University of Cape Town

School of Management Studies

Faculty of Commerce

Examining Personality Assessment in Asynchronous Video Interviews (AVIs): Convergence between Human Personality Judgements and Artificial Intelligence/Machine Learning Scoring

Jacobus Fouché Cronje

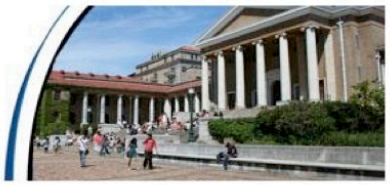
(CRNJAC009)

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Organisational Psychology.

Supervisor: Professor François S de Kock

February 2025

UNIVERSITY OF CAPE TOWN
FACULTY OF COMMERCE
Igniting Knowledge and Opportunity



Plagiarism Declaration

COMPULSORY DECLARATION:

1. This dissertation has been submitted to Turnitin (or equivalent similarity and originality checking software) and I confirm that my supervisor has seen my report and any concerns revealed by such have been resolved with my supervisor.
2. I certify that I have received Ethics approval (if applicable) from the Commerce Ethics Committee.
3. This work has not been previously submitted in whole, or in part, for the award of any degree in this or any other university. It is my own work. Each significant contribution to, and quotation in, this dissertation from the work, or works of other people has been attributed, and has been cited and referenced.

Student number	CRNJAC009
Student name	Jacobus Fouché Cronje
Signature of Student	Signed by candidate
Date:	10 June 2025

Acknowledgements

I would like to express my gratitude to my supervisor, Professor Francois de Kock, for his guidance, patience, and support throughout the year. I am deeply appreciative of his profound understanding of the field of organisational psychology, current advancements in personnel selection, as well as his feedback, encouragement, and interest in my research.

Secondly, I want to express my appreciation to Tristan Bell for his eagerness and dedication in assisting with the development of the AI algorithm used in this study. I am grateful that I was able to call upon his expertise and learn from him during this process. I would like to thank him for his professionalism, insight, patience, hard work, and passion for the fields of psychology and computer science.

I also extend my thanks to all the human evaluator participants in my study. Their valuable contributions, dedication, time management, and professionalism were essential to the success of my dissertation, and this work would not have been possible without their involvement.

Last but not least, to my family, friends, and colleagues: thank you for your interest in my studies and academic journey. Thank you for your love and support, every act of kindness, and every word of encouragement. I appreciate you more than words can express.

Abstract

The assessment of personality is an essential component of personnel selection due to its validity in predicting job performance. To assess personality, asynchronous video interviews (AVIs) scored using artificial intelligence (AI) algorithms are increasingly used, allowing candidates to record responses to interview prompts that are subsequently evaluated automatically by AI algorithms and/or human raters. As questions remain about the validity of AI-based AVI scoring approaches, this study examines the convergence between human- and AI-scored personality assessments. To measure personality, the study focuses on the HEXACO model, which measures Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience. Verbal responses were transcribed from videotaped AVIs of 161 mock interview candidates who answered five AVI questions. Responses were scored by 15 trained human raters and a closed-dictionary text-analysis keyword-counting AI algorithm developed for this study, respectively. The correlation between trait-level scores produced by human judges and AI scoring was tested both across traits and within traits (trait-level) to assess scoring convergence. Moreover, in addition to comparing score levels produced by the two scoring methods (AI vs. human raters), score spread (i.e., variability), rank-order stability, and rating reliability were evaluated. The findings revealed a moderately positive and significant overall convergence ($r = .29$, $p < .001$) across traits between human and AI evaluations, which suggests that AI scoring may potentially be useful as a replacement of human evaluations when general screening is desired. Trait-level convergence varied between scoring methods, with the scoring consensus between human raters and AI being higher for some traits than for others, suggesting that these methods rely on different information and/or may interpret interview responses differently. The research highlights the potential of AI to complement human-based scoring of AVIs used in recruitment, selection, and assessment while also identifying the limitations of algorithm-based scoring in capturing complex human behaviour in interviews. The findings may further contribute to understanding the role of AI in personality assessment and implications for organisational practices.

Keywords: recruitment, personnel selection, interviews, asynchronous video interviewing, personality, HEXACO, automation, AI, ML, human evaluation, convergence, correlation, industrial/organisational psychology.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background

1.1.1 Recruitment and Selection

Recruitment and selection are essential Human Resource Management (HRM) functions that enable organisations to align their workforce with strategic plans and organisational needs (Cascio & Aguinis, 2024; Lievens & Chapman, 2019; Ployhart et al., 2017). Recruitment involves attracting individuals to participate in a selection process. In contrast, selection is a predictive method that uses assessment measures to link an individual's knowledge, skills, abilities, and other characteristics (KSAOs) to their potential job performance (Anderson et al., 2001; Langer et al., 2023; Ployhart et al., 2017). Those individuals with the highest scores on predictor measures are typically hired (Ployhart et al., 2017). These predictor measures include a variety of assessments, among which interviews and psychometric assessments, such as personality assessments, are widely utilised (Cascio & Aguinis, 2024; Coetzee et al., 2021). Assessing personality during personnel selection processes is of increasing importance, as personality trait measures have become better predictors of employees' potential job performance (Anderson et al., 2001; Dupré & Wille, 2025; Plouffe et al., 2017; Salgado, 2017), including in South Africa (Coetzee et al., 2021; van Aarde et al., 2017).

1.1.2 Personality at Work: Defined

Personality refers to a person’s enduring natural tendencies and characteristic patterns of emotional states, thinking, desires and behaviour, collectively described as traits that help predict behaviour, emotional responses, and attitudinal outcomes related to workplace performance (Bergner, 2020; Christiansen & Tett, 2013; De Raad & Barelds, 2020; Funder, 2012; Heimann et al., 2021; Ones et al., 2025; Woods et al., 2013). As a multifaceted and complex concept, personality is shaped by various factors, including genetics, environment, and social dynamics (Bergh & Geldenhuys, 2013; Luthans et al., 2021). Various personality models can be used to understand, research, and apply personality traits in practice (De Raad & Barelds, 2020; Feher & Vernon, 2021). Rather than other models, such as the Five Factor or Big Five, this study specifically adopts the HEXACO model (Feher & Vernon, 2021). Widely used in personality research, this framework includes six traits: Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience (Ashton et al., 2014; Julian et al., 2022).

Although the Big Five is considered one of the leading personality trait models among researchers and practitioners (Anglim & O’Connor, 2019; Feher & Vernon, 2021; Romano et

al., 2023), the HEXACO model was preferred in this study because it includes a sixth trait, Honesty-Humility, which is not explicitly present in the Big Five (Ashton & Lee, 2020). Nonetheless, the traits of the HEXACO model correlate with the relevant traits in the Big Five, namely Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (Feher & Vernon, 2021; Lee et al., 2005). Furthermore, much of the predictive power of the HEXACO model can be attributed to the Honesty-Humility factor (Feher & Vernon, 2021), which provides valuable insights into ethical behaviour and interpersonal tendencies, including both prosocial and antisocial behaviours, aspects that are not fully captured by the Big Five (Ashton & Lee, 2020; Romano et al., 2023). The emphasis on the HEXACO model, therefore, provides a foundation for exploring the application of personality assessment in research and recruitment practices, particularly the measurement of personality through interviews, as discussed in the next section and applied in this study.

1.1.3 Interviews and Personality Assessment

Personality measurement is typically designed as self-report questionnaires; however, it is also frequently conducted through interviews (Hilgert et al., 2016). Interviews are among the most used selection methods (Speer et al., 2022). Traditionally, the interview is a communication process in which the organisation or selection panel learns more about the candidate in relation to the role and organisational characteristics, while the candidate also gains insight into the job and the organisation (Cascio & Aguinis, 2024). While this conversation can take many forms, ranging from unstructured to more structured, and, more recently, technology- and digitally-driven, interviews are especially predictive of job performance when they are structured (Baumgartner et al., 2024; Patel et al., 2025; Speer et al., 2022). One of the reasons structured interviews are so effective is that the interviewer establishes and applies predetermined criteria for questions, observations, and assessments, whether conducted face-to-face or through technology (Baumgartner et al., 2024; Patel et al., 2025).

1.1.3.1 Humans Evaluating Personality in Interviews

However, employment interviews are not only used to assess specific job- or organisation-related requirements but also serve as an effective means of evaluating personality (Trull et al., 1998; Van Iddekinge et al., 2005). While self-report personality questionnaires have traditionally been relied upon in selection, concerns about faking have led to interview-based judgments as alternatives, potentially offering better predictions of job performance (Hickman et al., 2024). It is well established that human observers can be accurate judges of personality and predictors of behaviour. Research in the personality

literature, dating back over 30 years, suggests that human ratings of personality (observer ratings) can predict behaviour as accurately as, and sometimes better than, self-reports (Connelly & Ones, 2010; Mount et al., 1994). However, the quality of these ratings depends on, and can be improved through, factors such as rater training (De Kock et al., 2020), which are discussed in more detail later in this study.

Currently, while humans can be accurate in their judgements, increasing reliance is being placed on technology to aid human evaluations in interviewing and subsequent personality evaluation, as explored in the following section (Langer et al., 2023; Patel et al., 2025).

1.1.4 The Role of Technology in Personality Assessment and Selection

Over 50 years ago, Wernimont and Campbell (1968) highlighted that technological advancements can potentially transform selection processes by enabling a deeper understanding of behavioural prediction (Auer et al., 2022). In recent years, the recruitment and selection landscape has undergone a significant technological transformation, driven by advancements in automation and artificial intelligence (AI), particularly in assessment practices (Alexander III et al., 2025; Langer et al., 2023; Nikolaou, 2021; Ployhart et al., 2017). While traditional assessment methods, such as psychometric tests and structured interviews, remain prominent (Coetzee et al., 2021), the adoption of asynchronous video interviewing (AVI) and artificial intelligence (AI) has introduced innovative ways to enhance organisational decision-making, efficiency, and global reach (Dunlop et al., 2022; Orji et al., 2025; Stevenor et al., 2024). This section explores two critical technological advancements, AVI and AI, highlighting their unique contributions to modern selection processes.

1.1.4.1 Asynchronous Video Interviewing (AVI)

Asynchronous video interviews have been growing increasingly popular in personnel selection and research contexts (Dunlop et al., 2022; Patel et al., 2025). AVIs provide a convenient alternative to traditional interviews by enabling candidates to record their responses to questions on an online platform, which evaluators can assess later, either manually or through automation, thereby eliminating the need for real-time interaction (Dunlop et al., 2022; Lukacik et al., 2022; Patel et al., 2025). This method of interviewing is attractive to organisations because it offers cost savings, standardisation, and customisation while also providing applicants and selection members with flexibility and convenience (Patel et al., 2025). Aside from assessing job-related competencies through AVIs, they are increasingly used to evaluate applicants' personalities, either by humans or computers (Holtrop et al., 2022; Liff et al., 2024). When these AVIs are scored not by humans but by

computerised automation, it is called automated video interviewing (Hickman et al., 2022; Liff et al., 2024). Thus, recruitment and selection processes continue to evolve, incorporating technological advancements such as AVIs and computerised personality evaluations, as explored in the following section, to improve the identification of the best-suited candidates for organisational needs.

1.1.4.2 Automation in Selection: The Role of AI and ML in AVIs

Technology has supported human resource management (HRM) practices for decades (Langer et al., 2019). More recently, an increasing number of organisations have adopted artificial intelligence (AI) and machine learning (ML) in their recruitment and personnel selection processes, automating tasks such as interpreting interview performance (Goretzko & Israel, 2022; Mirowska, 2020; Woo et al., 2024). AI and ML are also explicitly utilised in scoring asynchronous video interviews (AVIs), enabling the analysis of candidate responses with greater efficiency and consistency. While AI and ML are often used interchangeably to describe computerised automation, they differ in meaning (Kühl et al., 2022). AI broadly refers to algorithms that enable computers to perform tasks requiring human-like intelligence. ML, on the other hand, is a subset of AI where computers learn and adapt their behaviour based on the data they process (Campion & Campion, 2023; Fan et al., 2023; Riedl, 2019).

1.1.4.2.1 Deep learning (DL). Although beyond the scope of this study, deep learning (DL) is noteworthy as it is also utilised in AVI scoring (Lukacik et al., 2022). DL is an advanced machine learning form that imitates how a biological brain autonomously learns and processes data. It extracts and analyses progressively complex patterns through multiple layers of artificial neurons, commonly called "nodes" (Campion & Campion, 2023; Fan et al., 2023; Norton et al., 2024).

1.1.4.2.2 Natural language processing (NLP). A machine learning method particularly relevant to this study is Natural Language Processing (NLP), explored in more detail later. NLP involves analysing textual data by referencing dictionaries of words and their associations (Campion & Campion, 2023). This study focuses on a closed-vocabulary keyword-counting algorithm, which calculates personality trait scores based on the frequency of words associated with specific traits (Holtrop et al., 2022), as discussed throughout the analysis.

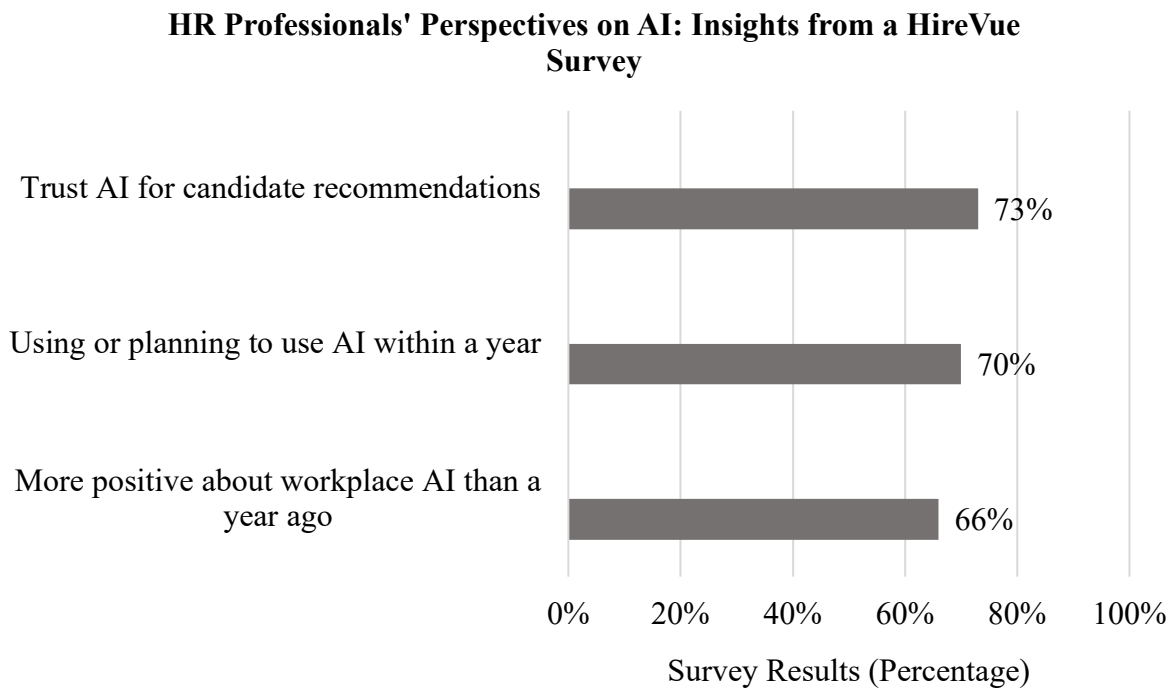
Upon discussing AI and ML scoring methods for AVIs, it becomes apparent that human-based scoring of interview responses can, and is now routinely, complemented and improved by AI algorithms, which appear to streamline the evaluation process (Hickman et al., 2022; Lukacik & Bourdage, 2025; Suen et al., 2019).

1.1.4.3 HR Perceptions of AI in Selection: Insights from HireVue's Survey

To further contextualise the study's focus on human and AI evaluations in selection, it could be helpful to examine broader trends in HR's adoption of AI tools. For example, HireVue, a well-known international company headquartered in South Jordan, Utah, and a leader in AI-driven hiring solutions and asynchronous video interviews (AVIs) (HireVue, n.d.-a; Patel, 2022; Roulin et al., 2021), conducted a survey exploring perceptions of AI in the selection process (HireVue, 2024). The survey included responses from a thousand HR practitioners, and as shown in Figure 1, the findings highlighted the growing acceptance of AI in recruitment processes among HR practitioners (HireVue, 2024).

Figure 1

HireVue 2024 Survey: HR Practioners' Perspectives on AI in Selection



Source: HireVue (2024)

Seventy-three per cent (73%) of HR professionals expressed trust in AI to recommend candidates, with seventy per cent (70%) currently using or planning to use AI within the coming year as part of the hiring process (HireVue, 2024). Additionally, sixty-six per cent (66%) reported increased enthusiasm about incorporating AI into workplace practices compared to the previous year (HireVue, 2024). These results may reflect the broader trend of integrating AI into HR practices, which appears increasingly recognised for its potential benefits (HireVue, 2024).

These survey results feed into a closer look at how AI and human evaluators perform and compare in selection processes, particularly in evaluating personality traits from AVIs. The following section delves deeper into this comparison.

1.1.4.4. Relevance and Current use of AVIs in South Africa

In South Africa, asynchronous video interviews (AVIs), personality assessments, and automated scoring may be highly relevant due to legislative frameworks such as the Employment Equity Act 55 of 1998 and its 2013 Amendment Act, the Protection of Personal Information Act of 2013, and the Health Professions Act 56 of 1974. These laws regulate the use of assessment tools and selection practices within the country (Coetzee et al., 2021). Despite this potential relevance, the adoption of technologies such as AVIs remains slow. Socio-economic challenges including crime, load-shedding, and limited digital skills among employees contribute to organisational hesitation in fully embracing these technologies (Naidu et al., 2025). Additionally, specific statistics on AVI usage in South African workplaces, such as those from HireVue (2024), are not readily available. This indicates a gap in research and organisational adoption of AVIs within the South African context.

1.1.5 Human and AI Evaluations in Selection

As both humans and machines are used to evaluate interview responses, a critical question arises: How do AI assessments compare to those made by humans? Several studies comparing the personality judgments made by AI models and human evaluators have found that computer assessments correlate with those made by humans, as described in more detail later (Hickman et al., 2022; Holtrop et al., 2022; Koutsoumpis et al., 2024; Youyou et al., 2015). This suggests that AI can predict personality traits automatically, without relying on human social-cognitive skills, and potentially as accurately as or better than humans, more quickly and at a lower cost (Hickman et al., 2022; Youyou et al., 2015). Consequently, personality assessment methods increasingly rely on AI evaluation techniques, such as text word counting and advanced natural language processing models, to score candidate responses during selection processes (Jayaraman et al., 2024; Stachl et al., 2020). One of the aims of scoring interview responses is evaluating personality, a strong predictor of applicants' potential job performance (Hickman et al., 2022; Holtrop et al., 2022; Salgado, 2017).

1.1.6 Convergence of Human and AI Personality Judgements

Researchers in this field, such as Hickman et al. (2022), Hinds and Joinson (2024), Holtrop et al. (2022), and Koutsoumpis et al. (2024), have studied the correlation, alignment and rank-order stability between AI and human judgement in personality assessment. AI tools often leverage machine learning techniques to analyse verbal cues transcribed from AVIs.

While advanced ML approaches are continually being developed, text-based analysis has been identified as the most promising and efficient technological advancement in automatically scoring personality traits from AVIs. In contrast, human raters rely on their professional judgement and a broader range of interpretive cues, including body language, facial expressions, tone of voice, and accents, alongside the spoken content. Therefore, the complementary potential of human and AI methods in personality evaluation has been indicated, with AI offering efficiency on one end and humans' professional, nuanced judgement on the other. Exploring the alignment between AI text-based analyses and human evaluators' judgements is instrumental in understanding the evolving role of computerised technology in advancing personality assessment practices. This convergence and its implications are explored in detail throughout the current study. It aims to examine the comparative judgements of human evaluators and AI in scoring HEXACO traits from AVI responses.

