In Part I, I conducted a systematic literature review of research on automated interviews, personality interviews, and employment interviews, to synthesize existing research, and identify key themes and gaps within and between these fields. In Part II, I developed and validated an AI chatbot-based personality assessment using a job-relevant behavior description interview. In Part III, I examined practical considerations for this AI-based interview in selection contexts, including user perceptions and engagement with the developed AI chatbot, adverse impact in machine-derived scores, and resistance to response distortion. In the following sections of Chapter 1, I discuss why personality matters in work contexts, how personality has traditionally been measured, the use of employment interviews to assess personality, and introduce AI and NLP to help address limitations of both traditional personality measures and employment interviews.

Personality and the Workplace

Personality traits conceptualize an individual's pattern of cognitions, emotions, and behaviors (Goldberg, 1990), making them critical factors in understanding work behavior. Based on the Five Factor Model (i.e., the Big Five; McCrae & Costa, 1997), personality is conceptualized through five latent factors: extraversion, associated with being sociable, assertive, and talkative; agreeableness, associated with being cooperative, flexible, tolerant, and forgiving; conscientiousness, associated with being dependable, achievement-striving, persevering, and orderly; emotional stability (conceptualized as the opposite of neuroticism), associated with

being calm, self-confident, and resilient; and openness to experience, associated with being curious, broad-minded, and original (Costa & McCrae, 1999; Goldberg, 1990). As such, personality has been linked to many work outcomes, including job performance (Barrick & Mount, 1991; Hurtz & Donovan, 2000), job satisfaction (Heller et al., 2009; Judge et al., 2002), subjective well-being (Costa & McCrae, 1980; Steel et al., 2008), and workplace behavior, such as organizational citizenship behaviors (OCBs; Chiaburu et al., 2011), counterproductive work behaviors (CWBs; Berry et al., 2007), and procrastination behaviors (Kim et al., 2017). Conscientiousness, in particular, has consistently been the strongest personality predictor of job performance (Sackett & Walmsley, 2014; Barrick et al., 2001; Salgado, 2003). Moreover, personality can even relate to workplace safety behaviors, for example Beus et al. (2015) found that conscientiousness and agreeableness were associated with safer workplace behaviors, whereas extraversion and neuroticism were positively associated with unsafe behaviors.

When understanding how personality influences workplace behavior, it's important to consider how personality traits are conceptualized. One perspective is the dispositional approach, wherein the focus is on broad characteristics of individual differences and suggests their stability across situations (Costa & McCrae, 1999). The other approach takes an interactionist approach and focuses on the interaction between a specific situation and the individual, such as the cognitive-affective personality (CAPS) system approach (Mischel & Shoda, 1995) and trait activation theory (TAT; Tett & Gutterman, 2000). Under TAT, behavior results from the interaction between the person and the situation, wherein situational cues activate certain traits, and specific traits are expressed in behavior when the situation and context allows (Woods et al., 2013). For example, in assessment centers, exercises that require social interaction and relational skills (e.g., leaderless group discussions) have higher trait activation potential for extraversion,

whereas activities that require organization and planning (e.g., in-basket and case analysis) have higher activation potential for conscientiousness (Lievens et al., 2006; Meriac et al., 2014; Speer et al., 2015). Thus, understanding how situational factors influence the expression of personality traits is crucial in determining why certain traits are better suited for specific jobs. For example, agreeableness and openness to experience are particularly relevant in roles that involve interpersonal interaction, such as teamwork and client-facing positions (Mount et al., 1998; Nikolaou, 2003). Similarly, in high-risk occupations, conscientiousness and agreeableness have been linked to safer workplace behaviors, while extraversion and neuroticism are associated with higher risk-taking and unsafe behaviors (Beus et al., 2015). Given that trait activation depends on specific contextual cues, traditional approaches to personality assessment, which often rely on Likert-style, self-report methods, may not fully capture how personality manifests in real-world job settings.

Traditional Approaches to Personality

In psychological research, Likert scales have long been the dominant method for assessing psychological constructs and have undergone significant methodological advancements over the past 30 years to enhance their psychometric validity (Jebb et al., 2021). While these measures are widely used and established as effective tools for capturing personality traits, they still come with notable limitations, particularly in selection contexts. One key concern is their susceptibility to response distortion, as candidates may be motivated to present as an ‘ideal’ version of themselves to increase Person-Organization fit (Roulin & Krings, 2020). Consider an item assessing conscientiousness from the Big Five Inventory (BFI-2; Soto & John, 2017): “*I am someone who tends to be lazy.*” When applying for a job, a candidate might engage in impression management by inflating their responses to appear more conscientious than they

actually are. Even personality assessments specifically designed for organizational use, such as the Hogan Personality Inventory (HPI; Hogan & Hogan, 1992), are not free from this concern. For example, websites like <https://personality-test-prep.com/test-providers/hogan/hpi/> provide not only an overview of the assessment but also example statements, suggested responses, and tips on how to create a "realistic profile" while maintaining subscale consistency, essentially guiding individuals on how to optimize their results rather than provide entirely candid responses.

Another concern is that personality inventories do not capture how personality traits manifest into actual workplace behaviors and rely on self-perceptions (Baumeister et al., 2007). In contrast to this, other reports have demonstrated value in personality assessment beyond self-reports (Connelly & Ones, 2010; Oh et al., 2011), as they offer insights into how individuals are perceived by others based on their behaviors and interactions (McAbee & Connelly, 2016; Funder & Colvin, 1988), which is especially important in organizational contexts. Moreover, other reports have been found to be stronger predictors of performance outcomes whereas self-reported personality were largely unrelated (Connelly et al., 2021). However, obtaining other reports of personality-relevant behavior in organizational settings is resource-intensive and, in selection contexts, often relies on candidate-nominated raters.

A third concern about self-report personality inventories is survey fatigue. These Likert-style methods often induce survey fatigue for participants, which can decrease the quality of responses (Porter, 2004; Huang et al., 2015). Comprehensive inventories like the IPIP-NEO (Goldberg, 1999), with its 300 items and 40–60-minute completion time, exemplify this issue. Longer surveys tend to increase dropout rates and contribute to careless responding (Galesic & Bosnjak, 2009). However, survey fatigue may be moderated by the salience of survey content,

meaning if individuals perceive the assessment as useful, engaging, or meaningful, they are more likely to complete it attentively (Heerwegh & Loosveldt, 2009).

Employment Interviews

An employment interview is defined as a personally interactive process in which one or more individuals ask questions to a job candidate and assess their responses to evaluate qualifications for making employment decisions (Levashina et al., 2014). Employment interviews are highly flexible assessment tools, in that they can vary in structure, content, and the competencies being evaluated. However, their design and implementation can impact their validity and effectiveness in selection contexts. Interviews range from unstructured to structured, generally structured interviews demonstrate greater validity than unstructured interviews (Huffcutt & Arthur, 1994; Judge et al., 2000), but both have their advantages and disadvantages (Mueller & Segal, 2015). In unstructured interviews, there are no constraints on the questions asked and no standardized guidelines for evaluating responses. In contrast, highly structured interviews use predetermined questions, with each response assessed using a defined rating scale, ensuring greater consistency and reliability.

Often, interviews aim to assess personality traits (Heimann et al., 2021; Salgado & Moscoso, 2002; Van Iddekinge et al., 2005). Meta-analyses found that personality traits, particularly facets of conscientiousness, were assessed through interview questions (Cortina et al., 2000; Huffcutt et al., 2001). Recently, there has been increased research on personality interviews, likely due to relationships with criteria (e.g., job performance) and low adverse impact (Levashina et al., 2014). Additionally, drawing from Heimann et al.'s (2021) theoretical framework for assessing personality through personality-based structured interviews, personality measures are better predictors of performance when they are contextualized, for interviews to

adequately predict performance the questions should be job-related, and interviews that feature an open response format place higher cognitive demands on applicants and decreases intentional response distortion compared to self-report measures (Van Iddekinge et al., 2005). Thus, personality-based employment interviews are a useful tool for predicting job performance.

Along with question design considerations, interviewer and rater training also impacts the validity and reliability of interviews (Conway et al., 1995; Macan, 2009; Powell & Bourdage, 2016). For instance, Roulin et al. (2019) found that interviewer training was important to ensure consistent use of questions and use of standardized evaluation benchmarks, but that the personality of interviewers is slightly related to inconsistencies. Such that interviewers higher in extraversion engaged in more rapport building behaviors including use of probing questions, suggesting that while training is beneficial, it does not ensure complete consistency across raters. Thus, it may be beneficial for researchers and practitioners to explore additional methods to increase consistency and structure during the interview process.

Artificial Intelligence and Natural Language Processing Applied to Personality

A Brief Introduction to NLP

AI is a technology that enables machines to simulate human intelligence. Machine Learning (ML), a sub-discipline of AI, uses algorithms to learn from large datasets, and NLP, a branch of ML, enables machines to understand and interpret human language (Manning & Schütze; 1999). In NLP, methods can be broadly categorized into supervised (using labeled data) and unsupervised techniques. For supervised methods, three primary approaches have emerged: dictionary-based, representation-based machine learning, and large language models (Feuerriegel et al., 2024).

Dictionary-based approaches, among the earliest NLP methods, rely on predefined word lists to classify text into categories. This represents a closed vocabulary approach, where data is passed through a predefined lexicon, or dictionary. A popular example is Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2001), which utilizes a dictionary of over 2,300 words and stems to analyze linguistic and psychological constructs. While dictionary-based methods offer high interpretability, they are limited in their ability to capture nuanced meaning, such as sarcasm or context-dependent shifts in language.

Representation-based machine learning extends beyond dictionary-based methods by transforming text into numerical formats that can be analyzed using statistical or machine learning models. This represents an open vocabulary approach, which is more data-driven. These representations range from bag-of-words models, which count word frequencies, to more sophisticated word embedding-based methods, which capture contextual relationships between words, allowing for more nuanced linguistic analysis. These representations can then be fed into machine learning models, including linear regression, decision trees, and neural networks, to predict linguistic or psychological constructs. A major shift in NLP came with the introduction of transformer models (Vaswani et al., 2017). Unlike traditional models, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), transformers utilize a self-attention mechanism that allows them to weigh the relative importance of different words within a sentence (Vaswani et al., 2017). This process enables transformers to capture context more effectively than previous approaches, thus making them a powerful tool in understanding language. Transformers are composed of encoders and decoders, which process input sequences (e.g., a sequence of words in a sentence) and transform them into vector representations that capture the contextual relationships between words (i.e., contextualized word embeddings). One

of the most popular transformers is BERT (Devlin et al., 2018), which is an encoder-only model that captures the bi-directional relationships among words. Encoder-only structures are particularly adept in producing deep, context-based representations of text input, and are effective for tasks such as named entity recognition and text classification (Patwardhan et al., 2023). In contrast, decoder-only architectures generate text in an autoregressive manner, meaning that the text is generated word-by-word, with each word prediction being made based on the previously generated sequence of words. These models are designed for text generation and language modeling tasks, with GPT-3 and GPT-4 being the most notable examples. Finally, encoder-decoder transformer models, such as DeBERTa (He et al., 2020), are effective for sequence-to-sequence tasks. For instance, Whisper (OpenAI, 2022) translates speech to text by using encoders to process audio input, then leverages decoders to generate the corresponding text sequence.

Lastly, large language models (LLMs) represent the most advanced NLP approach and are built upon the transformer architecture. LLMs are deep learning architectures trained on massive text corpora using self-supervised learning, enabling them to capture intricate semantic and syntactic patterns in language (Feuerriegel et al., 2024). Unlike earlier approaches, LLMs generalize across tasks, allowing them to be fine-tuned for specific applications (e.g., personality assessment, psychological profiling, and behavioral analysis). A notable strength of LLMs is their ability to perform zero-shot learning (where the model is given a task without labeled examples) or few-shot learning (where the model is given a small number of labeled examples) with minimal or no fine-tuning across a range of tasks. For example, the LLaMA 3 models have been shown to generalize effectively across tasks, achieving performance comparable to other widely used LLMs such as GPT-4 (Grattafiori et al., 2024).

Artificial Intelligence for Personality Assessment

These advancements in NLP and ML have allowed for the development of innovative methods to assess personality using textual responses gathered from social media and big data (Bleidorn et al., 2017; Park et al., 2015), chatbot-based open-ended questions (Fan et al., 2023; Sun et al. 2024; Jayaratne & Jayatilleke, 2020), and through automated video interviews (AVIs; Hickman et al., 2022). Early approaches in this area relied on text from social media, such as Facebook and Twitter posts. For example, Kern et al. (2014) analyzed Facebook posts by extracting linguistic features, specifically unigrams, bigrams, and trigrams, and correlating their frequencies with the Big Five personality dimensions. Their findings demonstrated links between words used in posts and self-reported personality; for instance, individuals with higher levels of neuroticism tended to use more negative expressions (e.g., “depression,” “I hate,” frequent swear words), whereas those with higher emotional stability were more likely to use positive expressions (e.g., “workout,” “success,” “beautiful day”). Similarly, Park et al. (2015) utilized Facebook status posts to infer personality by tokenizing the text from posts, applying LDA topic modeling (Blei et al., 2003) to identify latent themes, performing dimensionality reduction to condense features, and then regressed features on self-reported personality scores. Their approach achieved convergent correlations ranging from $r = .35$, for agreeableness and neuroticism, to $r = .43$ for openness. Kosinski et al. (2013) used digital records of users’ Facebook ‘Likes’ and to predict personal attributes, including gender, age, sexual orientation, personality, and drug use. Their results showed that Facebook ‘Likes’ could be used to predict gender, sexual orientation, smoking behavior, age, ethnicity, and relationship status, indicating these digital records contain rich data and signals about a person’s identity and behavior.

Though rich in data, these sources have raised significant privacy and ethical concerns (Oswald et al., 2020), as well as concerns regarding psychometric validity (Bleidorn & Hopwood, 2019; Tay et al., 2020). Some more recent and promising approaches to assessing personality include automated video interviews (AVIs; Hickman et al., 2022) and chatbot-based interviews (Fan et al., 2023; Sun et al., 2024). In automated video interviews, verbal, paraverbal, and nonverbal behaviors serve as inputs for machine learning models. Hickman et al. (2022) extracted verbal behavior using LIWC features and document-term matrices from transcribed text, paraverbal behavior from features like pitch, voice quality, and speech rate, and nonverbal behavior from features like facial expressions and head movements. They found that machine-derived scores demonstrated some mixed evidence for reliability, good convergent and discriminant validity, and predictive validity for academic outcomes. Their findings also indicated that verbal behavior contributed most to prediction across traits (with exception to extraversion) compared to paraverbal and nonverbal behavior, aligning with similar research on AVI research suggesting that nonverbal features increase subgroup differences without adding substantial additional predictive power in modeling (Booth et al., 2021; Chen et al., 2016; Hickman et al., 2024b).

Unlike AVIs, chatbot interviews rely solely on textual cues, which serve as indicators of personality traits, with specific words and phrases providing insights into trait-relevant features. Unlike human interviewers, AI-powered chatbots can conduct thousands of interviews daily, offering more objective and efficient personality assessments (Zhou et al., 2019). Machine learning algorithms extract trait-relevant features from text responses, using self-reported personality scores as ground truth for model training (Fan et al., 2023). While these machine-inferred personality scores demonstrate acceptable reliability and convergent validity, they

exhibit lower discriminant validity (Azucar et al., 2018; Fan et al., 2023; Sun, 2021; Tay et al., 2020). Additionally, chatbot-based assessments appear to be more resistant to applicant faking (Völkel et al., 2020) and receive similar ratings in terms of perceived fairness, satisfaction, and usability compared to traditional personality surveys (Yuan et al., 2024).

This dissertation aims to contribute to both research and practice by exploring how AI and NLP can enhance personality assessment in selection contexts. Conceptually, it advances our understanding of AI-driven selection assessments; methodologically, it introduces a transparent, validated AI-based personality assessment tool; and practically, it provides organizations with a cost-effective, scalable solution for evaluating candidates. To achieve these goals, the following studies systematically review existing research, develop and validate an AI chatbot-based personality interview, and examine key practical considerations such as user perceptions, fairness, and resistance to response distortion.

Chapter 2 - Part I: Systematic Literature Review of Interviews in Personality and Organizational Research

In today's fast-paced research environment, it is essential for scholars to keep up with the growing volume of literature, especially in interdisciplinary fields in which research spans multiple domains and artificial intelligence (AI) is increasingly integrated. Systematic literature reviews (SLRs) are particularly useful for providing a comprehensive overview of a research topic, identifying gaps and trends in the literature, and offering directions for future research (Paul & Barari, 2022). While traditional reviews often involve an unstructured and flexible process for compiling relevant literature, SLRs adopt a structured approach to ensure that the selection of literature is not solely dependent on the author's expertise. The SLR process generally involves three key phases: a planning phase, where a premise and protocol for the review are established; a conducting phase, which includes identifying and selecting primary studies, followed by the extraction of relevant information; and a reporting phase, where the methodology is transparently documented, and the results are presented (Kitchenham, 2004). While more rigorous and structured, the SLR process is labor-intensive, time-consuming, and error-prone due to the manual nature of collecting, identifying, and reviewing significant amounts of literature (Van Dinter et al., 2021).

Especially a well-conducted SLR, or even meta-analysis, will follow strict guidelines (e.g., PRISMA 2020; Page et al., 2021) to ensure methodological rigor, standardization, and transparency. Michelson and Reuter (2019) estimated that conducting one SLR in the pharmaceutical industry costs around \$140,000 and advocated for more automated solutions using machine learning tools. In a review on automating SLRs, van Dinter et al. (2021) proposed fully automating the article filtering process in conducting SLRs, but very few studies have

explored automation in the planning and reporting phases of SLRs. Though more recent AI-based tools have become available to semi-automate this process, usually focusing on the planning phase and conducting phases (Bolaños et al., 2024), but some offer support for the reporting phase by generating summaries from documents (e.g., Iris.ai). Addressing these limitations, Xiong et al. (2023) developed a data analytics-based SLR protocol that uses semi-automated web scraping tools, natural language processing (NLP), and text analytics to streamline the planning, conducting, and reporting phases, creating a more systematic, comprehensive SLR process.

Given the increased integration of AI into personnel selection, particularly in interviews and for personality assessment, an SLR is warranted and will serve as a useful overview of this interdisciplinary topic. By synthesizing research from organizational, psychological, and data sciences, this part of my dissertation aims to explore latent themes across different interview domains and differences in topic prevalence, as well as showcase how data analytics can be leveraged to conceptually integrate literature from diverse fields, such as psychology. In doing so, I aim to address the following research questions:

RQ1: What latent themes emerge across different domains of interview research (i.e., automated, personality, and employment interviews)?

RQ2: How does the prevalence of these themes differ across the domains of interview research?

RQ3: What gaps, or underrepresented themes, exist across the domains of interview research?

Method

To effectively gather a broad range of literature from organizational, psychological, and data science journals, I used Xiong et al.'s (2023) semi-automated SLR protocol¹ and web scraping tools. Web scraping tools were developed in Python (version 3.11.7), and text analysis was conducted in R (version 4.4.1; R Core Team, 2024) using the *stm* package (Roberts et al., 2019).

Defining and Refining the Search Scope

Similar to a traditional SLR, the first step involves defining the search scope through relevant keywords or phrases. I defined and tested keyword sets using Google Scholar searches, ensuring they generated a sufficient number of relevant articles. Initially, I used exact match search terms, including “personality interview”, “employment interview”, and “automated interview”, which produced 1250, 18000, and 733 results, respectively as of May 13, 2024. I then tested additional search terms including “automated video interviews”, “open-ended interview” AND “personality”, “job interview”, and “selection interview”. The term “automated video interviews” was too narrow, yielding 271 articles, many of which overlapped with “automated interviews”. Similarly, “open-ended interview” AND “personality” could be covered by broadening the search to non-exact matches, such as “personality+interview”. I decided to include the variations of “employment interview” (i.e., “job interview” and “selection interview”) since these terms are often used interchangeably in the literature and would not always be captured by non-exact match searches. To collect more relevant articles, the goal was

¹ This review did not strictly adhere to PRISMA guidelines (Page et al., 2021) since it follows Xiong et al.'s (2023) semi-automated data analytics-based protocol and does not have the same level of human coding and decision-making as a traditional SLR. However, all data extraction and filtering steps are reported in the Methods section, and a PRISMA checklist with notes on which steps were followed and applicable is included in Appendix Table A.1.

to narrow down search terms to titles and abstracts. While PsycNet's advanced search allows for this more targeted approach, Google Scholar advanced search allows search terms to appear anywhere in the text or in the title. This, combined with the piloting of keywords, led to broadening the search terms to use the Boolean operator AND instead of exact matches. The final search terms included “personality+interview”, “automated+interview”, and variations of selection-based interviews “employment+interview”, “job+interview”, and “selection+interview”.

Data Collection and Handling

Articles were collected from three databases: ScienceDirect, Springer, and PsycNet. ScienceDirect and Springer were selected since they are prominent publishers in a broad range of disciplines, including business and management, and publish interdisciplinary research, which relates to automated employment interviews. PsycNet was selected due to its comprehensive coverage of psychological research, since personality assessments and selection techniques are core areas of interest within Industrial/Organizational psychology. The automated data collection tools are based on Xiong et al. (2023) and an example can be found in Appendix A.

Initially, a total of 6,518 articles were extracted from ScienceDirect ($N = 2,002$), Springer ($N = 2,247$), and PsycNet ($N = 2,269$). Then duplicates were removed ($N = 475$) across the three datasets, followed by articles without valid abstracts ($N = 162$), and finally, using the *cld3* package in R (Ooms, 2024), the language of the abstract was classified and articles with non-English abstracts were excluded ($N = 115$). The next step involved refining the search results to focus on the most relevant articles. For data collection in ScienceDirect and Springer, the Boolean operator "AND" was used in the search queries (e.g., “personality+interview”), which broadened the results by returning articles where the search terms appeared anywhere in the document. While this approach allowed for capturing a wider range of articles, it also introduced

a considerable amount of noise. Many articles retrieved this way may only mention interviews briefly, such as in the introduction, methods, or discussion section, without making them a central focus of the study. Additionally, book chapters mentioning the interview process were removed by filtering out abstracts containing “this chapter” ($N = 26$). This phrase was selected after manually reviewing abstracts, which consistently began with phrases like “In this chapter...” or “This chapter [discusses/covers/etc.]...”

Differently, in PsycNet’s advanced search the same keywords but the search was limited to the article title and abstracts. To address this difference in search terms, articles that did not include the search terms in either the titles or abstracts were excluded, ensuring a closer alignment between the terms and the article’s main topic. Given the different scopes of the topic in the review, I employed tailored strategies for filtering the search results. For employment-related interviews, three specific search terms were used: “job interview,” “selection interview,” and “employment interview.” Since the substantial size of the sample for employment interviews, a more stringent criteria was applied by requiring exact matches of these phrases in the titles or abstracts. For personality and automated interviews, I used broader search terms: “personality+interview” and “automated+interview.” Recognizing the need for more flexibility in these areas, partial matches were allowed, meaning the terms could appear in various forms as long as the abstracts still dealt with these topics substantively. This filtering process removed 5,006 articles, leaving a final sample of 734 articles. Figure 1 includes a visualization of this filtering process and comparison with PRIMA guidelines (Page et al., 2020) can be found in Appendix A.

Analytic Strategy

To uncover latent themes within the filtered collection of abstracts, I used topic modeling, a machine learning method designed to identify latent topics based on patterns of word co-occurrence across documents, in our case, abstracts. To explore differences in topic prevalence based on the research domain, I employed structural topic modeling (STM), which extends the traditional LDA (Latent Dirichlet Allocation) topic modeling approach by incorporating metadata to explain topic prevalence, such as covariates (Roberts et al., 2019). LDA is a generative probabilistic model that helps uncover latent topics in a collection of documents (Blei et al., 2003) and assumes that each document is a mixture of multiple topics, and each topic is represented by a probability distribution over words that frequently co-occur. In topic modeling, each document is conceptualized as a combination of different topics, while each topic is characterized by a probability distribution over words (Barde & Bainwad, 2017). Structural topic modeling (STM) extends this approach by allowing researchers to incorporate document-level metadata, enabling an exploration of how topics vary across different contexts or research domains. This makes STM particularly useful for analyzing differences in topic prevalence across fields.

To achieve this, I used search term presence as three metadata fields in the model. These fields represented whether the article title or abstract contained specific search terms: "personality+interview," "automated+interview," or employment-related terms (including "employment+interview," "job+interview," or "selection+interview"). Each term was coded as 0 if it was not present and 1 if it was present. This allows for examination of how frequently topics appeared in each research domain and provided insights into the distribution of topics across personality-based, automated, and employment interviews. Despite having more stringent filter criteria for employment-related articles, the sample size of employment-related articles ($N = 496$)

was disproportional to personality-related ($N = 249$) and automated-related ($N = 34$). This disparity is discussed further in the limitations section.

To analyze the abstracts, the text was first preprocessed through the following steps: tokenizing (breaking the text into individual words), stemming (reducing words to their root form, e.g., “interviewer” to “interview”), removing general stopwords (common words such as “and,” “the,” and “or”), removing context-specific stopwords (including all words used in the search terms), and removing punctuation and numbers. The number of topics was selected based on topic exclusivity (how unique words are to a given topic), semantic coherence (how well words within a topic relate to each other), held-out likelihood (how well the model generalizes to unseen data), and residuals (unexplained variance). Based on these statistics (see Table 1), a 5-topic solution based on a balance between semantic coherence and decent exclusivity, ensuring topics were distinct and interpretable. To determine topic labels, I examined the top words based on the score metric (see Table 2), which divides the log frequency of the word in the topic by the log frequency of the word in other topics (Roberts et al., 2019). In the results section, I discuss each topic in terms of labeling, top associated words and articles, and results from STM (see Table 3). See Figure 2 for visualization of the topics and their top words. Data can be accessed at https://osf.io/4vs9n/?view_only=7352bc6ba2944c5f9f2797ba03375d9c.

Results

Topic 1 – Clinical Assessment

Topic 1 was labeled as *Clinical Assessment* as top words and most representative articles emphasized the use of interviews for clinical purposes, namely personality disorder diagnosis. Top words within this topic included “BPD” (borderline personality disorder), “DSM” (referring to the DSM-5), “disord,” and “diagnost”. The most representative articles focused on interviews

for a broad range of personality diagnosis and functioning. Most representative articles included studies on profiling narcissism using the DSM-5 (Fossati et al., 2017), assessing personality functioning related to complex posttraumatic stress disorder (Maples-Keller et al., 2021; Møller et al., 2021), how parents personality functioning play a role in youth personality development (Mackin et al., 2022), and future interpersonal functioning (Calabrese et al., 2024), and validating a personality functioning scale for assessing personality disorders (Zettl et al., 2020). Collectively, these articles highlight the use and importance of clinical interviews for personality disorder diagnosis and role in clinical settings. While this topic is closely related to personality disorders, results from STM did not show a significant positive relationship with the personality search term covariate, rather topic prevalence was negatively related to automated (Estimate = -.25, $p <.001$) and employment-related search terms (Estimate = -.41, $p <.001$). Meaning that articles focusing on clinical assessments and personality disorders are not prevalent in discussions surrounding employment or automated interviews. This finding is expected as clinical assessments are not appropriate tools for selection contexts. However, integrating AI-based tools and automated interviews could still prove beneficial in clinical settings in standardizing evaluations and potentially providing more efficient solutions to healthcare and clinical personality assessments (e.g., King, 2023).

Topic 2 – Bias and Faking

Topic 2 was labeled as *Bias and Faking* as top words and most representative articles emphasized how biases and individual differences can affect interview processes and outcomes. Top words within this topic included “applic”, “interviewe”, “fake”, “nonverb”, and “impress”. Several of the most representative articles examine how interviewer–interviewee demographic differences, particularly gender (Binning et al., 1988; Reis et al., 1999) and race (de Kock &

Hauptfleisch, 2020; Prewett-Livingston et al., 1996), can bias interview outcomes. These articles draw on frameworks such as social identity theory (Tajfel and Turner, 1979) and similarity effects, including same-race rating effect and sex-similarity attraction, suggesting that interviews may favor candidates who share similarities. Other articles focus on how nonverbal behaviors and initial impressions impact evaluations. For example, research on authenticity (Heimann & Schmitz-Wilhelmy, 2024) and on cues like smiling, hand gestures, and other nonverbal signals (DeGroot & Gooty, 2009; Goldberg & Rosenthal, 1986; Tescari & Bangerter, 2024) indicates that such cues play a significant role in forming impressions. Additional articles highlight how preinterview information can bias ratings subsequent questioning strategies in unstructured interviews (Binning et al., 1988) and demonstrates that impression management and rapport-building can further influence interviewer judgments (Ho et al., 2019; Wingate & Bourdage, 2019). Along these lines articles focused on how interviewers process information based on favorability, normativity, and relevance to the job (Constantin, 1976) and how timing and sequence of interviews can impact evaluations (Steiner & Rain, 1989; Wexley et al., 1972), citing recency and contrast effects demonstrating how interviewer ratings can be skewed.

Methodologically, studies employing experimental designs often use prerecorded interviews to assess rater error associated with individual differences. In contrast, research involving actual job candidates typically use structured interview formats with clear behaviorally anchored rating scales for scoring (e.g., DeGroot & Gooty, 2009; Heimann & Schmitz-Wilhelmy, 2024; Osborn et al., 1998; Prewett-Livingston et al., 1996). Collectively, this body of research suggests that, regardless of interview structure, interviewers rely on both nonverbal behaviors and preinterview cues to form their judgments.

Results of the STM indicated a positive relationship between topic prevalence and personality (Estimate = .21, $p < .001$), automated (Estimate = .19, $p = .010$), and employment-related search terms (Estimate = .43, $p < .001$). The positive relationships across personality, automated, and employment-related interviews suggest that interviewer biases and faking behaviors are challenges that appear across these three research domains. Expectedly, this relationship is stronger with employment-related search terms, indicating these are more heavily addressed in relation to employment interviews as they are usually tied to important decisions, such as hiring, thus faking and bias may have particularly significant consequences.

Topic 3 – Legal Approaches to Fairness

Topic 3 was labeled as *Legal Approaches to Fairness* as top words and most representative articles examined various aspects of employment discrimination, hiring practices, and inclusive policies, with an emphasis on the impact of disability, ethnicity, gender, age, and mental health. Top words within this topic included “cancer”, “ethnic”, “labour”, and “accommod.” Most representative articles explored disability-inclusive employment policies for workplace accommodations (Blanck, 2022; Burke, 2018) and how disclosing mental health issues may contribute to difficulties in job the market (Baert et al., 2016; Bjørnshagen, 2021; Bjørnshagen & Ugreninov, 2021). Representative articles also focused on discrimination based on ethnicity and gender. Field experiments in European countries examined that applicants with foreign names or minority backgrounds are less likely to be invited for interviews (Asali et al., 2018; Midtbøen, 2016; Zschirnt, 2020). This also extends to older workers, as they may be perceived as less adaptable and requiring more training due to having less technical skills (Baert et al., 2016; Van Borm et al., 2021), this difference is particularly pronounced in physically demanding and lower-level clerical roles (Oesch, 2020). Results of the STM indicated a negative

relationship between topic prevalence and personality (Estimate = -.17, $p < .001$) and automated search terms (Estimate = -.14, $p = .024$), meaning articles focusing on hiring discrimination and legal frameworks were less prevalent in research surrounding personality and automated interviews. However, as automated interview methods are relatively new and concerns about machine learning bias and fairness continue to grow, these discussions may expand the focus of discrimination in selection beyond human biases to include biases introduced by machine learning algorithms.