1.1.7 Research Scope

While personality judgements of others are frequently made by humans or computers based on various interactions or sources, such as conversations or digital data, the accuracy and reliability of these judgements become especially important in recruitment contexts (Funder, 2012; Goretzko & Israel, 2022; Hinds & Joinson, 2024). As explored by De Kock et al. (2020), effective evaluators, or 'good judges,' whether human or algorithmic, should produce high-quality and accurate ratings. Understanding the correlation between these methods, and their strengths and limitations, may aid in determining how humans and algorithms can collaborate most effectively in selection practices. In line with the concept of the good judge, an inherent requirement of psychological measurement is to ensure reliability and validity in accurately assessing constructs such as personality (Foxcroft & Roodt, 2013; Plouffe et al., 2017). Especially in the South African context, legislation such as the Employment Equity Act mandates using valid and reliable measures to ensure fairness in recruitment (Coetzee et al., 2021). However, although the automated scoring of AVIs is becoming increasingly popular, there is limited psychometric evidence to support its reliability and validity (Hickman et al., 2022). Therefore, the study explores aspects of construct validity, such as convergent (mono-trait) validity, as well as interrater reliability in human and AI personality evaluations from AVIs. This may contribute to the growing body of research in this area.

Convergent validity is assessed by comparing different methods of measuring the same construct and evaluating their agreement (Plouffe et al., 2017). In this study, human and

AI evaluations of the six HEXACO traits for the AVI participants were compared. Additionally, interrater reliability, in this context, refers to the consistency of ratings between human and AI evaluators, ensuring a reliable assessment of the same traits and supporting the accuracy of the HEXACO personality evaluation (Cascio & Aguinis, 2024).

Ultimately, the study investigates the convergence between human evaluations and AI text-to-personality algorithm ratings in assessing personality from asynchronous video interviews, emphasising on the potential efficacy of integrating AI into selection practices.

1.2 Research Problems

Advancements in artificial intelligence and machine learning practices have led to the development of asynchronous video interview scoring and personality assessment systems (Lukacik et al., 2022). These systems utilise AI algorithms to evaluate candidates' performance and personality traits in video-based interviews (Holtrop et al., 2022). However, as the use of such AI technologies increases in recruitment processes, questions may arise regarding the reliability and validity of these practices compared to human evaluations (Koutsoumpis et al., 2024). Holtrop et al. (2022) mention that clear guidelines on the reliability and validity of text-analysis tools and criteria for effective text-analysis solutions are needed to determine appropriate reliability and construct/criterion-related validity coefficients and improve generalisability and implementation in practice. While research and advancements are emerging in this domain, a critical gap exists in understanding the convergence between AI and human scoring methodologies.

1.3 Research Objectives

Building on the literature and background, AI/ML technology demonstrates promising potential in assessing personality traits during selection, predominantly text-based evaluations. This study aims to contribute to organisational research by investigating the effectiveness of integrating artificial intelligence, particularly text analysis, into personality evaluation practices within the context of asynchronous video interviews (AVIs) used in recruitment and selection. Specifically, the study examines the alignment between human evaluator ratings and AI-generated personality ratings from AVIs, aiming to shed light on the practical applications, validity, reliability and implications of AI in AVIs and personnel selection.

Additionally, this study aims to examine the convergence and alignment of scores between human and AI evaluations across the HEXACO traits rather than determining who is 'right' or 'wrong.' Traits with the least convergence may highlight areas where human and AI evaluations diverge significantly, potentially reflecting differences in how personality-related

information is processed after judgement. Conversely, traits with the highest convergence may indicate alignment in evaluations, suggesting an opportunity for AI to complement or, in some cases, substitute human scorers. Whether convergence is low, moderate, or high, these findings shed light on trait-specific patterns of convergence and divergence, contributing to a deeper understanding of the differences in evaluation processes. Such insights may inform the refinement of AI technologies and their integration into recruitment practices, supporting AI as a possible tool to complement or align with human judgement in personality evaluation through AVIs.

1.3.1 Alignment with and Contributions Beyond Prior Research

As discussed in the following section, this research aligns with and builds upon three prior studies on asynchronous video interviews and personality measurement using human and computerised scoring.

First, this study builds on the work of Hickman et al. (2022), who utilised machine learning models to predict Big Five personality traits based on verbal, paraverbal, and nonverbal data from AVIs. While their study demonstrated mixed reliability and validity depending on whether the models were trained on self-reports or interviewer reports, the current research diverges by focusing on the HEXACO model and employing a closed-vocabulary text-based or keyword analysis approach. This narrower scope does not focus on the multimodal machine learning models used by Hickman et al. but rather offers insights specific to text-based evaluation.

Second, the study draws on the findings of Holtrop et al. (2022), who introduced the HEXACO Text-to-Personality (HTTP) technique. This advanced keyword method considers grammar and moderator words, such as 'not,' to indicate the opposite of a specific trait. They compared the HTTP with a more basic keyword or dictionary approach, similar to the one used in the current study. They also compared open- and closed-vocabulary methods for assessing personality from AVIs. The current study used a simpler keyword-counting and closed-dictionary approach, as the advanced HTTP did not significantly improve the construct and criterion-related validity of the more basic approach. Furthermore, their study explored personality evaluations across self-reports and interviewer ratings, particularly focusing on Dutch-language interviews and manual transcription. By addressing the limitations highlighted in their work, the current study employed automatic transcription for interview responses originally in English while remaining grounded in the HEXACO model of personality. The transcription was performed using Microsoft Office 365, which may have helped mitigate the validity and quality issues associated with their manual transcription

(University of Queensland, n.d.). Furthermore, while Holtrop et al. (2022) examined general and trait-activating questions, this research focuses exclusively on general interview questions to evaluate the broader applicability of AI-driven scoring across diverse interview contexts.

Finally, the study builds on the contributions of Koutsoumpis et al. (2024), who explored human and machine learning assessments of Extraversion and Conscientiousness in AVIs using multimodal data (text, voice, and facial expressions). Their findings suggested more substantial validity for trait-relevant questions but highlighted concerns around test-retest reliability. Unlike Koutsoumpis et al. (2024), the present research investigates all six HEXACO traits from general rather than trait-specific questions. To maintain consistency with prior research, the methods Koutsoumpis et al. (2024) used to train human evaluators were adapted and incorporated into the current study, as explored and provided in subsequent sections.

Therefore, this study aims to advance research on personality assessment in asynchronous video interviews by applying the closed-vocabulary technique to English-language interviews with automatic transcription, addressing a limitation in Holtrop et al. (2022), who relied on manual transcription in Dutch. Furthermore, the current study utilises implicit or general interview questions instead of trait-activating questions to evaluate the applicability of AI scoring across a broader range of non-personality-specific questions. Additionally, unlike earlier studies that employed more advanced AI models to predict personality traits, this research examines the six HEXACO traits using only text-based analysis. By addressing the gaps and building on prior research, this study may provide additional insights into the strengths and limitations of AI-driven personality assessments compared to human scorers in AVI contexts.

1.4 Research Questions

This section outlines the primary research questions that guide the investigation into the relationship between human evaluations and AI-generated personality assessments in asynchronous video interviews. By exploring these questions, the study aims to clarify the alignment between human and AI scoring methods, aiming to contribute insights into the effectiveness and implications of using AI in personality evaluation within the context of AVIs.

Question 1: What is the relationship between human evaluator ratings of personality and Artificial Intelligence (AI)-generated personality scores in asynchronous video interviews?

Question 2: How does the consistency of AI-generated personality judgements compare to the inter-rater reliability of human evaluator judgements, and what implications do these findings have for the reliability and variability of personality assessment in asynchronous video interviews?

Question 3: Will human rater judgements of HEXACO (Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience) personality traits correlate positively with the AI/ML-scored personality trait scores?

Question 4: Among the HEXACO personality traits (Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience), which trait exhibits the lowest correlation between human and AI scorers in personality assessments?

1.5 Layout of Dissertation

The introductory chapter provided an overview of the study, emphasising the significance of investigating the convergence of human and AI evaluations of HEXACO personality traits in asynchronous video interviews. The following chapter defines key concepts, explores the theoretical frameworks underpinning the research, and conducts a comprehensive review of the relevant literature. Subsequently, the methods chapter, Chapter 3, details the research design, participants, data collection procedures and the analytical techniques used. Following this, the results chapter presents the findings of the study, including the correlation studies and additional statistical analyses conducted. The discussion chapter interprets these findings, examines their implications, acknowledges the limitations of the research, and offers recommendations for future studies. Finally, the dissertation concludes with a summary of the main insights gained from the research, followed by annexures supporting the study.

Chapter 2: Literature Review

Building on the concepts discussed in Chapter 1, this chapter begins by providing an overview of the study's main focus areas: personality evaluation from asynchronous video interviews (AVIs) within the recruitment and selection process, which can be scored by either humans or computers. Thereafter, the use of assessments in the personnel selection process is discussed, with particular emphasis on personality assessments and interviews as the primary methods. Personality theory and its evaluation, AI-driven personality assessment, and the HEXACO model are then explored. Traditional interviews and technology-driven interviews, namely AVIs, are examined, with particular attention to design factors, including scoring as a key design consideration. Core theoretical frameworks regarding the design and scoring of AVIs, as well as their relevance to the South African context, are described. Based on a comprehensive review of existing literature, the convergence between variables, specifically AI versus human scoring of personality in AVIs, is outlined. Plausible hypotheses are integrated and presented. This chapter concludes with a summary of the main concepts, hypotheses, and variables.

2.1 Introduction: The Present Study

As covered in the preceding chapter, personnel selection processes aim to predict job performance (Salgado, 2017). Central to the personnel selection process is the assessment of applicants, typically conducted through interviews, personality-related measures, and various other assessments, such as cognition, aptitude, interests, and integrity, all aligned with the organisation's needs (Cascio & Aguinis, 2024; Coetzee et al., 2021; Moerdyk, 2022). The present study is particularly concerned with personality evaluation through asynchronous video interviews (AVI), specifically comparing those scored by humans and computers. The approach to scoring, whether by humans, AI, or both, is essentially a key design consideration for AVIs.

The purpose of this chapter is to provide a more extensive review of the existing literature on human and AI personality scoring in the context of AVIs compared to the previous chapter. A significant intention of the literature review is to explore the impact of AVI design factors on personality assessment outcomes as a selection method. This study fundamentally investigates the impact of AVI design factors, specifically scoring, by focusing on two central frameworks: the modular approach proposed by Lievens and Sackett (2017) and the conceptual model of AVI design by Lukacik et al. (2022). With advancements in artificial intelligence and evolving interviewing practices, particularly in the context of AVIs, there is a significant opportunity for further research to enhance our understanding of

how different AVI design choices, such as scoring by AI versus human evaluators, influence personality measurements and scores.

While there have been notable strides in the field, substantial improvements are still required before automatically generated text-to-personality assessments can effectively replace or complement evaluations conducted by human interviewers (Holtrop et al., 2022). As stated earlier, to address these limitations, the present study will build on the work of Hickman et al. (2022), Holtrop et al. (2022), and Koutsoumpis et al. (2024) by conducting a similar study with humans and AI scoring of AVI-based HEXACO personality traits in English, rather than Dutch, using implicit rather than explicit trait-activating interview questions, employing a closed-dictionary text-analysis AI technique, and utilising automatic transcription of the AVIs. This chapter will review relevant literature to contextualise the significance of the frameworks, highlight the overlaps with similar research, and identify the gaps that this study seeks to fill in the current body of knowledge.

2.2 Personnel Assessment and Selection

The personnel selection process is integral to identifying and predicting potential job performance, thereby contributing to organisational success (Salgado, 2017). There is a growing need to understand how to evaluate applicants' knowledge, skills, abilities, and other attributes (KSAOs), determine the most effective assessment methods, attract highly suited candidates, and select those who demonstrate both high performance potential and a strong fit with the recruiting organisation (Potočnik et al., 2021). Interviews, including emerging AVIs, and personality assessments remain two of the most commonly used tools in the selection process for predicting such workplace behaviour (Cascio & Aguinis, 2024; Hickman et al., 2019; Connelly et al., 2022; Liff et al., 2024). Understanding the type of assessments used during personnel selection is important, as each may vary in how it measures individual differences (Casio & Aguinis, 2024). These assessments should be conducted in a consistent, dependable, and relatively error-free manner to ensure that the results are trustworthy and can be used effectively (Casio & Aguinis, 2024). Given that interviews and personality assessments are commonly used in selection contexts, the current study needs to explore how these methods are applied in practice, mainly through AVIs.

2.2.1 Personality in Personnel Selection

As discussed in Chapter 1, personality is a multifaceted psychological construct defined by characteristics in thinking, behaviour, and emotion. It is widely recognised as a key predictor of job performance, satisfaction, and intention to stay (Dai et al., 2022). Personality assessment is possible because of the proven principles that people are relatively

consistent in their behavioural patterns, underlined by traits, and that our behaviours differ, making these patterns measurable (Coetzee et al., 2021). Typical methods of personality assessment at work include interviews, behavioural observations, and a wide range of structured self-report inventories based on various theories, such as Cattell's 16 factors, Jung's personality types, and the Big Five or HEXACO trait models, to name a few (Foxcroft & Roodt, 2013; Cascio & Aguinis, 2024; Coetzee et al., 2021; Moerdyk, 2022). In both research and workplace settings, the HEXACO personality model, which encompasses six broad traits, is particularly relevant (Ashton & Lee, 2007). This model is used in the current study and is described in more detail in the subsequent section.

2.2.1.1 HEXACO Model Overview

The HEXACO model of personality is a six-dimensional framework that expands on the traditional five-factor model, namely the Big Five, with origins in lexical or 'language-personality' studies (Dai et al., 2022; Holtrop et al., 2022; van Kempen, 2024). Developed by Ashton and Lee (2007), it includes the dimensions of Honesty-Humility (H), Emotionality (E), eXtraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O). Honesty-Humility is the key addition, reflecting traits related to counterproductive work behaviour, such as sincerity, fairness, and modesty, which are not fully captured in the Big Five (Ashton & Lee, 2007; Holtrop et al., 2022).

The model has been increasingly used in various fields, including workplaces and research in organisational psychology, as it provides a comprehensive framework for understanding personality in various contexts (Ashton & Lee, 2020; Jayaraman et al., 2024). Consequently, both humans and computers can be trained to evaluate applicants using the HEXACO framework, particularly in the context of employment interviews, and the comparative analysis of their evaluations is attracting increasing interest (Dai et al., 2022; Holtrop et al., 2022; Koutsoumpis et al., 2024).

Although structured self-report inventories are primarily used in industrial/organisational psychology (Coetzee et al., 2021), personality assessments, primarily through interviews, have become common in personnel selection due to the impact of personality on responses to job-related questions (Hickman et al., 2022). In addition to traditional interviews, there is an increasing reliance on computerised technology, such as asynchronous video interviews and artificial intelligence, to support human evaluators in assessing candidates during recruitment and selection (Dunlop et al., 2022; Langer et al., 2023; Mori et al., 2024; Patel et al., 2025), as discussed in the following section.

2.2.2 Interviews in Personnel Selection

Interviews, as a widely used selection method, serve both as a predictive tool for potential job performance and as a means of data collection to enhance organisational success (Cascio & Aguinis, 2024; Coetzee et al., 2021; Cummings et al., 2020). Interviews are often considered an effective assessment tool, generally expected by applicants. They can be conducted either face-to-face, virtually or through an online video platform, engaging with the interviewees with or without a live interviewer (Coetzee et al., 2021; Cummings et al., 2020; Lukacik et al., 2022; Macan, 2009). While there are different types of interviews (Coetzee et al., 2021), this research uses and focuses on asynchronous video interviews in selection contexts. However, for context, traditional interview methods are also briefly explored.

2.2.2.1 Traditional Interviews

The traditional interview is a two-way communication process in which the applicant learns more about the job and the organisation, while the interviewer or organisation gains a better understanding of the applicant in relation to the role (Cascio & Aguinis, 2024). Traditional interviews have been favoured due to their perceived natural interaction compared to other selection methods (Coetzee et al., 2021). Generally, traditional interviewing methods have spanned two poles: structure, standardisation and uniformity on one end, and flexibility and openness on the other, which impacts the types of questions asked by interviewers and the expected responses from the interviewees (Coetzee et al., 2021; Macan, 2009; Platt, 2012). However, traditional real-time interviews are considered time- and resource-intensive, requiring training and human involvement, and are prone to bias (Coetzee et al., 2021; Platt, 2012). Considering these challenges, traditional job interviews are increasingly being complemented by asynchronous video interviews (Koutsoumpis et al., 2024).

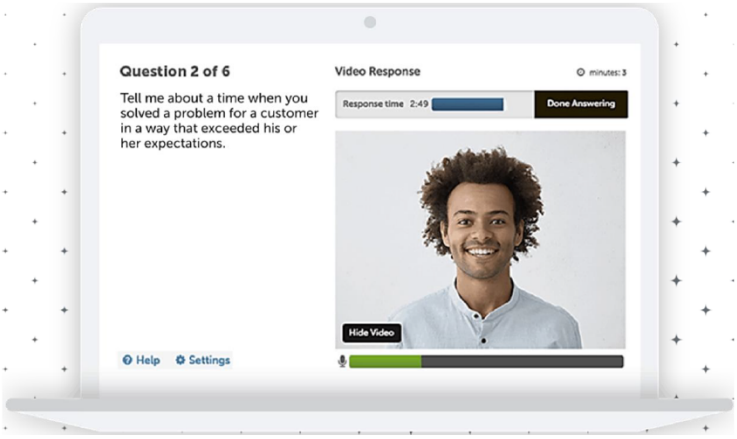
2.2.2.2 Asynchronous Video Interviewing (AVI) Overview

Asynchronous Video Interviewing (AVI) offers an alternative to conventional real-time interviews, providing a consistently available and easily accessible solution (Lukacik et al., 2022). In contrast to traditional interview processes, AVIs require candidates to access an online platform and record video responses to interview questions without any direct interaction with interviewers (Dunlop et al., 2022; Lukacik et al., 2022). These responses are then saved and assessed later, either by an interviewer or a computer algorithm (Dunlop et al., 2022; Lukacik et al., 2022). AVI technology continues to integrate into traditional selection processes, blending human and computerised evaluations (Koutsoumpis et al., 2024), with

companies already offering these solutions in practice. Currently, companies offering platforms for conducting asynchronous video interviews (AVIs) include HireVue, Modern Hire, myInterview, Spark Hire, amongst others (Patel, 2022). AVIs differ in design, and while they are used and offered in practice, they continue to face various challenges that are relevant to the current study, as discussed in the following sections. Figure 2 provides an example of an AVI platform for illustration purposes.

Figure 2

Example of an online or web-based AVI system



Source. HireVue (n.d.-b)

2.2.2.2.1 Approaches to AVI design and scoring. AVIs can be evaluated using both or either human-based and algorithm-based methods (Hickman et al., 2022). There is growing interest in algorithm-based methods (AI / ML) due to their time and cost efficiency, as well as the enhanced standardisation of the evaluation process (Oostrom et al., 2024). The computerised scoring models underpinning AVIs aim to enhance convergence with human ratings (Hickman et al., 2022). These concepts, human and AI scoring, are explored in greater depth further in this chapter and study.

2.2.2.2.2 Challenges in AVI utilisation and design. While AVIs offer practical advantages, further research is needed into the use of AVIs, particularly regarding their design, features, and impact on applicants (Dunlop et al., 2022; Lukacik et al., 2022). Generally, asynchronous media has been associated with candidates viewing the interview process less favourably (Suen et al., 2019). In light of this, prioritising reliability, validity, and the applicant experience in the use of AVIs is crucial to ensure the success thereof in practice (Dunlop et al., 2022). Whether AVI responses are evaluated by human scorers or algorithmically, more research is necessary to improve both applicant experiences and assessment outcomes (Dunlop et al., 2022). Therefore, in the development and refinement process of AVIs, it is

vital that human evaluators remain custodians of the process and uphold professionalism while leveraging the strengths of computerised systems (Hunkenschroer & Luetge, 2022; Xu, 2022).

Approaches to enhancing AVI processes and design features are discussed in the subsequent sections, particularly under the theoretical frameworks in Section 2.3. The next section delves deeper into effective personality measurement through AVIs, laying the foundation for further exploration later in the chapter.

2.2.2.2.3 Theory underpinning successful personality evaluation from AVIs. The effectiveness of personality evaluation from AVIs can be understood through a combination of Social Presence Theory and key principles from Funder's (1995) Realistic Accuracy Model (RAM). Together, these frameworks offer insights into how communication quality and personality judgements can be enhanced in AVI contexts, as discussed in this section.

A possible downside to the use of AVIs is that they typically lack the direct interaction of live interviews, which can make applicants view the assessment method less favourably (Basch et al., 2022; Patel, 2022; Suen et al., 2019). In this context, Social Presence Theory suggests that fostering a sense of connection enhances the quality of communication, thus influencing the quality of candidate responses and, in turn, the data available for scoring (Patel, 2022; Rizi & Roulin, 2024). This is important to consider in the current study when using and scoring personality traits from AVIs. However, this limitation can be mitigated by adding human features such as video-based introductions or engaging question formats, which enhance applicants' sense of engagement and create a more immersive and connected experience (Lukacik et al., 2022; Patel, 2022).

The Social Presence Theory aligns with Funder's RAM, which identifies four essential stages for accurate personality judgements: relevance, availability, detection, and utilisation of behavioural cues (De Kock et al., 2020; Funder, 1995). In the context of AVIs, RAM highlights the importance of both the "good judge" (evaluator) and the "good information" (available data) (De Kock et al., 2020). Human judges excel at interpreting a broad range of behavioural cues, including nonverbal signals, while automated systems, as used in this study, provide consistent, text-based evaluations but are constrained by their inability to process nonverbal behaviours (Cummings et al., 2020; Holtrop et al., 2022; Koutsoumpis et al., 2024). Therefore, as RAM and Social Presence Theory suggested, the interplay between evaluator type and data quality is crucial to the success of personality assessments in AVIs.

RAM is discussed further in this dissertation, specifically in section 2.3 on theoretical frameworks, focusing on its application to the human training of personality judgements. For now, its inclusion complements Social Presence Theory by illustrating how the quality of both the evaluator and the available data influences outcomes. Together, these frameworks emphasise the importance of designing AVIs that provide rich behavioural information and support practical evaluation, whether human or automated.

2.2.3 Summarising Personality, Interviews, AVIs, and AI in Assessments

Before delving into the theoretical frameworks for AVIs and comparing AI and human scoring, it is beneficial to summarise the key insights on personality, interviews, AVIs, and AI in assessments. These concepts will be explored further in the subsequent section on theoretical frameworks, providing a foundation for the discussion to follow. While traditional face-to-face interviews provide a more in-depth understanding of candidates (Salgado, 2017), they can be time-consuming and resource-intensive. This has led to a growing interest in technologies that automate the interpretation of candidate responses (Mirowska, 2020). With the rise of such technologies, methods like asynchronous video interviews (AVIs) and automatic scoring are becoming increasingly common (Basch et al., 2021; Hickman et al., 2022). Specifically, personality assessments from interviews, particularly AVI-based personality assessments (AVI-PA), are becoming more feasible, as personality traits and individual differences are often expressed through language, which both computers and humans can be trained to assess (Fast & Funder, 2008; Holtrop et al., 2022; Koutsoumpis et al., 2024). Automatic AVI-PA enables organisations to evaluate candidates more efficiently and potentially without constant human involvement (Hickman et al., 2022; Stevenor et al., 2024). As a result, there is a growing need for collaboration between traditional recruitment practices and the fields of computer and data science to thoroughly investigate and understand the impact of technology on the recruitment and selection lifecycle (Nikolaou, 2021). This theme is explored further in this study.

2.3 Theoretical Frameworks for AVIs: AI and Human Scoring

The theoretical underpinnings of this research draw on various frameworks that emphasise structured and modular approaches to the design and evaluation of asynchronous video interviews. Frameworks such as the modular approach and response evaluation consistency highlight how pre-interview decisions and standardised scoring methods shape the interview process and candidate responses (Lievens & Corstjens, 2018; Lievens & Sackett, 2017). Central to this standardisation is the training of human evaluators, which follows evidence-based models such as the Realistic Accuracy Model (Funder, 1995),

Brunswik's Lens Model (1956), and frame-of-reference (FOR) training (Bernardin & Buckley, 1981). As discussed in the next section, these models help ensure that evaluators share a common understanding of the assessment criteria and scoring, providing a foundation for accurate and consistent evaluations. The integration of these elements is further detailed in the subsequent section.

2.3.1 Training of Human Evaluators

Interviews are often conducted by individuals with minimal or no training, even though research shows that training improves the quality of this assessment method (Camp et al., 2011). The following research suggests that rater training should adopt a comprehensive approach, ensuring standardisation in evaluating personality using models such as HEXACO. Specifically, evaluator training typically incorporates frame-of-reference (FOR) training to align evaluators by defining performance assessment criteria, providing examples, and conducting practice sessions with feedback (Roch et al., 2012; Herde & Lievens, 2023). Supporting theories and approaches include Funder's (1995) Realistic Accuracy Model and Brunswik's (1956) Lens Model. These approaches are discussed further in the following sections.

2.3.1.1 Frame of Reference Training (FORT)

Frame of Reference Training (FORT), developed by Bernardin and Buckley (1981), aims to improve evaluators' assessment accuracy and teach new raters adaptable rating frameworks, making information processing more efficient (Roch & O'Sullivan, 2003; Martin, 2019). Establishing and following a standardised approach to observing behaviour helps raters identify effective and ineffective performance more easily while ensuring consistent evaluations (Bernardin & Buckley, 1981). The effectiveness of FORT lies in guiding raters to assess performance based on key behaviours, helping them compare ratings to established norms and reach consensus (Bernardin & Buckley, 1981). Generally, FOR-trained raters provide more accurate ratings than untrained ones, proving its effectiveness (Roch & O'Sullivan, 2003).

The primary objective of FORT is to prepare raters to effectively observe behaviour and evaluate performance (Lievens & Sanchez, 2007). The standard FORT approach adheres to the principles of exposure, practice, and feedback and can also be understood through the perspective of the schema-based theory of learning (Lievens & Sanchez, 2007; Martin, 2019). This theory explains that a schema, shaped by existing knowledge and beliefs, is a mental framework that prioritises key details, filters out irrelevant ones, and evolves by integrating new information (Gorman & Rentsch, 2009). Therefore, FORT aims to provide raters with

more specific and suitable schemata better aligned with their rating context than their existing mental frameworks (Lievens & Sanchez, 2007).

2.3.1.1.1. FORT overview. Initially, the multifaceted nature of the construct, such as personality, is highlighted, with each aspect (e.g., the HEXACO model and its traits) clearly defined (Martin, 2019). Subsequently, illustrative behavioural scenarios, called vignettes, are presented to represent each aspect (Martin, 2019; Powell & Goffin, 2009). In the exposure phase, trainers clarify how to assess each vignette accurately (Martin, 2019). During the practice stage, trainee raters can evaluate the vignettes (e.g., HEXACO traits) and justify their assessments (Martin, 2019; Powell & Goffin, 2009). Trainers then provide feedback on the accuracy of the trainee raters' evaluations (Martin, 2019; Powell & Goffin, 2009). As part of this feedback process, trainers communicate more accurate ratings for each vignette based on expert findings and explain the rationale behind these ratings (Martin, 2019; Powell & Goffin, 2009).

2.3.1.2 Realistic Accuracy Model

Complementing FORT, The Realistic Accuracy Model (RAM) highlights the complexity of personality assessment, extending beyond simple definitions to complex issues regarding the validity of personality traits and measurement thereof (Funder, 1995; Powell & Goffin, 2009). Successful judgements are accurate and result from a combination of a capable judge, an understandable target, sufficient accessible information and evident traits or characteristics that are readily noticeable (Chen et al., 2018). Therefore, according to RAM, accurate personality judgement entails the individual exhibiting relevant behaviour, which must be noticeable to the evaluator, who must then appropriately recognise and interpret the cues to infer the target's personality correctly (Powell & Goffin, 2009).

Therefore, the accuracy of personality evaluation relies on the relevance, accessibility, detection, and utilisation of behavioural cues (Funder, 1995).

2.3.1.3 Lens Theory

Additional to FORT and RAM, Brunswik's Lens Model or Theory has been widely used in studying cue relevance, utilisation, and judgement accuracy across domains, including personality traits (Martin, 2019). This theory suggests that when people judge personality, they rely on available cues from the target, which may vary in relevance (Martin, 2019). Accurate judgements occur when the perceived cue attributes are of high quality, match reality, and are relevant to the specific characteristics being studied (Martin, 2019; Mosier & Kirlik, 2004). Therefore, as described by Brunswik (1996) and in line with RAM, in the present study, the lens model may explain how decisions are made regarding

personality by interpreting the available cues, forming judgements, and evaluating their accuracy. Accurate personality evaluation relies on the target displaying relevant cues, detectable by the evaluator, and correctly utilised for assessment (Letzring et al., 2006).

2.3.1.4 Human Evaluator Training Frameworks Integrated

The training of human evaluators typically follows an integrated and standardised approach to ensure consistency and accuracy in personality assessments, which has been shown to improve the accuracy of ratings for most traits (Lievens & Sanchez, 2007; Powell & Goffin, 2009). Central to this process is Frame-of-Reference Training, which appears to enhance evaluators' accuracy by aligning assessment criteria, providing behavioural examples and feedback. This method is underpinned by schema-based learning theory, which equips evaluators with refined mental frameworks for processing and interpreting personality-related information. Complementing FORT, Funder's Realistic Accuracy Model (RAM) and Brunswik's Lens Theory aim to provide theoretical foundations to understand the complexities of personality evaluation. RAM emphasises the interplay of cue relevance, detectability, and utilisation in forming accurate judgements, while Lens Theory focuses on the evaluator's ability to interpret and apply observed behavioural cues effectively. Together, these frameworks could help and guide trainers and evaluators to achieve higher accuracy in personality assessment by addressing both the practical and theoretical aspects of the evaluation process.

2.3.2 Modular Framework

Lievens and Sackett (2017) suggest enhancing selection procedures by adopting a modular approach to the core components of the selection process. This approach, also called modularity, involves assessing the extent to which smaller elements of a system, such as selection processes, can be separated, reconstructed, and used to enhance the system's effectiveness (Lievens & Sackett, 2017). This approach aims to improve the selection process by examining the effects of its measurement components, such as personality assessments, on the results and validity of the process (Lievens & Sackett, 2017). By breaking selection procedures into components, it is possible to better identify shared elements and differences across various methods, as well as determine how and whether the process can be enhanced (Lievens & Sackett, 2017). Therefore, a modular approach allows for the development of new selection procedures by adapting, combining and integrating different components (Lievens & Corstjens, 2018). Central to this concept is predictor constructs and predictor methods, because they form the key components of selection procedures (Lievens & Sackett, 2017).

2.3.2.1 Predictor Constructs and Predictor Methods

Predictors in personnel selection refer to specific behavioural domains, from which information is gathered using a particular method, and can be understood through their components: predictor constructs and methods (Arthur & Villado, 2008).

As noted by Lievens and Sackett (2017), predictor constructs are the psychological traits or attributes assessed during a selection process. These constructs are often organised into frameworks, such as personality traits (Arthur & Villado, 2008; Lievens & Sackett, 2017). On the other hand, predictor methods are the specific techniques used to collect information about these constructs, such as questionnaires and interviews (Arthur & Villado, 2008; Lievens & Sackett, 2017). Therefore, while predictor constructs define what is measured, predictor methods focus on how that information is gathered (Arthur & Villado, 2008; Lievens & Sackett, 2017). These methods can be further broken down into smaller components known as "predictor method factors," such as the response format, scoring approach and response evaluation consistency, which help distinguish different selection tools (Lievens & Sackett, 2017). Predictor constructs, predictor methods and factors are important concepts within the modular approach to selection processes and beneficial for conceptualising the current study.

2.3.2.2 Relevance to the Current Study

In the context of the current study, which examines human and AI evaluations of personality in asynchronous video interviews, the modular framework supports the exploration of how specific components, such as scoring approaches (predictor method factor), affect personality (predictor construct) ratings from AVIs (predictor method). By isolating and studying these components, the study could help drive further innovation and refine the methods used in selection procedures. The insights gained from the scoring approach (human and AI) could potentially inform and improve one another, ultimately contributing to more effective and integrated selection practices, particularly in personality assessment from AVIs. In particular, factors related to predictor methods and scoring, as outlined in Annexure A, are summarised and adapted from Lievens and Sackett (2017). Notably, the predictor method factor, namely Response Evaluation Consistency, discussed in the following section, is particularly relevant to this study.

2.3.2.3 Response Evaluation Consistency

Within the Modular Framework, Response Evaluation Consistency refers to standardising scoring responses (Lievens & Corstjens, 2018). The scoring process can be either unconstrained, where evaluators do not have clear criteria or expected answers, or

more calibrated, where evaluators are trained to apply predetermined answers and/or evaluative criteria when assessing candidates (Lievens & Corstjens, 2018; Lievens & Sackett, 2017). In the context of this study, calibration is used, where human evaluators, through their training as described in Section 2.3.1 and Chapter 3, are provided with scoring guidelines, and AI is typically programmed to score in a predefined manner, such as the closed-dictionary approach (Holtrop et al., 2022). Automated scoring is expected to represent the highest level of standardisation in response evaluation and can improve the reliability and validity of the process (Lievens & Corstjens, 2018). However, the rating consistency of human raters can be enhanced through rubrics and training (Lievens & Sackett, 2017).

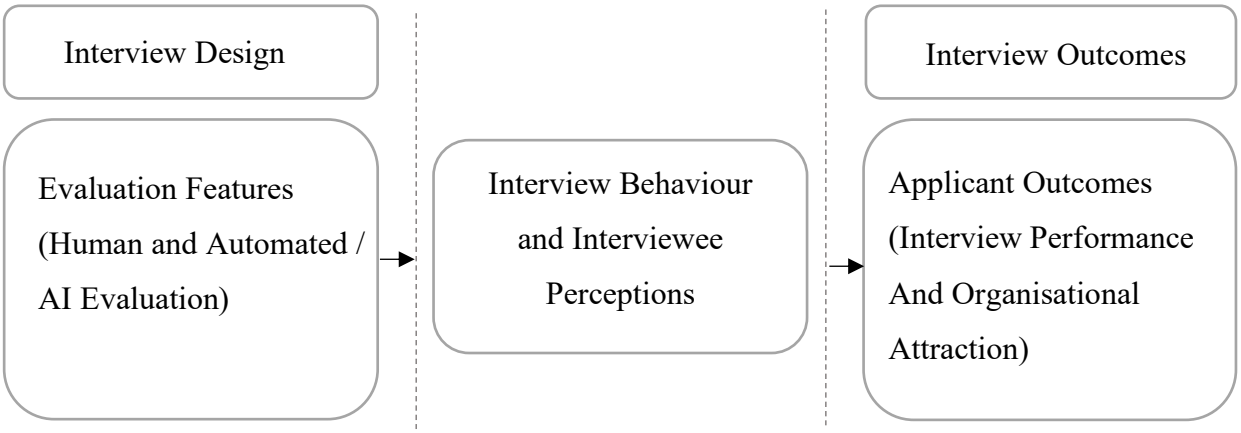
Therefore, these adjustments in interview scoring, such as automation, training, and rubrics, can enhance consistency and promote further innovation (Lievens & Corstjens, 2018). This is particularly relevant for examining the convergence between human evaluators and AI in AVIs on personality, as it may improve our understanding of these methods and the comparability and reliability of these assessments.

2.3.3 Alternative Conceptual Model of AVI Design

Lukacik et al. (2022) propose a framework for understanding factors related to AVI design, focusing on pre-interview design decisions and their impact on candidates' perceptions and interview outcomes. Like the modular approach, their AVI design model emphasises how choices made during the pre-interview phase, such as using human or computerised scoring and other technology-related decisions, can impact applicant outcomes and performance (Lukacik et al., 2022). Although not the primary focus of this study, such design decisions may also influence candidates' overall perceptions of the assessment method, which could affect their interview performance, results and organisational attraction. This process is shown in Figure 3 below, highlighting how the AVI evaluation method could impact the interview outcomes, as further discussed in section 2.4.

Figure 3

Conceptual or theoretical framework illustrating how pre-interview AVI design influences post-interview results.



Source. Adapted from Lukacik et al. (2022).

2.4 AVI Scoring: Human vs. AI Evaluators

Effective evaluators, the "good judge", produce high-quality ratings, defined by their accuracy in assessing attributes against specified criteria (De Kock et al., 2020). Only human evaluators have traditionally scored and evaluated interview performance, including asynchronous video interviews (Koutsoumpis et al., 2024). However, organisations increasingly incorporate AI technologies to complement and enhance decision-making in human resource (HR) practices, such as recruitment and selection (Hickman et al., 2024).

Currently, both human evaluators and artificial intelligence play essential roles in evaluating and scoring asynchronous video interviews (Suen et al., 2019), each influencing the assessment outcome, as illustrated in Figure 3, particularly about the evaluation features within AVI design (Lukacik et al., 2022). The ability to automatically infer personality traits from responses to job interview questions can replace traditional time-consuming personality assessments (Dai et al., 2022). However, AI and human evaluators have advantages and limitations, making it essential to consider their integration in the ongoing quest to find and establish 'the good judge' (De Kock et al., 2020; Xu, 2022). The following section elaborates on some of these advantages and limitations and the application of our two evaluators of interest: humans and artificial intelligence.

2.4.1 Human Judgements of Personality from Interviews

Currently, the primary decision-maker in recruitment and assessments remains human and will continue to be crucial in the foreseeable future (Hmoud & Várallyai, 2019; Langer et al., 2019). Research shows that human evaluators typically produce high-quality results,

detect subtle or nuanced personality traits and adapt well to new applications or contexts, compared to computer-based evaluations (Cummings et al., 2020; Prinzing et al., 2024). However, humans are also susceptible to biases that can distort data (Cummings et al., 2020).

Effective interviews require adequate training for both interviewers and evaluators (Holtrop et al., 2022; Koutsoumpis et al., 2024). Human involvement in interviews can be time-consuming, and individuals differ in their ability to predict outcomes and personality traits from interviews or written data (Holtrop et al., 2022). Therefore, it is clear that human involvement in interviews, though often preferred, come with challenges including time constraints and potential biases.

Since humans create algorithms, human biases can be embedded in AI algorithms during coding, potentially causing AI systems to exhibit those biases (Koutsoumpis et al., 2024; Soleimani et al., 2022). As per subsequent sections, AI may offer efficiency and impartiality in selection decisions, but still lacks genuine human understanding and judgement. AI/ML, as a supplementary tool to human decision-making in recruitment and selection, remains plausible and promising (Goretzko & Israel, 2022; Neumann et al., 2023).