Topic 4 – AI and ML Applications

Topic 4 was labeled as *AI and ML applications* as top words and most representative articles used ML techniques to capture aspects of interviewee behavior. Top words within this topic included “jit” (referring to virtual reality-job interview training [VR-JIT]), “virtual”, “user”, “skill”, “human”, and “system.” Most representative articles within this topic highlight aspects of AI and ML for employment interview contexts, such as using machine learning models to process verbal data (Qian et al., 2020; Yamamoto et al., 2023), detection of emotion and nonverbal behavior (Adiani et al., 2022; Bevilacqua et al., 2013; Chollet et al., 2014), and engagement tracking with virtual agents by using facial expressions and eye tracking data (Youssef et al., 2015; Yu et al., 2019). Notably some of these articles focus on developing AI-based tools to facilitate interview training (Adiani et al., 2023; Yu et al., 2019), which can also provide insights for refining model training approaches. While many articles focused on extracting nonverbal features, verbal data was processed using different NLP methods, including BoW (Ammari et al., 2012), Word2Vec embeddings (Qian et al., 2020), and dictionary-based approaches (Despotakis et al., 2013). Model training techniques included various forms of cross-validation, such as test-train splits (e.g., 80/20 or 60/40), k-fold cross-validation, and leave-one-

out cross-validation. Additionally, feature selection techniques were applied to reduce the number of features, retaining only the most relevant ones to improve model performance.

Results from STM indicated that topic prevalence was positively related to automated (Estimate = .16, p = .038), but negatively related to personality search terms (Estimate = -.13, p = .007). Meaning, articles focusing on AI and ML techniques were more prevalent when discussing automated interviews, but not for employment interviews compared to personality interviews. These findings suggest that articles discussing AI and machine learning techniques were more frequently associated with automated interviews but were less prevalent in the context of personality-based assessments. Additionally, the lack of a significant relationship with employment-related search terms indicates that discussions of AI-driven interview methods may not be fully integrated into broader employment research.

Topic 5 – Broad Applications of Interview Assessments

Topic 5 was labeled as *Broad Applications of Interview Assessments* as top words and most representative articles cover a wide range of uses for interview methods, they span from methodological discussions to psychological assessments to physiological responses to employment interview contexts. Top words within the topic included “rorschach”, “cortisol”, “tsst” (referring to the Trier Social Stress Test), “reliabi”, and “scale” (referring to automated video interview). Most representative articles discuss how evaluation techniques of the Rorschach test could be applied to clinical interviews (Gruen, 1955; Zubin et al., 1956), use of mock interviews to induce stress (Lee et al., 2023; Stanek et al., 2024; Young et al., 2001), development of automated systems for measuring health (Nease et al., 1996), and discussion on validity evidence of employment interviews (Huffcut & Arthur, 1994). Results from STM analysis indicate that topic prevalence was not significantly associated with any of the search

terms, suggesting that this topic is broadly distributed across the corpus rather than being dominated by a single research domain.

Discussion

In this study, I aimed to systematically review research on automated personality-based interviews by synthesizing literature from psychological, organizational, and data sciences. To do so, I explored latent themes and differences in topic prevalence across these domains, while showcasing how data analytics-based tools can be leveraged to integrate literature into highly interdisciplinary topics. The findings highlight emerging topics both within and across fields and raise important questions regarding the incorporation of AI, offering directions for future research.

In addressing RQ1, I identified five latent themes (i.e., topics) that emerged across different domains of interview research. These findings demonstrate the diverse applications of interviews, as well as important considerations of bias and fairness across contexts. Topic 1, *Clinical Assessment*, underscores the role of interviews in clinical settings, particularly for assessing personality functioning and diagnosing personality disorders. Interviews within this domain tend to be less structured than employment interviews, aiming to capture nuanced aspects of personality that may not be fully elicited with a highly structured interview. Topic 2, *Bias and Faking*, reflects concerns about how interviewee characteristics, such as race, gender, and personality traits, along with impression management strategies, can affect interview performance and the decision-making process. Topic 3, *Legal Approaches to Fairness*, reflects various aspects of employment discrimination and hiring practices, highlighting differential outcomes based on demographic and health factors. Topic 4, *AI and ML*, represents an emerging area in interview literature, driven by technological advancements as data science tools and AI

integrate with interview training and increasingly reshape the traditional interview process.

Lastly, Topic 5, *Broad Applications of Interview Assessments*, reflects how interviews are used broadly across domains, covering discussion on methodological considerations to using interviews for stress assessments.

For RQ2 and RQ3, I examined how topic prevalence varies across psychological, organizational, and data science domains, while identifying gaps and underrepresented areas in the literature. For Topic 1, *Clinical Assessments*, our results suggest a separation from employment-related contexts, which aligns with expectations, as clinical tools are designed for diagnostic purposes rather than evaluation or job suitability. However, the negative relationship with automated interviews demonstrates an underrepresented area in literature. For instance, a recent review highlighted the uses of machine learning models for clinical assessments of bipolar disorder, depression, and PTSD using audio and visual data (Low et al., 2020). These applications may not have emerged as a distinct theme in our analysis due to the emphasis toward employment and selection contexts, rather than clinical diagnostics. This underrepresentation, however, suggests an area that could be further explored and integrated into research on automated interviews. Many clinical assessments rely on a well-trained human interviewer to conduct semi-structured interviews, posing the question of how these interviews could be adapted to incorporate AI. For instance, replacing a human interviewer with an AI agent in diagnosing personality disorders could increase accessibility for individuals with limited resources for therapy. However, this substitution may lack the therapeutic benefits human interviewers provide through empathy and rapport-building in clinical settings (Szalai, 2021). Therefore, while AI may not yet fully replicate these complex interactions, there are promising avenues for using AI to enhance and partially automate the clinical interview process. However,

careful consideration is needed regarding the effectiveness of AI-based tools in capturing the nuances typically conveyed through human interaction and any potential impact on the content quality and depth of responses during the interview.

In contrast, Topic 2, *Bias and Faking*, was a prevalent topic across all domains, particularly within employment settings. This suggests that issues related to interviewer bias and impression management affect various interview types, including those designed for personality assessment. These challenges are particularly important for personality-based interviews, where accurately assessing an individual's personality traits is critical and has implications for hiring decisions. This also underscores a key point about how we conceptualize and mitigate bias in the interview process. Whereas human-based interviews typically focus on biases introduced by individual characteristics, such as differences in interviewer ratings, a data science perspective shifts the focus toward bias within machine learning algorithms. While machine learning models are often designed to enhance fairness through systematic processes, there is growing concern that these models may exacerbate inequalities (Landers & Behrend, 2023). Highlighting the need for AI-integrated theories, Hickman et al. (2024a) identify four key stages where algorithmic bias can emerge—data generation, model training, model testing, and deployment—each of which requires targeted strategies to mitigate potential biases effectively. Moreover, regarding impression management, applicants in selection contexts may try to present themselves as ideal candidates, often through techniques like authenticity cues (Heimann et al., 2024) or assertive and self-promotional behaviors (Ellis et al., 2002). In NLP and ML-based assessments, participants theoretically should be unable to fake responses since they are typically unaware of the specific features used in modeling. However, impression management may still be feasible during an AI-mediated interview, potentially through adjusting sentiments or language to appear

more positive (Fan et al., 2023). This raises an important question for future research: how might AI tools transform the way we conceptualize and measure impression management and response distortion in personality-based employment interviews? Insights into these behaviors could transform how we design personality assessments and address faking in AI-based interview settings, ensuring their validity and reliability in both employment contexts and broader personality assessment applications.

For Topic 3, *Legal Approaches to Fairness*, results suggest negative relationships with personality interviews and automated interviews. This is interesting as machine learning bias is an emerging area of research as AI tools become increasingly integrated into organizational contexts (e.g., Booth et al., 2021; Hickman et al., 2024a; Hickman et al., 2024b; Tay et al., 2022). However, Figure 2 visually demonstrates this topic is closely related to Topic 4, thus discussions on fairness and ML are closely related. Moreover, ongoing concerns about demographic-based discrimination in employment interviews underscore the need for fairness frameworks that address how the purpose and structure of interviews may influence assessment outcomes, particularly regarding equity across demographic groups. Although personality assessments are often used as part of a selection battery to mitigate subgroup differences (Ployhart & Holtz, 2008), personality-based employment interviews may still evoke issues of stress and anxiety, potentially impacting the measurement of personality traits. Furthermore, the integration of AI raises interesting questions about stress and interview performance. Specifically, could AI-based interviews, conducted through chatbots without audio or video, moderate the relationship between anxiety and interview ratings? For example, candidates may experience reduced anxiety due to the absence of direct social evaluation. Future research may examine how stress, health-related factors, and equity considerations intersect in personality-

based interviews, particularly considering emerging technologies, to ensure fairness across diverse populations.

For Topic 4, *AI and ML applications*, our findings reflect the growing trend of integrating technology into the employment interview process. This emphasizes the importance of incorporating insights from the field of data sciences to effectively leverage AI and ML in organizational contexts. As discussed in previous topics, the integration of AI requires a reframing of how we measure, perceive, and evaluate established theories and relationships between constructs. However, our analysis indicated this topic was not significantly prevalent in personality interview literature, revealing a notable gap in the application of these systems. Given the utility of personality interviews in selection contexts (e.g., Heimann et al., 2021; Van Iddekinge et al., 2005), further exploration of AI-driven personality assessments could provide practical advantages, offering more scalable and objective measures of personality traits. For example, recent research has demonstrated the validity of AI chatbots in inferring personality traits through semi-structured, narrative-based interviews (Fan et al., 2023; Sun et al., 2024). Adapting traditional personality assessment theories for AI-based tools, however, requires rethinking how trait-relevant cues are captured. Hickman et al.'s (2022) operational model for automated video interviews, based on Funder's (1995) realistic accuracy model, illustrates this shift. This model posits that personality traits are indirectly inferred through trait-relevant behaviors, and the accuracy of these measurements relies on sufficient detection and availability. In AI assessments, such as video or chatbot-based interviews, the ways in which trait-relevant cues are perceived vary significantly, from verbal behaviors in videos to textual cues in chatbot interviews. This underscores the importance of adapting traditional personality assessment frameworks to align with AI-based methods to detect and assess personality traits.

Lastly, for Topic 5, *Broad Applications of Interview Assessment*, our findings indicate that this topic was similarly prevalent across the literature on personality, automated, and employment interviews. Within this topic, the breadth of articles highlight the versatility of interviews as methods to evaluate job candidates, to assess health, and as experimental tools to induce and study stress. The broad scope of this topic suggests that interviews serve both practical and theoretical functions, ranging from selection assessments and clinical evaluations to behavioral research.

Conceptual and Practical Contributions

This study makes significant conceptual and practical contributions. First, this is the first study to systematically synthesize existing literature on personality interviews, employment interviews, and automated interviews. The findings highlight key areas of overlap across these broad domains, as well as underexplored and emerging research avenues. This integration allows for better understanding of the scope of these interview processes and identify where AI-based tools could benefit traditional approaches and methodologies, while providing a more comprehensive and interdisciplinary understanding of automated personality-based employment interviews. Second, this study identifies important gaps in the literature, particularly in the application of AI and machine learning for personality interviews. Although AI-based methods are becoming increasingly popular, the literature on adapting established personality assessment frameworks to integrate AI methods may still be emerging and overall underrepresented. Third, our findings emphasize important considerations for fairness, bias, and impression management across interview contexts. While AI and ML offer systematic and scalable alternatives to human-based interviews, they introduce unique sources of bias that require careful consideration, as

these biases differ fundamentally from those associated with human-based interviews in both conceptualization and assessment. Additionally, this study offers insights into how impression management behaviors, which have long been studied in traditional selection methods, may manifest within an AI-based interview, raising questions about job candidate behavior and implications in personality and automated interviews. Finally, methodologically, this study demonstrates how Xiong et al.’s (2023) SLR protocol can facilitate research across interdisciplinary fields by integrating findings from diverse sources. By employing semi-automated tools (e.g., web scraping) to search for and collect relevant articles, this approach enables a more comprehensive and unbiased overview of the literature. This methodology is particularly valuable for future research aiming to advance theory in rapidly evolving fields that intersect across psychology, organizational science, and data science.

Limitations

Despite the contributions, this study is limited in several ways. First, as the scope was limited to three databases, this may be neglecting works published outside of these databases and thus limiting our ability to fully capture themes. Future research could incorporate additional databases more relevant to each field, such as IEEE for data science or Business Source Complete through EBSCOhost for organizational research. Second, due to employment-related search terms expanding to “job” and “selection” terms, our samples of articles from the organizational sciences are quite skewed. Future research could incorporate additional psychological and clinical psychology terms and databases to gather more literature from this domain. Relatedly, I may have under-sampled the automated interview and personality interview categories as I did not use additional search terms to search for articles. While the search term “video+interview” was tested, it was too broad of a search. However, future research could

include alternative sets of keywords, such as “automatic interview”, “audio-video interview”, and “psychological interview”. Third, this study focused on identifying latent themes across the three domains; however, this approach primarily offers an overall description of the literature and is limited in its ability to uncover causal relationships or provide an in-depth exploration of the articles prevalent within each topic. Future research may benefit from applying the full range of Xiong et al.'s (2023) SLR protocol, including the final stage of citation network analysis. This stage assesses key articles and knowledge flows, offering valuable insights into trends, knowledge diffusion, and the evolution of theories within the field over time. Incorporating this stage could enrich understanding of how foundational ideas and newer concepts shape ongoing research trajectories. Additionally, AI-based tools have been developed to aid the SLR process, mainly to help screen and extract information and it is likely LLMs will be integrated in this process in the future (Bolaños et al., 2024). Though, LLMs have their own limitations since they are trained on general data and may not perform well in specialized fields, they may “hallucinate” and generate inaccurate or fabricate information, and the decision-making process is more complex and output can be inconsistent (Ji et al., 2023). As newer reasoning models come out (e.g., OpenAI o1 and o3-mini), this may help with consistency and understanding of the decision-making process, however it should not completely replace human judgement and decision-making.

Conclusion

This study conducted a systematic literature review on automated personality-based employment interviews by synthesizing research from psychological, organizational, and data sciences. Using web scraping tools, 734 articles were gathered from three bibliographic databases: ScienceDirect, Springer, and PsycNet. Through structural topic modeling, five latent

topics were identified and I examined how their prevalence differed across interview domains, highlighting both overlaps and critical gaps in the literature. This interdisciplinary approach demonstrates that AI integration into personality and employment interviews requires a shift in how we approach established theoretical frameworks and methodological practices. By leveraging semi-automated data analytics tools, this approach contributes to more efficient synthesis of interdisciplinary research, advancing our understanding of automated personality-based employment interviews. This study provides essential insights and lays a foundation for future research on integrating AI into personality assessments for organizational selection.

Chapter 3 - Part II: Development and Validation an AI Chatbot and Scoring Algorithm

In Part II of my dissertation, I aim to address 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, I adapt these methods to an automated format using an AI chatbot. I use methodology and archival data from Heimann et al. (2021) to guide my methodology² and serve as a baseline for the chatbot development and scoring procedures and collect additional data using an AI chatbot to administer the interview. The scoring algorithm uses word embeddings extracted using NLP transformers and zero-shot prompt engineering using a large language model (LLM). I 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 behavior description interviews,

² I thank the authors Anna L. Heimann and Pia V. Ingold for their collaboration in translating the interview questions and rating scales, for sharing their data and materials, and for their valuable review of this part of my dissertation.

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 behavior 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 (Gonzalez-Mulé et al., 2014; Poropat, 2009; Salgado & Moscoso, 2002), academic performance, and workplace behavior, making interviews a valuable method for assessing personality (Heimann et al., 2021; Van Iddekinge et al., 2005). 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 behavior description interview, interviewees are prompted to describe their actions in specific work-related scenarios (Janz, 1982). If candidates are asked to discuss their behavior in interactions with colleagues and teamwork, the situation should 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, behavior 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 (Heimann et al., 2021; Van Iddekinge et al., 2005).

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. The Realistic Accuracy Model (RAM) is grounded in psychological research and focuses on enhancing the diagnostic accuracy of personality assessments by outlining the processes through which observers form accurate judgments of others' personalities (Funder, 1995). According to RAM, accurate personality judgement depends on four aspects: *relevance* of behavioral cues, the extent to which these cues are *available* to observers, the extent to which these cues are *detected*, and the way in which these cues are *utilized*. Since personality traits 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 behavior 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, I 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, I 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 I 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 behavior description interviews accounted for significant incremental variance beyond verbal cognitive ability and self-reported personality in predicting types of OCBs (e.g., compliance, helping, and initiative). Beyond task and contextual performance, personality derived from

traditional self-report inventories have been significantly related to a wide range of work-related outcomes, including counterproductive work behaviors (CWBs; Berry et al., 2007; Jensen & Patel, 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, I 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.

For CWB, emotional stability, agreeableness, and conscientiousness have been consistently linked with lower levels of workplace deviance (Berry et al., 2007). Moreover, research has shown that individuals high in combinations of conscientiousness, emotional stability, and agreeableness exhibit the lowest levels of CWB (Jensen & Patel, 2011). Drawing from the theory of reasoned action (Fishbein & Ajzen, 1975), Jensen & Patel (2011) posit that CWB is more likely to occur when the individual believes detection is low and have low motivation to follow workplace norms, and individuals low in conscientiousness, agreeableness, or emotional stability are more likely to hold these beliefs. Moreover, the relationship between job stress and CWB is stronger for those low in agreeableness or conscientiousness (Zhou et al., 2014).

For job satisfaction, personality plays a critical role as individuals tend to seek out environments and occupations that align with their personality traits (i.e., the corresponsive principle, Roberts et al., 2003). Generally, emotional stability, extraversion, conscientiousness, and agreeableness have been modestly linked to greater job satisfaction (Judge et al., 2002). Moreover, job satisfaction tends to be higher when an individual's personality aligns with the

dominant traits of their occupational group (Törnroos et al., 2019). Thus, personality is a key factor in understanding job satisfaction.

For SWB, defined as the emotional and cognitive evaluation of one's life (Diener et al., 2003), personality plays an important role in shaping behavior and understanding how individuals experience life events. Neuroticism, specifically, has been a consistent negative correlated with SWB (Gutiérrez et al., 2005; Steel et al., 2008), as the tendency to experience negative emotions such as anxiety and stress can negatively impact perceptions of life satisfaction (Abdullahi et al., 2020). In addition to neuroticism, extraversion and conscientiousness have been strong positive predictors, followed by openness and agreeableness (Anglim et al., 2020). Given these associations, personality traits are expected to significantly correlate with SWB.

Similarly, personality has been significantly linked to stress, particularly the subjective appraisal of stress rather than actual physiological stress (Luo et al., 2023). Bolger & Zuckerman (1995) proposed a framework to understand how personality, particularly neuroticism, influences both exposure to and reactivity to stressful events, such that individuals high in neuroticism are more likely to generate or create stressful situations through maladaptive behaviors and coping strategies (Bolger & Schilling, 1991). While exposure to conflicts is important, their framework emphasizes that the way individuals react to and perceive conflict plays a critical role in determining stress outcomes. Thus, personality traits are expected to significantly relate perceptions of stress.

Methods

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 ($M_{\text{age}} = 30.56$, $SD = 7.51$; 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, I 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. I used OpenAI's Speech-to-Text Whisper API (OpenAI, 2022) with the large Whisper model to perform both transcription and translation³. 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

The 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 ($n = 8$).

The 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

³ I thank Pengda Wang for his collaboration in developing the code for transcribing and translating audio files, his contributions to designing zero-shot prompts, and his invaluable feedback on this section of my dissertation.

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>) and took participants approximately 35 minutes to complete ($M = 35$ minutes, $SD = 16$ minutes for Sample 2; $M = 34$ minutes, $SD = 14$ minutes for Sample 3). Appendix B includes a screenshot of an example conversation with the AI chatbot. 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), I 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. I 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 $\alpha = .86$. Example agreeableness item: “*I am someone who is respectful, treats others with respect.*” Internal consistency for agreeableness was Cronbach’s $\alpha = .76$. Example conscientiousness item: “*I am someone who is dependable, steady*” Internal consistency for conscientiousness was Cronbach’s $\alpha = .81$. Example emotional stability item: “*I am someone who is relaxed, handles stress well.*” Internal consistency for emotional stability was Cronbach’s $\alpha = .91$. Example openness item: “*I am someone who is original, comes up with new ideas*” Internal consistency for openness was Cronbach’s $\alpha = .84$.

For the working adult sample (Sample 3), I used a contextualized version of the BFI-2 (Soto & John, 2017). To contextualize this measure, I incorporated a combination of instructional contextualization, tagged contextualization, and complete contextualization.

Contextualization involves modifying measures to fit the relevant context, such organizational, academic, personal, and romantic contexts, and has been found to improve context-relevant predictions (Swift & Peterson, 2019; Shaffer & Postlethwaite, 2012).

Instructional contextualization involves modifying the instructions to fit the context. An example of this approach was changing “*You will be presented with a number of characteristics that may or may not apply to you. Read each statement carefully and decide the extent to which you agree or disagree*” to “*You will be presented with a number of characteristics that may or may not apply to you in the workplace. Please think specifically about your behavior in professional or work settings as you read each statement carefully. Consider how each statement describes you in the context of your workplace, and decide the extent to which you agree or disagree.*” Tagged contextualization involves adding contextual tags to the end of items and has been found to increase reliability and validity of personality inventories in predicting academic outcomes (Lievens et al., 2008). Moreover, tagged contextualization contributes more to predictive validity than instructional contextualization alone (Schlotzhauer et al., 2024). An example of this approach was changing “*Is relaxed, handles stress well*” to “*Is relaxed, handles stress well at work.*” Complete contextualization involves redesigning a statement to fit the context and has been found to increase criterion validity more than tagged contextualization (Holtrop et al., 2014). An example of this approach was changing “*Has an assertive personality*” to “*Has an assertive personality when engaging in workplace discussions.*” A combination of these three approaches has been shown to improve the criterion-related validity of forced-choice personality measures (Li et al., 2024).

See Appendix D for complete measure with contextualized items and instructions. 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 $\alpha = .89$. Example agreeableness item: “*I am someone who is respectful, treats others professionally at work.*” Internal consistency for agreeableness was Cronbach’s $\alpha = .88$. Example conscientiousness item: “*I am someone who is dependable, steady in fulfilling work responsibilities.*” Internal consistency for conscientiousness was Cronbach’s $\alpha = .89$. Example emotional stability item: “*I am someone who is relaxed, handles stress well at work.*” Internal consistency for emotional stability was Cronbach’s $\alpha = .89$. Example openness item: “*I am someone who is original, comes up with new ideas for work projects.*” Internal consistency for openness was Cronbach’s $\alpha = .91$.

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 Appendix B. To ensure consistency across the approach used by the interviewers in Heimann et al. (2021), the chatbot was designed 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

I 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., 2024c). In this study, text from each interview question, including both the question and the participant's response, was used to provide additional context for the participant's response, ultimately allowing for a deeper understanding of the data. The text from each question was then trained on its respective human-rated domain score. Specifically, the text from question 1, which assessed emotional stability, was trained on the human-rated emotional stability domain-level score, while the text from question 2, targeting conscientiousness, was trained on the human-rated conscientiousness domain-level score, and so forth. This process resulted in 15 models (three per trait). The final domain-level embedding-based ratings were then calculated by averaging the predictive values of the respective models. Consistent with earlier studies (e.g., Hickman et al., 2024c; Speer et al., 2021), the embeddings were trained using *k*-fold cross-

validation, a method that splits the data into sections (folds). Specifically, during model training I used 10-fold, which means in each iteration, 90% of the data is used for training, and the remaining 10% is used for testing and L2 regularization (i.e., Ridge regression)⁴.

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, I 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, I 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 the pilot, I 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 Appendix B. By structuring the prompt in this way, the LLM was guided to generate both a

⁴ I explored four versions of model training that manipulate the number of outer folds for cross validation, regularization, and dimension reduction techniques. Comparison and further discussion on model selection is included in Appendix B.

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, I 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 $\alpha = .83$.

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 $\alpha = .86$.

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 $\alpha = .84$.

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 $\alpha = .81$.

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 $\alpha = .88$.

Table 4 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 5 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; $r = .36, p < .001$). These correlations are comparable to, though weaker than, the archival data (Sample 1).

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, I screened the responses in all three samples. I 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, I examined the word count for each interview question in each sample. Table 6 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. Additionally, the chatbot interviews took a comparable amount of time, averaging 35 minutes ($SD = 15$ minutes) for Samples 2 and 3 combined, compared to 30 minutes in Sample 1. This discrepancy in response length and similar interview duration reflects differences in ease of verbal communication versus written communication, and also suggests that, to achieve quality responses comparable to human interviews, researchers should account for the additional time participants may need to provide detailed responses.

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, I 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 7 and 8 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 7) and Sample 3 (see Table 8).

Convergent validity was assessed using the convergence index (C_1), which represents the coverage of correlations for the same traits assessed across different methods (i.e., monotrait-heteromethod [MTHM] correlations). A large C_1 indicates that trait scores converge well across methods. C_1 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-monomeethod [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 much higher than previous studies (see Fan et al., 2023; Hickman et al., 2022) and 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, I conducted 30 hierarchical regression analyses (five personality domains × six relevant outcome variables) for Samples 2 and 3 separately⁵. 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

⁵ This approach does significantly increase the likelihood of a Type II error and typically all five personality trait scores would be entered simultaneously into the regression model at each step rather than separately. However, this method was chosen based on (1) the small sample size of the student (Sample 2) and working adult (Sample 3) samples, and (2) the poor discriminant validity of machine-derived scores. This approach was also adopted by Fan et al. (2023) to help isolate the effects of machine-derived scores on each criterion.

zero-shot scores) were added separately in Step 2. Tables 9 through 14 report the results of the regressions.

For OCB (see Table 9), 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 10), 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 11), 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 12), 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 ($\Delta R^2 = .05, p = .006$).

.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 13), 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 14), 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, this 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, my 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. In the human interviews, these were treated as developmental mock interviews, and they received feedback on their performance to help improve future employment interviews. In the chatbot interviews, they did not receive this development feedback, thus engagement and motivation were likely lower. 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. In the future, I will aim to recruit active job seekers and add development feedback to motivate participants to engage more deeply during the chatbot interview. Second, this study was designed with a highly structured chatbot interview format to be consistent with procedures from the archival data. The 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, which inflates the correlations between personality traits and outcomes, especially for measures of job performance. Though both of these were honest reporting conditions and there was no motivation to inflate responses, future research is encouraged to gather other, preferably supervisor reports.

Practical Implications

The findings of this study have significant implications for both researchers and practitioners seeking to develop automated, scalable personality assessment tools. The results demonstrate the feasibility of using AI-based personality assessments to reduce administrative burdens during the hiring process while maintaining predictive validity, particularly by minimizing the need for interviewer training and administration time. 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.

Chapter 4 - Part III: Practical Considerations for Selection Contexts

In selection contexts, the validity and fairness of assessment methods are critical for ensuring both legal defensibility and equitable hiring outcomes. As AI-driven personality assessments become more prevalent in employment decisions, it is essential to evaluate their practical implications. This section of my dissertation examines three key considerations through a utility lens: (1) user perceptions of AI-based assessments, (2) potential subgroup differences that may introduce bias, and (3) the risk of response distortion in selection contexts. By addressing these factors, I aim to provide a comprehensive understanding of the practical challenges and implications of integrating AI-driven personality interviews into personnel selection.

User Perceptions

As the use of AI tools in organizational contexts continues to grow (Tippens et al., 2021), understanding user perceptions and engagement becomes increasingly crucial. During the selection process, organizations should consider candidate reactions and perceptions of selection tools, as candidate perceptions of selection methods can impact their attitudes and behaviors throughout the selection process, such as withdrawing from the candidate pool, accepting the job offer, or pursuing legal action (Nikolaou & Georgiou, 2018; Gilliland & Steiner, 2012; Truxillo et al., 2004). Structured interviews are rated positively by candidates (Hausknecht et al., 2004), and the personableness, competence, and informativeness of an interviewer during the selection process can impact attitudes and perceptions of the job candidates (Nikolaou, 2011). Thus, by automating this process and removing the human component, it is especially important to examine how users and job candidates interact with and perceive the chatbot interview.

ML and NLP techniques offer several benefits in selection contexts by scoring previously underused textual data, improving criterion-related validity, and reducing subgroup differences (Campion & Campion, 2023), yet applicants generally hold negative views towards ML algorithms in selection contexts and perceive them as unfair, ineffective, and lacking in transparency (König & Langer, 2022; Zhang & Yencha, 2022). Recent research on highly automated interviews showed that participants perceive a lack of control in the process, likely due to a lack of social bandwidth, and reported them as being more unfair compared to videoconferencing interviews (Langer et al., 2019). The authors note this perception may be due to a lack of familiarity and transparency of the highly automated interview, organizations are encouraged to provide information about novel selection procedures and tools to mitigate negative candidate perceptions. This disconnect is quite interesting given human decision-making is more of a “black-box” than machine learning algorithms. Referring back to Topic 2 of bias and faking in Part I of this dissertation, this theme was almost exclusively dedicated to how interviewers can be biased by impression management, nonverbal cues, and information unrelated to suitability for the position. With traditional selection methods, research found that providing applicants with information on methods improve reactions (Truxillo et al., 2009). However, with ML methods being more transparent and can lead to decreased organizational attractiveness (Langer et al., 2018). But providing rationale on why certain information is used can potentially improve applicant reactions (Langer et al., 2020).

Differently, extant research found that users interacting with a chatbot reported higher levels of engagement, provided more authentic and higher quality responses, and demonstrated lower levels of survey fatigue compared to traditional scale-based methods (Zhou et al., 2019). Xiao et al. (2020) further enhanced user experience by integrating active listening practices into

chatbots, such as repeating each participant's answers and acknowledging their statements. This approach resulted in participants reporting an increased perception of connection with the chatbot. Additionally, users have been found to perceive chatbot interview methods similarly to traditional personality surveys in terms of how well they felt scores fit them, how satisfied they were with the assessment, and how useful they found it (Yuan et al., 2024). Moreover, Chu et al. (2024) found that from a career guidance chatbot participants found the experience rewarding, however participants rated it as more difficult to use and stay focused during the conversation compared to traditional personality scales. Thus, an AI chatbot may be an acceptable middle-ground between human-based interviews and highly automated interviews, however they may be more taxing than traditional Likert-style scales.

RQ1: How do users perceive an AI chatbot in terms of engagement compared to traditional scale-based methods?

RQ2: How do users perceive the use of an AI chatbot in selection contexts?

Subgroup Differences

In the US, the Four-Fifths rule is used to measure adverse impact in selection procedures, meaning that the proportion of selected candidates from one subgroup (which can be racial or gender-based) is less than 80% of the proportion of selected candidates from another group. Thus, on the prediction end, organizations must ensure machine-inferred scores do not demonstrate subgroup differences. To mitigate adverse impact in selection, one common method that has been used is to use predictors with smaller subgroup differences, such as personality assessments (Ployhart & Holtz, 2008). Additionally, one notable advantage of structured interviews is that research has demonstrated lower adverse impact compared to mental ability tests (Huffcutt & Roth, 1998; Ployhart & Holtz, 2008), and when the interview is carefully

designed, they can provide a high level of criterion-related validity (McDaniel et al., 1994). Moreover, recent research has demonstrated that the use of NLP to score candidate essay responses in conjunction with cognitive ability assessments showed better predictive validity while also reducing adverse impact (Campion et al., 2016; Campion et al., 2024).

Given these findings, an AI chatbot-driven, personality-based employment interview should, in theory, minimize subgroup differences in selection outcomes. However, despite these potential advantages, there is growing concern about algorithmic bias in machine learning-based selection systems (Landers & Behrend, 2023). Biases can emerge if ML models are trained on non-representative datasets that reflect historical inequities. Even when explicit demographic data is removed, ML algorithms can still learn and reinforce biased patterns present in the training data (Tay et al., 2020). For example, if past hiring decisions systematically disadvantage minority applicants, an AI model trained on those decisions may perpetuate similar biases (König & Langer, 2022). Thus, it's important to examine prediction bias from these NLP-based models and prompt engineering from textual data gathered from an AI chatbot.

RQ3: Do NLP-derived personality scores differ between participants based on gender, age, or race? How does this compare to traditional survey-based methods?

Faking

In selection contexts, applicants may distort their responses to appear as an “ideal” candidate, rather than responding based on their actual self-perceptions. This is often referred to as *faking*. Levashina & Campion (2007) have defined faking in employment interviews as the intentional distortion of responses to obtain a better score and/or create favorable perceptions. Differently, in personality inventories, faking has been conceptualized slightly differently, since faking involves endorsing an item differently than one’s actual self-perception (e.g., endorsing

an extraversion-related item higher, when the individual perceives themselves as more introverted), whereas in interviews faking includes a continuum of different tactics (Melchers et al., 2020). In their development of an interview faking behavior scale, Levashina & Campion (2007) identified four factors of faking: slight image creation (exaggerating while staying close to the truth), extensive image creation (fabricating information), image protection (omitting unfavorable job-related details), and ingratiation (trying to be likable regardless of actual performance).

From research on faking in personality inventories, job applicants tend to exaggerate personality trait scores in a positive direction (particularly higher conscientiousness and emotional stability scores; Hu & Connolly, 2021) and adapt them based on the perceived culture or type of job they are applying for (Birkeland et al., 2006; Roulin & Krings, 2020). Rosse et al. (1998) compared job applicants personality assessments scores to job incumbents and found significant inflation for extraversion, openness, agreeableness, and conscientiousness scores. Additionally, meta-analytic findings suggest that applicants report higher scores on extraversion, emotional stability, conscientiousness, and openness, and that they tailor these responses to align with the specific job (Birkeland et al., 2006). However, these behaviors may come with negative outcomes for the individual, such that faking during the application stage has been associated with poor job fit, which in turn is positively related to job stress and negatively related to affective well-being and employee engagement (Charbonneau et al., 2021).

For organizations, faking is an ongoing concern since distorted responses could adversely impact selection decisions, but the degree to which faking is present and detrimental to validity is still debated (Tett & Simonet, 2021). On one end, studies suggest the degree of response distortion is limited and not detrimental to the selection process (Ellingson et al., 2007) and that

faking can be systematically accounted for by other measures (e.g., social desirability; Hogan et al., 2007). On the other end, studies have found that applicant faking is both prevalent and negatively impacts the psychometric properties of the selection measure and quality of potential hiring decisions by the organizations (Donovan et al., 2013). In a recent meta-analysis on faking on personality assessments, Speer et al. (2025) conceptualize these opposing views as psychometric theory (Allen & Yen, 2001) and unified views of validity (Binning & Barrett, 1989) versus socioanalytic theory (Hogan & Bickle, 2013). Where the former posit that construct irrelevant variance will negatively impact validity, and the latter posit that faking is a sign of effective social skills. Speer et al. (2025) found that faking does negatively impact criterion-related validity of personality assessments and emphasized the importance of using assessments that are more faking resistant.

In NLP and ML, participants should be theoretically unable to fake since they are not aware of the key features used in modeling; however, impression management may still be possible during the interview with the chatbot by targeting more positive sentiments and language (Fan et al., 2023). Along these lines, Völkel et al. (2020) demonstrated that, when instructed to create a false personality profile to an AI chatbot, participants were able to distort responses to a small degree (up to 10%), more effectively for agreeableness and extraversion compared to conscientiousness and neuroticism. Though participants in this study reported numerous distortion tactics spanning syntax (e.g., grammar and spelling), word choice (e.g., curse words and keywords related to personality traits), emotional tone (e.g., politeness and positivity), and content (e.g., length and level of detail in responses), their overall trait distortion was relatively small compared to their baseline personality scores (Völkel et al., 2020).

However, the extent to which an AI chatbot is resistant to faking in a personality-based employment interview has yet to be empirically examined.

RQ4: Is an AI chatbot-based employment interview more faking resistant compared to self-report methods in assessing personality traits?

Additionally, previous research has shown that induced faking (i.e., where participants are directed to fake) can also increase personality scores on self-report measures (Viswesvaran & Ones, 1999). However, lab studies on faking have been critiqued based on lack of external validity and limited generalizability to actual selection contexts (Melchers et al., 2020). Though meta-analyses suggest that participants who are instructed to fake distort their scores more than those in actual selection context (Viswesvaran & Ones, 1999; Birkeland et al., 2006), indicating that scores from induced faking represents maximal response distortion and scores from actual selection contexts represents typical response distortion. In their recent meta-analysis, Speer et al. (2025) examined the faking differences between simulated and actual selection contexts and found faking was similarly problematic for criterion-related validity in both induced and motivated contexts. As these findings were based on faking on personality inventories, research has yet to examine these differences for NLP-based personality scores compared to scale-based methods.

RQ5: Are there differences in NLP-derived personality scores between participants that are motivated to fake (i.e., in a high-stakes context) and participants that are instructed to fake?

Methods

Participants

Data for this study was collected from a working adult sample recruited through Connect. To be eligible, participants had to be U.S.-based, work at least part-time (minimum of 30 hours

per week), and have at least one year of work experience. Initially, 146 participants completed the survey. However, six submissions were removed due to missing chatbot interview data ($n = 1$) or low-quality responses during the chatbot interview ($n = 5$). Participants who passed at least 8 out of the 9 quality checks were retained. Two remaining responses only passed 7 quality checks. Upon further review, their missed attention checks occurred during the outcome variable sections, likely due to the extended chatbot interview duration (ranging from 1 to 1.5 hours). Given that participants were required to submit their surveys within 2 hours to receive credit, these responses were flagged. However, since both participants provided effortful responses throughout the chatbot interview, only the data from the outcome variable sections were removed. For the motivated faking condition, I included a manipulation check at the end of the study and before debriefing asking if they believe the manipulation or not, six participants in the motivated faking condition were removed as they did not believe the manipulation, meaning 87% of participants reported believing in the manipulation. The final sample included 135 participants, 48 in the honest condition, 47 in the induced faking condition, and 40 in the motivated faking condition.

For demographics, 54% identified as female ($n = 73$), the average age was 39 years (SD = 11 years), and the majority of participants were White 71% ($n = 96$). Most participants held at least a bachelor's degree (70%; $n = 94$) and came from a range of industries, with the most common sectors being IT (18%, $n = 24$); Education (10%; $n = 14$), and healthcare (10%; $n = 14$).

Procedure

This survey included 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”). The first section included surveys on objective criteria, the

Big Five Inventory-2 (BFI-2; Soto & John, 2017), and user perceptions of the BFI-2. This provided a baseline for personality scores before experimental condition assignment. After the first section, participants were randomly assigned to one of three conditions: Honest, Motivated Faking, and Induced Faking. All conditions were presented with a scripted prompt that indicates the purpose of this study is to validate a new method to assess personality for job selection. The honest condition was given the chatbot interview and asked to respond honestly for research purposes. The induced faking condition was told to respond as if they were an actual candidate applying for the job. The motivated faking condition was told high-scoring participants would be offered additional opportunities for validating assessment tools in the future. See Appendix C for scripts for each condition. After condition assignment, participants completed the chatbot interview (developed from Part II), followed by user perception questions, and a separate self-reported personality measure (IPIP-50; Goldberg, 1992). Both the induced faking and motivated faking conditions received condition-specific reminders between the user perception questions and the self-reported personality measure, such that induced faking participants were instructed to continue responding as if they were an applicant and the motivated faking participants were reminded that their responses would be considered as part of their application. In the third section, participants completed questionnaires on outcome variable measures (e.g., job performance, SWB, OCB, etc.). At the end of the study, participants were debriefed on the study (with explicit clarification regarding the fake job condition).

Personality Measures

Baseline Self-Reported Personality

To collect a baseline for participants personality, I used the contextualized version of the BFI-2 (Soto & John, 2017) that was administered in Part II (i.e., a combination of instructional,

tagged, and complete textualization). Items were ranged on an agreement scale ranging from 1 (strongly disagree) to 5 (strongly agree). Internal consistency was high across all dimensions, with Cronbach's $\alpha = .90$ for Extraversion, .84 for Agreeableness, .91 for Conscientiousness, .89 for Neuroticism, and .90 for Openness.

Post-Condition Assignment Personality

For personality after condition assignment, I used a contextualized version of the 50-item International Personality Item Pool (IPIP-50; Goldberg, 1992), similar to the archival data from Part II and to the BFI-2 I contextualized for work using a combination of instructional, tagged, and complete contextualization (see Appendix D for complete measure). Items were ranged on an accuracy scale ranging from 1 (very inaccurate) to 5 (very accurate). Example extraversion item: "*At work, I feel comfortable around my colleagues.*" Internal consistency for extraversion was Cronbach's $\alpha = .89$. Example agreeableness item: "*At work, sympathize with colleagues' feelings.*" Internal consistency for agreeableness was Cronbach's $\alpha = .88$. Example conscientiousness item: "*At work, I like order.*" Internal consistency for conscientiousness was Cronbach's $\alpha = .82$. Example emotional stability item: "*At work, I get irritated easily.*" Internal consistency for emotional stability was Cronbach's $\alpha = .93$. Example openness/intellect item: "*At work, I have excellent ideas.*" Internal consistency for openness was Cronbach's $\alpha = .82$.

AI Chatbot Interview

Following Part II, there were two concerns with the chatbot interview that I addressed before conducting this study: low word count and some questions were perceived as less relatable. First, I added additional instructions before the chatbot interview encouraging participants to be detailed in their response and for each question to provide background on the situation, how they experienced it (e.g., how they felt and what happened), and how they

behaved. Second, I increased the minimum word count for each response from five to 30 words based on review of the chatbot transcripts from Part II and internal piloting. I reviewed the descriptive statistics of word count in cases where all three raters and the zero-shot prompt assigned the same rating of 1, 3, or 5 for each question (see Table C.1). This helped inform a minimum word count threshold based on responses with high rater agreement. On average, the minimum word count across ratings of 1, 3, and 5 was 27 words. Thus, I imposed a 30-word minimum for each question. After this, I piloted the interview with the new word count minimum and instructions to ensure the interview would not be too cumbersome for participants. When describing a situation, their experience, and subsequent behavior, meeting the 30-word threshold was not challenging.

Third, not every participant will have experienced every situation exactly, however some situations were more difficult for participants to recall a similar experience than others. For example, one question asked about participating in a workshop, but many people have not experienced that exact situation. This was re-worded to be more general and ask about a time when they attended a large event with people they did not know. This way the question still targets extraversion and can be used to rate responses with the original BARS, while being more relatable to workers that may not have had these opportunities.

Machine-Derived Personality Scores

Similar to the study in Part II, I used an embedding-based approach and prompt engineering to score textual data from the chatbot interview. For embedding-based scores, I first extracted word embeddings using the DistilBERT transformer (Sanh et al., 2019) for each question and then applied the trained models from Part II to these embeddings. The predicted

scores for each question were then treated as “items” and averaged to calculate the domain scores.

For prompt engineering, I implemented few-shot learning (FSL) where the model is given a few labeled examples, which facilitates human learning (Lake et al., 2015; Wang et al., 2020). Unlike the embedding-based models trained in Part 2, using LLMs do not preserve the context or parameters of the training data. Thus, by giving the model labeled examples from Part II this facilitates better alignment with human-rated examples and increases consistency in applying BARS. In theory, this process acts similarly to fine-tuning without updating the model’s weights. Instead, the model uses in-context learning to adapt its predictions based on the provided examples. To select examples, I identified cases where all three raters and the zero-shot prompt had the same score of either 1, 3, or 5 for each question. The selected examples were chosen based on alignment with the BARS and available cues. In the majority of cases, there was at least one option to choose from. However, in some cases there were no responses where there was perfect agreement across the three raters and LLM scores. In these cases, I prioritized human scores and looked for the closest LLM score. For example, Question 2 for a rating of 1 had no responses where all raters and the LLM scored it as a 1. Therefore, I sampled from cases where all human raters scored it as 1 and the LLM scored it as a 2. After labeled examples were identified, I added in justification as to why each example was given it’s corresponding rating. This helps the model better understand how to use trait-relevant cues and BARS. Appendix C includes an example of the few-shot prompt and labeled examples provided to the model.

User Engagement and Perceptions

User Engagement with the Personality Scale

For user engagement with the personality scale (i.e., the BFI-2), I used O'Brien et al.'s (2018) User Engagement Scale—Short Form (UES-SF). This instrument measures focused attention (feeling absorbed in the task), perceived usability (affect as a result of the interaction), aesthetic appeal (attractiveness or visual appeal of the tool), and reward factor (perceived value) with four items each. These items were tailored to pertain to the BFI-2 and rated on an agreement scale ranging from 1 (strongly disagree) to 5 (strongly agree). Example focused attention item: "*I was absorbed in this experience.*" Internal consistency for focused attention was Cronbach's $\alpha = .67$. Example perceived usability item: "*Using this questionnaire was taxing.*" Internal consistency for perceived usability was Cronbach's $\alpha = .83$. Example aesthetic appeal item: "*This questionnaire was aesthetically appealing.*" Internal consistency for aesthetic appeal was Cronbach's $\alpha = .89$. Example reward factor item: "*Using the questionnaire was worthwhile.*" Internal consistency for reward factor was Cronbach's $\alpha = .85$.

User Engagement with the AI Chatbot

At the end of the interview with the AI chatbot, participants were prompted to answer user perception questions on perceived difficulty, perceived validity, affect toward the measure, and concentration. Of the 135 participants, 116 elected to answer all the user engagement questions. Word count was measured by the total number of words provided during the chatbot interview, and interview duration was measured by the length of time that a participant spent on the chatbot interview.

Perceived difficulty was measured using three items designed to assess how difficult users perceived the interview. Items were rated on an extent scale ranging from 1 (*not at all*) to 5 (*to a great extent*). Example item: "*To what extent did you find it challenging to answer my questions during our conversation?*". Internal consistency was Cronbach's $\alpha = .67$.

Perceived validity was measured using three items designed to assess how useful users believed the interview with the chatbot could be in. Items were rated on an agreement scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Example item: “*Do you think the way you chatted with me could accurately reflect your personality?*”. Internal consistency was $\alpha = .76$.

Affect was measured using three items designed to assess how much users liked the interview with the chatbot. Items were rated on an extent scale ranging from 1 (*not at all*) to 5 (*to a great extent*). Example item: “*Did you enjoy interacting with me?*”. Internal consistency was $\alpha = .74$.

Concentration was measured using three items designed to assess how easily distracted users felt during the interview. One item was rated on a frequency scale ranging from 1 (*never*) to 5 (*always*). Example item: “*Did you find yourself losing interest or not paying attention while chatting with me?*”. Two items were rated on a five-point scale ranging from 1 (*not at all*) to 5 (definitely). Example item: “*Was it easy to stay focused on our conversation?*”. Internal consistency was $\alpha = .74$.

I also used the UES-SF (O’Brien et al., 2018) to assess user engagement with the chatbot interview and directly compare user engagement with the personality scale. These items were tailored to pertain to the chatbot interview and were rated on an agreement scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Example focused attention item: “*The time I spent using this chatbot just slipped away.*” Internal consistency for focused attention was Cronbach’s $\alpha = .81$. Example perceived usability item: “*I found this chatbot confusing to use.*” Internal consistency for perceived usability was Cronbach’s $\alpha = .83$. Example aesthetic appeal item: “*The chatbot was attractive.*” Internal consistency for aesthetic appeal was Cronbach’s $\alpha = .89$.

Example reward factor item: “*Using this chatbot was worthwhile.*” Internal consistency for reward factor was Cronbach’s $\alpha = .92$.

Fairness Perceptions of the AI Chatbot

For fairness perceptions I used two scales adapted from Lee et al. [under review] that measures methodological integrity and attitude towards that chatbot interview hiring decisions and two subscales from the Selection Procedural Justice Scale (SPJS; Bauer et al., 2001) that measure perceptions of justice, specifically change to perform and propriety of questions.

Methodological integrity was measured using five items assessing perceptions of fairness, bias, accuracy, ethics, and consistency. Attitude was measured using three items assessing perceptions of transparency, legal defensibility, and acceptability. Items were rated on extent scales ranging from 1 (extremely [negative descriptor]) to 5 (extremely [positive descriptor]). Example methodological integrity item: “*To what extent do you think it is fair to adopt the chatbot interview for making hiring decisions?*” Internal consistency for methodological integrity was Cronbach’s $\alpha = .91$. Example attitude item: “*To what extent do you think it is acceptable to adopt the chatbot interview for making hiring decisions?*” Internal consistency for attitude was Cronbach’s $\alpha = .82$.

Chance to perform was measured using four items assessing the extent to which participants felt they had the opportunity to demonstrate their skills and abilities. Propriety of questions was measured using three items assessing how appropriate they found the chatbot interview questions. Items were rated on agreement scales ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Example chance to perform item: “*I could really show my skills and abilities through this interview.*” Internal consistency for chance to perform was Cronbach’s $\alpha = .95$.

Example attitude item: “*The chatbot interview itself did not seem too personal or private.*”

Internal consistency for propriety of questions was Cronbach’s $\alpha = .59$.

Results

Table 15 reports the correlations between study variables for combined (all conditions merged) and the honest reporting condition. Table 16 reports correlations for induced and motivated faking conditions. Data were analyzed using R (version 4.4.2), embedding-based models were applied to out-of-sample data using the *text* package. To conduct few-shot learning I used the LLaMA 3.1 model via together.ai, as I did in Part II.

Response Quality in the Chatbot Interview

First, I examined the response quality to determine if changes in the chatbot instructions and minimum word count were effective. Notably, the word count was significantly greater in Part III compared to the chatbot conditions in Part II ($t(232.08) = -8.35, p <.001$), such that the average number of total words in the chatbot conversation in Part III had an average of 1095 words compared to Part II with an average of 709 words. Additionally, a qualitative review of responses indicated that participants provided more contextual information when answering each question.

I also reviewed the chatbot transcripts to identify instances where participants did not meet the 30-word minimum per response. When the responses were too short, the chatbot prompted participants with a follow-up message, such as: “*Could you please explain a bit more? A minimum of 30 words is required for this response. You need X more words.*” While some participants encountered this prompt, most did not, suggesting that the instructions set expectations for response length and elicited more information from participants. However,

survey fatigue appeared to be a factor toward the end of the interview, as some participants reported difficulties to recall 15 unique work-related events without being overly repetitive.

In terms of model performance, the embedding-based models did not show significant correlations with self-reported personality scores across the three conditions for baseline and post-condition personality scores. Though, the few-shot prompt-based scores demonstrated decent performance for assessing extraversion, as well as agreeableness when compared to post-condition personality score. This finding, while somewhat disappointing, is not overly surprising as a sample size of $N = 421$ is relatively small, thus the training model may likely be overfit and does not generalize well out-of-sample. In contrast, LLMs are trained on massive datasets, and had access to the BARS and labeled examples for each question. See Table 18 for descriptive statistics for personality trait scores and Table 19 for model performance correlations.

User Perceptions

To address RQ1, I conducted a series of paired t -tests comparing user engagement scores from the UES-SF between the AI chatbot and the BFI-2. See Table 20 for results of the paired t -tests and Figure 3 for visualization. Results indicated that participants found the AI chatbot easier to focus on compared to the BFI-2 ($t(134) = -2.36, p = .019$), but the chatbot was more frustrating to use than the BFI-2 ($t(134) = -8.03, p < .001$). There were no differences for aesthetic appeal or reward factor, meaning participants rated the chatbot and the Likert-style measure similarly in terms of visual design and the perceived value of effort spent.

To address RQ2, I examined the means, standard deviations, and distribution of chatbot user engagement and fairness perceptions (see Table 21). Participants in the faking conditions were instructed that their responses about the chatbot would not be included in any selection materials and were encouraged to respond honestly. To confirm that there were no differences

across conditions, I conducted a series of ANOVAs for each engagement and perception variable, with results indicating no significant differences across the honest, induced faking, and motivated faking groups. Regarding chatbot engagement, participants reported low perceived difficulty ($M = 2.31$, $SD = .88$) and generally positive ratings for validity ($M = 3.87$, $SD = .71$), affect ($M = 3.91$, $SD = .97$), and ease of concentration ($M = 4.25$, $SD = .78$). For fairness perceptions, participants on average felt that the chatbot's method integrity ($M = 3.49$, $SD = 1.00$) and its use in selection procedures ($M = 3.44$, $SD = 1.02$) were fair, though ratings varied considerably across individuals. From the SPJS, participants strongly agreed that the chatbot's interview questions were appropriate for selection contexts ($M = 4.29$, $SD = .56$). However, perceptions of whether the chatbot allowed a fair chance to perform and fully showcase skills were more mixed ($M = 3.62$, $SD = 1.01$), suggesting that while many participants felt the chatbot provided a reasonable opportunity, others were more uncertain. Overall, while perceptions of the chatbot as a tool were generally positive, its role in selection procedures elicited more varied responses, indicating that some participants may be more skeptical of its integration into hiring decisions.

Subgroup Differences

To address RQ3, I examined gender, age, and race-based differences in NLP-derived personality scores and compared them to self-reported personality scores. For gender (see Table 22), results indicate that in self-reported measures, neuroticism scores were higher for women compared to men in both the baseline (BFI-2) and post-condition (IPIP) assessments. However, this difference was not reflected in embedding-based or few-shot learning-derived scores. Conversely, agreeableness scores were higher for men compared to women in the baseline (BFI-2) self-reported measure, but this difference was not observed in the post-condition (IPIP) or

embedding-based scores. However, the gender difference in agreeableness was present in the few-shot learning-derived scores.

For age-based differences (see Table 22), I examined the relationships between age and personality traits using Spearman's rank-order since age was non-normally distributed. Results indicated that age was positively correlated with conscientiousness for the baseline (BFI) and post-condition (IPIP) measures and negatively correlated with neuroticism for both the baseline (BFI) and post-condition (IPIP) measures. These findings align with longitudinal personality development literature that as people age, they generally become more agreeable, conscientious and emotionally stable (Roberts & Nickel, 2021). These differences were not reflected in the NLP-derived scores.

For race-based differences (see Table 23), I examined differences in NLP-derived personality scores and compared them to self-reported personality scores between participants who identified as White versus those identifying as non-White or part-White⁶. Results indicate that the baseline scores (BFI-2) did not significantly differ between White and non-/part-White participants, however the post-condition (IPIP-50) scores for agreeableness tended to favor White participants. This difference was not reflected in embedding-based or few-shot learning-derived scores.

Faking Resistance

Lastly, to address RQ4 and RQ5, I examined the extent to which the AI chatbot-based NLP-derived scores are resistant to response distortion. For simplicity, and since the embedding-

⁶ Due to sample size restrictions, I was unable to conduct more subgroup-specific comparisons, since my second and third largest racial groups were Black or African American ($n = 19$) and Asian (which includes Chinese, Japanese, Korean, and Vietnamese; $n = 10$).

based scores were uncorrelated to self-reported personality scores, I only use the few-shot prompt-based scores for this. To assess faking resistance of the chatbot-based scores compared to the self-reported scores, I conducted a series of mixed-methods ANOVAs to compare the mean personality scores between the Honest, Motivated Faking, and Induced Faking conditions at the domain-level. The mixed-methods ANOVA results are presented in Table 24, and visualization of the personality trait scores by condition and method are presented in Figure 4.

Results indicated there was a significant interaction between condition and method for conscientiousness ($F(4, 264) = 2.50, p = .043$), such that in both induced and motivated faking conditions the chatbot-based scores were significantly greater than scores in the honest condition. This suggests that participants adjusted their responses to appear more conscientious when instructed or motivated to fake, compared to the post-condition measure (IPIP) which did not demonstrate significantly inflated scores. There were no significant interactions for the other traits. Additionally, there were consistently no differences in baseline (BFI) and post-condition (IPIP) scores for either faking condition, suggesting that they did not fake as expected. Notably, the baseline scores for agreeableness and conscientiousness were already high with mean values for all conditions above 4.

For RQ4, I did not find support for chatbot-based NLP-derived scores to be more faking resistant than traditional personality measures (i.e., IPIP). However, response distortion itself was not strongly confirmed. For RQ5, I found no significant differences between the induced (maximal) and motivated (typical) faking conditions.

Discussion

The growing integration of AI-based tools into personnel selection offers unique opportunities and challenges for researchers and practitioners. This part of my dissertation aimed

to examine a range of practical considerations when applying AI-based tools in organizational contexts. Specifically, I aimed to (1) examine how users engage with an AI chatbot-based employment interview and their perceptions on fairness in the hiring process, (2) whether the NLP-derived scores demonstrated subgroup differences, and (3) explore whether the AI chatbot interview was more or less faking resistant than traditional scale-based methods.

First, the results provide mixed support for the use of chatbot-based assessments in selection. While users generally had favorable feedback and perceptions in terms of engagement, there was less of a consensus around whether it would be appropriate for employment contexts. Aligning with similar chatbot-based studies (e.g., Chu et al., 2024) users found the chatbot more difficult to use than scale-based methods but still found the experience rewarding. This suggests that AI-based assessments may offer a more dynamic and engaging alternative to static, scale-based methods. However, in high-stakes selection contexts, it is critical to mitigate negative reactions and experiences for job applicants, as perceptions of fairness and procedural justice may influence candidate behavior, organizational attractiveness, and legal defensibility.

Second, for subgroup differences, NLP-derived personality trait scores showed minimal evidence of subgroup differences based on gender or age. While the embedding-based models generally demonstrated weak predictive validity, the few-shot prompt-based models were more predictive and did not reflect subgroup differences present in traditional self-report scales. Though, for age, the scale-based differences (i.e., older workers being more conscientious and emotionally stable) is not a concern of bias with self-report, but rather reflects the maturity principle of personality development (Roberts & Nickels, 2021).

Third, for faking resistance, while AI-based assessments should in theory be less “fakable” than scale-based assessments, this was not the case in this study. Findings indicated

participants did increase their chatbot-based conscientiousness scores in the motivated conditions. However, self-reported scores remained stable across conditions, raising questions of whether participants were actively engaging in faking behaviors as intended with the study design or if there was a ceiling effect for the faking conditions since their baseline scores were already above 4. Furthermore, the lack of significant differences between Induced (maximal) and Motivated (typical) faking groups suggests that greater motivation (in theory) to fake did not meaningfully inflate responses, at least within this study's design. Taken together, these findings suggest that while chatbot-based personality assessments offer promising new approaches to personnel selection, they may not necessarily be immune to response distortion and require careful implementation to ensure fairness and mitigate negative reactions from applicants.

Limitations and Future Directions

This study was limited in several ways that offer opportunities and challenges for future research. First, the manipulation and study design may not have been strong enough to elicit faking behaviors. As the fake job posting was originally designed for students, translating this to working adults through crowdsourcing platforms poses challenges. For instance, the incentive was based on being offered additional opportunities to complete surveys, but they can likely find dozens of other survey opportunities in minutes to earn extra compensation. In the future, this incentive should be tailored to ensure it is motivating enough to induce faking behaviors.

Second, I was only able to focus on gender and age-based differences since the sample I collected was not representative enough to examine differences based on race, industry, or other demographic variables. In the future, a more representative sample should be collected (e.g., using quota sampling from crowdsourcing sites) to conduct more thorough subgroup difference analyses. Third, there is a high potential that participants were using generative AI to complete

the chatbot interview, and recent research has demonstrated that LLMs can perform well in these interview-like situations (Canagasuriam & Lukacik, 2025). In the chatbot training data in Part II, participants who clearly used generative AI in their responses were removed and omitted from analyses. For example, when copying over the LLM response to an interview question, one participant left in the model's response of "Here's an example of how someone might experience and handle participating in a course or seminar where they meet a lot of new people, and how they connect with others during the event." However, this brings up an interesting point of whether to systematically search for and remove potential generative AI users in this sample. While traditional data cleaning methods would warrant excluding AI-generated responses, in a study examining applicant faking behavior where generative AI exists, it is likely that response distortion includes some form of AI-assisted faking. This also highlights the critical importance of developing tools and designing assessments in a way that deters or automatically detects AI-assisted faking (see Hickman, 2025).

Chapter 5 - General Discussion

As AI and NLP become integrated into organizational systems and processes, it is essential for researchers to conduct throughout investigations and processes when developing new tools. This dissertation aimed to make the following contributions: (1) conceptually, by advancing the understanding of how AI and NLP can be leveraged in selection assessments; (2) methodologically, by developing and validating an AI-based tool for personality assessment and scoring algorithms through transparent methods; and (3) practically, by providing organizations with a valid, systematic tool to assess personality that is cost-effective, scalable, and addresses concerns in the selection process.

In Part I, I presented a systematic literature review integrating research from organizational, psychological, and data sciences to identify latent themes and underrepresented areas of research. Using a data-analytics based SLR protocol, this study not only showcased how these tools can be leveraged for more effective processes, but also uncovered five themes within interview research: clinical assessments, bias and faking, legal approaches to fairness, AI and ML applications, and broad applications of interviews. This review highlighted how bias and faking are recurring and persistent challenges that are not solely confined to organizational research. Additionally, while AI and ML applications are gaining popularity through automated interviews, the integration with psychological sciences may still be lacking.

In Part II, I aimed to address this underrepresented area by developing and validating an AI chatbot to administer a personality-based employment interview. In this study, two NLP approaches were implemented to score textual data: word embeddings from the DistilBERT transformer and zero-shot prompting using Llama 3.1. The findings from this study demonstrated moderate convergence with self-reported personality scores, but lower discriminant validity,

which aligns with previous research. Additionally, these NLP-based scores demonstrated incremental validity in predicting OCBs, job satisfaction, and subjective well-being, suggesting that AI-based assessments can provide unique insights beyond self-report. However, challenges of modularizing components of the behavior description interview were present, with participants providing significantly shorter responses in the chatbot interviews than human-administered interviews, limiting the amount of trait-relevant cues that were available for raters. This also highlights a weakness of embedding-based models, and ML models in general, they are highly dependent on the data which they are trained on. Since cues were limited, this sets a challenging foundation for model training.