2.4.2 AI/ML Judgements of Personality from Interviews

It is clear that personality assessment can be conducted through various methods (Jayaraman et al., 2024), with artificial intelligence and machine learning emerging as some of the most significant innovations in personnel selection since the start of employment assessments (Campion & Campion, 2023). Words people use reveal rich insights into their personality, which are commonly analysed and studied through text, typically referred to as content coding, content analysis, or text analysis (Prinzing et al., 2024). More than eight years before the current study, Campion et al. (2016) demonstrated the potential of computerised scoring for competencies from written data, showing that computers can achieve reliability comparable to human raters, with evidence of construct validity and significant cost savings (Cascio & Aguinis, 2024). Currently, seventy-nine percent of employers already utilise some form of AI in their recruitment and selection processes (Mori et al., 2024), with more than two-thirds of international HR professionals reporting plans to use AI in 2025 (HireVue, 2024).

2.4.2.1 Text Analysis: Closed- and Open-Vocabulary

Text-analysis techniques, particularly the text-to-personality method, still hold considerable promise for automatically evaluating job-relevant psychological traits during interviews (Holtrop et al., 2022). These algorithms analyse text derived from AVIs, whether transcribed manually or through automated means (Holtrop et al., 2022). Automatic

transcription can be accomplished using speech-to-text software, such as Google Cloud's speech-to-text (Koutsoumpis et al., 2024) or the transcription facility in Microsoft Word (University of Queensland, n.d.).

The advancements in AI and ML have enhanced our understanding and methods of analysing candidates' personalities; however, they are yet to be perfect (Goretzko & Israel, 2022; Hinds & Joinson, 2024). Currently, open- and closed-vocabulary techniques are commonly used to obtain text-based personality scores from AVIs, with the closed-vocabulary approach relying more heavily on human judgement than the open-vocabulary method (Holtrop et al., 2022).

2.4.2.1.1 Closed-vocabulary algorithms. Closed-vocabulary algorithms allocate words to psychosocial categories, such as personality, to create dictionaries for scanning text, counting keyword frequencies, and generating scores (Eichstaedt et al., 2021; Holtrop et al., 2022).

2.4.2.1.2 Open-vocabulary algorithms. Open-vocabulary methods, on the other hand, are more advanced data-driven algorithms that identify clusters of words from large linguistic datasets without relying on predefined personality-related words and categories (Eichstaedt et al., 2021; Holtrop et al., 2022).

Typically, to optimise the efficacy of AI techniques, substantial and diverse datasets are necessary for personality evaluation (Holtrop et al., 2022). Machine learning algorithms developed for AVIs aim for maximal convergence by assigning weights to specific indicators to optimise alignment with human-reported personality evaluations (Hickman et al., 2022). AI/ML algorithms often demonstrate higher inter-judge agreement than human evaluators when assessing the same personalities (Youyou et al., 2015).

While AI offers exciting prospects for personnel selection, it is not without faults (Goretzko & Israel, 2022; Hickman et al., 2022). For example, AI systems can demonstrate biases like humans when they derive knowledge from potentially biased or incomplete data sources (Xu, 2022). Algorithms trained on such datasets may reflect these biases, impacting their predictions (Koutsoumpis et al., 2024). It is also highly likely that AI may miss certain subtleties if not adequately trained on the nuances that humans may detect and use in their evaluations, which is particularly relevant in the closed-dictionary approach (Cummings et al., 2020; Hickman et al., 2022; Holtrop et al., 2022; Prinzing et al., 2024). Only when trained effectively can AI/ML algorithms identify patterns within datasets that may elude human detection (Stachl et al., 2020).

Therefore, while the ability of AI to render judgements on par with or better than human evaluators seems promising, the training of such systems, particularly on large,

representative datasets, as well as further education and training of human resource personnel, appear essential to support HRM processes such as recruitment and to mitigate potential human shortcomings such as biases (Goretzko & Israel, 2022; Hinds & Joinson, 2024; Lukacik et al., 2022; Mirowska, 2020; Riedl, 2019). Language-based analysis of personality appears most promising in the quest to introduce and enhance AI in recruitment and selection (Holtrop et al., 2022; Koutsoumpis et al., 2024), as discussed in the following section.

2.4.2.2 Language Analysis and Natural Language Processing (NLP)

Language use reflects our personality and individual differences (Holtrop et al., 2022). Therefore, AVIs can be analysed to identify key phrases and evaluate people's specific language (Lukacik et al., 2022). Research indicates a strong connection between language use and personality traits, with many studies comparing linguistic patterns with self-reported or observed behaviours (Park et al., 2015). Consequently, a key aspect of this study is the application of text-to-personality language analysis techniques to assess personality traits.

2.4.2.2.1 Natural Language Processing (NLP). Natural Language Processing (NLP) is a field of AI that enables computers to understand, interpret, and generate human language, making it instrumental in analysing text data and facilitating fair assessments of candidates' responses (Campion & Campion, 2023). However, training with criterion data is required (Campion & Campion, 2023).

Several NLP approaches have been applied in both research and practice. Recently, transformer-based models, such as XLNet, have gained popularity for text-based personality prediction and other NLP tasks due to their superior contextual understanding (Jayaraman et al., 2024). Holtrop et al. (2022) developed the HEXACO text-to-personality technique (HTTP) based on the HEXACO personality assessment to evaluate personality traits during job interviews, demonstrating its validity and reliability. These text-analysis techniques represent promising advancements in assessing job-related psychological attributes during interviews (Holtrop et al., 2022; Koutsoumpis et al., 2024). However, Hickman et al. (2024) argue that while automated language-based personality inference can streamline the recruitment process, decision-makers should balance its convenience against its moderate accuracy in replicating human ratings and its potential to amplify existing biases.

2.4.2.2.2 Generative Pre-trained Transformers (GPTs). Although beyond the scope of the current study, recent psychological research has also examined Generative Pre-trained Transformers (GPTs) in text analyses and within the AVI context. Prinzing et al. (2024)

compared GPTs to humans in text analyses, with the models being untrained while the humans were trained, and found that GPT-based analysis is as valid and reliable as human-based text analysis. Another study found that GPTs, as a form of AI, may threaten the validity of AVIs as a selection assessment (Canagasuriam & Lukacik, 2025). Individuals who used AI assistance to 'cheat' in their interview responses achieved significantly higher overall performance scores than those who did not rely on AI (Canagasuriam & Lukacik, 2025).

While interesting and alarming, these studies highlight the potential and challenges of using language analysis in psychological assessment, and the need for further research to integrate AI effectively into human-dominated selection practices.

Drawing on the reviewed literature, the following hypothesis was proposed as a tentative answer to the first and second research questions:

H1a. A positive relationship exists between human evaluator ratings and AI algorithm ratings in evaluating each HEXACO personality trait from AVIs.

2.4.2.3 Assessment Legislation and the Relevance of AVIs and AI Scoring in South Africa

Artificial Intelligence (AI) involves using digital computerised technology to perform tasks that typically require human intelligence, facilitating data processing, pattern recognition, and task automation (Mori et al., 2024; Riedl, 2019). While AI may offer advantages over human evaluations, such as reducing human bias in scoring and providing timely feedback, it also raises ethical concerns, including privacy issues, algorithmic imperfections, and challenges related to transparency (Hunkenschroer & Luetge, 2022). Given the limited international regulation on the use of AI, Hunkenschroer and Luetge (2022) recommend that companies establish internal standards for its ethical use in recruitment, ensuring compliance with privacy laws, maintaining transparency, and incorporating human oversight.

In South Africa, interviews, AVIs and personality assessments, as well as their scoring, whether human or AI-based, are highly relevant due to regulations such as the Health Professions Act 56 of 1974 (HPA), the Employment Equity Act 55 of 1998, the Employment Equity Amendment Act of 2013, and the Protection of Personal Information Act of 2013 (POPIA). Unlike international contexts, the HPA mandates that only qualified and registered psychological professionals may administer, score, interpret, and communicate assessment findings, ensuring ethical standards and participant protection (HPCSA, 2008). Complementing this, the Employment Equity Act requires that selection methods demonstrate validity, reliability, and fair application, ensuring no group or individual is

unfairly disadvantaged (Foxcroft & Roodt, 2013). Additionally, POPIA governs the lawful processing of personal data, requiring consent, transparency, and robust data security measures to protect individuals' privacy throughout the assessment process (Coetzee et al., 2021). Internationally, Booth et al. (2021) also highlighted the importance of accuracy, fairness, and bias elimination when developing AI or ML models for high-stakes recruitment scenarios.

Therefore, regarding AI use in South Africa, clear guidelines for AI in assessments are not yet available. The Department of Communications and Digital Technologies (DCDT) has developed the South African National Artificial Intelligence Policy Framework (Towards the Development of South Africa National Artificial Intelligence Policy), which outlines the country's approach to and appropriate use of AI (Department of Communications and Digital Technologies, 2024). This policy framework highlights key areas of focus, particularly the use of AI to improve employment opportunities and foster economic growth. However, its implementation in organisational areas such as recruitment and selection remains unclear. Nonetheless, it may be important for South African practitioners to continuously ensure that whichever assessment methods are used in recruitment and selection align with existing legislation and regulations. As extensively discussed by Oosthuizen (2022), Industrial / Organisational (I/O) Psychologists, as psychological practitioners in the workplace and in line with the Health Professions Council of South Africa's (HPCSA) scope of practice, find themselves at the forefront of the rise and revolutionisation of AI in the workplace and must ensure its effective, legitimate, and reliable application. This brings the focus back to the alignment of human and AI scoring, as explored in the following section.

2.4.3 Convergence, Interrater Agreement, and Methodological Consistency

In evaluating the alignment between human and AI-based scoring, concepts such as construct validity, convergence, rank-order stability, interrater agreement, and methodological consistency emerge.

Construct validity refers to the degree to which a test or instrument accurately measures the theoretical construct it was designed to assess (Foxcroft & Roodt, 2013; Serapio-García et al., 2023). In the current study context, it ensures that the personality assessments, whether conducted by human raters or AI, accurately reflect the specific traits defined by the HEXACO model. Convergence, as a facet of construct validity, refers to the extent to which different methods, such as assessments made by human raters and those made by AI, produce similar results when evaluating the same constructs, specifically the HEXACO personality traits in this study (Cheung et al., 2024). Additionally, rank-order

stability in this context also refers to the consistency or correlation of personality traits over time or across different evaluators (Henry et al., 2024; Wortman et al., 2012). Notably, from previous research, the strongest correlation is observed between human and AI measures of extraversion, likely due to extraversion being the most recognisable personality trait within the HEXACO framework (Holtrop et al., 2022). It was therefore proposed that:

H1b. Convergence between AI and observer personality ratings will be strongest for extraversion, compared to other personality traits.

Interrater agreement, as a measure of interrater reliability, emphasises the level of consistency or similarity between the scores provided by different raters, which is important for ensuring that the methods are comparable (Cascio & Aguinis, 2024). Therefore, it is important to highlight the importance of producing comparable outcomes despite differing scoring methodologies. By examining these aspects, this study aims to assess the validity of AI as an alternative to human scoring.

2.5 Conclusion and Research Hypotheses

In conclusion, this study explores the alignment between human evaluator ratings and AI-based assessments of personality traits in asynchronous video interviews (AVIs), recognising the crucial role of personality in predicting job performance. As AI becomes increasingly integrated into personnel selection, mainly through AVIs, understanding the potential for consistency between AI algorithms and human evaluations is vital. This research focuses on assessing that alignment, explicitly emphasising the HEXACO personality traits. Therefore, as explored further in the following chapter, to analyse the alignment between the scoring methods, the independent variable is the human evaluators' scores. In contrast, the dependent variable is the AI algorithm scores derived from AVI text analysis. The following hypotheses were also proposed earlier in the chapter as important for testing in the study.

Null Hypothesis (H_0): There is no significant relationship between human evaluator ratings and AI algorithm ratings in evaluating each HEXACO personality trait from AVIs.

Alternative Hypothesis (H_{1a}): A positive relationship exists between human evaluator ratings and AI algorithm ratings in evaluating each HEXACO personality trait from AVIs.

Alternative Hypothesis (H_{1b}): The convergence between AI and observer ratings of personality will be strongest for extraversion compared to other personality traits.

Chapter 3: Research Methods

This chapter outlines the methods and steps the researcher used to gather empirical data for testing the hypotheses. It includes an overview of the research design, study components, population and sample, data collection process, integrated ethical considerations, data capture, data preparation, and statistical analysis.

3.1 Research Design

This study adopted a quantitative approach, utilising a within-subjects post-test only experimental design. In this design, also known as a repeated measures design, each participant is exposed to all independent variables and conditions under investigation (Charness et al., 2012; Rosenthal & Rosnow, 2008). Specifically, participants' AVIs are evaluated for personality by human scorers and an AI text-to-personality algorithm developed for this study. This design enables the comparison of scores across different measurement conditions (Rosenthal & Rosnow, 2008). The primary focus of the research is to compare the ratings provided by human evaluators with those generated by the AI algorithm. This necessitates a within-subjects design that allows each participant to be evaluated using both scoring methods. The human evaluators score the participants' HEXACO personality traits based on the AVI text transcripts. At the same time, the AI algorithm also evaluates the same transcripts according to the HEXACO model and its traits. These study components are discussed in greater detail in the following sections.

3.2 Components of the Study

The components of the study, outlined below, include two approaches or methods to evaluating personality from AVI transcripts. The first component involves human evaluators, trained in personality assessment using the HEXACO model. The second component involves an AI algorithm to analyse the AVI transcripts and score or evaluate HEXACO personality traits. These two approaches will be further elaborated upon in the following sections.

3.2.1 Human Evaluators

The human evaluators consist of post-graduate students in Industrial and Organisational Psychology at the University of Cape Town (UCT), who received extensive training in AVI personality assessment (AVI-PA) techniques relevant to this study. While the original intention was to include only honours degree students, the decision was later made to extend participation to master's degree students to enhance the diversity of perspectives and expertise within the evaluator group. The evaluation group comprised 15 individuals with varying expertise and exposure in Industrial and Organisational (I/O) Psychology. Thirteen of

the evaluators were honours degree students specialising in I/O Psychology, while the remaining two were enrolled in a professional master's degree programme in I/O Psychology. The latter were registered as student psychologists with the Health Professions Council of South Africa (HPCSA) at the time of evaluation, with one evaluator additionally holding a Doctorate-level degree (PhD) in I/O Psychology. Although there is no clear indication in the literature of whether rater experience or age significantly impacts rating quality (De Kock et al., 2020), this combination of academic qualifications and experience may have provided a strong foundation for insightful AVI personality evaluations.

As indicated, postgraduate students were chosen as the human evaluators because, according to the literature, they can perform well with training and may even outperform some recruiters (De Kock et al., 2020; Petersheim et al., 2022; Powell, 2008). While experienced recruiters may excel in certain aspects, such as deception detection, students demonstrate nearly comparable proficiency in evaluating personality traits and even surpass recruiters in traits like extraversion, conscientiousness, and openness (Schmid Mast et al., 2011). Therefore, well-trained students, as Powell (2008) guided, can accurately assess overall personality characteristics based on interview data. The primary approach to rater training is frame-of-reference (FOR) training, which aims to standardise evaluations and reduce rater bias by aligning raters' understanding of assessment criteria and enhancing the quality and accuracy of scores (De Kock et al., 2020; Herde & Lievens, 2023; Powell & Goffin, 2009).

Student human evaluators were therefore trained on how to conduct personality assessments using the HEXACO model and given comprehensive training in understanding AVIs and how to evaluate these, following a frame-of-reference (FOR) training technique, provided in more detail later. This type of training typically provides evaluators with consistent benchmarks and definitions to improve their scoring accuracy (Herde & Lievens, 2023; Powell, 2008). The evaluators scored each interview based on specific criteria to identify key personality traits from the AVIs. Training materials, including examples, practice evaluations, and feedback sessions, are provided and included in Annexures E and F, primarily acquired and adapted from Koutsoumpis et al. (2024) to ensure that the evaluators' training and ratings align with previous research, the expected evaluation process and personality dimensions. Another study component is the AI algorithmic scoring approach, which is discussed below.

3.2.2 AI Text-to-Personality Algorithm

The AI HEXACO text-to-personality algorithm used in this study builds on the work of Holtrop et al. (2022) and Koutsoumpis et al. (2024). It applies the HEXACO personality model as the assessment framework to estimate personality traits from AVI transcripts. The algorithm assesses all HEXACO traits: Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience by analysing the content (text or words) of the transcripts, aligning with the closed-dictionary approach (Eichstaedt et al., 2021; Holtrop et al., 2022). The algorithm is theoretically grounded, as previous research shows correlations with self-reported and interviewer-rated personality evaluations (Holtrop et al., 2022; Koutsoumpis et al., 2024; Stevenor et al., 2024). See Annexure B for a demonstration and explanation of the AI algorithm.

3.2.2.1 Text Analysis with Python

The text analysis was coded and conducted using Python, which is widely preferred for data analysis and automation due to its ease of use, versatility, and extensive library (additional tools) support (Soliev et al., 2023). Specifically, the Regular Expression ('re') package and the Python built-in text processing functions were used for tasks such as string manipulation (or text preparation) and data preprocessing (Chapman & Stolee, 2016; Stenegren, 2023). These were used alongside the OpenPyXL and Pandas Python libraries (or tools), which are software packages that provide a wide range of functions to simplify data handling and analysis, enabling faster and more accurate management and processing of data (McKinney, 2022; Stenegren, 2023). These libraries and packages automated tasks such as reading and extracting data from text files, performing text analysis (e.g., using Regular Expressions) to score HEXACO traits, and outputting the resulting scores into Excel files for further analysis and review (Chapman & Stolee, 2016; McKinney, 2022; Stenegren, 2023). Python's variety of libraries, built-in tools, and integration with Microsoft Excel enable seamless data handling, analysis, and visualisation (McKinney, 2022; Soliev et al., 2023), making it particularly useful for the algorithm in this study, which relies on text data and Excel for both input and output.

3.2.2.2 Specific Design of the Closed-Dictionary Algorithm

The closed-dictionary algorithm was developed to evaluate the participant transcripts against a predefined HEXACO personality trait dictionary stored in Excel. The algorithm assessed each word in the transcripts, assigning scores based on factor loadings outlined by Holtrop et al. (2022). The algorithm automated the scoring process by matching words in the

transcript with corresponding traits in the dictionary. The resulting scores were aggregated for each HEXACO personality trait and automatically exported in Excel for further analysis.

3.2.2.2.1 Scoring methodology. The closed-dictionary approach calculated scores by analysing the frequency and types of words used, comparing them to a predefined trait dictionary in Excel format. The AI algorithm compares the text to predetermined dictionaries of personality-related words linked to each HEXACO trait. The internal scoring process of the algorithm considers the frequency and context of these words to assign scores for each personality trait. Scores were computed on an interval scale ranging from -1 to +1, where -1 indicated a negative presence of the trait, 0 represented an equal balance of negative and positive presence, and +1 reflected a positive presence of the trait. As Holtrop et al. (2022) outlined, this interval-based scoring method offered a structured and standardised framework for evaluating personality traits. Annexure B provides more details on how the algorithm functions, including examples of how it scores specific interview responses.

3.2.3 HEXACO Text-to-Personality AI Algorithm Repository

The HEXACO text-to-personality AI algorithm has been made publicly available on GitHub to promote transparency and facilitate future research. This repository provides access to the algorithm and related resources, ensuring that researchers and practitioners can explore, replicate, and build upon the methodology used in this study. The GitHub repository can be accessed at the following link: https://github.com/TB-UCT/HEXACO_Text_to_Personality_Project.

3.3 Independent and Dependent Variables

Both groups of evaluators (human and AI) generated scores on the exact six HEXACO dimensions. However, the method of evaluation differs, making the comparison between these two evaluation methods the key focus of the research. Therefore, considering the research design, the independent variable represents the human evaluator's scores, while the dependent (criterion) variable represents the AI algorithm evaluation based on the AVI text analysis.

3.4 Data Collection

3.4.1 Source of Data

The AVIs used in the study are considered secondary data obtained from previous studies (Ebrahim, 2022; Patel, 2022). The interviews are based on a position at a fictional organisation named "Hooper", and each participant's interview included five questions, as provided in Table 1, directly obtained and provided by Patel (2022) and Ebrahim (2022).

Table 1

Interview Questions

Question Number	Question
1	Describe a time when you had to collaborate with others to succeed at a task. What was the task you had to accomplish? What made the collaboration successful? What was your role or contribution?
2	Describe a situation where you had to evaluate the risks, benefits, and potential outcomes of a decision. For example, buying something important, investing in something, starting a new project, etc. How did you handle it? And what was the outcome?
3	Describe a time when you took the lead on a group project. What was the project, how did you behave as a leader, and what was the outcome?
4	How would you handle a situation where your work colleagues ignore your ideas and input?
5	How do you manage your time and prioritise tasks?

Source. Ebrahim (2022) and Patel (2022).

3.4.1.1 Fictitious Company and Interview Setup

Participants were presented with a fictitious company introduction and asked to imagine they were applying for a job of their choice at the company (Ebrahim, 2022; Patel, 2022). A general company introduction was used instead of a specific job description to avoid linking the interview to any particular role (Ebrahim, 2022; Patel, 2022). This approach was taken to prevent participants from disengaging or dropping out of the study if they disliked a particular job description (Ebrahim, 2022; Patel, 2022). Additionally, it was intended to prevent participants from forming a negative perception of the AVI process based on a disinterest in the job role (Ebrahim, 2022; Patel, 2022).

3.4.1.2 Online AVI Platform and Recording Settings

Development and presentation of the AVIs are discussed in Ebrahim (2022) and Patel (2022). The AVIs were conducted using the online video interviewing platform, myInterview. The platform was set up to give participants 60 seconds to prepare before the video recording began. Moreover, participants had unlimited time to respond to each question, unlimited chances to re-record their answers, and the option to complete the interview in segments. These settings were kept constant to minimise confounding variables,

such as reconsideration and performance. Participants were given a practice round before the official interview to familiarise themselves with the platform. The practice setup mirrored the interview configuration. All participants were instructed to respond in English.

3.4.1.3 Video and Audio Specifications

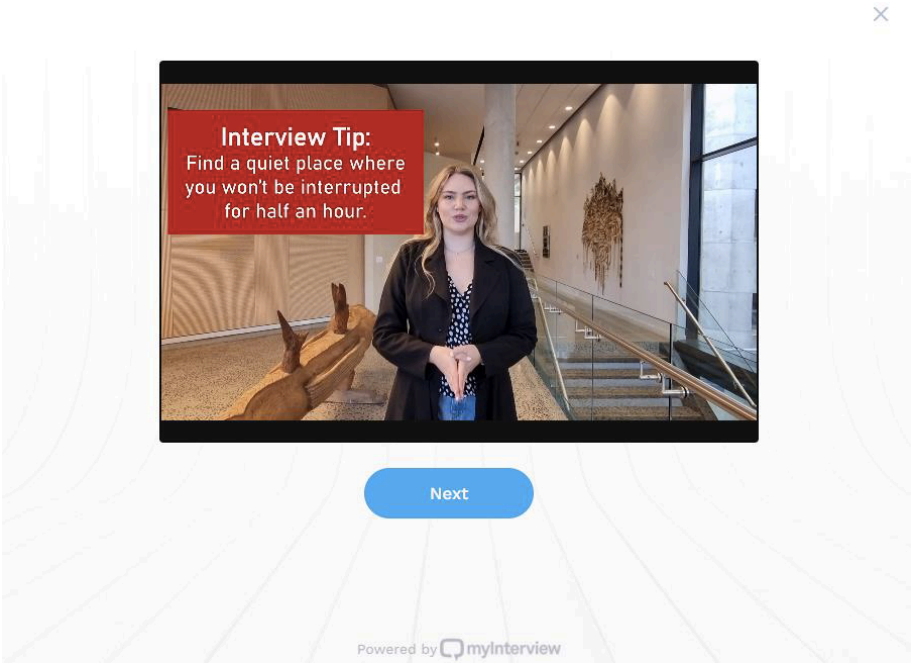
High-quality welcome videos and pre-recorded interview questions in 1080p (full HD) resolution at 60 frames per second (FPS) were used (Ebrahim, 2022; Patel, 2022). The audio was recorded at a 48kHz sampling rate (Ebrahim, 2022; Patel, 2022). The videos were filmed in a foyer to resemble a lobby setting, with a white female interviewer in her mid-twenties (Ebrahim, 2022; Patel, 2022). In line with Social Presence Theory (Chapter 2), fostering connection enhances communication quality, influencing the quality of candidate responses and the data used for scoring (Patel, 2022; Rizi & Roulin, 2024). To align with this theory, human features, such as video introductions and engaging question formats, are incorporated to enhance engagement and create a more immersive experience, ultimately intended to help improve the response quality (Lukacik et al., 2022; Patel, 2022). The AVIs were also edited using Adobe Premiere Pro 2021 to be presented with a company logo, instructions, and tips for completing the AVI (Ebrahim, 2022; Patel, 2022), as illustrated in Figure 4.

3.4.1.4 Welcome Video and Interview Questions

The welcome video lasted approximately four minutes and 35 seconds, briefly introducing the fictitious company and providing guidance for completing the AVI (Ebrahim, 2022; Patel, 2022). As per Table 1, interview questions included three behavioural and two situational questions, addressing communication, interpersonal skills, leadership, critical thinking, and time management (Ebrahim, 2022; Patel, 2022). These dimensions were chosen because they are commonly used in interviews and relevant to various job roles. (Ebrahim, 2022; Patel, 2022) A screenshot of how the questions were displayed is shown in Figure 5.

Figure 4

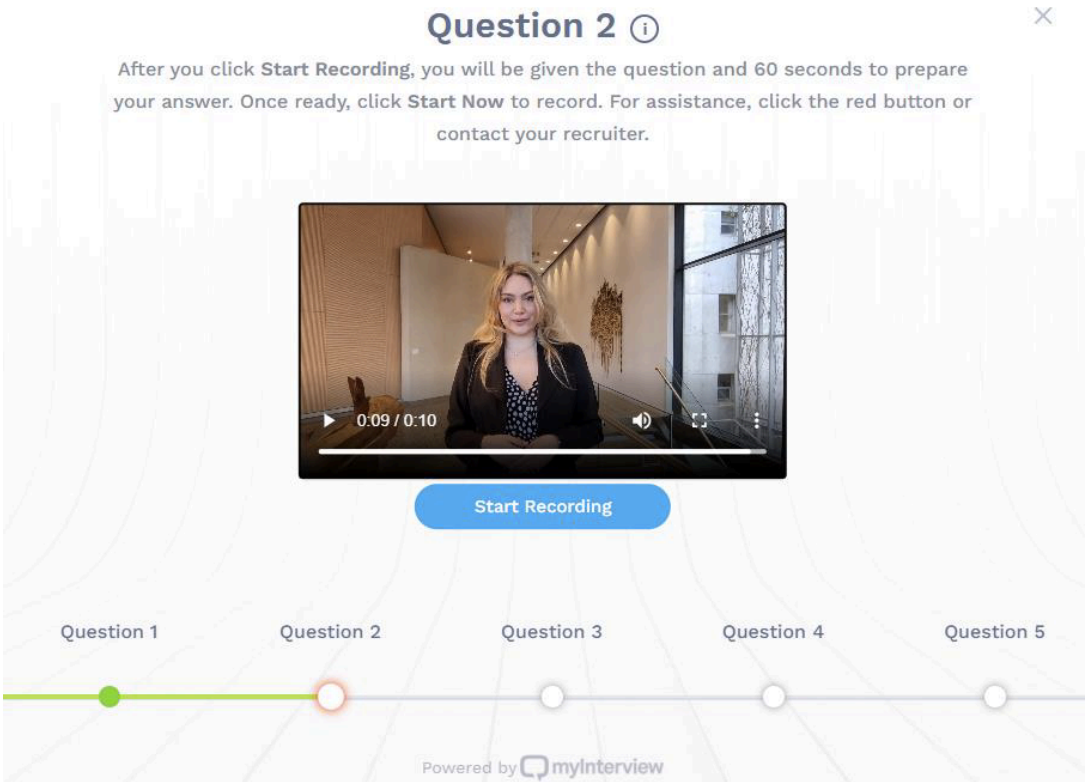
AVI Welcoming Video Screenshot



Note. Screenshot obtained from Patel (2022)

Figure 5

AVI Question 2 Screenshot

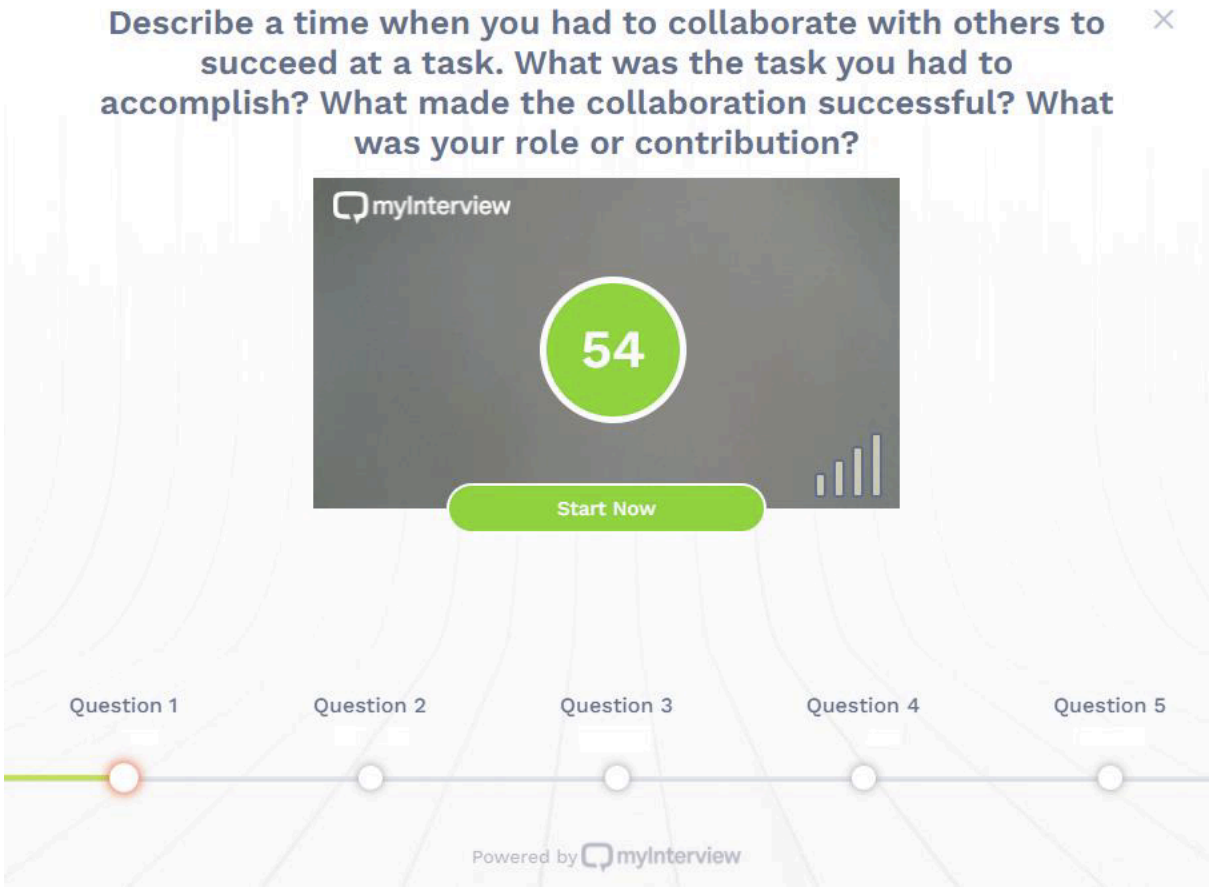


Note. Screenshot obtained from Patel (2022)

After watching the video that presented the question, participants clicked "Start Recording" and were directed to a recording page similar to the one shown in Figure 3 (Ebrahim, 2022; Patel, 2022). As illustrated in Figure 3, the video-based AVI questions were accompanied by text-based versions of the questions (Ebrahim, 2022; Patel, 2022). The welcome video and interview questions were all delivered in English (Ebrahim, 2022; Patel, 2022), as per Table 1. As discussed in the following section, the recorded video-based answers per question were transcribed into text-based answers for evaluation purposes.

Figure 6

AVI Question 1: Text-Based



Note. Screenshot obtained from Patel (2022)

3.4.2 Textual Data

All recorded interviews were transcribed using a consistent transcription process and Microsoft Office 365's built-in transcription service, as advised by the University of Queensland (n.d.), to ensure uniformity in text analysis. Consistency is a key criterion for data quality; therefore, the human evaluators and the AI algorithm drew from the same set of information (Cong et al., 2007; Koutsoumpis et al., 2024). This study used textual data from the transcribed interviews to evaluate the AVI participants' personalities. This textual data is

essential for human evaluators and the AI algorithm, as it allows for an unbiased comparison of how each scoring system interprets personality traits from the same source material. Although human evaluators were provided with video recordings as background and supplementary information, their training emphasised that they should not rely on visual cues to evaluate personality. Instead, the textual data from the transcribed interviews was primarily used in this study to evaluate the AVI participants' personalities by human and AI evaluators. It was outside the scope of this study to analyse video data.

Much like the algorithmic scoring approach to assess the AVIs that had to be developed for the study, as outlined in preceding discussions, humans were also trained to score the AVI data. The following section describes the training process undertaken by the human evaluators.

3.5 Training of Human Evaluators

The two components or variables of the study were explicitly geared towards evaluating personality from the AVIs. The algorithm was programmed (as detailed in Section 3.2.2), while the human evaluators, the second component, received comprehensive training focused on assessing the HEXACO personality model from AVI responses. Guided by prominent studies, the training process primarily followed frame-of-reference (FOR) training, which aligns evaluators by defining performance assessment criteria, providing examples, and offering practice sessions with feedback (Herde & Lievens, 2023; Roch et al., 2012). Additionally, the process incorporated the Realistic Accuracy Model proposed by Funder (1995) and Brunswik's (1956) Lens Theory. Before the training session, the human evaluators were thoroughly oriented to the process through an invitation letter (Annexure C), welcome emails, a consent form (Annexure D), pre-training readings (Annexure E), and a two-hour in-person training session, as outlined in Annexure F and discussed in more detail in the following section.

3.5.1 Frame of Reference Training (FORT)

The frame-of-reference training (FORT), which formed the basis of the training approach, was developed by Bernardin and Buckley (1981) to improve the precision of evaluators' assessments and to train new raters in adaptable rating frameworks, thereby facilitating smoother and more efficient information processing (Martin, 2019). FORT has been proven successful in improving the accuracy of evaluators' ratings (Roch & O'Sullivan, 2003). FORT aims to establish a standardised approach to observing behaviour and developing a common reference point for identifying effective and ineffective performance, ensuring that raters' standards align with the purpose (Bernardin & Buckley, 1981).

Typically, trained raters using the Frame-of-Reference (FOR) method provide more accurate ratings than untrained raters, confirming the success of FOR training (Roch & O'Sullivan, 2003), as detailed in the next section.

3.5.1.1 Frame of Reference Training (FORT) Approach

The primary objective of FORT is to prepare raters to effectively use a specific system in observing behaviour and evaluating performance (Lievens & Sanchez, 2007). This approach is grounded in the schema-based theory of learning, which describes how a mental framework, shaped by existing knowledge and beliefs, prioritises key details, filters out irrelevant ones, and evolves with the integration of new information (Gorman & Rentsch, 2009; Lievens & Sanchez, 2007).

3.5.1.1.1 Principles of FORT. The standard FORT approach adheres to three principles: exposure, practice, and feedback (Martin, 2019). This process aims to equip raters with schemata that is more suitable than their existing mental frameworks (Lievens & Sanchez, 2007). The FORT process is well guided by the works of Martin (2019) and Powell and Goffin (2009), as discussed below**.**

In the exposure phase, the multifaceted nature of the construct, such as personality, is introduced, with each aspect clearly defined. Trainers present illustrative behavioural scenarios (vignettes) representing each construct aspect. During the practice stage, trainee raters can evaluate the dimensions portrayed in the vignettes. They are also required to provide justifications for their assessments. In the feedback process, trainers review and provide feedback on the accuracy of the trainee raters' evaluations. The feedback includes communicating the correct ratings for each vignette based on normative data derived from expert raters' 'true scores'. Trainers also explain the rationale behind these true scores to ensure understanding.

FORT aims to establish a reliable framework for consistent and accurate performance evaluations. This study complements it with the Realistic Accuracy Model and Lens Theory, described in the following section.

3.5.2 Realistic Accuracy Model

The Realistic Accuracy Model (RAM) highlights the complexity of personality assessment and the need to move beyond simple definitions thereof, instead delving deeply into complex issues regarding the validity of personality traits and their measurement (Funder, 1995; Powell & Goffin, 2009). Successful judgements, marked by accuracy, result from a combination of a capable judge, an understandable target, sufficient accessible information and evident traits or characteristics that are readily noticeable (Chen, et al.,

2018). Accurate personality judgement entails the individual exhibiting relevant behaviour, which must be noticeable to the evaluator, who must then correctly recognise and interpret the cues to accurately infer the target's personality (Powell & Goffin, 2009). Thus, the accuracy of personality evaluation relies on the relevance, accessibility, detection, and utilisation of behavioural cues (Funder, 1995). Aligning with the work of De Kock et al. (2020), the Realistic Accuracy Model accentuates the significance of having both a 'good judge' (evaluator) and access to 'good information' (available data). Therefore, RAM suggests that to assess personality effectively, it is important to consider how well the traits are expressed and how accurately others or raters can observe them. Building on these concepts is the Lens Theory, as described below.

3.5.3 Lens Theory

Brunswik's Lens Theory has been widely used in studying cue relevance, utilisation, and judgement accuracy across domains, including personality traits (Martin, 2019). This theory suggests that when people judge personality, they rely on available cues from the target, which may vary in relevance (Martin, 2019). Accurate judgements occur when the perceived cue attributes are high quality, match reality, and are relevant to the specific characteristics studied (Martin, 2019; Mosier & Kirlik, 2004). Therefore, as described by Brunswik (1995) and in line with RAM, in the present study, the lens model may explain how decisions are made regarding personality by interpreting the available cues, forming judgements, and evaluating their accuracy. Accurate personality evaluation relies on the target displaying relevant cues detectable by the evaluator and correctly utilised for assessment (Letzring et al., 2006). Therefore, these aspects were incorporated into the training of human evaluators.

3.5.4 Training Method

As explored in the previous sections, the Realistic Accuracy Model and Lens Theory were considered in designing and implementing the frame-of-reference training stages: exposure, practice, and feedback. These concepts are further contextualised in the following section, which covers the pre-training and training materials.

3.5.4.1 Pre-Work Training Material

It was decided that the training session should be complemented with preparatory work, estimated to take approximately 30 to 60 minutes. Pre-work activities for training sessions typically serve as instructional support, helping individuals engage more effectively with the material and fulfilling several key roles: introducing relevant content, strengthening understanding, and enabling more meaningful discussions (Koszalka et al., 2021). The pre-work, provided in Microsoft PowerPoint slides, introduced key topics to the human evaluators, including the study's objectives, the HEXACO personality model, and Asynchronous Video Interviewing (AVI). The content was structured into three sections: AVI as interview approach, the HEXACO model, and Scoring Criteria, ensuring evaluators gained familiarity with the essential concepts required for the training session. This preparatory work was valuable as it established a foundational knowledge base upon which the training could be built, assuming evaluators had a basic understanding of these topics.