In Part III, I examined three critical considerations for integrating AI-based assessments in organizational contexts: user perceptions and engagement, subgroup differences, and resistance to faking behaviors. In this stage, I was also able to address design challenges apparent in Part II, by adding further instructions for participants and setting word count minimums to encourage participants to provide more trait-relevant features. I was also able to leverage quality, labeled examples from Part II to help inform better LLM-based scoring. The findings from this study had promising outcomes in that users could navigate through the AI chatbot interview with minimal issues and generally positive perceptions of the tool. I anticipate that applicant reactions will remain a central focus in future discussions and become an increasingly prevalent area in AI research. Additionally, as generative AI tools continue to shape various stages of the selection process, methods for detecting faking behaviors will need to evolve accordingly. Moving forward, these tools must continue to undergo rigorous evaluation not only for bias and psychometric validity but also for transparency.

As AI and ML methods become increasingly popular in organizational settings, it is important that selection assessments undergo rigorous development and validation. For researchers, these findings contribute to our conceptual understanding of how interviews can be integrated with advanced technology and applied in organizational contexts. For practitioners, the development of automated and engaging personality assessment tools has the potential to streamline selection processes, reducing administrative burdens while enhancing candidate experience. Beyond hiring, these tools could serve as training resources, preparing job candidates for structured interviews. However, as AI-based tools become more prevalent, issues of transparency, fairness, and user engagement remain paramount. While AI chatbots present an innovative solution in selection contexts, their design, scoring algorithm, and user experience must be continually refined to maintain validity, reliability, and fairness.

References

- Abdullahi, A. M., Orji, R., Rabiu, A. M., & Kawu, A. A. (2020). Personality and Subjective Well-Being: Towards Personalized Persuasive Interventions for Health and Well-Being. *Online Journal of Public Health Informatics*, 12(1), e1.
<https://doi.org/10.5210/ojphi.v12i1.10335>
- Adiani, D., Nguyen, E. T., Urban, J., Fadler, M., Alam, A., Garcia-Alamilla, J., Sarkar, N., & Sarkar, M. (2023, July). Design of an interview script authoring tool for a job interview training simulator using graph transformations. In H. Mori & Y. Asahi (Eds.), *Human interface and the management of information*. HCII 2023. Lecture notes in computer science (pp. 413-431). Springer, Cham. https://doi.org/10.1007/978-3-031-35129-7_30
- Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.
- Ammari, A., Dimitrova, V., & Despotakis, D. (2012). Identifying relevant YouTube comments to derive socially augmented user models: A semantically enriched machine learning approach. In *Advances in User Modeling: UMAP 2011 Workshops, Girona, Spain, July 11-15, 2011, Revised Selected Papers 19* (pp. 71-85). Springer Berlin Heidelberg.
https://doi.org/10.1007/978-3-642-28509-7_8
- 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>
- Anglim, J., Horwood, S., Smillie, L. D., Marrero, R. J., & Wood, J. K. (2020). Predicting psychological and subjective well-being from personality: A meta-analysis. *Psychological Bulletin*, 146(4), 279–323. <https://doi.org/10.1037/bul0000226>

- Asali, M., Pignatti, N., & Skhirtladze, S. (2018). Employment discrimination in a former Soviet Union Republic: Evidence from a field experiment. *Journal of Comparative Economics*, 46(4), 1294-1309.
- 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>
- Baert, S., De Visschere, S., Schoors, K., Vandenberghe, D., & Omey, E. (2016). First depressed, then discriminated against?. *Social Science & Medicine*, 170, 247-254. <https://doi.org/10.1016/j.socscimed.2016.06.033>
- Barde, B. V., & Bainwad, A. M. (2017). An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 745-750). IEEE.
- Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1), 1-26. <https://doi.org/10.1111/j.1744-6570.1991.tb00688.x>
- Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next?. *International Journal of Selection and Assessment*, 9(1-2), 9-30. <https://doi.org/10.1111/1468-2389.00160>
- Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology*, 54(2), 387-419. <https://doi.org/10.1111/j.1744-6570.2001.tb00097.x>

- Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior?. *Perspectives on Psychological Science*, 2(4), 396-403. <https://doi.org/10.1111/j.1745-6916.2007.00051.x>
- Ben Youssef, A., Chollet, M., Jones, H., Sabouret, N., Pelachaud, C., & Ochs, M. (2015). Towards a socially adaptive virtual agent. In *Intelligent Virtual Agents: 15th International Conference, IVA 2015, Delft, The Netherlands, August 26-28, 2015, Proceedings* 15 (pp. 3-16). Springer International Publishing. https://doi.org/10.1007/978-3-319-21996-7_1
- Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology*, 92(2), 410–424. <https://doi.org/10.1037/0021-9010.92.2.410>
- Beus, J. M., Dhanani, L. Y., & McCord, M. A. (2015). A meta-analysis of personality and workplace safety: Addressing unanswered questions. *Journal of Applied Psychology*, 100(2), 481–498. <https://doi.org/10.1037/a0037916>
- Bevilacqua, V., Guccione, P., Mascolo, L., Pazienza, P.P., Salatino, A.A., Pantaleo, M. (2013). First Progresses in Evaluation of Resonance in Staff Selection through Speech Emotion Recognition. In: Huang, DS., Jo, KH., Zhou, YQ., Han, K. (Eds.) *Intelligent Computing Theories and Technology*. ICIC 2013. Lecture Notes in Computer Science, (pp. 658-671). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39482-9_76
- Binning, J. F., Goldstein, M. A., Garcia, M. F., & Scattaregia, J. H. (1988). Effects of preinterview impressions on questioning strategies in same- and opposite-sex

employment interviews. *Journal of Applied Psychology*, 73(1), 30–37. <https://doi.org/10.1037/0021-9010.73.1.30>

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, 14(4), 317-335. <https://doi.org/10.1111/j.1468-2389.2006.00354.x>

Bjørnshagen, V. (2021). The mark of mental health problems. A field experiment on hiring discrimination before and during COVID-19. *Social Science & Medicine*, 283, 114181.

Bjørnshagen, V., & Ugreninov, E. (2021). Labour market inclusion of young people with mental health problems in Norway. *ALTER. European Journal of Disability Research*, (15-1), 46-60.

Blanck, P. (2022). Disability-inclusive employment, cancer survivorship, and the Americans with Disabilities Act. *Journal of Cancer Survivorship*, 16(1), 142-151.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Bleidorn, W., & Hopwood, C. J. (2019). Using Machine Learning to Advance Personality Assessment and Theory. *Personality and Social Psychology Review*, 23(2), 190-203. <https://doi.org/10.1177/1088868318772990>

Bleidorn, W., Hopwood, C. J., & Wright, A. G. (2017). Using big data to advance personality theory. *Current Opinion in Behavioral Sciences*, 18, 79-82.

<https://doi.org/10.1016/j.cobeha.2017.08.004>

Bolaños, F., Salatino, A., Osborne, F., & Motta, E. (2024). Artificial intelligence for literature reviews: opportunities and challenges. *Artificial Intelligence Review*, 57, 259.

<https://doi.org/10.1007/s10462-024-10902-3>

Bolger, N. & Schilling, E. A. (1991). Personality and the problems of everyday life: The role of neuroticism in exposure and reactivity to daily stressors. *Journal of Personality*, 59(3), 355-386.

Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021, October). Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 international conference on multimodal interaction* (pp. 268-277). <https://doi.org/10.1145/3462244.3479897>

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>

Burke, T.B. (2018). Does the ADA Discriminate Against Deaf People?. In: Boonin, D. (Eds.) The Palgrave Handbook of Philosophy and Public Policy. Palgrave Macmillan, Cham.
https://doi.org/10.1007/978-3-319-93907-0_30

Calabrese, W. R., Emery, L. T., Evans, C. M., & Simms, L. J. (2024). Diagnostic and Statistical Manual of Mental Disorders, personality disorders and the alternative model: Prediction of naturally observed behavior, interpersonal functioning, and psychiatric symptoms, 1 year later. *Personality Disorders: Theory, Research, and Treatment*, 15(5), 361-370.

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>

Campion, E. D., & Campion, M. A. (2024). Impact of machine learning on personnel selection. *Organizational Dynamics*, 53(1), 101035.

Campion, E. D., Campion, M. A., Johnson, J., Carretta, T. R., Romay, S., Dirr, B., Deregle, 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. A., & Campion, E. D. (2023). Machine learning applications to personnel selection: Current illustrations, lessons learned, and future research. *Personnel psychology*, 76(4), 993-1009.

Campion, M. A., Campion, J. E., & Hudson, J. P., Jr. (1994). Structured interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology*, 79(6), 998–1002. <https://doi.org/10.1037/0021-9010.79.6.998>

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>

Charbonneau, B. D., Powell, D. M., Spence, J. R., & Lyons, S. T. (2021). Unintended consequences of interview faking: Impact on perceived fit and affective outcomes. *Personnel Assessment and Decisions*, 7(1), 6.

<https://doi.org/10.25035/pad.2021.01.006>

- Chen, L., Feng, G., Martin-Raugh, M. P., Leong, C. W., Kitchen, C., Yoon, S. Y., Lehman, B., Kell, H., & Lee, C. M. (2016, September). Automatic Scoring of Monologue Video Interviews Using Multimodal Cues. In *INTERSPEECH* (pp. 32-36).
- 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>
- Chollet, M., Ochs, M., Pelachaud, C. (2014). From Non-verbal Signals Sequence Mining to Bayesian Networks for Interpersonal Attitudes Expression. In: Bickmore, T., Marsella, S., Sidner, C. (Eds.) *Intelligent Virtual Agents*. IVA 2014. Lecture Notes in Computer Science, (pp. 120-133). Springer, Cham. https://doi.org/10.1007/978-3-319-09767-1_15
- Chu, C., Sun, T., Zhang, B., & Rounds, J. (2024, June 29). Assessing Vocational Interests through Chat: Development and Validation of the Career Guidance Chatbot (CGC-bot). <https://doi.org/10.31234/osf.io/upx5q>
- 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>
- Connelly, B. S., McAbee, S. T., Oh, I.-S., Jung, Y., & Jung, C.-W. (2022). A multirater perspective on personality and performance: An empirical examination of the trait-reputation-identity model. *Journal of Applied Psychology*, 107(8), 1352–1368. <https://doi.org/10.1037/apl0000732>

- Constantin, S. W. (1976). An investigation of information favorability in the employment interview. *Journal of Applied Psychology*, 61(6), 743–749. <https://doi.org/10.1037/0021-9010.61.6.743>
- Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, 80(5), 565–579. <https://doi.org/10.1037/0021-9010.80.5.565>
- 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>
- Costa, P. T., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. *Journal of Personality and Social Psychology*, 38(4), 668–678. <https://doi.org/10.1037/0022-3514.38.4.668>
- Costa, P. T., & McCrae, R. R. (1999). A five-factor theory of personality. In John, P. O., & Robins, R. W. (Eds.) *Handbook of personality: Theory and research* (pp. 159-181)
- 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>
- de Kock, F. S., & Hauptfleisch, D. B. (2018). Reducing racial similarity bias in interviews by increasing structure: A quasi-experiment using multilevel analysis. *International Perspectives in Psychology*, 7(3), 137-154. <https://doi.org/10.1037/ipp0000091>

DeGroot, T., & Gooty, J. (2009). Can nonverbal cues be used to make meaningful personality attributions in employment interviews?. *Journal of business and psychology*, 24, 179-192.

<https://doi.org/10.1007/s10869-009-9098-0>

Despotakis, D., Dimitrova, V., Lau, L., & Thakker, D. (2013). Semantic aggregation and zooming of user viewpoints in social media content. In *User Modeling, Adaptation, and Personalization: 21th International Conference, UMAP 2013, Rome, Italy, June 10-14, 2013 Proceedings* 21 (pp. 51-63). Springer Berlin Heidelberg.

https://doi.org/10.1007/978-3-642-38844-6_5

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding.

<https://doi.org/10.48550/arXiv.1810.04805>

Diener, E., Oishi, S., & Lucas, R. E. (2003). Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life. *Annual review of psychology*, 54(1), 403-425. <https://doi.org/10.1146/annurev.psych.54.101601.145056>

Donovan, J. J., Dwight, S. A., & Schneider, D. (2014). The impact of applicant faking on selection measures, hiring decisions, and employee performance. *Journal of Business and Psychology*, 29, 479-493. <https://doi.org/10.1007/s10869-013-9318-5>

Ellingson, J. E., Sackett, P. R., & Connelly, B. S. (2007). Personality assessment across selection and development contexts: Insights into response distortion. *Journal of Applied Psychology*, 92(2), 386–395. <https://doi.org/10.1037/0021-9010.92.2.386>

Ellis, A. P. J., West, B. J., Ryan, A. M., & DeShon, R. P. (2002). The use of impression management tactics in structured interviews: A function of question type? *Journal of Applied Psychology*, 87(6), 1200–1208. <https://doi.org/10.1037/0021-9010.87.6.1200>

Evans, J.R. and Mathur, A. (2005). The value of online surveys. *Internet Research*, 15(2), 195-219. <https://doi.org/10.1108/10662240510590360>

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>

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., Netzer, O., Siegel, A. A., Plank, B. & Van Bavel, J. J. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, 4, 96-111.

<https://doi.org/10.1038/s44159-024-00392-z>

Fishbein, M. & Ajzen, Icek. (1975). *Belief, attitude, intention and behaviour: An introduction to theory and research*.

Fossati, A., Somma, A., Borroni, S., Pincus, A. L., Markon, K. E., & Krueger, R. F. (2017). Profiling pathological narcissism according to DSM-5 domains and traits: A study on consecutively admitted Italian psychotherapy patients. *Psychological Assessment*, 29(11), 1400.

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>

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology*, 55(1), 149–158. <https://doi.org/10.1037/0022-3514.55.1.149>

Furnham, A., Petrides, K. V., Jackson, C. J., & Cotter, T. (2002). Do personality factors predict job satisfaction?. *Personality and individual Differences*, 33(8), 1325-1342.

[https://doi.org/10.1016/S0191-8869\(02\)00016-8](https://doi.org/10.1016/S0191-8869(02)00016-8)

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73(2), 349-360.

<https://doi.org/10.1093/poq/nfp031>

Gilliland, S. W., & Steiner, D. D. (2012). Applicant reactions to testing and selection. *The Oxford Handbook of Personnel Assessment and Selection, Oxford Library of Psychology*.

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, 26-42.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. J. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, the Netherlands: University Press.

Goldberg, S., & Rosenthal, R. (1986). Self-touching behavior in the job interview: Antecedents and consequences. *Journal of nonverbal behavior*, 10, 65-80.

<https://doi.org/10.1007/BF00987206>

Gonzalez-Mulé, E., DeGeest, D. S., McCormick, B. W., Seong, J. Y., & Brown, K. G. (2014).

Can we get some cooperation around here? The mediating role of group norms on the

relationship between team personality and individual helping behaviors. *Journal of Applied Psychology*, 99(5), 988–999. <https://doi.org/10.1037/a0037278>

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>

Gruen, A. (1955). The relation of dancing experience and personality to perception. *Psychological Monographs: General and Applied*, 69(14), 1–16. <https://doi.org/10.1037/h0093701>

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>

Heerwegh, D., & Loosveldt, G. (2009). Explaining the intention to participate in a web survey: a test of the theory of planned behaviour. *International Journal of Social Research Methodology*, 12(3), 181-195. <https://doi.org/10.1080/13645570701804235>

Heimann, A. L., & Schmitz-Wilhelmy, A. (2024). Observing Interviewees' Inner Self: How Authenticity Cues in Job Interviews Relate to Interview and Job Performance. *Journal of Business and Psychology*, 1-20. <https://doi.org/10.1007/s10869-024-09949-4>

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>

Hess, R. (2022). *Applicant Reactions to Artificial Intelligence in Selection*.
<https://www.proquest.com/dissertations-theses/applicant-reactions-artificial-intelligence/docview/2676599021/se-2>

Hickman, L. (2025). All Your Base Are Belong to Us: The Urgent Reality of Unproctored Testing in the Age of LLMs. *International Journal of Selection and Assessment*, 33(2), e70005. <https://doi.org/10.1111/ijsa.70005>

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., Herde, C. N., Lievens, F., & Tay, L. (2023). Automatic scoring of speeded interpersonal assessment center exercises via machine learning: Initial psychometric evidence and practical guidelines. *International Journal of Selection and Assessment*, 31(2), 225-239. <https://doi.org/10.1111/ijsa.12418>

Hickman, L., Huynh, C., Gass, J., Booth, B., Kuruzovich, J., & Tay, L. (2024a). Whither bias goes, I will go: An integrative, systematic review of algorithmic bias mitigation. *Journal of Applied Psychology*. Advance online publication. <https://doi.org/10.1037/apl0001255>

Hickman, L., Langer, M., Saef, R. M., & Tay, L. (2024b). Automated speech recognition bias in personnel selection: The case of automatically scored job interviews. *Journal of Applied Psychology*. Advance online publication. <https://doi.org/10.1037/apl0001247>

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. (2024c). 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>

Ho, J. L., Powell, D. M., & Stanley, D. J. (2021). The relation between deceptive impression management and employment interview ratings: A meta-analysis. *Canadian Journal of Behavioural Science / Revue Canadienne des Sciences du Comportement*, 53(2), 164–174. <https://doi.org/10.1037/cbs0000223>

Ho, J. L., Powell, D. M., Barclay, P., & Gill, H. (2019). The influence of competition on motivation to fake in employment interviews. *Journal of Personnel Psychology*, 18(2), 95–105. <https://doi.org/10.1027/1866-5888/a000222>

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology*, 92(5), 1270–1285. <https://doi.org/10.1037/0021-9010.92.5.1270>

Hogan, R., & Blickle, G. (2013). Socioanalytic theory. In *Handbook of personality at work* (pp. 53-70). Routledge.

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory (HPI)* [Database record]. APA PsycTests. <https://doi.org/10.1037/t02029-000>

Holtrop, D., Born, M. P., de Vries, A., & de Vries, R. E. (2014). A matter of context: A comparison of two types of contextualized personality measures. *Personality and Individual Differences*, 68, 234-240. <https://doi.org/10.1016/j.paid.2014.04.029>

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>

Hu, J. & Connelly, B. S. (2021). Faking by actual applicants on personality tests: A meta-analysis of within-subjects studies. *International Journal of Selection and Assessment*, 29(3-4), 411-426. <https://doi.org/10.1111/ijsa.12338>

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, 30, 299-311. <https://doi.org/10.1007/s10869-014-9357-6>

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., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, 79(2), 184–190. <https://doi.org/10.1037/0021-9010.79.2.184>

- Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology*, 83(2), 179–189. <https://doi.org/10.1037/0021-9010.83.2.179>
- Huffcutt, A. I., & Woehr, D. J. (1999). Further analysis of employment interview validity: A quantitative evaluation of interviewer-related structuring methods. *Journal of Organizational Behavior*, 20(4), 549-560. [https://doi.org/10.1002/\(SICI\)1099-1379\(199907\)20:4<549::AID-JOB921>3.0.CO;2-Q](https://doi.org/10.1002/(SICI)1099-1379(199907)20:4<549::AID-JOB921>3.0.CO;2-Q)
- 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>
- Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology*, 85(6), 869–879. <https://doi.org/10.1037/0021-9010.85.6.869>
- 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>
- Jayaratne, M., & Jayatilleke, B. (2020). Predicting personality using answers to open-ended interview questions. *IEEE Access*, 8, 115345-115355. <https://doi.org/10.1109/ACCESS.2020.3004002>
- Jebb, A. T., Ng, V., & Tay, L. (2021). A review of key Likert scale development advances: 1995–2019. *Frontiers in Psychology*, 12, 637547. <https://doi.org/10.3389/fpsyg.2021.637547>

Jensen, J. M., & Patel, P. C. (2011). Predicting counterproductive work behavior from the interaction of personality traits. *Personality and Individual Differences*, 51(4), 466-471.

<https://doi.org/10.1016/j.paid.2011.04.016>

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Eshii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12), 1-38. <https://doi.org/10.1145/3571730>

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology*, 87(4), 765.

Judge, T. A., Cable, D. M., & Higgins, C. A. (2000). The employment interview: A review of recent research and recommendations for future research. *Human Resource Management Review*, 10(4), 383-406. [https://doi.org/10.1016/S1053-4822\(00\)00033-4](https://doi.org/10.1016/S1053-4822(00)00033-4)

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>

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., Kosinski, M., Ramones, S. M., & Seligman, M. E. P. (2014). The online social self: An open vocabulary approach to personality. *Assessment*, 21(2), 158-169.

<https://doi.org/10.1177/1073191113514104>

Kim, S., Fernandez, S., & Terrier, L. (2017). Procrastination, personality traits, and academic performance: When active and passive procrastination tell a different story. *Personality and Individual differences*, 108, 154-157. <https://doi.org/10.1016/j.paid.2016.12.021>

King, M. R. (2023). The future of AI in medicine: A perspective from a Chatbot. *Annals of Biomedical Engineering*, 51(2), 291-295. <https://doi.org/10.1007/s10439-022-03121-w>

- Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004), 1–26.
- Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., Speer, A., Hardy III, H. J., Gibson, C., Frost, C., Liu, M., McNeney, D., Capman, J., Lowery, S., Kitching, M., Nimbkar, A., Boyce, A., Sun, T., Guo, F., ... & Newton, C. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. *Personnel Psychology*, 76(4), 1061-1123.
<https://doi.org/10.1111/peps.12608>
- 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>
- König, C. J., & Langer, M. (2022). Machine learning in personnel selection. In *Handbook of research on artificial intelligence in human resource management* (pp. 149-167). Edward Elgar Publishing.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15), 5802-5805. <https://doi.org/10.1073/pnas.1218772110>
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.
- Landers, R. N., & Behrend, T. S. (2023). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*, 78(1), 36.
<https://doi.org/10.1037/amp0000972>
- Langer, M., König, C. J., & Fitili, A. (2018). Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies

for personnel selection. *Computers in Human Behavior*, 81, 19-30.

<https://doi.org/10.1016/j.chb.2017.11.036>

Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly automated job interviews:

Acceptance under the influence of stakes. *International Journal of Selection and Assessment*, 27(3), 217-234. <https://doi.org/10.1111/ijsa.12246>

Langer, M., König, C. J., Sanchez, D. R. P., & Samadi, S. (2020). Highly automated interviews:

Applicant reactions and the organizational context. *Journal of Managerial Psychology*, 35(4), 301-314. <https://doi.org/10.1108/JMP-09-2018-0402>

Lee, B. M., Kang, C. Y., Li, L., Rami, F. Z., & Chung, Y. C. (2023). Physiological responses to the virtual reality-based Trier social stress test in patients with psychosis. *Virtual Reality*, 27(4), 3115-3123. <https://doi.org/10.1007/s10055-023-00857-1>

Lee, J., Voss, N. M., Stoffregen, S. A., Giordano, F. B., Warren, C., Chlevin-Thiele, C., & Klos, L. S. (Under review). Interpretable Integration of Predictions by Multiple Machine Learning Methods: Enhancing Integrity for Personnel Decisions.

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., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology*, 91(2), 247–258. <https://doi.org/10.1037/0021-9010.91.2.247>
- 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>
- Lievens, F., Harris, M. M., Van Keer, E., & Bisqueret, C. (2003). Predicting cross-cultural training performance: The validity of personality, cognitive ability, and dimensions measured by an assessment center and a behavior description interview. *Journal of Applied Psychology*, 88(3), 476–489. <https://doi.org/10.1037/0021-9010.88.3.476>
- Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1), 96-116. <https://doi.org/10.1002/lio2.354>
- 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>

- Mackin, D. M., Finsaas, M. C., Nelson, B. D., Perlman, G., Kotov, R., & Klein, D. N. (2022). Intergenerational transmission of depressive and anxiety disorders: Mediation via youth personality. *Journal of Psychopathology and Clinical Science*, 131(5), 467-478.
<https://doi.org/10.1037/abn0000759>
- Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
- Maples-Keller, J. L., Hyatt, C. S., Sleep, C. E., Stevens, J. S., Fenlon, E. E., Jovanovic, T., Rothbaum, B. O., Ressler, K. J., Carter, S., Bradley, B., Fani, N., Powers, A., & Michopoulos, V. (2021). DSM-5 alternative model for personality disorders trait domains and PTSD symptoms in a sample of highly traumatized African American women and a prospective sample of trauma center patients. *Personality Disorders: Theory, Research, and Treatment*, 12(6), 491–502. <https://doi.org/10.1037/per0000477>
- McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: The trait-reputation-identity model. *Psychological Review*, 123(5), 569–591.
<https://doi.org/10.1037/rev0000035>
- McAbee, S. T., & Oswald, F. L. (2013). The criterion-related validity of personality measures for predicting GPA: A meta-analytic validity competition. *Psychological Assessment*, 25(2), 532–544. <https://doi.org/10.1037/a0031748>
- McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist*, 52(5), 509–516. <https://doi.org/10.1037/0003-066X.52.5.509>

- 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>
- McLellan, E., MacQueen, K. M., & Neidig, J. L. (2003). Beyond the Qualitative Interview: Data Preparation and Transcription. *Field Methods*, 15(1), 63-84.
<https://doi.org/10.1177/1525822X02239573>
- Melchers, K. G., Roulin, N., & Buehl, A-K. (2020). A review of applicant faking in selection interviews. *International Journal of Selection and Assessment*, 28(2), 123-142.
<https://doi.org/10.1111/ijsa.12280>
- Meriac, J. P., Hoffman, B. J., & Woehr, D. J. (2014). A conceptual and empirical review of the structure of assessment center dimensions. *Journal of Management*, 40(5), 1269-1296.
<https://doi.org/10.1177/0149206314522299>
- Meta AI (2024). *Meta Llama 3.1*. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1
- Michelson, M., & Reuter, K. (2019). The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary Clinical Trials Communications*, 16, 100443.
<https://doi.org/10.1016/j.conctc.2019.100443>
- Midtbøen, A. H. (2016). Discrimination of the second generation: Evidence from a field experiment in Norway. *Journal of International Migration and Integration*, 17, 253-272.
<https://doi.org/10.1007/s12134-014-0406-9>

- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102(2), 246.
- Møller, L., Meisner, M. W., Søgaard, U., Elklist, A., & Simonsen, E. (2021). Assessment of personality functioning in ICD-11 posttraumatic stress disorder and complex posttraumatic stress disorder. *Personality Disorders: Theory, Research, and Treatment*, 12(5), 466–474. <https://doi.org/10.1037/per0000491>
- 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>
- Mount, M. K., Barrick, M. R., & Stewart, G. L. (1998). Personality predictors of performance in jobs involving interaction with others. *Human Performance*, 11(3), 145-166.
- Mueller, A. E., & Segal, D. L. (2015). Structured versus semistructured versus unstructured interviews. *The Encyclopedia of Clinical Psychology*, 1-7.
<https://doi.org/10.1002/9781118625392.wbecp069>
- 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>
- Nease, R. F., Tsai, R., Hynes, L. M., & Littenberg, B. (1996). Automated utility assessment of global health. *Quality of Life Research*, 5, 175-182. <https://doi.org/10.1007/BF00435983>

- Nikolaou, I. (2003). Fitting the person to the organisation: examining the personality-job performance relationship from a new perspective. *Journal of Managerial Psychology*, 18(7), 639-648. <https://doi.org/10.1108/02683940310502368>
- Nikolaou, I. (2011). Core processes and applicant reactions to the employment interview: An exploratory study in Greece. *The International Journal of Human Resource Management*, 22(10), 2185-2201. <https://doi.org/10.1080/09585192.2011.580187>
- Nikolaou, I., & Georgiou, K. (2018). Fairness reactions to the employment interview. *Journal of Work and Organizational Psychology*, 34(2), 103-111.
<https://doi.org/10.5093/jwop2018a13>
- Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, 93(1), 116–130. <https://doi.org/10.1037/0022-3514.93.1.116>
- O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, 112, 28-39.
- Oesch, D. (2020). Discrimination in the hiring of older jobseekers: Combining a survey experiment with a natural experiment in Switzerland. *Research in Social Stratification and Mobility*, 65, 100441. <https://doi.org/10.1016/j.rssm.2019.100441>
- Oh, I.-S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology*, 96(4), 762–773. <https://doi.org/10.1037/a0021832>

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology*, 60(4), 995-1027.

<https://doi.org/10.1111/j.1744-6570.2007.00099.x>

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, 78(4), 679–703. <https://doi.org/10.1037/0021-9010.78.4.679>

Ooms, J (2024). *cld3: Google's Compact Language Detector 3*. R package version 1.6.1 <https://docs.ropensci.org/cld3/> <https://ropensci.r-universe.dev/cld3>

OpenAI. (2022). Whisper API. <https://openai.com/index/whisper/>
Osborn, S.M., Field, H.S. & Veres, J.G. Introversion-Extraversion, Self-Monitoring, and Applicant Performance in a Situational Panel Interview: A Field Study. *Journal of Business and Psychology* 13, 143–156 (1998). <https://doi.org/10.1023/A:1022965721229>

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in industrial-organizational psychology and human resource management: Forward progress for organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior*, 7(1), 505–533. <https://doi.org/10.1146/annurev-orgpsych-032117-104553>

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E.,

- McDonald, S., ... Moher, D. (2021). The Prisma 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372. <https://doi.org/10.1136/bmj.n71>
- Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6), 934–952. <https://doi.org/10.1037/pspp0000020>
- Patwardhan, N., Marrone, S., & Sansone, C. (2023). Transformers in the real world: A survey on NLP applications. *Information*, 14(4), 242. <https://doi.org/10.3390/info14040242>
- Paul, J., & Barari, M. (2022). Meta-analysis and traditional systematic literature reviews—What, why, when, where, and how?. *Psychology & Marketing*, 39(6), 1099-1115. <https://doi.org/10.1002/mar.21657>
- Pennebaker, J. W., Francis, L. E., & Booth, R. J. (2001). *LIWC: Linguistic inquiry and word count*.
- Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, 61(1), 153-172. <https://doi.org/10.1111/j.1744-6570.2008.00109.x>
- 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>
- Porter, S. R., Whitcomb, M. E., & Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, 2004(121), 63-73. <https://doi.org/10.1002/ir.101>

- Powell, D. M., & Bourdage, J. S. (2016). The detection of personality traits in employment interviews: Can “good judges” be trained?. *Personality and Individual Differences*, 94, 194–199. <https://doi.org/10.1016/j.paid.2016.01.009>
- Prewett-Livingston, A. J., Feild, H. S., Veres, J. G. III, & Lewis, P. M. (1996). Effects of race on interview ratings in a situational panel interview. *Journal of Applied Psychology*, 81(2), 178–186. <https://doi.org/10.1037/0021-9010.81.2.178>
- Qian, Y., Ubale, R., Lange, P., Evanini, K., Ramanarayanan, V., & Soong, F. K. (2020). Spoken language understanding of human-machine conversations for language learning applications. *Journal of Signal Processing Systems*, 92(8), 805–817. <https://doi.org/10.1007/s11265-019-01484-3>
- He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Reis, S.B., Young, I.P. & Jury, J.C. Female administrators: A crack in the glass ceiling. *Journal of Personnel Evaluation in Education*, 13, 71–82 (1999). <https://doi.org/10.1023/A:1008002221974>
- Roberts, B. W., & Nickel, L. B. (2021). Personality development across the life course: A neo-socioanalytic perspective. In O. P. John & R. W. Robins (Eds.), *Handbook of personality: Theory and research* (4th ed., pp. 259–283). The Guilford Press.
- Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science*, 2(4), 313–345. <https://doi.org/10.1111/j.1745-6916.2007.00047.x>

- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, 91(2), 1–40.
<https://doi.org/10.18637/jss.v091.i02>
- Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, 83(4), 634–644. <https://doi.org/10.1037/0021-9010.83.4.634>
- Roulin, N., & Krings, F. (2020). Faking to fit in: Applicants' response strategies to match organizational culture. *Journal of Applied Psychology*, 105(2), 130–145. <https://doi.org/10.1037/apl0000431>
- 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., & Walmsley, P. T. (2014). Which personality attributes are most important in the workplace?. *Perspectives on Psychological Science*, 9(5), 538-551.
<https://doi.org/10.1177/1745691614543972>
- 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>
- Schlotzhauer, A.E., Ng, M.A. & Su, S. How to frame the frame of reference: A comparison of contextualization methods. *Journal of Business and Psychology*, 40, 385–403 (2025). <https://doi.org/10.1007/s10869-024-09953-8>
- Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology*, 65(3), 445-494. <https://doi.org/10.1111/j.1744-6570.2012.01250.x>
- Siocon, G. 2023. Ways AI is changing HR departments. Business News Daily. <https://www.businessnewsdaily.com/how-ai-is-changing-hr>
- Smith, M. B., Wu, I-H. (R.), Holmes, R. M., Jr., & Hodge, A. M. (2024). An integrative conceptual review of multiperspective frameworks in personality research and a roadmap for extended applications in organizational psychology. *Journal of Applied Psychology*. Advance online publication. <https://doi.org/10.1037/apl0001195>

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., Bauer, J. A., & Fox, S. (2010). Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior: Do we know what we think we know? *Journal of Applied Psychology*, 95(4), 781-790.