The PowerPoint slides included text, visuals and the researcher's narration. The narrated slides were designed following Mayer's Cognitive Theory of Multimedia Learning and principles from Social Agency Theory. Mayer's Cognitive Theory of Multimedia Learning emphasises the integration of visuals and audio narration to reduce cognitive load and enhance information retention (Mayer & Moreno, 2003; Mayer, 2021). This approach aimed at presenting information clearly, helping human evaluators process complex material effectively. Furthermore, principles from Social Agency Theory were incorporated, using a human voice in the narration to foster a sense of connection and engagement, making the learning experience more personable and relatable (Atkinson et al., 2005). Overall, the pre-work provided a structured, engaging, and more personal foundation, preparing the evaluators for the upcoming training process.

3.5.4.2. Training Process

During the initial phases of the training session, evaluators were introduced to the study, reminded of key ethical considerations when evaluating personality and completed a brief knowledge test based on the pre-work to assess their understanding. Including a knowledge test in training is important as it enhances retention and understanding of the material while identifying areas that require further clarification (Bae et al., 2019; Roediger et al., 2011). Evaluators were encouraged to note any uncertainties from the pre-work and bring these questions to the training session, ensuring a comprehensive, collective, and shared understanding of the material. This structured approach aimed to enhance the evaluators'

readiness and effectiveness during the training session, following the FORT approach, provided below.

3.5.4.2.1 Exposure. The researcher gave a brief lecture and recapitulation on AVIs and the HEXACO personality model for approximately one hour. The exposure phase covered the details of the AVIs, the five interview questions, the rating scale, each HEXACO scale, and the corresponding behavioural cues. The cues or descriptions were mainly derived from the study by Koutsoumpis et al. (2024).

3.5.4.2.2 Practice. The practice phase consisted of 'example tasks' and 'practice tasks.' The researcher guided the human evaluators through the 'example tasks,' which involved two interview responses discussed as a group. Afterwards, the researcher provided scores and the reasons for them, which were also discussed as a group. When a specific HEXACO cue was identified, the researcher explained it and opened up the discussion to the group on (a) why it constituted a HEXACO cue and (b) which HEXACO trait it represented, in line with and ensuring effective cue utilisation (Funder, 1995).

Subsequently, the human evaluators individually completed the 'practice tasks,' which involved watching and analysing transcribed AVIs from two participants, accompanied by the HEXACO traits, definitions, and rating scales. Each AVI was analysed for each of the six HEXACO traits. After each AVI, the human evaluators assessed the extent to which each candidate demonstrated the various HEXACO traits on the provided rating scale, ranging from 1 to 5. A score of 1 indicated a very low indication of the specific trait. In contrast, 5 indicated a very high indication, as outlined in Table B3, Annexure B, and the training material in Annexures E and F. The human evaluators shared their ratings with a nearby peer upon completion. Any significant or persistent discrepancies were brought up for group discussion to ensure a standard for scoring and the interpretation of responses in accordance with the provided scoring and interpretation material.

3.5.4.2.3 Feedback. The researcher reviewed the practice task with the human evaluators to ensure accurate detection and utilisation of relevant HEXACO cues while providing a rationale for the ratings. They were given a 15-minute break, after which the process of exposure, practice, and feedback was repeated.

The training programme aimed to ensure that the students or human evaluators understood how to score each personality trait based on the interview transcripts. Examples and descriptors of high, medium, and low trait scores were provided to ensure the human evaluators had a consistent benchmark when assessing the HEXACO personality traits. The following steps in the evaluation process were clearly discussed with the human evaluators,

including when to expect access to the participants' videos, which were uploaded to a secure OneDrive folder. Access was provided via the human evaluators' student UCT email addresses, granting them access to their assigned participants for rating. This included the relevant rating sheet with instructions. Annexures E and F provide these steps and the complete training material.

3.6 Population and Sample

The target population includes working-age individuals applying for various jobs through AVIs as selection tools, with AI algorithms assessing their performance and personality traits. The AVI set included 198 pre-recorded AVIs from the USA, drawn from previous studies by Ebrahim (2022) and Patel (2022). The sample characteristics are detailed in Table 2, presenting the USA AVI samples from the studies by Ebrahim (2022) and Patel (2022), which utilised the same AVIs but differed in sample sizes. The current study utilised these secondary data samples, selecting 161 out of the 198 videos provided in the previous studies as adequate. Each participant's videos (comprising five questions per participant) were manually inspected by the researcher for completion, clarity, and duration. A minimum length of approximately 10 seconds per response was considered acceptable, although more extended responses were preferred.

Table 2
Demographic Statistics for the USA Sample of AVI Participants

Demographic Category	Ebrahim (2022) (N = 162)	Patel (2022) (N = 169)
Gender		
Female	99 (61.1%)	104 (61.50%)
Male	62 (38.3%)	64 (37.90%)
Other	1 (0.6%)	1 (0.60%)
Employment Status		
Employed	116 (71.6%)	99 (58.60%)
Unemployed	26 (16.0%)	26 (15.40%)
Student	16 (9.9%)	17 (10.10%)
Student and working part-time	3 (1.9%)	3 (1.80%)
Missing values	1 (0.60%)	2 (1.20%)
Device Used		
Laptop	115 (71.0%)	150 (88.76%)

Desktop	29 (17.9%)	-
Smartphone	14 (8.6%)	18 (10.65%)
Tablet	3 (1.9%)	-
Missing values	1 (0.6%)	1 (0.60%)
Resolution	-	
High Definition (720p)	-	109 (64.50%)
Full HD and above	-	60 (35.50%)
AVI Experience		
Yes (asynchronous)	42 (25.9%)	8 (4.73%)
Yes (synchronous)	26 (16.0%)	35 (20.71%)
Yes (unsure)	20 (12.3%)	20 (11.83%)
No	73 (45.1%)	79 (46.75%)
Missing values	1 (0.6%)	1 (0.59%)

Note. Sample information obtained and adapted from Ebrahim (2022) and Patel (2022).

3.7 Data Collection, Analysis and Procedure

3.7.1 Ethical Approval and Registration of the Study

Before the commencement of the research, the researcher obtained ethical approval from the Ethics in Research Committee at the University of Cape Town (Annexure G) and DSA100 approval from the Department of Student Affairs to involve UCT students in the research study (Annexure H). After ethical approval was granted, the researcher pre-registered the study on AsPredicted (available at https://aspredicted.org/k4g2-3yy2.pdf), as outlined in Annexure I. Preregistration was conducted to demonstrate a strong commitment to preventing questionable research practices, avoiding the formation of hypotheses after results are known (a practice termed HARKing), and promoting open science (Buchanan et al., 2021; Ebrahim, 2022; Field, 2018).

3.7.2 Data Collection

Data collection involved obtaining secondary data comprising asynchronous video interviews, scoring from human evaluators, and results from the AI text-to-personality algorithm.

3.7.2.1 Secondary Data: AVIs

According to previous studies from which the AVIs originate (Ebrahim, 2022; Patel, 2022), the AVI data was collected over six weeks between September and October 2021. The study involved a sample from the United States of America (USA). Prolific, an online

participant recruitment platform (Palan & Schitter, 2018), was used to recruit participants for the USA sample. Once participants clicked on the study link, they were provided with a description of the study, informed that their anonymity would be protected, and advised of their right to withdraw from the study at any stage, as participation was voluntary. Participants were compensated upon completion. These videos were securely stored within the section of organisational psychology at UCT, with access granted to the researcher following ethical clearance.

3.7.2.2 Human and AI Evaluations

The data collection for human and AI evaluations followed a structured timeline. Human evaluators were initially given until 15 September 2024 to submit their final ratings, with each evaluator assessing approximately ten participants, though this varied slightly in some cases. However, an extension was granted to accommodate additional time needs, and the final human ratings were collated at the end of October 2024, all recorded in Microsoft Excel format. Concurrently, by the end of October 2024, the final AI-based text-to-personality algorithm was ready for use, generating scores automatically, which were also produced in Microsoft Excel format. These datasets were thus prepared by the end of October 2024 for further analysis.

3.7.3 Data Capture, Preparation, and Analysis

Several steps were undertaken to prepare the dataset for analysis, following recommendations by Field (2018) and Tabachnick and Fidell (2019). All data was initially captured in Microsoft Excel before being imported into SPSS, also known as the Statistical Package for the Social Sciences (Grotenhuis & Visscher, 2014), for further analysis.

3.7.3.1 Data Cleaning and Preparation

The dataset underwent an initial visual inspection in Excel, a crucial step for identifying anomalies and ensuring accuracy (Tabachnick & Fidell, 2019). After transferring the data to SPSS, Missing values were assessed using SPSS's Missing Values Analysis (MVA) tool, which evaluates patterns and provides options for managing missing data, such as listwise deletion (George & Mallery, 2019). Outliers were detected through visualisations like histograms, box plots, and normal probability plots, enabling decisions on whether to retain, transform, or remove extreme values based on their potential influence on findings (Field, 2018).

Additionally, variable coding was standardised to distinguish between human and AI ratings of HEXACO traits. Human evaluations were labelled with the suffix "_Human" (e.g., E_Human), while AI ratings used "_AI" (e.g., E_AI), facilitating clarity in subsequent

analyses (Arkkelin, 2014). Both human and AI ratings, originally on different scales, were standardised to a uniform range of 1 (Very Low) to 5 (Very High), reflecting HEXACO trait presence. Subsequent analyses focused primarily on correlations and supporting analyses, such as inter-rater reliability, briefly outlined below but discussed in more detail in Chapter 4.

3.7.3.2 Descriptive Statistics and Correlation Analysis

Descriptive statistics were computed to examine central tendencies, dispersion, and assumptions for analysis (Ebrahim, 2022; George & Mallery, 2019; Grotenhuis & Visscher, 2014). A correlation matrix was generated to explore associations between HEXACO traits and the scoring methods, namely human vs AI. Spearman correlation, chosen for its robustness with ordinal data, offered a parsimonious method to identify the strengths of associations without inferring causality (Field, 2018). This approach aligned with the study's goal of examining the convergence between human and AI evaluations in HEXACO personality assessments.

3.7.3.3 Inter-Rater Reliability

To ensure consistency among human evaluators, inter-rater reliability was assessed using SPSS's Intraclass Correlation Coefficient (ICC), based on ten videos rated on five questions by all evaluators, guided by Field (2018). This step aimed to inspect the reliability of human scores and to minimise the risk of inconsistencies arising from their subjective judgement (Foxcroft & Roodt, 2013). In addition, the variability of the AI algorithm was analysed as an indication of consistency. More details on this analysis and its findings are provided in Chapter 4.

Chapter 4: Results and Data Analysis

This chapter details the statistical findings of this study and is structured as follows. First, the data cleaning process is detailed, ensuring the dataset is accurate and free from inconsistencies, missing values, and to assess whether outliers are present. Following this, the descriptive statistics of the dataset are presented, providing an overview of the dataset as rated by both humans and AI. This is followed by a detailed analysis that explores the correlations between human and AI ratings for each HEXACO trait to determine the extent of convergence. Additionally, the examination of inter-rater reliability among human raters, assessing the consistency of their evaluations and those of the AI, is provided. Finally, the chapter concludes with a hypothesis testing section, summarising the findings about the stated research hypotheses.

4.1 Introduction

This chapter presents the study's findings, focusing on evaluating HEXACO personality traits from AVI responses using human raters and an AI-based evaluation method. The AI method was developed specifically for this study, employing the closed-dictionary or vocabulary approach guided by Holtrop et al. (2022) and their HEXACO Text-to-Personality method. The main objectives of this chapter are to examine the level of convergence between human and AI ratings of HEXACO personality traits derived from text-based asynchronous video interviews and to assess the reliability and consistency of these ratings, with an additional focus on the interrater reliability of human raters.

The analyses conducted in this chapter are aligned with the study's hypotheses, explored in Chapter 2, which are as follows:

Null Hypothesis (H_0): There is no significant correlation between human evaluator ratings and AI algorithm ratings in evaluating each HEXACO personality trait from asynchronous video interviews.

Alternative Hypothesis (H_{1a}): A positive relationship exists between human evaluator ratings and AI algorithm ratings in evaluating each HEXACO personality trait from AVIs.

Alternative Hypothesis (H_{1b}): The convergence between AI and observer ratings of personality will be strongest for extraversion, compared to other personality traits.

A series of quantitative analyses were conducted to address these hypotheses. After data cleaning, these analyses included descriptive statistics, Spearman correlation analyses, inter-rater reliability assessments using the Intraclass Correlation Coefficient (ICC), and

additional analyses to evaluate variance and the degree of alignment between human and AI ratings, as outlined in the following sections.

4.2 Data Cleaning and Preparation

The initial phase of the analysis involved data cleaning to ensure the dataset's integrity before conducting statistical analyses, which is essential for accurate analysis and reliable decision-making (Chu et al., 2016). Following the best practices outlined by Field (2018) and Tabachnick and Fidell (2019), data preparation was conducted using Microsoft Excel and SPSS. This process included several key steps, which are detailed in the following sections: (1) Data Inspection and Conversion, (2) Format Adjustment for Analysis, (3) Variable Coding, (4) Handling Missing Values, (5) Assessment of Normality, and (6) Outlier Detection and Management.

4.2.1 Data Inspection and Conversion

The researcher first performed an initial baseline visual inspection in Excel, as reviewing the original data carefully is the most effective way to ensure a data file's accuracy (Tabachnick & Fidell, 2019) before proceeding with the statistical analysis in SPSS. A specific and important step was the conversion of AI ratings to the 1-to-5-point scale, as used with human raters. While the AI ratings were initially scored on a scale ranging from -1 to +1, the data had to be converted to ensure that both human and AI evaluations were assessed on a scale of 1 (Very Low) to 5 (Very High), indicating the presence of the HEXACO trait. The AI scoring and conversion process is described in more detail in Annexure B: AI Scoring Process.

4.2.2 Format Adjustment for Analysis

The final dataset used for analysis comprised ratings from human evaluators and the AI technique, both on the same scale, and was formatted in two ways: long and wide formats, as each allowed for different types of analysis (Field, 2018), as described in this chapter.

4.2.3 Variable Coding

The researcher ensured that variables were consistently labelled in data coding to differentiate between human and AI ratings for the HEXACO traits. This involved renaming columns in the dataset to follow a clear and structured naming convention. Specifically, human ratings were labelled with the suffix "_Human" (e.g., E_Human for the Emotionality trait rated by a human evaluator), while AI-generated ratings were labeled with the suffix "_AI" (e.g., E_AI). This clear and systematic approach to variable naming helped avoid confusion. It maintained clarity when working with the dataset, as Arkkelin (2014) guided, ensuring that human and AI ratings were easily identifiable for subsequent analyses.

4.2.4 Handling Missing Values

After converting scores, transferring them to SPSS and variable coding, the first step in data cleaning focused on identifying and addressing missing values within the dataset. Using SPSS, a comprehensive check for missing data was conducted by generating frequency tables and descriptive statistics. This allowed for identifying any patterns or occurrences of missing values across the key variables (Tabachnick & Fidell, 2019), particularly the HEXACO trait ratings from both human and AI evaluations.

Upon a thorough review, as per Table 3 and Annexure J, no missing values were found across the key variables, namely the HEXACO trait ratings provided by human evaluators and the AI method. As a result, no further action, such as listwise deletion or imputation, was necessary (George & Mallery, 2019; Tabachnick & Fidell, 2019). This step ensured that the dataset was fully intact, allowing for further analysis without concerns of data loss or potential biases related to missing data handling.

Table 3

Descriptive Statistics Verification of Completeness for HEXACO Trait Ratings

Variable	*N*	Mean (*M*)	Std. Deviation (*SD*)	Missing Count	% Missing
Human					
H	161	3.74	.58	0	.00
E	161	2.95	.52	0	.00
X	161	3.45	.46	0	.00
A	161	3.65	.49	0	.00
C	161	3.94	.55	0	.00
O	161	3.42	.50	0	.00
AI					
H	161	3.08	.10	0	.00
E	161	3.00	.08	0	.00
X	161	3.07	.07	0	.00
A	161	3.01	.09	0	.00
C	161	3.15	.08	0	.00
O	161	2.96	.04	0	.00

Note. H: Honesty-Humility, E: Emotionality, X: Extraversion, A: Agreeableness, C: Conscientiousness, O: Openness to Experience.

4.2.5 Assessment of Normality

4.2.5.1 Correlation Studies and Normality

Correlation studies are among the most frequently used statistical methods, assuming the data are normally distributed (Das & Imon, 2016; Tabachnick & Fidell, 2019). Therefore, assessing the normality of the HEXACO trait distributions was key to informing the choice of statistical methods.

4.2.5.2 Methods to Evaluate Normality

Several approaches were explored to evaluate normality and decide on a statistical method (see Das & Imon, 2016; Field, 2018; Tabachnick & Fidell, 2019). These approaches included scatter and residual plots, histograms, stem-and-leaf plots, box plots, and the data description and distribution function of SPSS. Specifically, quantitative measures of normality and distribution shape (including skewness and kurtosis values) and their depictions were inspected (Tabachnick & Fidell, 2019), provided in Annexure J. Skewness reflects data asymmetry, while kurtosis indicates extreme values through the "tailedness" of the distribution (Field, 2018). However, as these measures alone do not conclusively determine normality, additional tests were conducted (Das & Imon, 2016).

4.2.5.3 Statistical Tests for Normality

As per Table 4, the Kolmogorov-Smirnov (K-S) test with a Lilliefors significance correction and the Shapiro-Wilk test were applied to assess deviations from normality. These are two standard empirical distribution function (EDF) tests for normality (Das & Imon, 2016; Field, 2018).

The Shapiro-Wilk and Kolmogorov-Smirnov (K-S) tests both assess normality; however, the Shapiro-Wilk test is generally regarded as more powerful (Field, 2018) and was prioritised in the analyses. A significant result on the test ($p < .05$) indicates that the data distribution deviates significantly from normality, meaning it is non-normal (Field, 2018).

For the human trait ratings, all tests returned p-values less than .05, indicating significant deviations from normality. In contrast, the AI-generated ratings showed a mixed result. Some traits (E_AI, X_AI, C_AI, O_AI) did not significantly differ from normality ($p > .05$), suggesting approximate normality for these traits. However, Honesty-Humility (H_AI) and Agreeableness (A_AI) deviated significantly from normality.

4.2.5.4 Statistical Method Selection

While many parametric statistical methods, such as Pearson’s correlation, require data to be normally distributed (Arkkelin, 2014), Spearman's rank-order correlation was selected for this analysis. Spearman’s correlation does not assume normality, making it appropriate for

data exhibiting non-normal distributions (Field, 2018). Thus, the significant deviations in some of the human ratings from normality do not invalidate the use of this method.

4.2.5.5 Addressing Non-Normality and Outliers

Non-normality is more common than normality in personality measurement (Bishara & Hittner, 2015; Pek et al., 2018). Nonetheless, additional steps were taken to assess the impact of outliers, their exclusion, and transformations to normalise the data, reduce skewness, and mitigate the influence of extreme values, as outlined in Annexure K. Although the original dataset was retained for primary analysis, these steps were implemented proactively to prevent biases in correlation studies involving skewed data (Bishara & Hittner, 2015; Feng et al., 2014). Further details regarding these approaches are provided in the outlier management section.