<http://doi.org/10.1037/a0019477>

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>

Speer, A. B., Christiansen, N., & Honts, C. (2015). Assessment of personality through behavioral observations in work simulations. *Personnel Assessment and Decisions*, 1(1), 6.

<https://doi.org/10.25035/pad.2015.006>

Stanek, M. L., Boaz, K. M., Cordes, C. N., Niese, T. D., Long, K. E., Risner, M. S., ... & Zoladz, P. R. (2024). Social evaluative stress enhances central detail memory, reduces false memory, and results in intrusive memories that last for days. *Neurobiology of Learning and Memory*, 209, 107906. <https://doi.org/10.1016/j.nlm.2024.107906>

Stanek, M. L., Boaz, K. M., Cordes, C. N., Niese, T. D., Long, K. E., Risner, M. S., Blasco, J. G., Suzelis, K. N., Siereveld, M. S., Rorabaugh, B. R., & Zoladz, P. R. (2024). Social

evaluative stress enhances central detail memory, reduces false memory, and results in intrusive memories that last for days. *Neurobiology of Learning and Memory*, 209, 107906. <https://doi.org/10.1016/j.nlm.2024.107906>

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>

Steiner, D. D., & Rain, J. S. (1989). Immediate and delayed primacy and recency effects in performance evaluation. *Journal of Applied Psychology*, 74(1), 136–142. <https://doi.org/10.1037/0021-9010.74.1.136>

Stryker, C., & Holdsworth, J. (2024, August 11). *What is natural language processing?* IBM.
<https://www.ibm.com/think/topics/natural-language-processing>

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>

Swift, V., & Peterson, J. B. (2019). Contextualization as a means to improve the predictive validity of personality models. *Personality and Individual Differences*, 144, 153-163.

<https://doi.org/10.1016/j.paid.2019.03.007>

Szalai, J. (2021). The potential use of artificial intelligence in the therapy of borderline personality disorder. *Journal of Evaluation in Clinical Practice*, 27(3), 491-496.
<https://doi.org/10.1111/jep.13530>

Tajfel, H., & Turner, J. (1979). An integrative theory of intergroup conflict. In J. A. Williams & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Belmont, CA: Wadsworth.

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>

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in psychological assessment. *Advances in Methods and Practices in Psychological Science*, 5(1), 25152459211061337. <https://doi.org/10.1177/25152459211061337>

Tescari, M.-E., & Bangerter, A. (2025). Nonverbal behavior in selection interviews: Relation to communion, agency, and interview performance. *Journal of Personnel Psychology*, 24(1), 23–32. <https://doi.org/10.1027/1866-5888/a000350>

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>

Tett, R. P., Simonet, D. V., Walser, B., & Brown, C. (2013). Trait activation theory: Applications, developments, and implications for person–workplace fit. In *Handbook of personality at work* (pp. 71-100). Routledge.

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>

Tippins, N. T., Oswald, F. L., & McPhail, S. M. (2021). Scientific, legal, and ethical concerns about AI-based personnel selection tools: a call to action. *Personnel Assessment and Decisions*, 7(2), 1. <https://doi.org/10.25035/pad.2021.02.001>

Truxillo, D. M., Bodner, T. E., Bertolino, M., Bauer, T. N., & Yonce, C. A. (2009). Effects of explanations on applicant reactions: A meta-analytic review. *International Journal of Selection and Assessment*, 17(4), 346-361. <https://doi.org/10.1111/j.1468-2389.2009.00478.x>

Truxillo, D. M., Steiner, D. D., & Gilliland, S. W. (2004). The importance of organizational justice in personnel selection: Defining when selection fairness really matters. *International Journal of Selection and Assessment*, 12(1-2), 39-53.

<https://doi.org/10.1111/j.0965-075X.2004.00262.x>

Van Borm, H., Burn, I., & Baert, S. (2021). What does a job candidate's age signal to employers?. *Labour Economics*, 71, 102003.

Van Dinter, R., Tekinerdogan, B., & Catal, C. (2021). Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, 136, 106589. <https://doi.org/10.1016/j.infsof.2021.106589>

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>

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, 59(2), 197-210. <https://doi.org/10.1177/00131649921969802>

Völkel, S. T., Haeuslschmid, R., Werner, A., Hussmann, H., & Butz, A. (2020, April). How to Trick AI: Users' strategies for protecting themselves from automatic personality assessment. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-15).

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>

Watson, R. T. (2015). Beyond being systematic in literature reviews in IS. *Journal of Information Technology*, 30, 185-187. <https://doi.org/10.1057/jit.2015.12>

Wexley, K. N., Yukl, G. A., Kovacs, S. Z., & Sanders, R. E. (1972). Importance of contrast effects in employment interviews. *Journal of Applied Psychology*, 56(1), 45–48. <https://doi.org/10.1037/h0032132>

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>

- Williamson, L. G., Campion, J. E., Malos, S. B., Roehling, M. V., & Campion, M. A. (1997). Employment interview on trial: Linking interview structure with litigation outcomes. *Journal of Applied Psychology*, 82(6), 900–912. <https://doi.org/10.1037/0021-9010.82.6.900>
- Wingate, T. G., & Bourdage, J. S. (2019). Liar at first sight? Early impressions and interviewer judgments, attributions, and false perceptions of faking. *Journal of Personnel Psychology*, 18(4), 177. <https://doi.org/10.1027/1866-5888/a000232>
- 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>
- Woo, S. E., Tay, L., Jebb, A. T., Ford, M. T., & Kern, M. L. (2020). Big data for enhancing measurement quality. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychological research* (pp. 59–85). American Psychological Association. <https://doi.org/10.1037/0000193-004>
- Woods, S. A., Lievens, F., De Fruyt, F., & Wille, B. (2013). Personality across working life: The longitudinal and reciprocal influences of personality on work. *Journal of Organizational Behavior*, 34(S1), S7-S25. <https://doi.org/10.1002/job.1863>
- 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>

Xiong, R. R., Liu, C. Z., & Choo, KK.R. (2023). Synthesizing Knowledge through A Data Analytics-Based Systematic Literature Review Protocol. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-023-10432-3>

Yamamoto, K., Inoue, K., & Kawahara, T. (2023). Character expression for spoken dialogue systems with semi-supervised learning using Variational Auto-Encoder. *Computer Speech & Language*, 79, 101469. <https://doi.org/10.1016/j.csl.2022.101469>

Yang, Q., Wang, P., Plonsky, L. D., Oswald, F. L., & Chen, H. (2024). From Babbling to Fluency: Evaluating the Evolution of Language Models in Terms of Human Language Acquisition. <https://doi.org/10.48550/arXiv.2410.13259>

Young, E. A., & Nolen-Hoeksema, S. (2001). Effect of ruminations on the saliva cortisol response to a social stressor. *Psychoneuroendocrinology*, 26(3), 319-329. [https://doi.org/10.1016/S0306-4530\(00\)00059-7](https://doi.org/10.1016/S0306-4530(00)00059-7)

Yu, Z., Ramanarayanan, V., Lange, P., & Suendermann-Oeft, D. (2018, August). *An open-source dialog system with real-time engagement tracking for job interview training applications*. In M. Eskenazi, L. Devillers, & J. Mariani (Eds.), *Advanced social interaction with agents: 8th International Workshop on Spoken Dialog Systems. Lecture notes in electrical engineering* (Vol. 510, pp. 199-207). Springer, Cham. https://doi.org/10.1007/978-3-319-92108-2_21

Yuan, L(I.), Sun, T., Dennis, A. R., & Zhou, M. (2024). Perception is reality? Understanding user perceptions of chatbot-inferred versus self-reported personality traits. *Computers in*

Human Behavior: Artificial Humans, 100057.

<https://doi.org/10.1016/j.chbah.2024.100057>

Zettl, M., Volkert, J., Vögele, C., Herpertz, S. C., Kubera, K. M., & Taubner, S. (2020).

Mentalization and criterion a of the alternative model for personality disorders: Results from a clinical and nonclinical sample. *Personality Disorders: Theory, Research, and Treatment*, 11(3), 191–201. <https://doi.org/10.1037/per0000356>

Zhang, L., & Yencha, C. (2022). Examining perceptions towards hiring algorithms. *Technology in Society*, 68, 101848. <https://doi.org/10.1016/j.techsoc.2021.101848>

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).

Zhou, Z. E., Meier, L. L., & Spector, P. E. (2014). The role of personality and job stressors in predicting counterproductive work behavior: A three-way interaction. *International Journal of Selection and Assessment*, 22(3), 286-296. <https://doi.org/10.1111/ijsa.12077>

Zschirnt, E. (2020). Evidence of hiring discrimination against the second generation: results from a correspondence test in the Swiss labour market. *Journal of International Migration and Integration*, 21(2), 563-585. <https://doi.org/10.1007/s12134-019-00664-1>

Zubin, J., Eron, L. D., & Sultan, F. (1956). Current status of the Rorschach test: Symposium, 1955: 1. A psychometric evaluation of the Rorschach experiment. *American Journal of Orthopsychiatry*, 26(4), 773–782. <https://doi.org/10.1111/j.1939-0025.1956.tb06221.x>

Table 1. Model Selection Indices for Structural Topic Modeling

K	Exclusivity	Semantic Coherence	Held-out Likelihood	Residual
3	8.05	-59.21	-6.92	2.63
4	8.40	-60.78	-6.90	2.45
5	8.76	-66.67	-6.87	2.31
6	8.97	-76.31	-6.87	2.20
7	9.07	-77.98	-6.85	2.12
8	9.15	-78.58	-6.85	2.05
9	9.19	-77.49	-6.82	1.96
10	9.25	-76.10	-6.85	1.91

Notes. Exclusivity and semantic coherence are presented for models with varying numbers of topics (K). Higher exclusivity values indicate better separation between topics, while higher semantic coherence values indicate more interpretable topics. A 5-topic solution was selected since it offers a balance between semantic coherence and exclusivity.

Table 2. Overview of Topic Information

Topic	Label	Top Words	Proportion	Exclusivity	Coherence
1	Clinical Assessment	bpd, dsm, disord, diagnost, symptom, pds, ampd, borderlin, clinic, patholog	15%	9.00	-55.42
2	Bias and Faking	applic, interviewe, fake, nonverb, impress, pharmacist, tactic, decept, videotap, judgment	28%	8.97	-53.48
3	Legal Approaches to Fairness	cancer, applic, ethnic, labour, hire, sector, accommod, covid, januari, career	17%	8.65	-99.67
4	AI and ML	jit, virtual, user, skill, human, game, learn, system, learner, automat	21%	8.30	-69.33
5	Stress and Psychological Testing	rorschach, cortisol, tsst, reliabl, book, scale, avi, perceptu, mmi, psychopathi	20%	8.45	-61.25

Notes. The top words for each topic are based on the score metric, which weights words by their overall frequency and exclusivity to the topic (Roberts et al., 2019). Top words are shown in stemmed form, a process that reduces words to their base form and helps group different variations of a word together. For example, “interviewer” and “interviewees” get grouped together as “interviewe” through stemming. Proportion represents the percentage of each topic’s occurrence across all documents. Exclusivity indicates how distinct topics are from one another. Coherence indicates how semantically related words are within each topic.

Table 3. STM Results of Search Term Effect on Topic Prevalence

Covariate	Topic 1		Topic 2		Topic 3		Topic 4		Topic 5	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	.45**	.04	-.11	-.06	.24**	.05	.19**	.05	.22	.05
Personality	.00	.04	.21**	.05	-.16**	.05	-.13*	-.05	.08	.04
Automated	-.27**	.06	.21*	.08	-.14*	.07	.17*	.08	.03	.07
Employment	-.42**	.04	.44**	.06	.00	.05	.07	.05	-.09	-.05

Notes. Bolded values indicate significance at the $p < .05$ level. * $p < .05$, ** $p < .001$. Covariates were dummy coded as follows: 0 = document did not contain search term in title or abstract; 1 = document contained search term in title or abstract. *Personality* search term is based on the search term “personality+interview”, *Automated* is based on the search term “automated+interview”, *Employment* is based on the search terms “employment+interview”, “selection+interview”, or “job+interview”.

Table 4. Means, Standard Deviations, and Intercorrelations of Part II Study Variables for Chatbot Interview Samples

Variables	Students (Sample 2)	Working Adults (Sample 3)										
	M (SD)	M (SD)	1	2	3	4	5	6	7	8	9	10
<i>Self-reports</i>												
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
<i>Human ratings</i>												
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**	-
<i>Embedding</i>												
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**
<i>Zero-shot</i>												
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**
<i>Outcome variables</i>												
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. $N_{\text{Sample 2}} = 130$; $N_{\text{Sample 3}} = 88$. * $p < .05$ and ** $p < .001$.

Table 4 continued

	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
<i>Self-reports</i>																
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**
<i>Human ratings</i>																
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
<i>Embedding</i>																
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
<i>Zero-shot</i>																
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
<i>Outcome variables</i>																
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**	-

Table 5. Correlations between Self and Rater Personality Scores

	Human interview			Chatbot interview					
	Sample 1 (working adults)		<i>N</i> = 203	Sample 2 (students)			Sample 3 (working adults)		
	Self-report	Human interview rating		Self-report	Human interview rating	Self-report	Human interview rating		
	<i>M</i> (<i>SD</i>)	<i>M</i> (<i>SD</i>)	<i>r</i>	<i>M</i> (<i>SD</i>)	<i>M</i> (<i>SD</i>)	<i>r</i>	<i>M</i> (<i>SD</i>)	<i>M</i> (<i>SD</i>)	<i>r</i>
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 6. Means and Standard Deviations of Word Count by Interview Question

	Human interview		Chatbot interview
	Sample 1	Sample 2	Sample 3
	(working adults)	(students)	(working adults)
<i>Extraversion</i>			
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)
<i>Agreeableness</i>			
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)
<i>Conscientiousness</i>			
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)
<i>Emotional Stability</i>			
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)
<i>Openness</i>			
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 7. 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 <i>r</i>		
Heterotrait-monomehtod (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-heteromehtod (HTHM)	.132	.111	.292	.116	.233	.361	.207
Monotrait-heteromehtod (MTHM)	.232	.136	.553	.203	.680	.619	.404
Variance Partitioning	C1 .404	D1 .196	D2 .055	MV .142			

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

Table 8. 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 <i>r</i>		
Heterotrait-monomehtod (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-heteromehtod (HTHM)	.136	.075	.531	.154	.433	.568	.316
Monotrait-heteromehtod (MTHM)	.171	.044	.661	.221	.725	.739	.427
Variance Partitioning	C1 .427	D1 .110	D2 -.134	MV .245			

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

Table 9. 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	.28**	.08	.27**	.08	.44**	.09	.43**	.09	.42**	.09
Extraversion (M)			.42	.39	.11	.07			.46	.29	.08	.07
R ²	.10**		.11**		.12**		.23**		.25**		.24**	
ΔR ²			.02		.02				.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		.04	
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_{Sample 2} = 130; N_{Sample 3} = 88.

Table 10. 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		
	<i>B</i>	<i>SE</i>	Embedding		Zero-shot		<i>B</i>	<i>SE</i>	Embedding		Zero-shot		
			<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>			<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>	
Extraversion (S)	-.04	.07	-.04	.07	-.05	.08	-.20*	.10	-.20*	.10	-.20	.10	
Extraversion (M)			.05	.38	<.01	.07			.19	.30	.01	.08	
<i>R</i> ²	<.01		<.01		<.01		.05*		.05	.05			
ΔR^2									.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	
<i>R</i> ²	.21**		.21**		.21**		.41**		.46**	.43**			
ΔR^2									.05*				
Conscientiousness (S)	-.38**	.09	-.38**	.09	-.38**	.09	-.62**	.10	-.61**	.09	-.63**	.10	
Conscientiousness (M)			.08	.23	-.01	.06			.34	.22	.06	.06	
<i>R</i> ²	.13**		.13**		.13**		.33**		.35**	.34**			
ΔR^2									.02				
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
<i>R</i> ²	.08*		.08*		.09*		.16**		.18**	.17**			
ΔR^2									.02				
Openness (S)	-.13	.08	-.13	.08	-.12	.08	-.28*	.09	-.28*	.09	-.31*	.09	
Openness (M)					.01	.24	-.06	.06		.29	.21	.15*	.07
<i>R</i> ²	.03		.03		.03		.09*		.11*	.13*			
ΔR^2									.02				

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

Table 11. 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	
	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 12. 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	
	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*			.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			.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		.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		.31**		.31**		.33**	
ΔR ²			.00		.03				.00		.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		.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 13. 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	
	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 14. 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_{Sample 2} = 130; N_{Sample 3} = 88.

Table 15. Correlations for Study Variables for Part III for Combined and Honest condition.

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Baseline																		
1. Extraversion	.17	.64**	.38*	.52**	.91**	.57**	.26	.45*	.53**	-.04	.16	-.18	-.09	.01	.29*	.29*	.07	
2. Agreeableness	.32**	.47**	.37*	.39*	.21	.61**	.54**	.30*	.44*	-.02	.14	-.08	-.05	.09	.03	.26	-.03	
3. Conscientiousness	.51**	.39**	.41*	.39*	.54**	.43*	.41*	.78**	.41*	-.08	.10	-.06	.01	.09	.36*	.29*	.17	
4. Emotional Stability	.47**	.41**	.48**	.39*	.26	.23	.87**	.31*	.39*	-.05	.08	-.13	-.07	-.10	-.01	.03	-.06	
5. Openness	.50**	.42**	.48**	.44**	.51**	.68**	.39*	.17	.88**	.07	.19	-.01	.14	.18	.17	.24	.05	
Post-Condition																		
6. Extraversion	.81**	.26*	.23*	.32**	.39**	.58**	.18	.34*	.55**	-.06	.14	-.20	-.01	.07	.30*	.24	.09	
7. Agreeableness	.50**	.66**	.28**	.27*	.53**	.54**	.26	.24	.76**	.07	.27	-.08	.00	.16	.23	.42*	.00	
8. Conscientiousness	.32**	.41**	.34**	.82**	.33**	.31**	.34**	.30*	.33*	-.19	.01	-.22	-.19	-.18	-.09	.11	-.12	
9. Emotional Stability	.39**	.27*	.72**	.35**	.30**	.24*	.27*	.41**	.20	-.04	.11	-.01	-.03	-.01	.36*	.24	.21	
10. Openness	.51**	.39**	.32**	.35**	.75**	.52**	.65**	.35**	.32**	.01	.07	-.10	.03	.15	.23	.22	-.04	
Embedding																		
11. Extraversion	.05	-.04	.02	.07	.14	.00	.11	-.04	.02	.08		.79**	.80**	.64**	.72**	.46*	.39*	.41*
12. Agreeableness	.09	.03	.04	.09	.08	.12	.18*	.05	.05	.06	.65**		.61**	.55**	.61**	.53**	.52**	.39*
13. Conscientiousness	-.09	-.09	.01	-.07	.01	-.08	.00	-.14	.00	-.02	.67**	.65**		.72**	.74**	.48**	.29*	.55**
14. Emotional Stability	-.06	-.09	.02	.01	.02	-.05	-.01	-.08	.00	-.02	.59**	.58**	.68**		.72**	.41*	.28	.54**
15. Openness	.02	-.03	-.01	-.04	.08	.04	.12	-.11	-.05	.13	.69**	.64**	.76**	.69**		.46**	.33*	.32*
Zero-shot																		
16. Extraversion	.28**	.02	.16	.14	.14	.32**	.21*	.09	.19*	.23*	.57**	.46**	.50**	.42**	.50**	.64**	.56**	
17. Agreeableness	.23*	.17*	.22*	.08	.17*	.18*	.32**	.14	.26*	.20*	.43**	.58**	.39**	.39**	.41**	.52**	.39*	
18. Conscientiousness	.12	-.09	.11	.05	.05	.12	.04	.05	.18*	.04	.40**	.56**	.63**	.57**	.50**	.51**	.51**	
19. Emotional Stability	.11	.03	.12	.14	.11	.09	.06	.07	.10	.08	.28*	.42**	.37**	.63**	.36**	.36**	.35**	
20. Openness	.13	.04	.07	.05	.14	.14	.16	.06	.10	.21*	.50**	.55**	.57**	.52**	.68**	.59**	.48**	
User Engagement (S)																		
21. Focused Attention	.05	.02	.04	-.01	.25*	.08	.11	-.03	-.07	.11	.20*	.23*	.19*	.09	.18*	.13	.20*	.15
22. Perceived Usability	-.13	-.30**	-.43**	-.27*	-.20*	-.07	-.11	-.32**	-.36**	-.14	-.01	.04	.03	-.01	.08	-.19*	-.17*	.00
23. Aesthetic Appeal	.18*	.09	.25*	.25*	.31**	.09	.09	.11	.14	.11	.31**	.36**	.29**	.22*	.27*	.29**	.32**	.22*
24. Reward Factor	.32**	.31**	.46**	.34**	.54**	.20*	.31**	.23*	.27*	.38**	.19*	.18*	.15	.12	.15	.20*	.23*	.16
User Engagement (Ch)																		
25. Focused Attention	.09	.08	.19*	.13	.17*	.10	.17*	.06	.12	.14	.22*	.11	.16	.08	.05	.14	.12	.05
26. Perceived Usability	-.33**	-.23*	-.39**	-.29**	-.22*	-.28*	-.20*	-.28*	-.37**	-.18*	-.10	-.04	-.08	-.04	-.04	-.25*	-.24*	-.15
27. Aesthetic Appeal	.21*	.16	.28*	.15	.18*	.12	.24*	.12	.22*	.12	.29**	.27*	.16	.06	.17*	.22*	.28**	.10
28. Reward Factor	.28*	.23*	.39**	.34**	.33**	.18*	.28*	.24*	.25*	.29**	.17*	.13	.14	.07	.11	.23*	.21*	.11
29. Difficulty	-.26*	-.05	-.18	-.14	-.05	-.25*	-.03	-.04	-.07	-.04	-.13	-.06	-.13	-.16	-.14	-.24*	-.12	-.17
30. Validity	.28*	.19*	.38**	.17	.18	.13	.19*	.09	.24*	.05	.04	.05	.03	-.03	.05	.14	.18*	.01
31. Affect	.21*	.19*	.34**	.26*	.22*	.12	.16	.16	.19*	.10	.15	.12	.08	.11	.06	.17	.31**	.12
32. Concentration	.28*	.20*	.35**	.27*	.28*	.21*	.26*	.30*	.34**	.23*	.00	-.02	.02	.05	-.04	.14	.19*	.12
33. Duration (mins)	-.15	-.01	.05	-.06	.08	-.19*	-.01	-.03	.04	-.02	.12	.01	.14	.09	.12	.01	.11	.02
34. Word Count	-.17	-.02	.09	-.05	.00	-.17*	.07	-.04	.11	.03	.51**	.46**	.64**	.56**	.60**	.28**	.29**	.31**
Fairness Perceptions																		
35. Method Integrity	.24*	.14	.29**	.24*	.16	.14	.16	.15	.12	.07	.09	.15	.05	-.01	.07	.09	.10	.05
36. Attitude to ML	.15	.12	.27*	.24*	.07	.07	.10	.16	.15	.03	.07	.11	.08	-.03	.06	.08	.06	.02
37. Chance to Perform	.29**	.15	.32**	.28*	.25*	.18*	.21*	.16	.21*	.19*	.12	.04	.12	-.05	.02	.15	.10	.05
38. Propriety of Qs	.12	.18*	.14	.11	.20*	.09	.23*	.06	.10	.22*	.11	.09	.14	.14	.09	.09	.10	

Notes. Correlations for the combined data ($N = 135$) are shown below the diagonal and correlations for the honest condition ($N = 48$) are shown above the diagonal. * $p < .05$ and ** $p < .001$.

Table 15 continued

Variables	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	
Baseline																					
1. Extraversion	.24	-.02	.17	-.06	.12	.32*	.09	-.35*	.32*	.30*	-.20	.22	.27	.37*	-.34*	-.42*	.25	.12	.17	.23	
2. Agreeableness	.18	.22	.14	-.47**	.22	.42*	.22	-.22	.25	.38*	.06	.09	.21	.29	-.14	-.09	.29*	.27	.18	.24	
3. Conscientiousness	.39*	.18	-.03	-.43*	.16	.34*	.03	-.55**	.29*	.31*	-.35*	.32*	.33*	.41*	-.26	-.18	.30*	.26	.33*	.27	
4. Emotional Stability	.14	-.02	.05	-.22	.32*	.45*	.14	-.21	.18	.53**	.05	.16	.25	.36*	-.22	-.28	.44*	.46*	.44*	.43*	
5. Openness	.30*	.17	.27	-.09	.38*	.55**	.14	-.20	.25	.44*	.00	.11	.28	.35*	-.05	-.05	.17	.04	.25	.36*	
Post-Condition																					
6. Extraversion	.23	.03	.27	-.01	.10	.29*	.10	-.31*	.24	.23	-.19	.14	.21	.34*	-.24	-.40*	.19	.07	.09	.21	
7. Agreeableness	.23	.16	.32*	-.10	.30*	.49**	.30*	-.16	.47**	.44*	.03	.26	.23	.39*	-.11	-.09	.31*	.20	.27	.34*	
8. Conscientiousness	.12	.02	.08	-.44*	.35*	.46*	.17	-.27	.25	.53**	.03	.22	.24	.41*	-.10	-.27	.44*	.41*	.47**	.31*	
9. Emotional Stability	.40*	.12	-.20	-.38*	.26	.19	-.15	-.51**	.36*	.17	-.36*	.32*	.38*	.28	-.20	-.19	.22	.21	.33*	.15	
10. Openness	.16	.12	.23	-.04	.24	.48**	.14	-.13	.21	.42*	-.01	.08	.17	.38*	-.02	-.08	.11	.04	.20	.42*	
Embedding																					
11. Extraversion	.23	.48**	.35*	.12	.25	.06	.24	.13	.18	.03	-.11	-.09	.07	-.23	.04	.56**	-.05	-.09	-.05	.23	
12. Agreeableness	.37*	.54**	.40*	-.05	.41*	.26	.22	-.03	.32*	.16	-.11	-.04	.15	-.08	-.09	.40*	.12	.05	.04	.09	
13. Conscientiousness	.31*	.45*	.27	.04	.22	.11	.09	.06	.07	-.03	-.10	-.21	-.01	-.21	.10	.74**	-.21	-.15	-.05	.06	
14. Emotional Stability	.61**	.48**	.16	.02	.13	.09	.04	-.01	-.04	-.10	-.15	-.22	-.02	-.15	-.05	.53**	-.24	-.25	-.20	.25	
15. Openness	.37*	.59**	.31*	.07	.28	.16	.12	-.04	.19	.03	-.12	-.02	.03	-.20	.04	.61**	-.07	-.06	-.06	.17	
Zero-shot																					
16. Extraversion	.44*	.59**	.27	-.24	.29*	.20	.05	-.34*	.21	.18	-.38*	.06	.31*	.17	.06	.30*	-.03	-.07	.02	.13	
17. Agreeableness	.43*	.49**	.32*	-.39*	.30*	.23	.21	-.19	.27	.16	-.15	.16	.24	-.01	.06	.14	.07	-.04	-.01	.02	
18. Conscientiousness	.51**	.33*	.23	-.19	.17	.18	.03	-.22	.06	.01	-.22	-.17	.11	.02	-.04	.26	-.09	-.12	-.08	.18	
19. Emotional Stability	.59**	.27	-.24	.29*	.20	.05	-.34*	.21	.18	-.38*	.06	.31*	.17	.06	.30*	-.03	-.07	.02	.13		
20. Openness	.42**	.15	-.28	.23	.24	.06	-.36*	.17	.08	-.26	-.03	.28	.09	-.25	.10	-.03	-.10	-.04	.12		
User Engagement (S)																					
21. Focused Attention	.09	.10		.16	.31*	.43*	.64**	.13	.19	.32*	.03	-.10	.16	.16	.09	.12	.26	.15	.13	.10	
22. Perceived Usability	-.15	-.14	.06		-.13	-.18	.12	.47**	-.01	-.17	.10	-.09	-.20	-.40*	-.14	-.09	.00	-.01	-.03	.08	
23. Aesthetic Appeal	.27*	.26*	.23*	-.22*		.75**	.18	-.14	.61**	.64**	-.13	.31*	.42*	.25	-.04	.04	.53**	.46*	.57**	.16	
24. Reward Factor	.19*	.11	.33**	-.29**		.68**	.31*	-.21	.46**	.78**	-.10	.07	.38*	.46*	-.09	-.04	.54**	.49**	.56**	.23	
User Engagement (Ch)																					
25. Focused Attention	.10	.06	.51**	-.05	.30**	.33**		-.06	.27	.53**	-.08	.18	.45*	.39*	.13	.06	.44*	.43*	.18	.22	
26. Perceived Usability	-.19*	-.17*	.07	.40**	-.23*	-.32**	-.25*		-.19	-.33*	.42*	-.40*	-.60**	-.59**	-.06	.04	-.16	-.26	-.20	-.21	
27. Aesthetic Appeal	.12	.15	.10	-.05	.61**	.48**	.39**	-.35**		.61**	-.20	.48*	.41*	.39*	-.18	-.05	.62**	.48**	.50**	.25	
28. Reward Factor	.14	.14	.19*	-.24*	.59**	.69**	.59**	-.52**		.68**	-.21*	-.31**	-.30	-.40*	-.39*	-.04	.02	-.09	-.15	-.16	
29. Difficulty	-.19*	-.18	-.11	.14	-.24*	-.18*	-.28*	-.48**		-.21*	-.31**	-.30	-.40*	-.39*	-.04	.02	-.09	-.15	-.16		
30. Validity	.01	.12	.14	-.17	.47**	.37**	.39**	-.47**		.57**	.57**	-.35**	.62**	.41*	-.11	-.18	.52**	.54**	.44*	.32*	
31. Affect	.19*	.19*	.16	-.19*	.42**	.39**	.52**	-.62**		.51**	.70**	-.53**	.61**	.57**	-.15	-.16	.55**	.56**	.40*	.30*	
32. Concentration	.10	.09	.17	-.33**	.33**	.41**	.42**	-.58**		.36**	.61**	-.33**	.41**	.52**	-.08	-.09	.42*	.43*	.40*	.42*	
33. Duration (mins)	.08	.08	.20*	-.08	.12	.15	.09	-.04	.04	.03	-.01	.11	.00	.07	.46*	-.27	-.13	-.20	-.15		
34. Word Count	.17	.44**	.07	-.12	.13	.08	.10	-.14	.11	.07	-.04	.06	.04	.05	.48**	-.36*	-.24	-.22	.00		
Fairness Perceptions																					
35. Method Integrity	.01	.03	.22*	-.10	.46**	.48**	.42**	-.30**		.57**	.62**	-.24*	.65**	.46**	.32**	.08	.00	.91**	.72**	.31*	
36. Attitude to ML	.02	.01	.16	-.14	.46**	.45**	.41**	-.36**		.50**	.61**	-.19*	.64**	.41**	.34**	.13	.08	.90**	.69**	.28	
37. Chance to Perform	.02	.10	.08	-.15	.52**	.50**	.38**	-.35**		.53**	.66**	-.31**	.57**	.47**	.39**	.00	-.05	.66**	.64**	.41*	
38. Propriety of Qs	.00	.08	.13	-.04	.15	.26*	.24*	-.29**		.22*	.35**	-.27*	.40**	.35**	.31**	.11	.16	.38**	.34**	.32**	

Notes. Correlations for the combined data ($N = 135$) are shown below the diagonal and correlations for the honest condition ($N = 48$) are shown above the diagonal. * $p < .05$ and ** $p < .001$.