Table 4

Tests of Normality for Human and AI Ratings of HEXACO Traits

Trait	Kolmogorov-Smirnov Statistic	Kolmogorov-Smirnov p-value	Shapiro-Wilk Statistic	Shapiro-Wilk p-value
H_Human	.11	< .001	.97	.00
E_Human	.12	< .001	.98	.01
X_Human	.11	< .001	.97	.00
A_Human	.10	< .001	.98	.01
C_Human	.08	.02	.98	.01
O_Human	.14	< .001	.96	< .001
H_AI	.07	.09	.97	< .001
E_AI	.05	.20*	.99	.37
X_AI	.05	.20*	.99	.65
A_AI	.06	.20*	.98	.02
C_AI	.05	.20*	.99	.27
O_AI	.04	.20*	1.00	.93

Note. Kolmogorov-Smirnov test results were evaluated with a Lilliefors Significance Correction. An asterisk (*) indicates that the result for the Kolmogorov-Smirnov test is a lower bound of the true significance.

Note. H: Honesty-Humility, E: Emotionality, X: Extraversion, A: Agreeableness, C: Conscientiousness, O: Openness to Experience.

4.2.6 Outlier Detection and Management

Outliers, which can skew results and introduce bias, were identified and managed using a three-step approach: range-based detection to confirm scoring validity, z-score analysis to flag extreme deviations and log transformations to address skewness, as detailed in Annexure K, following the guidance of Field (2018) and Tabachnick and Fidell (2019). Since the range-based detection method did not reveal any values outside the scoring range, it was decided to retain all values and preserve the complete dataset. Additional methods for outlier management, as outlined in Annexure K, included:

Z-Score Analysis: Using a ±3 threshold, outliers were flagged, examined and excluded if they fell outside this range.

Transformations: Square root and Logarithmic Transformations were applied to address skewness and ensure normality.

Despite these efforts, some variables remained non-normal, as per Appendix K. This may accentuate the complexity of personality measurement and the importance of retaining natural variance in psychological research. As Aguinis et al. (2013) provided, outliers can offer valuable insights and contribute to new theoretical perspectives. Decisions about including or excluding outliers can significantly affect conclusions (Aguinis et al., 2013), impacting the insights regarding whether human and AI personality scoring align and the strength of their convergence. Therefore, considering that the extreme values were still within the 1–5 scoring range, no erroneous scores were identified, and all data were retained for further analysis. Figures J1 to J6 in Annexure J present box plots comparing AI and Human scores for each trait, illustrating the distribution, spread, mean scores, score differences between rater types, and ‘outliers’.

4.3 Exploratory Correlation Analysis

The correlation analysis used Spearman's correlation coefficient, also known as Spearman’s rho, a non-parametric measure of association or correlation (Field, 2018). This method is chosen because it reduces the impact of outliers and does not require the assumptions of normality and linearity needed for parametric tests, making it particularly suited for the current dataset that does not align with these assumptions (Field, 2018).

4.3.1 Spearman's Rank Correlation Overview

As explored during the data cleaning steps, the normality tests indicated that all human-rated HEXACO scores deviate from a normal distribution, as evidenced by significant results on the Shapiro-Wilk test for the human-rated traits, as well as for Honesty-Humility and Agreeableness rated by AI. In contrast, the other AI-rated scores generally align more

closely with a normal distribution, as reflected by the non-significant Shapiro-Wilk test results for the remaining HEXACO traits, as shown in Table 3.

Given this mixed distribution profile and guided by Field (2018), a non-parametric correlation method, such as Spearman's rank correlation, was deemed suitable. While the decision was made to proceed with the original complete dataset, a comprehensive assessment of the convergence between human and AI ratings was conducted under all three conditions outlined above and in Annexure K: inclusion of all data used in the primary analysis (Condition 1), exclusion of outliers (Condition 2), and transformation (Condition 3).

However, the primary correlation analysis and reporting were performed on the raw dataset (Condition 1) to preserve its integrity without imposing any conditions. Since data cleaning revealed no values outside the expected rating range (1–5), and psychological data is known to exhibit diversity due to inherent human differences (Field, 2018), all original scores were retained.

4.3.2 Main Correlation Analysis

As mentioned earlier, the main study and analysis were based on the dataset that retained all data points, including outliers, referred to as Condition 1. This approach differs from the other two, where outliers were either removed (Condition 2) or adjusted (Condition 3), as it preserves the entire dataset. Following the guidance of Aguinis et al. (2013) and Field (2018), these outliers were not treated as errors since they fell within the possible range of values. Instead, they were considered 'interesting outliers,' which often occur in psychological research. The decision to retain them was based on their potential to provide valuable insights, particularly in personality assessments, where extreme responses may reflect genuine variability in participant traits. Therefore, this central section outlines the steps taken to analyse the data while retaining the outliers. The alternative analyses that excluded and transformed the data are available in Annexure K.

Table 5 presents the results of the Spearman correlation analysis conducted on the dataset, further discussed and elaborated on in Chapter 5. Significant positive correlations were observed for H_Human and H_AI (r = .21, p = .01) as well as C_Human and C_AI (r = .16, p = .048). Additionally, other significant yet unexpected correlations emerged, such as H_Human and C_AI (r = .17, p = .03), A_AI and X_Human (r = .21, p = .01), A_Human and H_AI (r = .18, p = .02), C_AI and H_Human (r = .19, p = .02), X_AI and C_Human (r = .16, p = .04), and O_Human and H_AI (r = .18, p = .02). These findings may reveal similarities in how these traits are perceived and rated by the two methods.

In addition to the individual trait-level correlation analysis, this study was particularly interested in this study was the examination of the overall correlations across traits, as detailed in the next section.

Table 5

Spearman's rho Correlation Coefficients: Condition One

Variable	H_AI	E_AI	X_AI	A_AI	C_AI	O_AI
H_Human	**.207**	-.034	.007	.059	**.169**	-.013
p-value	.008	.672	.930	.456	.032	.872
E_Human	-.004	.118	.142	.133	-.140	.068
p-value	.963	.136	.073	.094	.077	.392
X_Human	**.295**	.042	.088	**.207**	.037	-.041
p-value	<.001	.600	.269	.008	.643	.610
A_Human	**.179**	-.025	.133	.053	**.186**	-.028
p-value	.023	.752	.091	.503	.018	.725
C_Human	.142	-.060	**.162**	-.017	**.156**	.019
p-value	.073	.453	.039	.827	.048	.816
O_Human	**.182**	.039	.136	.141	.106	.046
p-value	.021	.621	.085	.075	.179	.565

N = 161.

Note. The values in this table are rounded to three decimal places due to lower values for some traits. However, when reported outside the table, they are rounded to two decimal places, except for the C trait correlation, which is rounded to three decimal places as it is near significance ($p < .05$) and of particular interest in the study on convergence.

Note. H: Honesty-Humility, E: Emotionality, X: Extraversion, A: Agreeableness, C: Conscientiousness, O: Openness to Experience.

4.3.3 Correlation Across All Ratings

An additional analysis was conducted on the aggregated dataset to further explore the relationship between human evaluator ratings and AI algorithm ratings, considering all ratings rather than individual HEXACO traits.

4.3.3.1 Overall Correlation Results (Across Traits)

As shown in Table 6, the Spearman correlation for all ratings was significant and moderately positive ($r = .29$, $p < .001$), with a 95% confidence interval of [.23, .35], indicating a stronger relationship than the trait-specific correlations. There is a 95% chance that this confidence interval contains the true population correlation coefficient between human-rated and AI-rated scores of .29.

Table 6

Overall Spearman's Correlation Between Human and AI Evaluators: All HEXACO Traits

Variable Pair	Spearman's Correlation	p-value	Sample Size (N)
Human vs AI	.29	< .001	966

4.3.3.2 Summarised Correlation Findings

Table 7 summarises the correlation coefficients for each HEXACO trait and the overall correlation. Figure 7 further illustrates the scatterplots for the six individual traits at the trait level and the combined overall (across traits) correlation.

Table 7

Summarised Correlation Coefficients Between AI and Human Scoring: HEXACO Traits

Trait	*N*	Correlation (*r*)	*p*
Overall (Across Traits)	966	.29	< .001**
Honesty-Humility (H)	161	.21	.01*
Emotionality (E)	161	.12	.14
Extraversion (X)	161	.09	.27
Agreeableness (A)	161	.05	.50
Conscientiousness (C)	161	.16	.048*
Openness to Experience (O)	161	.05	.57

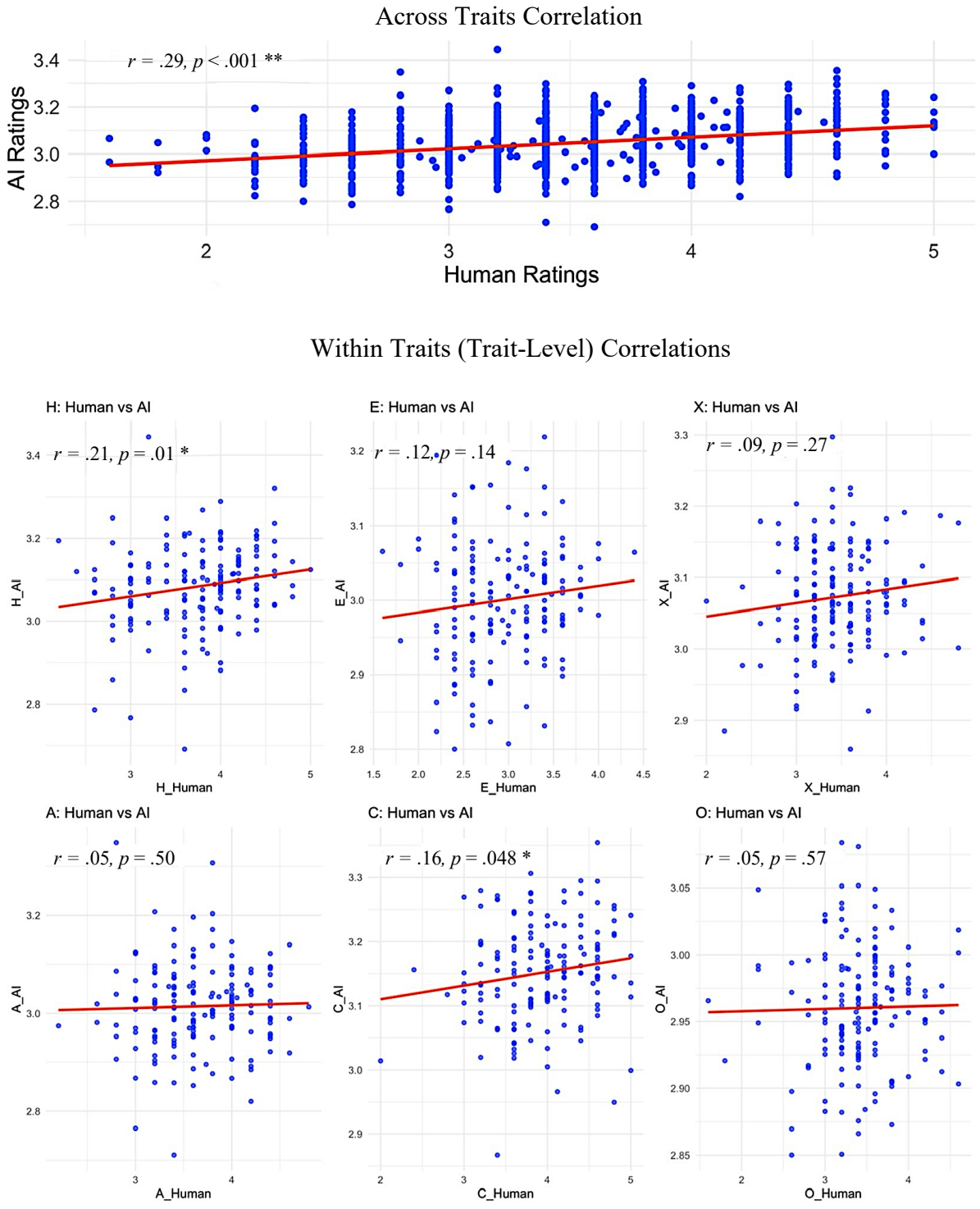
Note. * $p < .05$. ** $p < .001$. The C trait correlation is intentionally rounded to three decimal places to highlight its near significance ($p < .05$).

4.3.3.3 Key Observations from Scatterplots

As shown in Figure 7, the scatterplots illustrate the combined correlation (across traits) and six trait correlations individually. Overall, human and AI scores showed positive covariation, with a significant positive correlation for Honesty-Humility (H) and Conscientiousness (C), but weaker and non-significant correlations for the other traits (Extraversion, Agreeableness, and Openness).

Figure 7:

Scatterplots for correlations between Human and AI evaluations for each HEXACO trait and the overall correlation across traits. Each scatterplot includes a regression line, Spearman correlation coefficient and indication of significance.



Note. * $p < .05$. ** $p < .001$.

4.3.3.4 Correlation Outcomes and Hypotheses

As shown in Tables 5, 6 and 7, statistical tests were conducted to assess the relationships between human and AI-rated HEXACO traits, retaining outliers. Spearman's rho correlations were used to evaluate these relationships, as this non-parametric measure is less sensitive to the effects of outliers than Pearson's correlation (Field, 2018). The correlation matrices revealed several significant relationships between variables, offering insight into the degree of convergence between human and AI scoring methods.

In line with findings from a similar study by Holtrop et al. (2022), correlations between human and AI scores for the HEXACO traits were generally small to moderate, with most Spearman's rho values falling below .30. Notably, a significant overall correlation was observed ($r = .29$, $p < .001$), along with specific trait correlations for Honesty-Humility ($r = .21$, $p = .01$) and Conscientiousness ($r = .16$, $p = .048$). Other unexpected correlations involving traits that are not typically paired were also observed, as discussed earlier and presented in Table 6. These results suggest some convergence between the scoring methods. However, the relationships for the remaining traits were weak and not statistically significant.

4.4 Hypotheses Exploration and Testing

This section examines the hypotheses by analysing the relationships between human and AI ratings across the HEXACO traits with a significance threshold of $p < .05$. The null hypothesis (H0) posits no significant relationships between human evaluator ratings and AI algorithm ratings for the HEXACO traits. Two alternative hypotheses were proposed:

H_{1a}: A positive relationship exists between human and AI ratings for each HEXACO trait.

H_{1b}: Convergence between AI and human ratings will be more substantial for Extraversion than other traits.

Building on the correlation findings discussed earlier, the following section delves into hypothesis testing to evaluate the alignment between human and AI ratings for the HEXACO traits.

4.4.1 Hypotheses Testing: Trait-Level and Overall Convergence

Hypothesis testing is conducted at both the trait level, examining each HEXACO trait individually and at the overall level (across traits) by assessing correlations or convergence in the aggregated correlation data.

4.4.1.1 Trait-Level Hypotheses Testing

Honesty-Humility (H). The correlation between human and AI ratings for Honesty-Humility was statistically significant ($r = .21$, $p = .01$), with a small to medium effect size

(Cohen, 1988). These results support H_{1a} and lead to the rejection of H_0 for Honesty-Humility.

Emotionality (E). A weak positive but non-significant correlation ($r = .12$, $p = .14$) was found, suggesting no meaningful convergence between human and AI ratings for Emotionality. Consequently, H_0 is not rejected for this trait, and H_{1a} is not supported.

Extraversion (X). The correlation for Extraversion was weak and non-significant ($r = .09$, $p = .27$). H_{1a} is not supported, and H_0 is not rejected. Additionally, H_{1b} is not supported, as Extraversion did not demonstrate the most substantial convergence between human and AI ratings; Honesty-Humility exhibited the most substantial alignment.

Agreeableness (A). A weak and non-significant correlation ($r = .05$, $p = .50$) was observed, providing no evidence of notable alignment between human and AI ratings for Agreeableness. This result fails to reject H_0, and H_{1a} is not supported.

Conscientiousness (C). A small but significant positive correlation ($r = .16$, $p = .048$) was found, suggesting some alignment between human and AI ratings for this trait. H_{1a} is supported (and H_0 is rejected).

Openness to Experience (O). The correlation for Openness to Experience was very weak and non-significant ($r = .05$, $p = .57$), providing no evidence of meaningful convergence. Thus, H_0 is not rejected, and H_{1a} is not supported for this trait.

4.4.1.2 Overall Correlation (Across Traits)

An overall correlation was calculated to complement the trait-specific analyses to assess the general relationship between human and AI ratings across all HEXACO traits. The results revealed a statistically significant positive correlation ($r = .29$, $p < .001$), with a 95% confidence interval of [.23, .35], suggesting moderate convergence between the two scoring methods. Therefore, H_{1a} is supported, and H_0 is rejected.

4.4.2 Further Observations

Several incidental correlations were observed between AI and human ratings across different traits. This suggests a potential overlap in how these traits are assessed by the two scoring methods, which warrants further exploration in future research. For example, AI-scored Honesty-Humility exhibited a significant positive correlation with human-rated Extraversion ($r = .30$, $p < .001$), an unexpected finding. Additionally, AI-rated Honesty-Humility was significantly positively correlated with human-rated Agreeableness ($r = .18$, $p = .02$) and Openness to Experience ($r = .18$, $p = .02$), revealing further unexpected covariation between scoring approaches for these different traits.

Further significant correlations include AI-rated Extraversion vs. human-rated Conscientiousness (r = .16, p = .04), AI-rated Agreeableness and human-rated Extraversion (r = .21, p = .01), AI-rated Conscientiousness vs. human-rated Honesty-Humility (r = .17, p = .03), and AI-rated Conscientiousness vs. human-rated Agreeableness (r = .19, p = .02). These findings highlight expected and surprising correlations between traits and raters, suggesting evidence of convergence between AI and human assessment of personality from AVIs.

4.4.3 Summary of Hypotheses Testing

Overall, the results partially support H_{1a}, with significant positive correlations observed for Honesty-Humility and Conscientiousness. These mixed findings suggest some alignment between human and AI ratings, although the lack of covariation for other traits limits the overall support for H_{1a}. The significant overall correlation, indicating a moderate and positive relationship, provides additional evidence that human and AI evaluations align when considered in aggregate, even though individual trait correlations were sometimes weaker or non-significant. H_{1b}, however, was not supported. Contrary to expectations, Extraversion did not demonstrate the most substantial convergence between human and AI ratings. Instead, Honesty-Humility exhibited the most substantial alignment.

4.5 Additional Analyses

As per the following section, the reliability, consistency, and variability in HEXACO ratings across human and AI scoring were examined, along with MANOVA and ANOVA, to further assess the alignment between human and AI scoring in support of the main analysis.

4.5.1 Consistency and Variability in Human and AI HEXACO Ratings

This section examines the consistency of human raters in evaluating HEXACO personality traits and the variability in trait ratings between human evaluations and AI predictions. Consistency, or reliability, in this context refers to the extent to which a measurement produces stable and consistent outcomes when assessing the same construct under similar conditions (Brandmaier et al., 2018). The analysis focuses on two main aspects: the inter-rater reliability of human evaluations, measured using the Intraclass Correlation Coefficient (ICC), and the variability in ratings generated by human raters compared to the AI algorithm. While the single-score-per-participant AI design limited its reliability assessment, examining its variability and standard deviations around mean scores compared to human evaluations offered valuable insights into the consistency of both approaches.

4.5.1.1 Human Rater Reliability Analysis

To assess inter-rater reliability among the fifteen human raters who evaluated the same ten participants, the Intraclass Correlation Coefficient (ICC) was calculated (Field,

2018; Leyland et al., 2014). Following the guidance from Koutsoumpis et al. (2024), the ICC was used to evaluate the level of agreement between raters for personality ratings from the AVIs. The ICC ranges from 0 (no agreement) to 1 (perfect agreement) (Leyland et al., 2014).