Table 16. Correlations for Study Variables for Part III for Induced and Motivated Faking conditions.

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Baseline																		
1. Extraversion		.33*	.63**	.54**	.66**	.77**	.57**	.42*	.57**	.74**	.14	.06	-.07	-.04	.07	.41*	.46*	.21
2. Agreeableness	.50**		.33*	.44*	.46*	.28	.64**	.34*	.25	.42*	-.21	-.18	-.09	-.20	-.10	-.02	.00	-.14
3. Conscientiousness	.29*	.43*		.54**	.54**	.26	.31	.35*	.80**	.42*	.02	.04	.07	.07	-.04	.13	.49*	.19
4. Emotional Stability	.52**	.41*	.49**		.55**	.56**	.43*	.86**	.43*	.42*	.06	.09	-.04	-.01	-.04	.33*	.24	.03
5. Openness	.35*	.43*	.52**	.41*		.57**	.65**	.54**	.53**	.77**	.10	-.01	-.09	-.19	-.09	.15	.23	.07
Post-Condition																		
6. Extraversion	.71**	.30*	-.13	.16	.00		.64**	.56**	.29	.67**	.06	.17	-.10	-.02	.09	.43*	.21	.16
7. Agreeableness	.40*	.77**	.16	.21	.26	.44*		.39*	.27	.65**	.05	-.03	-.04	-.17	.00	.29	.07	.03
8. Conscientiousness	.33*	.37*	.32*	.76**	.09	.24	.40*		.42*	.44*	.07	.06	-.11	-.04	.02	.33*	.21	.04
9. Emotional Stability	.16	.27	.62**	.34*	.18	.05	.35*	.53**		.45*	.05	-.03	-.06	.03	-.05	.14	.48*	.19
10. Openness	.26	.33*	.12	.27	.53**	.29*	.52**	.30*	.30*		.08	-.06	-.13	-.12	.07	.33*	.19	.06
Embedding																		
11. Extraversion	.12	.03	.11	.18	.25	.02	.09	.01	.04	.15		.55**	.53**	.46*	.57**	.64**	.44*	.39*
12. Agreeableness	.04	.04	.00	.11	.04	.07	.16	.07	.06	.14	.55**		.56**	.50*	.56**	.28	.45*	.57**
13. Conscientiousness	-.01	-.11	.03	-.04	.12	.07	.04	-.10	.07	.15	.61**	.73**		.66**	.68**	.42*	.33*	.66**
14. Emotional Stability	-.03	-.08	.00	.11	.01	-.12	.04	.01	.01	-.02	.60**	.66**	.66**		.60**	.46*	.49*	.54**
15. Openness	.01	-.12	-.08	.02	.09	-.03	.03	-.13	-.12	.15	.69**	.70**	.82**	.71**		.45*	.35*	.68**
Zero-shot																		
16. Extraversion	.22	.02	.08	.14	.13	.26	.07	.04	.04	.12	.63**	.50**	.55**	.42*	.57**		.41*	.44*
17. Agreeableness	.01	.20	-.03	.04	.06	.08	.35*	.09	.01	.17	.46*	.76**	.53**	.48**	.54**		.43*	.59**
18. Conscientiousness	.15	-.12	.05	.16	.07	.15	.06	.17	.12	.11	.41*	.73**	.69**	.66**	.59**	.48**	.57**	
19. Emotional Stability	.12	.01	.18	.33*	.23	.01	.01	.14	.06	.08	.42*	.43*	.41*	.65**	.36*	.28	.31*	.59**
20. Openness	.18	-.06	-.12	.08	.10	.21	.12	.05	-.03	.26	.52**	.71**	.67**	.55**	.75**	.52**	.47**	.65**
User Engagement (S)																		
21. Focused Attention	-.20	-.24	.01	-.27	.25	-.20	-.22	-.37*	-.13	-.03	.03	.00	.17	.02	.14	-.01	.02	.08
22. Perceived Usability	-.12	-.26	-.40*	-.17	-.31*	-.08	-.23	-.16	-.33*	-.24	-.11	.06	.10	.00	.13	-.10	-.01	.11
23. Aesthetic Appeal	.21	.06	.30*	.12	.37*	-.07	-.17	-.18	-.12	-.11	.30*	.23	.29*	.25	.24	.19	.31*	.23
24. Reward Factor	.21	.29*	.57**	.15	.59**	-.13	.06	-.12	.14	.14	.16	.03	.13	.16	.10	.04	.11	.13
User Engagement (Ch)																		
25. Focused Attention	-.15	-.14	.02	-.14	.02	-.10	-.10	-.22	.01	-.12	.08	-.13	.10	-.01	-.10	-.02	-.20	-.10
26. Perceived Usability	-.24	-.32*	-.23	-.33*	-.23	-.26	-.30*	-.30*	-.21	-.22	-.19	.03	-.09	.08	.09	-.05	-.12	.05
27. Aesthetic Appeal	.07	.21	.28	.03	.16	-.11	.04	-.09	.00	-.20	.25	.18	.20	.16	.08	.01	.24	.04
28. Reward Factor	.12	.25	.34*	.18	.36*	-.01	.20	.00	.07	.09	.18	.01	.18	.14	.07	.05	.07	.03
29. Difficulty	-.32*	.03	.03	-.20	-.10	-.32*	.06	-.02	.19	-.06	-.14	.06	-.05	-.06	.00	-.03	-.03	.03
30. Validity	.29	.23	.50**	.06	.30	.05	.14	-.12	.27	-.09	.14	.09	.16	.05	.07	.10	.09	.08
31. Affect	-.04	.09	.24	.15	.13	-.11	.05	.05	-.01	-.02	.24	.13	.08	.17	.08	-.06	.27	.04
32. Concentration	.12	.07	.20	.14	.11	.06	.09	.22	.24	-.04	.05	-.12	.02	.04	-.04	-.01	.10	-.01
33. Duration (mins)	-.06	.15	.13	.01	.19	-.22	-.02	-.05	.13	.05	-.05	-.06	.04	.12	.08	-.18	.01	-.04
34. Word Count	-.08	.03	.10	.02	.01	-.05	.18	.04	.23	.14	.45*	.53**	.63**	.60**	.61**	.29	.38*	.33*
Fairness Perceptions																		
35. Method Integrity	.17	.11	.25	.00	.19	.00	.01	-.15	.01	-.17	.02	.03	.15	.07	.07	-.05	-.09	.05
36. Attitude to ML	.18	.13	.29*	.00	.15	.05	.07	-.13	.06	-.08	.02	-.03	.17	.01	.05	.01	-.07	.03
37. Chance to Perform	.37*	.12	.22	.10	.27	.19	.02	-.17	-.08	.00	.10	-.06	.10	.00	-.04	-.04	-.12	-.03
38. Propriety of Qs	.00	.03	.01	-.10	.07	.02	.09	-.12	.01	-.01	.02	.14	.14	.13	.17	.11	.24	.09

Notes. Correlations for the induced faking condition ($N = 47$) are shown below the diagonal and correlations for the motivated faking condition ($N = 40$) are shown above the diagonal. * $p < .05$ and ** $p < .001$.

Table 16 continued

Variables	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
Baseline																				
1. Extraversion	-.13	.34*	.20	-.26	.27	.52**	.38*	-.43*	.23	.42*	-.27	.39*	.45*	.32	-.01	.04	.30	.16	.39*	.12
2. Agreeableness	-.18	-.09	.20	-.12	-.09	.18	.18	-.14	-.03	.02	-.31	.28	.34	.26	-.17	-.06	-.03	-.12	.16	.33*
3. Conscientiousness	-.12	.19	.19	-.47*	.34*	.56**	.47*	-.44*	.30	.50**	-.24	.39*	.48*	.50*	.25	.31	.33*	.30	.41*	.16
4. Emotional Stability	-.10	.11	.27	-.48*	.36*	.50**	.43*	-.34*	.25	.32*	-.32	.42*	.42*	.32	.04	.11	.31	.29	.30	.03
5. Openness	-.29	.15	.25	-.24	.16	.51**	.34*	-.24	.12	.17	-.06	.13	.23	.36*	.10	.06	.11	.02	.24	.15
Post-Condition																				
6. Extraversion	-.02	.22	.13	-.12	.26	.45*	.27	-.26	.18	.29	-.26	.24	.26	.15	-.12	-.09	.22	.11	.27	.03
7. Agreeableness	-.17	.09	.23	-.02	.01	.28	.28	-.18	.12	.17	-.25	.20	.27	.23	.05	.00	.14	-.01	.30	.23
8. Conscientiousness	-.11	.10	.24	-.38*	.16	.41*	.22	-.27	.22	.20	-.16	.23	.20	.25	.08	.09	.23	.21	.20	.04
9. Emotional Stability	-.21	.16	.08	-.41*	.26	.47*	.41*	-.38*	.25	.46*	-.01	.11	.18	.54*	.16	.28	.13	.18	.30	.13
10. Openness	-.04	.27	.09	-.19	.13	.47*	.36*	-.23	.26	.29	-.06	.20	.16	.22	-.14	.01	.25	.12	.32*	.20
Embedding																				
11. Extraversion	.16	.49*	.24	-.08	.39*	.34*	.36*	-.35*	.45*	.35*	-.24	.17	.19	.42*	.48*	.52**	.36*	.36*	.32*	.06
12. Agreeableness	.48*	.28	.33*	.18	.46*	.21	.25	-.20	.29	.24	-.21	.24	.14	.26	.23	.40*	.37*	.42*	.13	-.01
13. Conscientiousness	.39*	.55**	.12	-.08	.38*	.18	.29	-.28	.20	.31	-.34	.22	.26	.38*	.33*	.54**	.25	.29	.32*	.20
14. Emotional Stability	.64**	.55**	.09	-.08	.34*	.07	.24	-.31	.11	.25	-.37*	.28	.31	.51*	.26	.56**	.25	.33*	.12	-.02
15. Openness	.36*	.70**	.08	.04	.29	.11	.11	-.25	.24	.28	-.46*	.25	.16	.30	.25	.53**	.28	.27	.19	.05
Zero-shot																				
16. Extraversion	.29	.63**	.08	-.26	.40*	.34*	.37*	-.40*	.42*	.47*	-.36*	.29	.31	.27	.18	.27	.40*	.34*	.44*	.01
17. Agreeableness	.20	.41*	.18	-.06	.34*	.28	.27	-.52**	.32*	.40*	-.24	.39*	.51*	.62**	.27	.38*	.38*	.34*	.41*	.01
18. Conscientiousness	.35*	.59**	.06	.08	.29	.12	.21	-.36*	.22	.32*	-.40*	.12	.24	.45*	.16	.34*	.27	.19	.25	.03
19. Emotional Stability	.26	.12	.01	.25	-.05	.26	-.10	.11	.23	.15	.20	.21	.26	-.05	.15	.23	.25	.05	-.08	
20. Openness	.33*	.11	-.10	.42*	.24	.27	-.35*	.31	.39*	-.45*	.37*	.37*	.47*	.15	.30	.29	.23	.41*	.01	
User Engagement (S)																				
21. Focused Attention	-.05	-.07		-.14	.17	.28	.42*	-.13	-.01	.10	-.24	.32	.31	.44*	.18	.08	.06	.03	-.03	.04
22. Perceived Usability	-.13	.06	.09		-.45*	-.62**	-.40*	.50*	-.22	-.52**	.23	-.29	-.30	-.22	-.21	-.26	-.31	-.39*	-.28	-.12
23. Aesthetic Appeal	.33*	.11	.17	-.20		.64**	.62**	-.49*	.62**	.69**	-.46*	.62**	.51*	.38*	.26	.34*	.54**	.67**	.57**	.02
24. Reward Factor	.29	-.10	.26	-.19		.62**		.56**	-.62**	.50*	.72**	-.54*	.58**	.54*	.49*	.27	.29	.51**	.52**	.57**
User Engagement (Ch)																				
25. Focused Attention	-.01	-.16	.42*	.08	.16	.13		-.46*	.51**	.69**	-.45*	.63**	.75**	.73**	.14	.24	.44*	.41*	.60**	.24
26. Perceived Usability	-.07	.03	.17	.21	-.17	-.25	-.30*		-.51**	-.71**	-.62**	-.71**	-.79**	-.61**	-.17	-.37*	-.62**	-.57**	-.61**	-.37*
27. Aesthetic Appeal	.05	-.01	.09	.06		.59**	.49**	.41*		.69**	-.25	.56**	.44*	.32	.10	.21	.67**	.65**	.67**	.18
28. Reward Factor	.14	-.06	.13	-.05	.45*	.59**	.57**	-.56**	.76**	.73**	-.49*	.62**	.73**	.59**	.13	.25	.64**	.61**	.79**	.35*
29. Difficulty	-.13	-.04	-.14	.10	-.26	-.13	-.41*	.43*	-.22	-.35*	-.71**	-.68**	-.30	-.36*	-.40*	-.49*	-.40*	-.53*	-.42*	
30. Validity	-.09	.04	.22	-.19		.57**	.62**	.46*	-.42*	.68**	.69**	-.17	.81**	.33	.16	.33	.74**	.73**	.74**	.34
31. Affect	.08	.01	.01	-.08	.42*	.38*	.43*	-.51**	.72**	.72**	-.53**	.49*	.58**	.16	.18	.58**	.47*	.82**	.43*	
32. Concentration	-.03	-.11	-.02	-.34*	.41*	.29	.23	-.57**	.34*	.48*	-.30	.50**	.44*	.13	.32	.20	.18	.42*	.13	
33. Duration (mins)	.01	.01	.31*	.04	.16	.25	.01	.03	.20	.07	.21	.24	.02	.01	.01	.65**	.19	.26	.14	.19
34. Word Count	.23	.49**	.05	-.09	.10	.02	.04	-.19	.18	.10	.09	.13	.13	.01	.41*	.23	.39*	.12	.24	
Fairness Perceptions																				
35. Method Integrity	-.09	.01	.32*	-.05	.35*	.41*	.37*	-.20	.44*	.48**	-.21	.76**	.30	.29	.28	.11	.91**	.68**	.25	
36. Attitude to ML	-.02	-.02	.28	-.11	.31*	.36*	.39*	-.31*	.39*	.49**	-.16	.71**	.22	.31*	.27	.16	.89**	.59**	.21	
37. Chance to Perform	.05	.02	.13	-.18	.42*	.37*	.36*	-.29*	.41*	.55**	-.33*	.61**	.26	.34*	.07	-.06	.61**	.63**	.38*	
38. Propriety of Qs	-.08	.16	.24	-.13	.23	.21	.27	-.32*	.22	.24	-.27	.53**	.34*	.31*	.24	.21	.56**	.53**	.16	

Notes. Correlations for the induced faking condition ($N = 47$) are shown below the diagonal and correlations for the motivated faking condition ($N = 40$) are shown above the diagonal. * $p < .05$ and ** $p < .001$.

Table 17. Means and Standard Deviations of Word Count by Interview Question

	Combined (N = 135)	Honest (N = 48)	Induced Faking (N = 47)	Motivated Faking (N = 40)
<i>Extraversion</i>				
Question 5	85.92 (49.73)	79.10 (33.85)	93.66 (56.07)	85.00 (57.27)
Question 7	75.87 (37.07)	70.50 (29.21)	85.38 (42.02)	71.13 (38.00)
Question 11	64.41 (29.61)	61.83 (23.49)	72.09 (40.33)	58.48 (17.72)
<i>Agreeableness</i>				
Question 4	76.73 (43.76)	78.90 (54.19)	82.11 (41.09)	67.83 (30.44)
Question 10	71.20 (35.04)	69.98 (25.27)	78.28 (40.09)	64.35 (37.99)
Question 13	63.21 (28.22)	64.35 (23.13)	67.47 (37.89)	56.85 (18.25)
<i>Conscientiousness</i>				
Question 2	80.81 (44.41)	81.38 (44.50)	81.70 (47.69)	79.10 (41.28)
Question 9	67.44 (31.57)	65.46 (30.24)	72.79 (37.69)	63.53 (24.35)
Question 14	61.11 (30.78)	57.02 (21.26)	66.32 (43.41)	59.80 (20.51)
<i>Emotional Stability</i>				
Question 1	102.78 (59.98)	110.71 (73.93)	103.00 (52.84)	93.00 (47.99)
Question 8	68.19 (30.33)	63.94 (24.95)	77.15 (38.75)	62.65 (21.95)
Question 15	66.19 (32.67)	63.19 (23.38)	70.55 (43.80)	64.58 (26.40)
<i>Openness</i>				
Question 3	76.39 (42.94)	72.40 (36.58)	85.36 (56.47)	70.65 (28.49)
Question 6	76.04 (39.44)	74.65 (34.22)	82.70 (37.90)	69.90 (46.29)
Question 12	60.98 (28.79)	57.55 (20.78)	69.11 (40.64)	55.45 (15.60)
Combined	1095.37 (459.14)	1065.92 (394.30)	1187.66 (565.50)	1022.28 (378.10)

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.

Table 18. Means and Standard Deviations of Personality Trait for Part III

	Combined (N = 135)		Honest (N = 48)		Induced Faking (N = 47)		Motivated Faking (N = 40)	
	M	SD	M	SD	M	SD	M	SD
Baseline								
Extraversion	3.34	0.80	3.40	0.90	3.33	0.77	3.29	0.70
Agreeableness	4.06	0.57	4.04	0.57	4.07	0.62	4.06	0.52
Conscientiousness	4.15	0.65	4.25	0.53	4.14	0.61	4.04	0.80
Neuroticism	3.63	0.74	3.67	0.70	3.62	0.81	3.60	0.73
Openness	3.64	0.75	3.67	0.80	3.65	0.68	3.61	0.78
Post-Condition								
Extraversion	3.23	0.83	3.26	0.90	3.23	0.75	3.21	0.86
Agreeableness	4.09	0.68	3.97	0.70	4.28	0.65	4.01	0.67
Conscientiousness	3.69	0.93	3.63	0.87	3.68	1.01	3.78	0.93
Neuroticism	4.16	0.59	4.11	0.55	4.16	0.50	4.22	0.73
Openness	3.93	0.58	3.90	0.64	3.97	0.51	3.91	0.60
Embedding								
Extraversion	3.41	0.17	3.39	0.17	3.46	0.17	3.39	0.15
Agreeableness	3.49	0.22	3.46	0.24	3.53	0.24	3.47	0.18
Conscientiousness	3.43	0.21	3.40	0.21	3.47	0.22	3.41	0.20
Neuroticism	3.33	0.22	3.31	0.25	3.36	0.22	3.32	0.17
Openness	3.42	0.24	3.39	0.25	3.49	0.25	3.38	0.20
Few-shot								
Extraversion	3.55	0.80	3.38	0.79	3.67	0.80	3.63	0.80
Agreeableness	3.61	0.72	3.40	0.77	3.75	0.65	3.70	0.68
Conscientiousness	3.51	0.75	3.35	0.71	3.55	0.83	3.64	0.69
Neuroticism	3.75	0.66	3.64	0.73	3.79	0.63	3.83	0.60
Openness	3.32	0.74	3.15	0.75	3.47	0.72	3.34	0.72

Table 19. Correlations Between Machine-derived and Self-reported Personality Trait Scores for Part III.

	Honest (N = 48)				Induced Faking (N = 47)				Motivated Faking (N = 40)			
	Embedding		Few Shot		Embedding		Few Shot		Embedding		Few Shot	
	r	p	r	p	r	p	r	p	r	p	r	P
Baseline												
Extraversion	-.04	.778	.29	.049	.12	.421	.22	.144	.14	.399	.41	.009
Agreeableness	.14	.335	.26	.073	.04	.804	.20	.176	-.18	.257	.00	.987
Conscientiousness	-.06	.663	.17	.251	.03	.839	.05	.740	.07	.651	.19	.236
Emotional Stability	-.07	.649	.14	.345	.11	.463	.33	.022	-.01	.961	-.10	.521
Openness	.18	.218	.17	.262	.09	.554	.10	.490	-.09	.599	.15	.342
Post-Condition												
Extraversion	-.06	.674	.30	.036	.02	.885	.26	.082	.06	.732	.43	.006
Agreeableness	.27	.064	.42	.003	.16	.296	.35	.017	-.03	.836	.07	.671
Conscientiousness	-.01	.959	.21	.154	.07	.654	.12	.412	-.06	.694	.19	.245
Emotional Stability	-.19	.192	.12	.403	.01	.970	.14	.350	-.04	.825	-.11	.506
Openness	.15	.320	.12	.416	.15	.325	.26	.081	.07	.658	.27	.094

Notes. r = Pearson's correlations between self-reported and machine-derived score (embedding or few shot). Bolded values indicate significance at the $p < .05$ level.

Table 20. Paired *t*-Tests User Engagement Scores Between the Scale and the AI Chatbot

	BFI		Chatbot		<i>t</i> (134)	CI LL	CI UL	<i>p</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>				
Focused attention	3.08	0.87	3.27	0.97	-2.36	-0.34	-0.03	.019
Perceived usability	1.51	0.74	2.17	0.96	-8.03	-0.82	-0.50	.000
Aesthetic appeal	3.18	0.96	3.29	0.97	-1.58	-0.26	0.03	.117
Reward factor	3.73	0.73	3.70	0.95	0.54	-0.08	0.15	.588

Notes: Paired *t*-tests. Bolded values indicate significance at the *p* <.05 level.

Table 21. Means, Standard Deviations, and Comparisons of User Engagement and Perceptions Across Conditions.

	Combined			Honest		Induced Faking		Motivated Faking		<i>F</i>	<i>p</i>
	<i>M</i>	<i>SD</i>	Density	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Difficulty Score	2.31	0.88		2.25	0.91	2.40	0.85	2.28	0.90	0.33	.723
Validity Score	3.87	0.71		3.82	0.74	3.80	0.75	4.03	0.60	1.11	.332
Affect Score	3.91	0.97		3.94	0.95	3.79	0.97	4.04	1.02	0.60	.549
Ease of concentration	4.25	0.78		4.19	0.92	4.30	0.68	4.26	0.72	0.23	.792
Focused attention	3.27	0.97		3.18	0.99	3.32	0.89	3.33	1.06	0.32	.725
Perceived usability	2.17	0.95		2.13	1.01	2.22	0.95	2.16	0.92	0.12	.888
Aesthetic appeal	3.29	0.97		3.23	1.05	3.38	0.87	3.27	1.00	0.32	.730
Reward factor	3.70	0.95		3.69	0.97	3.71	0.88	3.71	1.01	0.01	.992
Method Integrity	3.49	1.00		3.52	1.01	3.53	1.02	3.43	0.99	0.14	.872
Attitude towards ML	3.44	1.02		3.39	1.10	3.46	0.98	3.49	0.98	0.12	.888
Chance to perform	3.62	1.01		3.56	1.03	3.67	0.89	3.64	1.14	0.15	.858
Propriety of questions	4.29	0.56		4.27	0.55	4.32	0.57	4.28	0.57	0.11	.900

Notes. Combined ($N = 135$) reflects scores from the scores from the combined Honest ($n = 48$), Induced Faking ($n = 47$), and Motivated Faking ($n = 40$) conditions. The *F* and *p* values reflect results from a one-way ANOVA to assess whether scores differ significantly across conditions. For all conditions, they were told feedback on the chatbot would not be included in any application materials, thus group differences are not expected. The density column shows the density distribution of combined responses.

Table 22. Comparisons of Personality Traits by Gender and Age Across Different Measures

	Men (n = 61)		Women (n = 73)		<i>d</i>	CI	Age	
	M	SD	M	SD			<i>p</i>	<i>p</i>
Baseline								
Extraversion	3.48	.77	3.23	.81	-.31	[-0.65, 0.03]	.16	.062
Agreeableness	4.18	.51	3.95	.60	-.41	[-0.76, -0.07]	.14	.110
Conscientiousness	4.27	.58	4.05	.69	-.33	[-0.68, 0.01]	.27	.002
Neuroticism	2.18	.75	2.53	.71	.49	[0.14, 0.83]	-.19	.030
Openness	3.73	.75	3.58	.74	-.21	[-0.55, 0.14]	.13	.150
Post-Condition								
Extraversion	3.33	.82	3.15	.85	-.22	[-0.56, 0.12]	.02	.807
Agreeableness	4.19	.62	4.00	.72	-.29	[-0.63, 0.05]	.07	.433
Conscientiousness	4.25	.52	4.08	.64	-.29	[-0.63, 0.05]	.26	.003
Neuroticism	2.08	.87	2.52	.94	.48	[0.14, 0.83]	-.22	.009
Openness	3.99	.64	3.88	.54	-.20	[-0.54, 0.14]	-.03	.777
Embedding								
Extraversion	3.43	.19	3.40	.15	-.17	[-0.51, 0.17]	.10	.261
Agreeableness	3.50	.25	3.47	.20	-.13	[-0.47, 0.21]	-.02	.793
Conscientiousness	3.43	.23	3.42	.20	-.05	[-0.39, 0.29]	-.06	.513
Neuroticism	2.66	.23	2.68	.21	.07	[-0.27, 0.41]	.06	.505
Openness	3.44	.27	3.40	.20	-.14	[-0.48, 0.20]	-.08	.333
Few-shot								
Extraversion	3.71	.88	3.57	.74	-.19	[-0.53, 0.15]	-.05	.579
Agreeableness	3.77	.77	3.57	.69	-.37	[-0.71, -0.03]	.06	.525
Conscientiousness	3.54	.76	3.46	.73	-.08	[-0.42, 0.26]	.05	.595
Neuroticism	2.12	.60	2.23	.69	.22	[-0.13, 0.56]	.06	.530
Openness	3.41	.73	3.25	.78	-.25	[-0.59, 0.09]	-.02	.781

Notes. *M* = Mean, *SD* = Standard deviation, *d* = Cohen's *d* (effect size for gender differences), CI = 95% Confidence Interval for Cohen's *d*, *p* = Spearman's correlation coefficient, *p* = *p*-value for Spearman's correlation. *M*_{age} = 39 years; Median = 37 years; *SD* = 11 years. Negative *d* values indicate higher scores for women, while positive *d* values indicate higher scores for men. Significant effects for Cohen's *d* are determined by confidence intervals that do not include zero.

Table 23. Comparisons of Personality Traits by Race Across Different Measures for Samples from Part III

	White (n = 95)		Non/Part-White (n = 35)		<i>d</i>	CI
	M	SD	M	SD		
Baseline						
Extraversion	3.39	0.78	3.32	0.80	-.09	[-0.48, 0.29]
Agreeableness	4.09	0.55	3.96	0.64	-.23	[-0.61, 0.16]
Conscientiousness	4.15	0.66	4.26	0.54	.17	[-0.22, 0.56]
Neuroticism	2.34	0.73	2.34	0.73	-.01	[-0.39, 0.38]
Openness	3.63	0.70	3.74	0.79	.15	[-0.24, 0.54]
Post-condition						
Extraversion	3.30	0.81	3.14	0.83	-.19	[-0.58, 0.20]
Agreeableness	4.19	0.61	3.89	0.76	-.48	[-0.87, -0.08]
Conscientiousness	4.20	0.60	4.15	0.52	-.08	[-0.46, 0.31]
Neuroticism	2.27	0.91	2.31	0.93	.05	[-0.34, 0.44]
Openness	3.95	0.54	3.95	0.62	.00	[-0.39, 0.39]
Embedding						
Extraversion	3.42	0.16	3.39	0.19	-.18	[-0.56, 0.21]
Agreeableness	3.50	0.21	3.47	0.27	-.11	[-0.50, 0.28]
Conscientiousness	3.44	0.20	3.41	0.24	-.10	[-0.49, 0.29]
Neuroticism	2.67	0.22	2.67	0.23	.00	[-0.39, 0.39]
Openness	3.42	0.21	3.44	0.30	.06	[-0.33, 0.44]
Few-shot						
Extraversion	3.61	0.77	3.45	0.83	-.20	[-0.59, 0.18]
Agreeableness	3.67	0.66	3.45	0.87	-.31	[-0.70, 0.08]
Conscientiousness	3.54	0.74	3.42	0.81	-.16	[-0.55, 0.23]
Neuroticism	2.28	0.69	2.15	0.60	-.19	[-0.58, 0.19]
Openness	3.34	0.74	3.30	0.76	-.04	[-0.43, 0.34]

Notes. M = Mean, SD = Standard deviation, *d* = Cohen's *d* (effect size for race-based differences), CI = 95% Confidence Interval for Cohen's *d*, ρ = Spearman's correlation coefficient, p = *p*-value for Spearman's correlation. Significant effects for Cohen's *d* are determined by confidence intervals that do not include zero.

Table 24. Mixed Design ANOVA Results and Pairwise Comparisons

	DF	Extraversion		Agreeableness		Conscientiousness		Neuroticism		Openness	
		F	p	F	p	F	p	F	p	F	p
Condition	(2, 132)	.12	.891	2.60	.078	.21	.813	.24	.785	.66	.520
Method	(1, 132)	10.48	<.001	36.02	<.001	62.46	<.001	1.22	.296	40.12	<.001
Condition*Method	(2, 132)	1.63	.168	2.08	.084	2.50	.043	.65	.625	1.31	.266

Notes. This table presents the results of mixed-design ANOVAs examining the effects of Condition (Honest, Induced Faking, and Motivated Faking; between-subjects) and Method (Baseline [BFI], Post-condition [IPIP], chatbot [few shot]; within-subjects) for each of the Big Five domains. Bolded values indicate significance at the $p < .05$ level.

Figure 1. Visualization of SLR Article Filtering

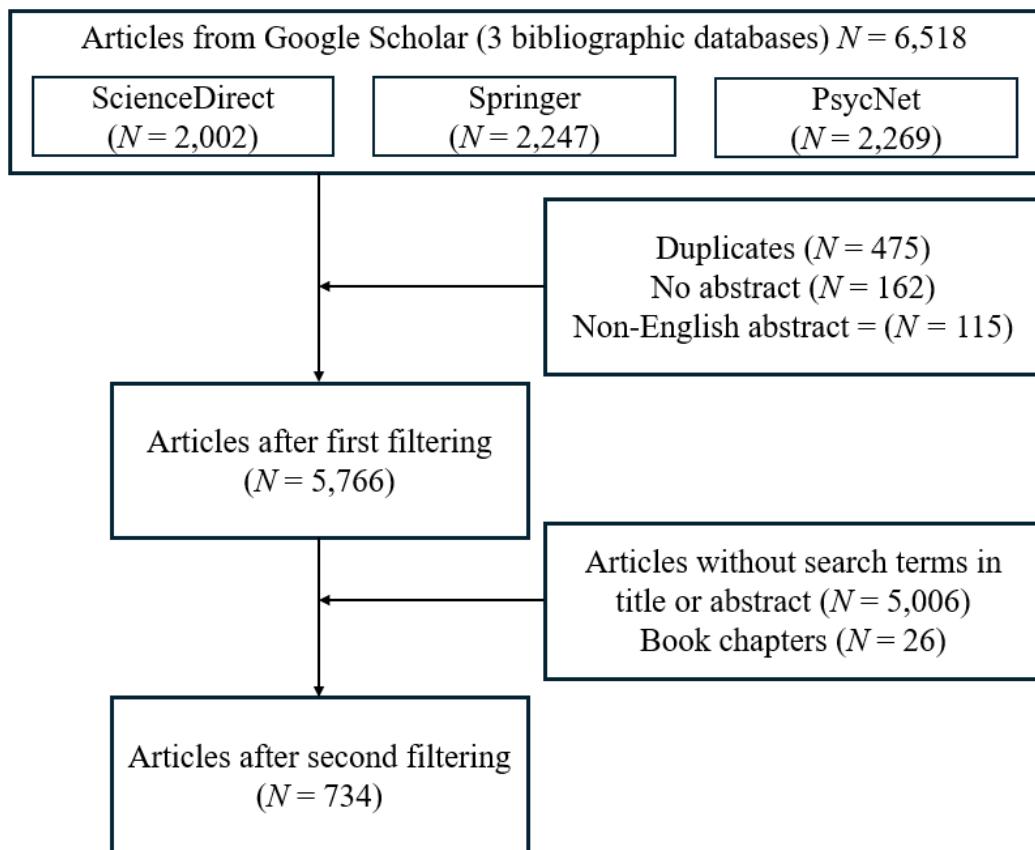
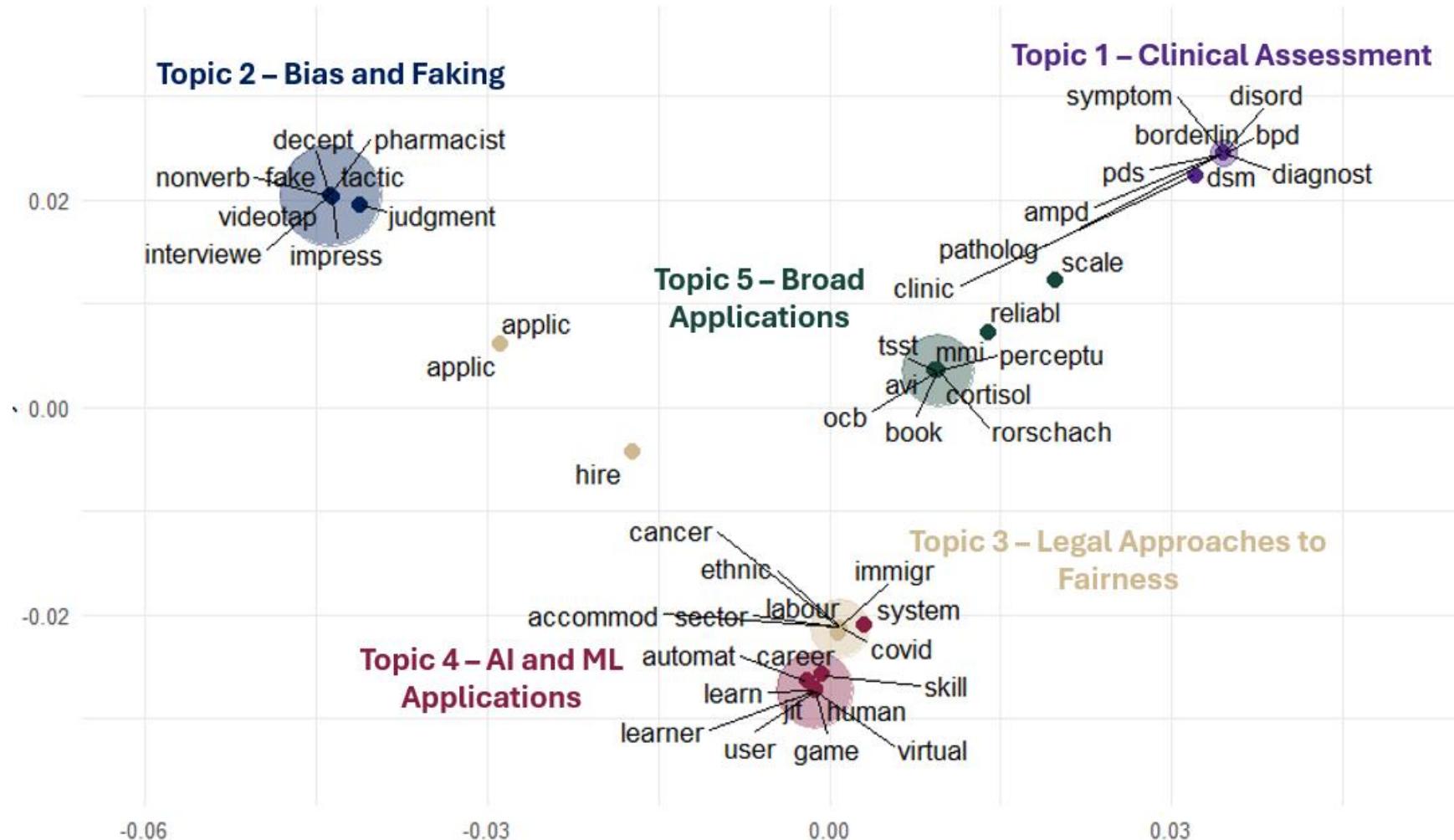


Figure 2. Visualization of Topics and Top Words



Notes. This figure displays a two-dimensional representation of topics derived from topic modeling. Each colored circle represents a topic, its size is based on topic proportion, and its position was calculated based on the Euclidean distance between the topic-word distributions. The featured words are the top 10 words from each topic determined by the score metric.

Figure 3. User Engagement Comparisons Between the AI Chatbot and the Personality Measure

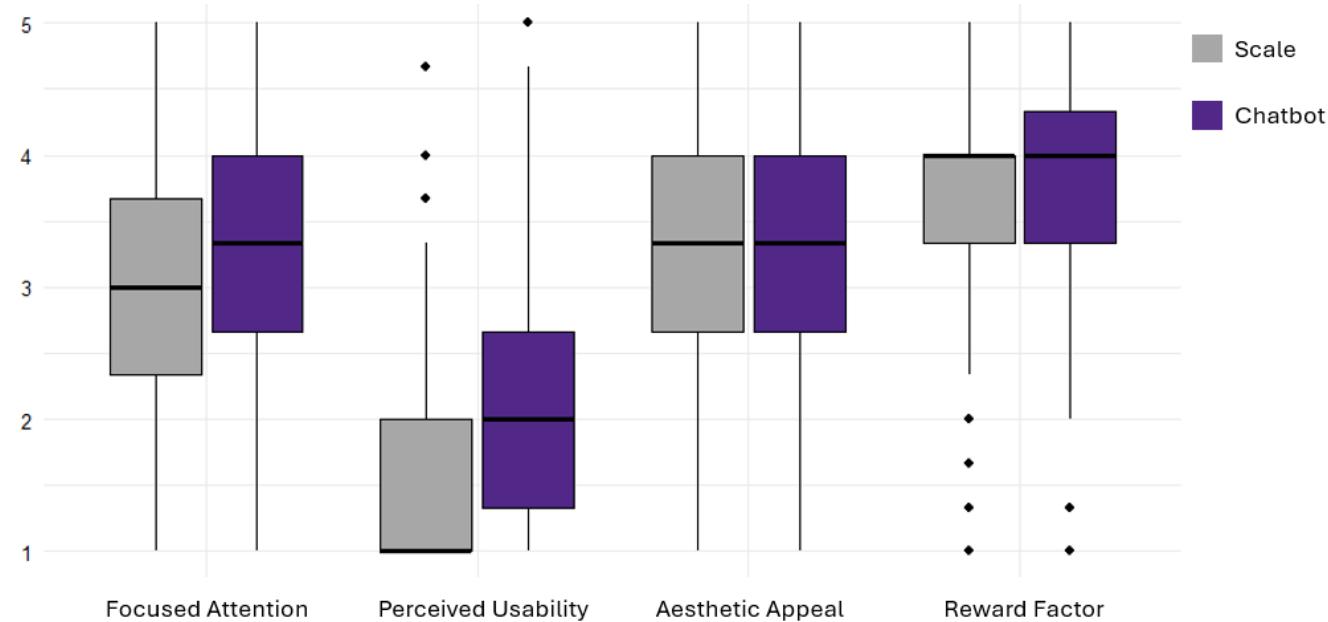
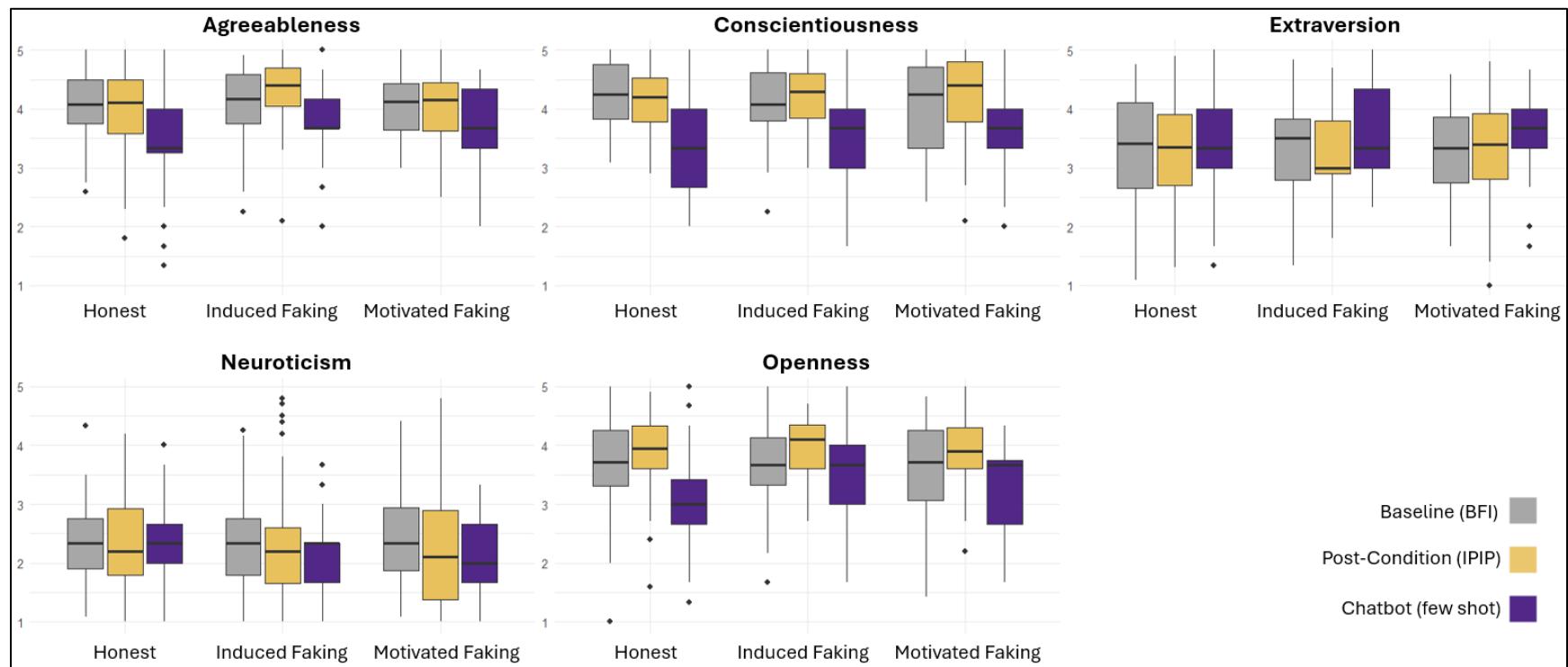


Figure 4. Boxplots of Personality Trait Scores by Condition and Assessment Method



Appendix A - Supplementary Materials for Part I

Example Web Scraper for Bibliographic Database

```
# Set up the environment
import time
import os
from selenium import webdriver
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.support.ui import WebDriverWait as wait
from selenium.common.exceptions import NoSuchElementException
import pandas as pd

# Chrome driver
driver = webdriver.Chrome()

# Create pagination for Google Scholar
#loop click each webpage
links = []
for i in range(0, 100, 10):
    url =
"https://scholar.google.com/scholar?start="+str(i)+"&q="personality+interview"+site:springer.co
m&hl=en&as_sdt=0,10"
    page = driver.get(url)
    driver.implicitly_wait(30)

    for i in driver.find_elements('xpath','//div[@class="gs_r gs_or gs_scl"]/div/h3/a'):
        link = i.get_attribute('href')
        links += [link]
print (len(links))

# List of links
links[0:50] # modify based on number of links gathered

project_info = {'Title': [], 'Author': [], 'Publisher':[],'Pages':[],
'DOI':[], 'Abstract':[]}

for i in links[0:10]: # modify based on links and how much the site allows
    driver.get(i)
    time.sleep(30)

try:
    Title=driver.find_element("xpath",'//*[@id="main"]/section/div/div/div[1]/h1').get_attrib
ute('textContent').strip()
```

```

except NoSuchElementException:
    Title = 0

try:
    Date=driver.find_element("xpath",'//*[@id="article-info-
content"]/div/div[2]/ul/li[1]/p/span[2]/time').get_attribute('textContent').strip()
except NoSuchElementException:
    Date = 0

try:
    Cite=driver.find_element("xpath",'//*[@id="article-info-
content"]/div/div[2]/p[1]').get_attribute('textContent').strip()
except NoSuchElementException:
    Cite = 0

try:
    Publisher=driver.find_element('xpath','//*[@id="main"]/section/div/div/div[2]/a[1]/span')
        .get_attribute('textContent').strip()
except NoSuchElementException:
    Publisher = 0

try:
    Author=driver.find_element("xpath",'//*[@id="main-
content"]/main/div[1]/header/ul').get_attribute('textContent').strip()
except NoSuchElementException:
    try: # alternative xpath
        Author = driver.find_element("xpath",'//*[@id="main-
content"]/main/div[2]/header/ul').get_attribute('textContent').strip()
    except NoSuchElementException:
        try: # alternative xpath
            Author = driver.find_element("xpath", '//*[@id="main-
content"]/main/article/div[1]/header/div/ul').get_attribute('textContent').str
ip()
        except NoSuchElementException:
            Author = 0

try:
    Pages=driver.find_element("xpath",'//*[@id="main"]/section/div/div/div[1]/ul[2]/li[1]').t
ext.strip()
except NoSuchElementException:
    Pages = 0

try:
    DOI = driver.find_element("xpath", '//*[@id="article-info-
content"]/div/div[2]/ul[1]/li[4]/p/span[2]').text.strip()
except NoSuchElementException:

```

```

try: # alternative xpath
    DOI = driver.find_element("xpath", '//*[@id="article-info-
        content"]/div/div/ul[1]/li[5]/p/span[2]').get_attribute('textContent').strip()
except NoSuchElementException:
    try: # alternative xpath
        DOI=driver.find_element("xpath",'//*[@id="chapter-info-
            content"]/div/div[2]/ul[2]/li[1]/p/span[2]').get_attribute('textContent').strip()
    except NoSuchElementException:
        DOI = 0

try:
    Abstract = driver.find_element("xpath",'//*[@id="Abs1-
        content"]').get_attribute('textContent').strip()
except NoSuchElementException:
    Abstract = 0

project_info['Title'].append>Title)
project_info['Author'].append(Author)
project_info['Publisher'].append(Publisher)
project_info['Pages'].append(Pages)
project_info['DOI'].append(DOI)
project_info['Abstract'].append(Abstract)

print(project_info)

```

PRISMA 2020 Checklist

Appendix Table A.1. PRISMA 2020 Checklist

Section and Topic	Checklist item	Location where item is reported and notes
Title	Identify the report as a systematic review	p. 14
Abstract	See the PRISMA 2020 for Abstracts checklist.	NA
Introduction		
<i>Rationale</i>	Describe the rationale for the review in the context of existing knowledge.	Chapter 1
<i>Objectives</i>	Provide an explicit statement of the objective(s) or question(s) the review addresses.	p. 15
Methods		
<i>Eligibility criteria</i>	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	p. 17-18
<i>Information sources</i>	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	p. 16-17
<i>Search strategy</i>	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	p. 16-17
<i>Selection process</i>	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	NA
<i>Data collection process</i>	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	p. 17-18
<i>Data items</i>	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	NA

	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Search terms used as metadata; p. 19
<i>Study risk of bias assessment</i>	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	NA
<i>Effect measures</i>	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	NA
<i>Synthesis methods</i>	<p>Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).</p> <p>Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.</p> <p>Describe any methods used to tabulate or visually display results of individual studies and syntheses.</p> <p>Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.</p> <p>Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).</p> <p>Describe any sensitivity analyses conducted to assess robustness of the synthesized results.</p>	p. 17-18 Text pre-processing described in p.19-20 Table 2, Table 3, Figure 2 p. 16, 18-19 NA NA
<i>Reporting bias assessment</i>	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	NA
<i>Certainty assessment</i>	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	NA
Results		
<i>Study selection</i>	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Figure 1
	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	NA, studies that did not meet filtering criteria were omitted

<i>Study characteristics</i>	Cite each included study and present its characteristics.	NA
<i>Risk of bias in studies</i>	Present assessments of risk of bias for each included study.	NA
<i>Results of individual studies</i>	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	NA
<i>Results of syntheses</i>	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	NA
	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Table 2 and Table 3
	Present results of all investigations of possible causes of heterogeneity among study results.	NA
	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	NA
<i>Reporting biases</i>	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	NA
<i>Certainty of evidence</i>	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	NA
Discussion	Provide a general interpretation of the results in the context of other evidence.	p. 25-31
	Discuss any limitations of the evidence included in the review.	p. 32
	Discuss any limitations of the review processes used.	p. 32-33
	Discuss implications of the results for practice, policy, and future research.	p. 31-32
Other Information		
<i>Registration and protocol</i>	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	NA
	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	NA
	Describe and explain any amendments to information provided at registration or in the protocol.	NA

<i>Support</i>	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	<i>NA</i>
<i>Competing interests</i>	Declare any competing interests of review authors.	Not included
<i>Availability of data, code and other materials</i>	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	OSF link on p. 20

Appendix B - Supplementary Materials for Part II

AI Chatbot Interview Example



Ava

Throughout this interview I will be asking you about common situations you may have experienced in your current or previous work. Please describe in detail the situations you have experienced and your behavior in those situations. The idea is that you refer to situations from your personal work experience. If you do not have relevant work experience, you can refer to situations you have experienced during your studies, internships, volunteer work, etc.



Ava

Let's get started!



Ava

Sometimes you have to wait longer for something than you originally thought. Think of a work-related situation that ended with you not getting the information you desperately needed even after asking several times. Please describe exactly how you experienced that situation and how you acted given the lack of information.



Participant1100001

Once I was waiting on data from a colleague to complete an urgent report for my boss. I asked them for updates multiple times but they just said it would take another day to two. It was frustrating, but I remained calm and informed my boss of the delay. I ended up working on completing other tasks while I waited for the data. Eventually I got the data and was able to send the report to my boss. Since I kept them updated, they were understanding of the delay.



Ava

Keeping your boss informed seems to have helped manage expectations well. [AI synthesized]



Ava

Everyone has different preferences and needs when it comes to organizing their workspace. Think of a situation in which you were working on a task where you needed a lot of materials or documents, and therefore a lot of space. In this situation, you should experience minimal time pressure. Please describe exactly how you experienced this situation and how you managed the setup of your workplace.

Type text, hit return key to send

0



**Ava**

Sometimes you have to wait longer for something than you originally thought. Think of a work-related situation that ended with you not getting the information you desperately needed even after asking several times. Please describe how you experienced that situation and how you acted given the lack of information.

Ashley

What do you mean?

**Ava**

I cannot provide additional information or elaborate further. Please answer the question to the best of your ability thinking of situations you experienced at work (or during your studies if you do not have work experience). I will repeat the question.

0



Example Behavior Description Interview Questions and Scoring Guide

from Heimann et al. (2021)

Example Behavior Description Interview Question for Extraversion

Sometimes you meet a lot of new people. Think of a situation in which you participated in a one or two-day course/seminar/workshop and you did not know a large portion of participants but quickly found some of them likeable. Please describe exactly how you experienced this situation and how you behaved towards the other seminar participants.

1	3	5
Experiences the contact situation with unknown persons as uninteresting or unpleasant; behaves reservedly towards the other participants; does not try to make new contacts; avoids interactions outside the seminar situation; surrounds himself only with known persons or remains alone	Experiences the possibility of contact with selected persons as interesting in principle; accepts offers of contact; possibly needs a concrete occasion or impetus from outside to approach unknown persons; surrounds himself mostly with known persons	Experiences enthusiasm for the opportunity to make contact with new people; actively seeks contact with many participants; approaches different people at every opportunity; surrounds self with familiar and unfamiliar people

Example Behavior Description Interview Question for Agreeableness

Sometimes you notice mistakes in others. Think of a situation in which you pointed out to a work or study colleague that they had made a mistake. Please describe exactly how you experienced this situation and how you acted when talking to this colleague.

1	3	5
Does not experience consideration for the feelings of the interlocutor as important; may not be aware of the emotional aspects of the situation; points out the mistake very directly; may reproach the person or be insulting	Experiences consideration of interlocutor's feelings as important; is polite; but points out error directly; does not make extra effort to emphasize friendliness to error	Experiences consideration for the feelings of the interlocutor as a priority; proceeds very gently; does not make any accusations; possibly points out the mistake; emphasizes friendliness towards mistakes

Example Behavior Description Interview Question for Conscientiousness

Everyone has a slightly different style of working and different demands in their own work. Think of a situation where you were engaged in the final steps of completing a large project, such as your thesis or another major written work. Please describe exactly how you experienced this situation and how you behaved in completing the work.

1	3	5
Experiences striving for good performance as of little importance; proceeds in an unstructured and unsystematic manner; does not plan work steps beforehand; has no schedule; does not check work before submission	Experiences striving for good performance as important; tends to proceed in a structured and systematic manner; plans steps of work; has a schedule; roughly checks work before turning it in	Experiences striving for good performance as paramount; takes a clearly structured and systematic approach to completing work; schedules buffer time for contingencies; checks work several times

Example Behavior Description Interview Question for Emotional Stability

Sometimes you have to wait longer for something than you originally thought. Think of a work-related situation that ended with you not getting the information you desperately needed even after asking several times. Please describe exactly how you experienced that situation and how you acted given the lack of information.

1	3	5
Experiences feelings of anger in relation to the lack of information; feels personally attacked or victimized; reacts very emotionally and with little control towards those responsible; is not capable of constructive solution-seeking	Experiences the lack of information as frustrating; feels unfairly treated; but remains polite to those in charge; sometimes constructively seeks alternatives to obtain the information	Experiences little inner tension regarding the lack of information; does not feel personally attacked; behaves politely toward those in authority; calmly and constructively seeks alternatives to obtaining the information

Example Behavior Description Interview Question for Intellect/Openness

There are some topics you are particularly interested in. Think of a situation in your studies or job where you had the opportunity and the time to deal in depth with a new topic beyond a work assignment. Please describe exactly how you experienced this situation and how you acted in terms of having the opportunity to engage with the new topic.

1	3	5
Does not experience the need to deal with the new topic in depth; is hardly willing to open up to new ideas; has difficulty being inspired and led to new topics	Experiences the in-depth examination of certain topics as interesting; is willing to open up to selected ideas to a certain degree; in principle allows himself to be inspired and led to new topics	Experiences great enthusiasm for exploring many different new topics in depth; completely opens up to new ideas; can inspire self and get excited about new topics

Embedding-based Model Selection Performance

For embedding-based models, I explored four variations of training, manipulating the number of outer folds for cross-validation, regularization (mixture parameter), and dimensionality reduction. In Version 1, I used 10 outer folds and 3/4 inner folds for cross-validation and regularization was set to 0, meaning Ridge regression was applied. These are also the default settings for the *textTrain()* function in the *text* package (Kjell et al., 2023). In Version 2, I retained the same cross-validation folds as Version 1, but I incorporated grid search to test multiple regularizations inside the inner loop (0 to 1 in increments of .1) through the mixture parameter, which controls the balance between L1 (Lasso) and L2 (Ridge) regularization. By testing multiple mixture values, the model selects the optimal regularization setting, improving generalizability across different responses. In Version 3, I reduced the number of outer folds from 10 to 5, which gives each split ~84 rows for test and ~337 rows for training. In Version 4, I added dimension reduction through PCA to reduce the dimensionality of word embeddings, which can be particularly helpful given the DistilBERT embeddings contain 768 dimensions. For this I specified a grid search for different variance retention levels (.80 to .95 in increments of .05), where .80 is more aggressive and focuses more on training the most important components, whereas .95 keeps almost all of the variance. The final model is selected using the best PCA and regularization, then the model is trained on the full outer training set using these values. This is repeated for each outer loop and then at the end the final overall model is selected based on the best-performing PCA setting across outer folds, the most frequently selected mixture values, and the entire dataset is trained on the entire dataset using these settings.

Appendix Table B.2 presents correlations between each model's predicted values and benchmarks (self, rater, and zero-shot), while Appendix Figure B.1 visualizes the results.

Performance across all versions was similar, with Versions 1 and 3 performing the best overall. Notably, in Version 3, 13 out of the 15 models selected Ridge regression (mixture = 0). Thus, for simplicity I selected Version 1 as the final model.

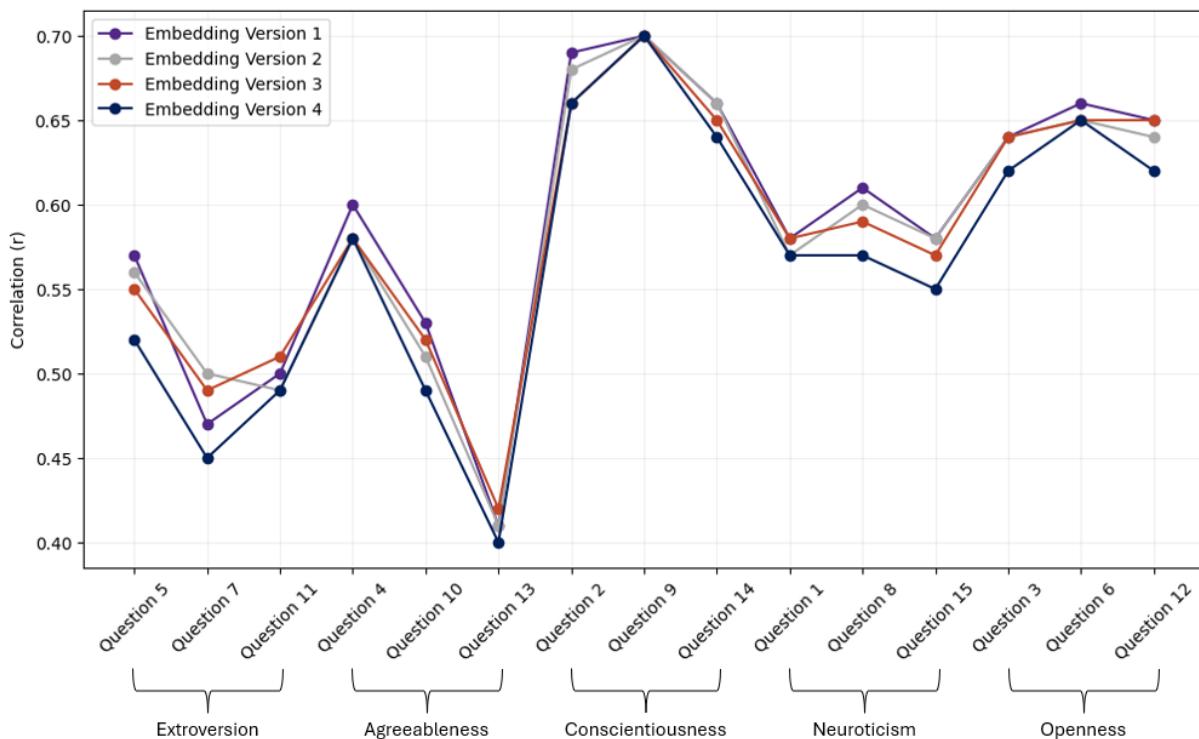
Appendix Table B.1. Domain- and Question-Level Correlations of Embedding-Based Model Predictions

	Version 1			Version 2			Version 3			Version 4			
	Validation Split			Outside: 10 Inside: 3/4			Outside: 10 Inside: 3/4			Outside: 5 Inside: 3/4			
Regularization	0			{0, .1, .2, .3, .4 ... 1}			{0, .1, .2, .3, .4 ... 1}			{0, .1, .2, .3, .4 ... 1}			
	PCA Preprocessing			None			None			None			
	Self	Rater	Zero	Self	Rater	Zero	Self	Rater	Zero	Self	Rater	Zero	
Domain													
Extraversion	.32	.58	.57	.31	.58	.58	.33	.60	.58	.33	.58	.56	
Agreeableness	.09	.61	.50	.09	.60	.49	.08	.60	.49	.08	.61	.50	
Conscientiousness	.28	.73	.59	.28	.73	.59	.28	.73	.59	.27	.73	.59	
Emotional Stability	.38	.64	.42	.38	.64	.43	.38	.65	.44	.38	.64	.43	
Openness	.27	.71	.72	.27	.70	.72	.27	.71	.72	.28	.70	.72	
Questions													
1	.39	.58	.45	.39	.57	.45	.39	.58	.47	.38	.57	.46	
2	.28	.69	.39	.28	.68	.38	.26	.66	.37	.27	.66	.40	
3	.26	.64	.54	.26	.64	.54	.28	.64	.54	.26	.62	.54	
4	.13	.60	.69	.13	.58	.68	.13	.58	.67	.13	.58	.67	
5	.28	.57	.50	.28	.56	.51	.27	.55	.50	.25	.52	.49	
6	.25	.66	.64	.25	.65	.63	.23	.65	.63	.26	.65	.64	
7	.27	.47	.50	.27	.50	.49	.28	.49	.48	.28	.45	.43	
8	.27	.61	.27	.28	.60	.27	.27	.59	.28	.24	.57	.30	
9	.25	.70	.66	.25	.70	.66	.25	.70	.66	.24	.70	.65	
10	.07	.53	.47	.06	.51	.45	.06	.52	.46	.06	.49	.48	
11	.29	.50	.47	.28	.49	.46	.32	.51	.47	.32	.49	.44	
12	.22	.65	.64	.23	.64	.63	.23	.65	.63	.23	.62	.62	
13	.03	.41	.14	.03	.41	.14	.01	.42	.18	.01	.40	.15	
14	.24	.66	.43	.24	.66	.43	.25	.65	.43	.23	.64	.44	
15	.38	.58	.39	.37	.58	.41	.37	.57	.40	.38	.55	.39	

Notes. The values in this table show the correlations between the embedding-based models'

predicted scores with self-reported scores (self), human rated scores (rater; i.e., the “ground truth” used for training), and zero-shot prompt-based scores (zero).

Appendix Figure B.1. Visualization of model performance for each question across Personality Traits



Example Zero-shot Prompt

****Task**:** Analyze and evaluate the provided answer according to the scoring instructions outlined below. You will play the role of a scoring expert and assess the answer based on the given behavioral anchors.

****Question**:**

{question}

****Answer**:**

{answer}

****Scoring Instructions**:**

1. ****Rating Criteria**:**

- Use a 5-point scale to evaluate the response, with the following behavioral anchors:
{instruction}
- Scores 2 and 4 reflect responses falling between the anchors for scores 1 and 3, or 3 and 5, respectively.

2. ****Evaluation Steps**:**

- Identify key behavioral cues in the answer that correspond to the behavioral anchors.
- Assign a numerical score based on how closely the response aligns with the descriptions of the anchors.

****Output Format**:**

- ****Score**:**
 - [Provide the numerical score here.]
- ****Reasoning**:**
 - [Provide a brief explanation of how the behavioral cues in the answer justify the assigned score.]

Appendix C - Supplementary Materials for Part III

Appendix Table C.1. Descriptive Statistics for Word Count in Responses with Perfect Agreement Across Human Raters and LLM from Part II

Q	Rating of 1				Rating of 3				Rating of 5			
	N	M	SD	min	N	M	SD	min	N	M	SD	min
1	1	25	NA	25	12	90	81	33	5	119	24	95
2	0	-	-	-	15	39	17	13	3	24	22	6
3	10	86	43	47	17	55	22	26	5	85	22	57
4	8	35	24	8	26	44	21	11	5	92	42	47
5	11	39	14	18	8	52	22	16	12	101	37	46
6	16	38	20	13	4	47	7	38	0	-	-	-
7	7	26	11	15	20	51	21	13	3	82	51	42
8	6	29	19	8	9	31	20	7	8	77	52	32
9	4	22	9	10	8	42	15	23	2	60	21	45
10	9	28	17	8	14	33	16	14	2	58	1	57
11	3	18	18	3	6	41	14	16	4	89	6	80
12	11	24	8	13	5	32	10	19	8	73	30	46
13	7	47	28	19	15	23	13	7	9	104	45	40
14	4	22	10	12	3	43	17	28	4	59	44	22
15	9	34	22	7	2	33	0	33	4	47	18	27

Notes. This table presents the descriptive statistics for response word counts in cases where human raters and the LLM assigned the same rating of 1, 3, or 5. Q = Question number, which corresponds to the interview questions provided in Appendix B (*AI Chatbot Interview Flow*); N = Number of cases where all human raters and the LLM assigned the same scores; M = Mean word count of responses; SD = Standard deviation of word count; min = Minimum word count from responses. On average, the minimum word count across responses rated 1, 3, and 5 was 27 words.

Selected Examples from Part II for Few Shot Learning

Appendix Table C.2. Examples for Few Shot Learning

Trait	Score 1	Score 3	Score 5
Extraversion	<p>Example: I can think of a time where I participated in an in person webinar, but I didn't interact with anyone. I'm pretty socially anxious in real life, so I kept to myself and focused more on the presentation.</p> <p>Reason: Showed no interest in engaging with others, exhibited withdrawn behavior, and avoids interactions with others.</p>	<p>Example: In my Kinesiology class I don't know anyone but I sat next to this one girl and I recognized her and I found out that we are in the same sorority so it's comforting knowing that I have someone I know personally in such a large class</p> <p>Reason: Showed interest in interacting with others, is in theory willing to accept offers of contact, but mostly interacted with someone familiar instead of branching out.</p>	<p>Example: I don't attend seminars, but in my line of work I meet new patients probably every day with some exceptions (some days are all familiar patients.) I approach them with a smile and a relaxed attitude. I generally love people, and I will start a conversation with anyone. Even grumpy patients I understand need to feel like they are being attended to. This is my expertise and probably my greatest asset in the work that I do. I am not afraid of talking to people or introducing myself at all and I love the connection!</p> <p>Reason: Shows great enthusiasm in meeting new people, proactively engages with them and initiates conversation.</p>
Agreeableness	<p>Example: I pointed out one of my employees was not following the company standardized recipes. They got very offended and hated me. I explained to them on several occasions it is my job to insure customers are receiving the correct, standardized drink and that employees are conforming. She still hated me, but the other employees quickly learned through their own experiences this girl is not fun and quite frankly, a bitch.</p> <p>Reason: Shows no consideration for feelings of others, dismisses employee's reaction and insults them, shows lack of empathy and does not handle the situation gently.</p>	<p>Example: I wasn't condescending and I politely told them they made an error. They thanked me and fixed it.</p> <p>Reason: Shows some consideration for others' feelings, but is still direct and does not make extra effort to cushion the correction.</p>	<p>Example: During a recent group project at work, I observed that a colleague had made an error in calculating some critical data for our presentation. Recognizing the importance of accuracy with our deadline approaching, I considered it essential to address the mistake promptly. I handled the situation with sensitivity. Rather than confronting them in public, I waited for a private moment to suggest discussing the data together. I emphasized that my goal was to achieve the best outcome for our team rather than criticize their work. By beginning with praise for their efforts on the project, I set a positive tone for our conversation. I gently highlighted the specific miscalculation and explained how I arrived at my conclusion. I made sure to present it as a team effort, suggesting that we could double-check everything together to improve the final presentation. My colleague appreciated this approach, acknowledged the oversight, and together we resolved it without any hard feelings. This experience</p>

			<p>highlighted the crucial role of effective communication and support within a team. Ultimately, we not only rectified the mistake but also enhanced our working relationship.</p> <p>Reason: Shows great empathy in addressing the colleague, takes extra effort to discuss with them privately and making sure they feel comfortable, worked together with the colleague to resolve the error.</p>
Conscientiousness	<p>Example: Once I finally was almost done with my essay and was editing it, I was so ready to be done that I completed it haphazardly.</p> <p>Reason: Did not express striving for good performance, took an unstructured and haphazard approach when completing the essay, just wanted to submit it.</p>	<p>Example: In the final steps of a project, I usually try to be as thorough as possible. I put the finishing touches to the work and mostly just review it.</p> <p>Reason: Approaches finishing the project in a somewhat careful manner, checks work before turning it in, but does not mention a highly structured approach or striving for good performance as very important.</p>	<p>Example: I tend to overthink large projects, completely sort them in my head until I am comfortable with my plan, then methodically proceed with each piece. I usually finish too early so that I can spend time on revisions. Big projects are just a series of steps.</p> <p>Reason: Demonstrates care for good performance as they overthink things, has a plan and takes a structured approach to completing the project, tries to finish early to save time for revisions, checks before submitting.</p>
Emotional Stability	<p>Example: During this experience I felt unheard and un-appreciated. Due to this acted like I did not care and didn't want to be at the job</p> <p>Reason: Expresses feeling of victimization (e.g., feels unappreciated), does not seek constructive solution, rather disengages fully</p>	<p>Example: Recently I was trying to get a job, but the person that interviewed me was not getting the information I needed to start the job, so I emailed her and went into her office to get information on any updates but she told me that she has not called the people she needed to get into contact with. I began to become frustrated, but I was able to fix all of the problems that occurred on my side of the paper work and now I am waiting for new information to come.</p> <p>Reason: Expressed frustration but remains polite and does not show feelings of anger. Took a proactive approach to solve the problem.</p>	<p>Example: I used to work with an author who would just disappear, sometimes when he hadn't yet sent what we needed to publish that day. One time, something changed in the markets that went against what he had written for that day. We needed to change the day's article to reflect this market change. I don't know much about markets, though. When I didn't get any answers from the author, I pivoted and asked an analyst and my department head for their insights. I then used their insights to add a note to the original article</p> <p>Reason: Expresses no frustration or anger, remained polite and sought out solutions from other sources to solve the problem.</p>

Openness	<p>Example: I feel like if its something that I didnt initially want to do or that I am not getting paid for I dont want to do it</p> <p>Reason: Shows no interest in trying new things unless required.</p>	<p>Example: I had the opportunity to assist with some marketing material from a different department which was outside my work requirements. I found the work engaging and interesting as it allowed me to use some creativity vs my usual work which is more technical. I acted in a professional and grateful for the opportunity to try something new.</p> <p>Reason: Willing to work on projects outside of their department, found the task engaging and interesting but did not indicate further engagement beyond the task. Generally had a positive attitude towards the opportunity.</p>	<p>Example: I got the chance to investigate behavioural psychology and its influence on customer behaviour during a slower time at work. I thought the subject was fascinating even if it had nothing to do with my assignments. I attended webinars, took online classes, and read books about marketing decision-making and psychological triggers. Using concepts like social proof and scarcity, I experimented various message tactics to put what I had learnt into practice. My team and I were able to improve our efforts and start new conversations because to this in-depth analysis. It was a fulfilling opportunity to broaden my knowledge and add originality to my work.</p> <p>Reason: Showed great enthusiasm for this new subject, engaged deeply with the topic by taking extra classes and obtaining additional resources, showed excitement and self-motivation to try new things.</p>
----------	--	---	---

Notes. Examples were chosen based on convergence between human raters and the LLM zero shot scores. Meaning for identifying an example for the score of 1, responses where all human raters and the LLM rated it as 1. From these options, the best example that contains the best trait-relevant cues is selected. In the majority of cases, there were at least 1 option to choose from. In some cases, there were no cases where there was perfect agreement across human raters and the LLM. For these questions, I prioritized cases where human scores had perfect agreement and looked for scores where the LLM scored within 1 point.

Example Few-shot Prompt

Task: Analyze and evaluate the provided answer according to the scoring instructions outlined below. You will play the role of a scoring expert and assess the answer based on the given behavioral anchors.

Question:

{question}

Answer:

{answer}

Scoring Instructions:

1. **Rating Criteria**:

- Use a 5-point scale to evaluate the response, with the following behavioral anchors: {instruction}
- Scores 2 and 4 reflect responses falling between the anchors for scores 1 and 3, or 3 and 5, respectively.

2. **Evaluation Steps**:

- Identify key behavioral cues in the answer that correspond to the behavioral anchors.
 - Assign a numerical score based on how closely the response aligns with the descriptions of the anchors.
1. **Read the Question**: Identify the context and key points that the question asks about.
 2. **Review the Answer**: Determine whether the answer addresses the question's context, the actions taken, the thought process, or the results.
 3. **Match to Behavioral Anchors**: Compare the answer's content with the rating scale to see where it best fits.
 4. **Decide on a Score**: Choose a score from 1 to 5. In ambiguous cases, compare neighboring scores (e.g., 2 vs. 3, or 3 vs. 4) and clarify your choice.
 5. **Write the Justification**: In the "Reasoning" section, briefly highlight behaviors or cues that led you to your rating, and explain any missing elements.

Below are some examples for the anchors for scores of 1, 3, and 5:

- Example response that received a score of 1: {Example1}
 - Reason for score of 1: {reason1}
-
- Example response that received a score of 3: {Example3}
 - Reason for score of 1: {reason3}
-
- Example response that received a score of 5: {Example5}
 - Reason for score of 1: {reason5}

Output Format:

- **Score**:
 - [Provide the numerical score here.]
- **Reasoning**:

[Provide a brief explanation of how the behavioral cues in the answer justify the assigned score.]

Condition Assignment Scripts

Honest Condition Script

The Psychology Department is working on developing assessment tools for their employee selection process. Based on this pilot test, they plan to start recruiting some student positions next academic year to assist with administrative duties and course development. The pre-employment assessment consists of two major parts, interview by a virtual chatbot and an online questionnaire.

It is extremely important for you to **provide honest responses to both the chatbot and when responding to the questionnaire**. Truthful responding will help ensure that the results from the chatbot and the results from the questionnaire align.

Induced Faking Condition Script

The Psychology Department is working on developing assessment tools for their employee selection process. Based on this pilot test, they plan to start recruiting some student positions next academic year to assist with administrative duties and course development. The pre-employment assessment consists of two major parts, interview by a virtual chatbot and an online questionnaire.

To ensure our methods are viable for actual selection contexts, we would like for you to **respond as if you were an actual candidate** applying for the position.

Candidates should be from a **variety of backgrounds who have ambitious career goals and are self-disciplined, initiative-taking, and persistent in carrying out job duties**. They must be **organized, determined, and able to display planned rather than spontaneous behavior**. Qualified students should also have a **broad knowledge background and possess some basic skills that are useful in office settings**.

Motivated Faking Condition Script

The Psychology Department at Kansas State University is working on developing assessment tools for their student worker selection process. Based on this pilot test, they plan to start recruiting some student positions next academic year to assist with administrative duties and course development.

As an additional benefit to piloting and validating their assessment tools, **we want to offer high-scoring participants additional opportunities for validating our assessment tools in the future.**

However, taking the assessment does not mean you are required to complete the additional surveys. You can always opt out if we contact you with another survey offer. If you are interested, you can let us know in the Qualtrics survey and we will consider your assessment results for future opportunities. Your demographic data will not be disclosed during the selection process.

The Psychology Department is looking for **candidates from a variety of backgrounds who have ambitious career goals and are self-disciplined, initiative-taking, and persistent in carrying out job duties.** They must be **organized, determined, and able to display planned rather than spontaneous behavior.** Qualified candidates should also have a broad knowledge background and possess some basic skills that are useful in office settings.

Below is the job posting that will become available for the upcoming semester.

As a participant in this study, if you perform well on the assessments you will be considered for additional survey opportunities for validating our assessment tools.

Student Intern

Department: Psychological Sciences

Employment type: Student, Part-Time (Fall 2025)

Location: Manhattan, KS

Worksite: Remote

About This Role:

The position will support one Psychological Sciences faculty member in developing a new curriculum for the PSYCH 110 course. The Psychology Department is looking for students from a variety of backgrounds who have ambitious career goals and are self disciplined, initiative-taking, and persistent in carrying out job duties. They must be organized, determined, and able to display planned rather than spontaneous behavior. Qualified students should also have broad knowledge background and possess some basic skills at office settings.

Worksite Options:

This position can be performed completely remote. Work hours are flexible based on applicant's preference and schedule.

Preferred Qualifications:

- Excellent written communication skills.
- Ability to work independently.
- Strong organizational skills.
- Proficient in the Microsoft Office suite, such as Word, PowerPoint, and Excel.

Anticipated Hiring Pay Range:

\$18.00 per hour

Equal Employment Opportunity:

Kansas State University is an Equal Opportunity Employer. All applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender, gender identity, age, national origin, disability or status as a protected veteran.

Would you like to be considered for additional survey opportunities? (Note: Taking the assessment does not necessarily mean you are required to complete future surveys)

Yes

No

Appendix D - Survey Measures

Contextualized Personality

Big Five Inventory-2 (Soto & John, 2017; JPSP)

Next, you will be presented with a number of characteristics that may or may not apply to you in the workplace. Please think specifically about your behavior in professional or work settings as you read each statement carefully. Consider how each statement describes you in the context of your workplace, and decide the extent to which you agree or disagree. Then, accurately select the answer option that best represents you.

(1=Strongly disagree, 2=Disagree, 3=Neither agree nor disagree, 4=Agree, 5=Strongly agree)

At work, I am someone who...

1. Is outgoing, sociable with colleagues.
2. Is compassionate, has a soft heart toward coworkers.
3. Tends to be disorganized with tasks and responsibilities at work.
4. Is relaxed, handles stress well at work.
5. Has few creative and artistic interests at work.
6. Has an assertive personality when engaging in workplace discussions.
7. Is respectful, treats others professionally at work.
8. Tends to be lazy in work activities.
9. Stays optimistic after experiencing a setback at work.
10. Is curious about many different topics/projects at work.
11. Rarely feels excited or eager for workplace activities.
12. Tends to find fault with others' work.
13. Is dependable, steady in fulfilling work responsibilities.
14. Is moody, has up and down mood swings at work.
15. Is inventive, finds clever ways to solve work-related challenges.
16. Tends to be quiet at work.
17. Feels little sympathy for coworkers.
18. Is systematic, likes to keep their workspace in order.
19. Can be tense at work.
20. Is fascinated by aesthetic or creative aspects of work.
21. Is dominant, acts as a leader in the workplace.
22. Starts arguments with other colleagues.
23. Has difficulty getting started on work tasks.
24. Feels secure, comfortable with their professional abilities.
25. Avoids intellectual, philosophical discussions in the workplace.
26. Is less active than other colleagues.
27. Has a forgiving nature towards colleagues.
28. Can be somewhat careless with work responsibilities.
29. Is emotionally stable, not easily upset by challenges at work.
30. Has little creativity in approaching work tasks.
31. Is sometimes shy, introverted at work.
32. Is helpful and unselfish with colleagues.
33. Keeps the workspace neat and tidy.

34. Worries a lot about work-related issues.
35. Values art and beauty in the workplace.
36. Finds it hard to influence colleagues.
37. Is sometimes rude towards colleagues.
38. Is efficient, gets work tasks done.
39. Often feels sad while at work.
40. Is complex, a deep thinker about work.
41. Is full of energy throughout the workday.
42. Is suspicious of colleagues' intentions.
43. Is reliable, can always be counted on to complete work tasks.
44. Keeps their emotions under control at work.
45. Has difficulty imagining things for work.
46. Is talkative at work.
47. Can be cold and uncaring towards colleagues.
48. Leaves a mess, doesn't clean up at work.
49. Rarely feels anxious or afraid at work.
50. Thinks discussions on poetry and plays with colleagues are boring.
51. Prefers to have other colleagues take charge.
52. Is polite, courteous to colleagues.
53. Is persistent, works until the task is finished.
54. Tends to feel depressed, blue at work.
55. Has little interest in abstract ideas at work.
56. Shows a lot of enthusiasm for work.
57. Assumes the best about colleagues.
58. Sometimes behaves irresponsibly at work.
59. Is temperamental, gets emotional easily at work.
60. Is original, comes up with new ideas for work projects.

IPIP-50 (Goldberg, 1992)

In this part, you will be presented with a number of characteristics that may or may not apply to you in the workplace.

Please read each statement carefully and indicate the degree to which each statement is an accurate description of you with respect to how you behave at work.

At work, I...

1. Am the life of the party.
2. Feel little concern for my colleagues.
3. Am always prepared.
4. Get stressed out easily.
5. Use a rich vocabulary.
6. Don't talk a lot.
7. Am interested in my colleagues.
8. Leave my belongings around my workspace.
9. Am relaxed most of the time.
10. Have difficulty understanding abstract ideas.
11. Feel comfortable around my colleagues.

12. Insult my colleagues.
13. Pay attention to details.
14. Worry about things.
15. Have a vivid imagination.
16. Keep in the background.
17. Sympathize with colleagues' feelings.
18. Make a mess of things.
19. Seldom feel blue.
20. Am not interested in abstract ideas.
21. Start conversations with colleagues.
22. Am not interested in other colleagues' problems.
23. Get work tasks done right away.
24. Am easily disturbed.
25. Have excellent ideas.
26. Have little to say.
27. Have a soft heart.
28. Often forget to put things back in their proper place.
29. Get upset easily.
30. Do not have a good imagination.
31. Talk to a lot of different people at work events.
32. Am not really interested in my colleagues.
33. Like order.
34. Change my mood a lot.
35. Am quick to understand things.
36. Don't like to draw attention to myself.
37. Take time out for colleagues.
38. Shirk my work duties.
39. Have frequent mood swings.
40. Use difficult words.
41. Don't mind being the center of attention.
42. Feel colleagues' emotions.
43. Follow a schedule.
44. Get irritated easily.
45. Spend time reflecting on things.
46. Am quiet around newcomers.
47. Make colleagues feel at ease.
48. Am exacting in my work.
49. Often feel blue.
50. Am full of ideas.

User Perceptions and Engagement

Perceived difficulty

1. Did you find my questions confusing?
(1=Not at all, 2=Very little, 3=Neutral, 4=Somewhat, 5=To a great extent)
2. To what extent did you find it challenging to answer my questions during our conversation?
(1=Not at all, 2=Very little, 3=Neutral, 4=Somewhat, 5=To a great extent)
3. Did you think answering my questions was mentally demanding?
(1=Not at all, 2=Very little, 3=Neutral, 4=Somewhat, 5=To a great extent)

Perceived validity

4. Do you think the way you chatted with me could accurately reflect your personality?
(1=Strongly disagree, 2=Disagree, 3=Neither agree nor disagree, 4=Agree, 5=Strongly agree)
5. Do you think it would be useful to have assessment components of this format (like chatting with a bot) for settings like personnel selection?
(1=Strongly disagree, 2=Disagree, 3=Neither agree nor disagree, 4=Agree, 5=Strongly agree)
6. Would you agree that an assessment conducted this way (via interactions with a chatbot) has value in real life?
(1=Strongly disagree, 2=Disagree, 3=Neither agree nor disagree, 4=Agree, 5=Strongly agree)

Affect Toward the Measure

7. Did you enjoy interacting with me?
(1=Not at all, 2=Very little, 3=Neutral, 4=Somewhat, 5=To a great extent)
8. Was chatting with me frustrating?
(1=Not at all, 2=Very little, 3=Neutral, 4=Somewhat, 5=To a great extent)
9. To what extent would you avoid interacting with chatbots like me in the future?
(1=Not at all, 2=Very little, 3=Neutral, 4=Somewhat, 5=To a great extent)

Concentration

10. Did you find yourself losing interest or not paying attention while chatting with me?
(1=Never, 2=Rarely, 3=Sometimes, 4=Frequently, 5=Always)
11. Was it easy to stay focused on our conversation?
(1=Not at all, 2=Probably not, 3=Neutral, 4=Somewhat, 5=Definitely)
12. Did you find it easy to get distracted when engaging in interactions like this?
(1=Not at all, 2=Probably not, 3=Neutral, 4=Somewhat, 5=Definitely)

User Engagement Scale – Short Form (for BFI-2; O'Brien et al., 2018)

(FA = focused attention, PU = perceived utility, AE = aesthetic appeal, RW = reward factor)

The following statements ask you to reflect on your experience of completing the questionnaire. For each statement, please use the following scale to indicate what is most true for you.

1. FA-S.1 I lost myself in this experience.

2. FA-S.2 The time I spent filling out this questionnaire just slipped away.
3. FA-S.3 I was absorbed in this experience.
4. PU-S.1 I felt frustrated while completing the questionnaire.
5. PU-S.2 I found this questionnaire confusing to use.
6. PU-S.3 Using this questionnaire was taxing.
7. AE-S.1 This questionnaire was attractive.
8. AE-S.2 This questionnaire was aesthetically appealing.
9. AE-S.3 This questionnaire appealed to my senses.
10. RW-S.1 Using the questionnaire was worthwhile.
11. RW-S.2 My experience was rewarding.
12. RW-S.3 I felt interested in this experience.

User Engagement Scale – Short Form (for chatbot; O'Brien et al., 2018)

(FA = focused attention, PU = perceived utility, AE = aesthetic appeal, RW = reward factor)

The following statements ask you to reflect on your experience of engaging with the chatbot. For each statement, please use the following scale to indicate what is most true for you.

1. FA-S.1 I lost myself in this experience.
2. FA-S.2 The time I spent using this chatbot just slipped away.
3. FA-S.3 I was absorbed in this experience.
4. PU-S.1 I felt frustrated while using this chatbot.
5. PU-S.2 I found this chatbot confusing to use.
6. PU-S.3 Using this chatbot was taxing.
7. AE-S.1 This chatbot was attractive.
8. AE-S.2 This chatbot was aesthetically appealing.
9. AE-S.3 This chatbot appealed to my senses.
10. RW-S.1 Using this chatbot was worthwhile.
11. RW-S.2 My experience was rewarding.
12. RW-S.3 I felt interested in this experience.

Methodological Integrity of Chatbot Interview for Hiring Decisions (adapted from Lee et al., [under review])

1. To what extent do you think it is fair to adopt the chatbot interview for making hiring decisions?
(1 = Extremely unfair; 5 = Extremely fair)
2. To what extent do you think it is bias-free to adopt the chatbot interview for making hiring decisions?
(1 = Extremely Biased, 5 = Extremely Bias-free)
3. To what extent do you believe that the adoption of the chatbot interview for making hiring decisions will result in accurate hiring decisions?
(1 = Extremely Inaccurate, 5 = Extremely Accurate)
4. To what extent do you think it is ethical to adopt the chatbot interview for making hiring decisions?
(1 = Extremely Unethical, 5 = Extremely Ethical)
5. To what extent do you believe that the adoption of the chatbot interview for making hiring decisions will result in consistent hiring decisions?
(1 = Extremely Inconsistent, 5 = Extremely Consistent)

Attitude Towards the Chatbot Interview for Hiring Decisions (adapted from Lee et al., [under review])

1. To what extent do you believe that the adoption of the chatbot interview for making hiring decisions is transparent and easily understandable?
(1 = Extremely Disagree, 5 = Extremely Agree)
2. To what extent do you believe that the adoption of the chatbot interview for making hiring decisions is legally defensible?
(1 = Extremely Not Defensible, 5 = Extremely Defensible)
3. To what extent do you think it is acceptable to adopt the chatbot interview for making hiring decisions?
(1 = Extremely Unacceptable, 5 = Extremely Acceptable)

Chance to Perform (from the Selection Procedural Justice Scale; Bauer et al., 2001)

(1=Strongly disagree, 2=Disagree, 3=Neither agree nor disagree, 4=Agree, 5=Strongly agree)

1. I could really show my skills and abilities through this interview.
2. This interview allowed me to show what my job skills are.
3. This interview gives potential applicants the opportunity to show what they can really do.
4. I was able to show what I can do on this interview.

Propriety of Questions (from the Selection Procedural Justice Scale; Bauer et al., 2001)

(1=Strongly disagree, 2=Disagree, 3=Neither agree nor disagree, 4=Agree, 5=Strongly agree)

1. The content of the chatbot interview did not appear to be prejudiced.
2. The chatbot interview itself did not seem too personal or private.
3. The content of the chatbot interview seemed appropriate.

Outcome Variables

Overall job performance

(1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree)

1. I am very competent at what I do.
2. I get my tasks done very effectively.
3. I perform my responsibilities well.

Subjective well-being (Diener et al., 1985)

Please indicate the degree to which you agree with the following statements on a 5-point scale.

(1=Strongly disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly agree)

1. In most ways my life is close to my ideal
2. The conditions of my life are excellent
3. I am satisfied with my life
4. So far I have gotten the important things I want in life
5. If I could live my life over, I would change almost nothing

Organizational Citizenship Behavior (Spector et al., 2010)

Recall that in the past year, how often have you done the following things?

(1=Never, 2=Once or twice, 3=Once or twice per month, 4= Once or twice per week, 5=Every day)

1. Took time to advise, coach, or mentor a co-worker.
2. Helped co-worker learn new skills or shared job knowledge.
3. Helped new employees get oriented to the job.
4. Lent a compassionate ear when someone had a work problem.
5. Offered suggestions to improve how work is done.
6. Helped a co-worker who had too much to do.
7. Volunteered for extra work assignments.
8. Worked weekends or other days off to complete a project or task.
9. Volunteered to attend meetings or work on committees on own time.
10. Gave up meal and other breaks to complete work.

Counterproductive work behavior (Spector et al., 2010)

Recall that in the past year, how often have you done the following things?

(1=Never, 2=Once or twice, 3=Once or twice per month, 4= Once or twice per week, 5=Every day)

1. Purposely wasted your employer's materials/supplies
2. Complained about insignificant things at work
3. Told people outside the job what a lousy place you work for
4. Came to work late without permission
5. Stayed home from work and said you were sick when you weren't
6. Insulted someone about their job performance
7. Made fun of someone's personal life
8. Ignored someone at work
9. Started an argument with someone at work
10. Insulted or made fun of someone at work

Job satisfaction in general (AJIG; Russell, Spitzmuller, Lin, Stanton, Smith, & Ironson, 2004)

Think of your current school or professional work in general. All in all, what is it like most of the time?

(1=Yes, it describes my work, 2=Cannot decide, 3=No, it does not describe my work)

1. Good
2. Undesirable
3. Better than most
4. Disagreeable
5. Make me content
6. Excellent
7. Enjoyable
8. Poor

Perceived stress (PSS-10; Cohen et al., 1983)

In the past month, how often have you...

(1=Never, 2=Almost never, 3=Sometimes, 4=Fairly often, 5=Very often)

1. been upset because of something that happened unexpectedly?
2. felt that you were unable to control the important things in your life?
3. felt nervous and “stressed”?
4. felt confident about your ability to handle your personal problems?
5. felt that things were going your way?
6. found that you could not cope with all the things that you had to do?
7. been able to control irritations in your life?
8. felt that you were on top of things?
9. been angered because of things that were outside of your control?
10. felt difficulties were piling up so high that you could not overcome them?

Task Performance (from the Individual Work Performance Questionnaire; Koopmans et al., 2012)

At work, in the past 3 months...

(1=Seldom, 2=Sometimes, 3=Frequently, 4=Often, 5=Always)

1. I managed to plan my work so that it was done on time.
2. My planning was optimal.
3. I kept in mind the results that I had to achieve in my work.
4. I was able to separate main issues from side issues at work.
5. I was able to perform my work well with minimal time and effort.
6. Collaboration with others was very productive.

ProQuest Number: 31937913

INFORMATION TO ALL USERS

The quality and completeness of this reproduction is dependent on the quality
and completeness of the copy made available to ProQuest.



Distributed by

ProQuest LLC a part of Clarivate (2025).

Copyright of the Dissertation is held by the Author unless otherwise noted.

This work is protected against unauthorized copying under Title 17,
United States Code and other applicable copyright laws.

This work may be used in accordance with the terms of the Creative Commons license
or other rights statement, as indicated in the copyright statement or in the metadata
associated with this work. Unless otherwise specified in the copyright statement
or the metadata, all rights are reserved by the copyright holder.

ProQuest LLC
789 East Eisenhower Parkway
Ann Arbor, MI 48108 USA