Table 8 summarises the ICC and Cronbach's Alpha values for each HEXACO trait. These metrics were identical in this study, as the ICC was calculated using the two-way mixed consistency method in SPSS (IBM, 2020). This alignment is consistent with the study design, where all fifteen raters evaluated the same participants on the same HEXACO traits.

Table 8

Interrater Reliability for HEXACO Traits

Trait	ICC (Average Measures)	α
Honesty-Humility (H)	.67	.67
Emotionality (E)	.66	.66
eXtraversion (X)	.90	.90
Agreeableness (A)	.89	.89
Conscientiousness (C)	.90	.90
Openness to Experience (O)	.83	.83

4.5.1.1.1 Reliability results. The reliability of ratings across the six HEXACO personality traits was evaluated using the Intraclass Correlation Coefficient (ICC, Average Measures), as per Table 8. The results demonstrated varying levels of interrater agreement, with ICC values ranging from .66 to .90. These findings reflect the degree of consistency among the fifteen raters when assessing the ten participants across each trait. Desirable Cronbach's alpha values typically range between .70 and .80, as suggested by Field (2018); however, in psychology, where diversity is inherent, values as low as .50 may be sufficient in early-stage research.

4.5.1.1.2 Trait-specific analysis. Honesty-Humility and Emotionality exhibited moderate or sufficient agreement, with ICCs of .67 and .66, respectively. While slightly lower than others, these values were sufficient for early-stage personality research (Field, 2018; Leyland et al., 2014). In contrast, Extraversion, Agreeableness, Conscientiousness and Openness to Experience demonstrated even higher than desired reliability (Field, 2018), with ICC and alpha values of .90, .89, .90, and .83 respectively. This indicates a very high level of agreement among raters for these traits, suggesting consistency in their evaluation of the HEXACO traits.

4.5.1.2 AI and Human Variability in HEXACO Trait Ratings

The variability in HEXACO ratings between human evaluators and AI predictions was examined. While Chapter 5 discusses this in more detail, Table 9 displays the standard deviations for the HEXACO trait ratings, revealing notable differences in variability between the human and algorithmic scores.

4.5.1.2.1 Human rating variability. The human ratings exhibit standard deviations ranging from .46 to .58, indicating some (or moderate) variability in how different human raters assessed the same traits. This suggests that human evaluators may bring subjective judgement and interpretation to the assessment, which could lead to a wider distribution of scores.

4.5.1.2.2 AI rating consistency. In contrast, the standard deviations for the algorithm ratings are much more minor, ranging from .04 to .10. Compared to humans, this lower variance indicates that the algorithmic predictions are more consistent and clustered around the mean. However, this also suggests a limitation in the algorithm's ability to differentiate between participants' personalities as effectively as human raters (Allik et al., 2010; Moerdyk, 2022).

Table 9

Variability in AI and Human Scores: HEXACO Trait Ratings

Variable	*N*	*M*	*SD*
Human			
H	161	3.74	.58
E	161	2.95	.52
X	161	3.45	.46
A	161	3.65	.49
C	161	3.94	.55
O	161	3.42	.50
AI			
H	161	3.08	.10
E	161	3.00	.08
X	161	3.07	.07
A	161	3.01	.09
C	161	3.15	.08
O	161	2.96	.04

Note. H: Honesty-Humility, E: Emotionality, X: Extraversion, A: Agreeableness, C: Conscientiousness, O: Openness to Experience.

4.5.2 Further Comparison of Human and AI Scoring of HEXACO Traits

To further assess the alignment between human and AI evaluations of HEXACO traits, additional analyses to Spearman correlations were conducted, using MANOVA, and ANOVA, following the guidance of Field (2018) and Tabachnick and Fidell (2019). These analyses aimed to determine the strength of association between human and AI ratings and the presence of systematic differences in trait assessments. MANOVA aimed to identify overarching evaluator differences, followed by ANOVA, which assessed trait-specific variations between evaluators. Further details on these analyses, including complete statistical outputs, are available in Annexure L.

4.5.2.1 Multivariate Effects (MANOVA)

Multivariate Analysis of Variance (MANOVA) examined significant differences between human and AI scoring across the six HEXACO traits. Pillai's Trace was chosen as the primary test statistic because it remains reliable even if the statistical assumption about similar variances (homogeneity of variance-covariance matrices assumption) is not perfectly met (Field, 2018). The results revealed significant multivariate effects for evaluator type (Pillai's Trace = .67, $F = 329.06$, $p < .001$, $\eta^2 = .67$), trait (Pillai's Trace = .71, $F = 77.62$, $p < .001$, $\eta^2 = .71$), and the evaluator x trait interaction (Pillai's Trace = .60, $F = 46.98$, $p < .001$, $\eta^2 = .60$), as per Table L2. These results suggest that not only do human and AI evaluators score traits differently, but the magnitude of these differences depends on the specific trait assessed. In addition to Pillai's Trace, Wilks' Lambda was also inspected for the Evaluator effect, as it provides a more conservative estimate of multivariate differences (Field, 2018; Tabachnick & Fidell, 2019), as shown in Table L2. The value of .33 ($p < .001$) indicates a significant divergence between human and AI scores across all six HEXACO traits. A large effect size ($\eta^2 = .67$) further suggests that evaluator type (human or AI) plays a key role in determining the trait ratings.

4.5.2.2 Trait-Specific Analyses (Repeated Measures ANOVA)

Spearman correlations, as reported in Table 7, provided a measure of convergence between human and AI ratings. However, to further explore potential systematic differences in how the two assessment methods measure the HEXACO traits, repeated measures ANOVA was performed with evaluator type (human vs AI) as the within-subjects factor. As per Table L3, Mauchly's test indicated violations of sphericity for Trait ($x^2 = 83.18$, $p < .001$) and Evaluator x Trait interaction ($x^2 = 63.14$, $p < .001$), which could lead to inaccurate statistical inferences if not corrected (Field, 2018). Therefore, Greenhouse-Geisser corrections were applied to adjust for violations of sphericity, ensuring accurate F-tests ($\varepsilon =$

.81 and .85, respectively). Hereafter, the within-subjects effects of evaluator type (human vs. AI), trait, and their interaction were examined to identify significant differences in HEXACO trait scoring between the two evaluators. The results, as per Table L4, revealed a significant main effect of evaluator type ($F(1,160) = 329.06$, $p < .001$, $\eta^2 = .67$), indicating substantial differences in trait ratings between human and AI evaluators. The main effect of trait was also significant ($F(4.03, 160) = 128.37$, $p < .001$, $\eta^2 = .45$), suggesting variability in how different traits were rated. The interaction effect between evaluator type and trait was significant ($F(4.26, 160) = 77.06$, $p < .001$, $\eta^2 = .33$). As shown in Table 9, human evaluations exhibited greater variability (higher standard deviations), whereas AI ratings were more compressed around lower means. Moreover, the MANOVA and ANOVA findings confirm that differences between human and AI ratings varied across traits and are explored further in the next chapter and in greater detail in Annexure L.

Chapter 5: Discussion

This chapter integrates the study's findings with existing literature to further analyse the convergence and divergence between human and AI evaluations of HEXACO personality traits derived from asynchronous video interviews (AVIs). Building on prior research (e.g., Hickman et al., 2022; Holtrop et al., 2022; Koutsoumpis et al., 2024), this study offers theoretical insights and practical implications for understanding the correlation between AI and human scoring of personality in technology-driven interviews (AVIs) within personnel selection contexts. The results reveal patterns of convergence and divergence, with reliable yet more significant variability in scores from human evaluators compared to the consistent AI evaluation. These findings contribute to ongoing discussions on the comparative and potentially complementary strengths and limitations of human and AI evaluation methods in personnel selection contexts.

5.1 Interpretation of Findings

5.1.1 Human-AI Convergence in Evaluation

Overall, the findings reveal a correlation between human and AI evaluations of HEXACO personality traits from AVIs. Small to moderate and significant convergence was observed for two HEXACO traits, Honesty-Humility and Conscientiousness, likely due to observable linguistic markers that AI could capture in comparison with human observations. In contrast, the other traits, Emotionality, Extraversion, Agreeableness, and Openness to Experience, demonstrated positive yet weaker and non-significant alignment. This likely highlights the algorithm's limitations in interpreting personality expressions beyond the dictionary and the human ability to detect a broader range of linguistic and non-verbal markers of personality for these traits. Interestingly, the Honesty-Humility trait in the current study ($r = .21$, $p = .01$) aligns most closely with the results reported by Holtrop et al. (2022), who reported similar statistics for the HTTP ($r = .20$, $p < .05$) and keyword counting ($r = .17$, $p < .05$) techniques, indicating a degree of alignment in how this trait is assessed. Nonetheless, in the current study, the overall correlation across traits was stronger and more significant than any individual trait-level correlation. The overall correlation, unexpected correlations and potential differences in scoring methods are explored further in the subsequent section.

5.1.1.1 Factors Influencing Across Traits Correlation

The correlation findings, presented in Table 6 and summarised in Table 7, highlight the distinction between evaluating specific traits and examining overall patterns. The overall across-trait correlation is higher and more significant than the correlations observed for

individual traits. This discrepancy may arise from several factors, including the aggregation effect and the influence of additional data points on the correlation coefficient.

5.1.1.1.1 The aggregation effect and macro-level bias. Clark and Avery (1976) noted that aggregated data primarily focuses on macro-level correlations. This approach can inflate coefficients compared to those calculated at the micro or trait level. However, this inflation may obscure the variability and nuances found in individual trait assessments. Therefore, caution is needed when interpreting macro-level findings, as they may not fully capture the complexities of micro-level relationships, which can lead to bias by inflating coefficients when compared to those at the micro or trait level.

5.1.1.1.2 Aggregation reducing errors and amplifying correlations. Personality research, dating back more than 40 years, has shown that aggregating correlations across traits, raters, and items reduces errors, leading to higher overall correlations (Cheek, 1982). Combining traits, such as the HEXACO, mitigates measurement error and reduces noise, yielding stronger correlations (Ostroff, 1993). This study's aggregated ratings likely reduced measurement error and amplified shared variance across traits, resulting in a higher overall correlation. This finding suggests that while individual traits may highlight nuanced differences between human and AI evaluations, aggregated ratings reflect greater convergence in overall evaluative patterns.

5.1.1.1.3 Influence of sample size and data points. Research has also demonstrated that correlation coefficients are influenced by the number of data points and larger sample sizes (Schönbrodt & Perugini, 2013). This study used fewer data points at the trait level, whereas the macro-level or overall correlations included more data points. This difference likely contributed to the observed higher correlation for overall ratings.

5.1.1.1.4 The broader evaluative tendencies of AI. The results also agree that AI algorithms are more effective at capturing broad evaluative tendencies than trait-specific nuances. This is evidenced by the lower variance and standard deviation observed in AI ratings compared to human ratings (Table 3 and 9). Thus, the stronger correlation in aggregated ratings likely reflects agreement between human and AI evaluators on overarching patterns.

Therefore, aggregation of ratings, when the overall correlation across all HEXACO traits was considered, likely enhanced convergence by reducing random noise and context-specific variability, consistent with prior research (Cheek, 1982; Schönbrodt & Perugini, 2013).

5.1.1.2 Unexpected Correlations

Several notable correlations emerged between AI and human ratings across different traits, suggesting some overlap in their assessment methods, as identified in Chapter 4. AI-rated Honesty-Humility correlated with human-rated Extraversion, Agreeableness, and Openness to Experience. Other significant relationships included AI-rated Extraversion with human-rated Conscientiousness, AI-rated Agreeableness with human-rated Extraversion, and AI-rated Conscientiousness with both human-rated Honesty-Humility and Agreeableness.

To some extent, these findings align with prior research on HEXACO trait intercorrelations. Ashton and Lee (2009) found positive correlations between Honesty-Humility and Agreeableness (r = .25 to .26) as well as between Extraversion and Openness to Experience (r = .26), consistent with broader personality research, including Big Five studies. While their study focused on trait relationships rather than cross-method correlation, the parallels may suggest potential underlying similarities in how personality dimensions manifest. However, research on AI-human scoring correlations remains limited, particularly for unexpected trait pairings. Future studies could explore why and how AI and human raters converge on specific traits.

5.1.1.3 Differences in Evaluation Methods

Differences were observed between human and AI evaluations of the HEXACO traits. Human raters exhibited greater variability, reflecting their sensitivity to contextual and interpersonal nuances. In contrast, AI ratings were more uniform but likely constrained by the predefined algorithm and closed dictionary approach. Furthermore, systematic differences between evaluation methods were noted, as confirmed by the trait-specific correlations in the main analysis, as well as the MANOVA and ANOVA results. Subsequent sections will elaborate on the variability, consistency, correlations of ratings, and trends observed in this study's evaluator methods.

5.1.2 Human and AI Scoring Differences

Significant differences in rating patterns emerged between human and AI evaluations. In addition to high interrater reliability, humans demonstrated more variability in their scores than AI, as reflected in the standard deviations around the mean in Table 3. Human raters' higher variability suggested their ability to interpret nuanced behaviours and non-verbal cues, while the consistency of AI stemmed from its reliance on predefined dictionaries and algorithms. Although this may lead to consistency in the AI evaluation method, which may aid replicability, it may overlook critical context-dependent elements required for a more

accurate yet refined personality assessment. The potential explanations or reasons for the high consistency and lower variability observed in AI evaluations are summarised below.

5.1.2.1 Closed-Dictionary vs. Open-Vocabulary Approaches.

The closed-dictionary AI method used in the study relies on a predefined set of words, limiting its ability to capture a broader range of linguistic nuances, as described in Chapters 1 and 2. This restricted vocabulary reduces the range of scores and predictive power, resulting in a more uniform set of ratings. In contrast, open-vocabulary AI models, which are more flexible, tend to generate a broader range of scores by capturing a wider array of linguistic features, providing a more nuanced understanding of personality (Eichstaedt et al., 2021). Therefore, although closed-dictionary methods are less flexible, they tend to generate consistent and reliable results when supported by a sufficiently large dictionary, although at the cost of missing subtleties in the language (Eichstaedt et al., 2021; Holtrop et al., 2022).

5.1.2.2 Influence of Text Length on Rating Variability

Another factor influencing the ratings' variance is the length of the text obtained from the AVIs. More extended text responses provide more opportunities for the algorithm to identify relevant linguistic connections with the HEXACO traits, which is particularly important when using closed-dictionary approaches, as these perform better with longer text (Holtrop et al., 2022). Thus, the algorithm's consistency depends on the amount of text provided. More extensive text data allows for more meaningful personality insights.

5.1.2.3 Improving Algorithm Variability

Expanding the dictionary and increasing the length of text responses would be beneficial to enhance the variability and reliability of the algorithm's predictions. While the AI dictionary used in this study contained 3,337 words, dictionaries with more extensive vocabularies, such as those containing 3,500 or more words, are recommended to improve the reliability of closed-dictionary measures (Holtrop et al., 2022). Open-dictionary machine learning approaches, trained on larger datasets, produce more varied and accurate results than closed-dictionary methods (Eichstaedt et al., 2021; Hickman et al., 2024).

5.1.2.4 Comparison of Human and Algorithmic Approaches

While machine learning (ML) approaches can detect personality indicators that humans may overlook, the closed-dictionary approach used in this study limits the algorithm's ability to gather a comprehensive range of information. On the other hand, human evaluators were provided with video data and transcripts, unlike the algorithm, which was limited to text-based information. Although text-based evaluations were the intended method for personality assessment, human evaluators had the advantage of considering a broader

range of data, including visual and non-verbal cues, when making their evaluations, which may have enabled them to capture a broader set of personality signals or cues. This flexibility may explain the more significant variation observed in human ratings, as human evaluators can consider a broader range of factors or cues in their assessments (Hickman et al., 2022; Hickman et al., 2024; Stachl et al., 2020).

5.1.2.5 Method Effects in Human and AI Personality Scoring

The observed discrepancies between human and AI HEXACO scores may also reflect a predictor method factor related to differences in scoring rather than true variations in personality traits. As explored in Chapter 2, Arthur and Villado (2008) emphasise the need to distinguish between construct-related and method-related variance in personnel selection, highlighting that differences in assessment methods can influence results. Similarly, Lievens and Sackett (2017) describe that predictor method factors play a role in selection outcomes, supporting the notion that the scoring method itself can have an impact on personality ratings.

In line with the Realistic Accuracy Model (RAM), human raters may have been more cue-sensitive than the algorithm. By incorporating visual, verbal, and textual cues in their evaluations, human raters can achieve greater attentiveness to the subtle expression of personality traits in cue detection and utilisation (De Kock et al., 2020; Funder, 1995). As indicated by the MANOVA/ANOVA results, systematic differences exist in how HEXACO traits were measured by the two rating methods. AI ratings tend to be more compressed, whereas human evaluations show greater variability across traits. These findings align with expectations, as AI evaluations relied solely on text-based analysis, whereas human raters had access to broader contextual cues.

However, these findings may demonstrate the possible complementary strengths of human raters and closed-dictionary algorithms in personality assessment. While the algorithm may offer predictive consistency, human evaluators may bring nuanced interpretive abilities that enhance the depth and variability of personality judgements (Cummings et al., 2020; Holtrop et al., 2022). The reliability of human ratings is discussed in the next section.

5.1.3 Rater Training and Interrater Reliability

The interrater reliability statistics (ICC) obtained in the current study closely align with those reported by Koutsoumpis et al. (2024) and demonstrate higher agreement for conscientiousness. Their study, which analysed only extraversion and conscientiousness, reported an ICC of .91 for extraversion and .77 for conscientiousness. In comparison, the current study achieved an ICC of .90 for extraversion and .90 for conscientiousness, indicating more substantial alignment among raters for the latter trait. This high reliability is

likely due to the comprehensive training and orientation process implemented in the present study, which followed and adapted the approach of Koutsoumpis et al. (2024), as described in Chapters 2 and 3. The training also incorporated principles from FORT, Lens Model Theory, and the Realistic Accuracy Model (RAM) to enhance raters' calibration and evaluation consistency. The full table of ICC values is presented in Table 8, with values ranging from .66 to .90 across the measured traits.

5.2 Theoretical and Methodological Considerations

The findings from this study contribute to the ongoing exploration of AI-driven personality assessment by highlighting its limitations and potential when compared to human evaluations. While basic AI-based approaches, such as closed-dictionary keyword counting techniques, offer consistency and scalability, they still face challenges in capturing complex personality traits as effectively as humans. This section examines key theoretical considerations, including the constraints of closed-dictionary methods and the benefits of combining AI with human evaluation to develop a more comprehensive assessment methodology.

5.2.1 Limitations of Closed-Dictionary Approaches

Compared to human ratings, the findings from the study demonstrate the inherent constraints of closed-dictionary or vocabulary algorithms, which rely on fixed vocabularies, keywords, and predefined rules. While these tools may provide consistent evaluations, they often fail to capture the richness of personality expressions, particularly for traits that require the interpretation of subtle cues (Holtrop et al., 2022). However, the consistency of this method may make it suitable as a screening tool or for rank-ordering candidates, after which humans can be involved to provide more nuanced and perceptive evaluations. Considering the lower variance associated with this AI approach, slight variations in personality scores from the mean should be interpreted as more significant than they would be with human evaluations. Thus, slight variations in the scores produced by the algorithm may indicate more substantial differences in the specific trait than the raw AI score might initially suggest.

5.2.2 Complementary Assessment Methods

The study highlights the potential of hybrid approaches that combine the consistency of AI with human evaluators' contextual sensitivity. As applied in this study, AI may offer significant value in screening and ranking candidates in selection contexts. This aligns with the findings of Goretzko and Israel (2022), who suggest that computers excel in screening processes by integrating data and filtering large candidate pools. Human evaluation and self-reports, individually or in combination, can provide deeper and richer insights into

personality, as explored later in this chapter. These methods could help mitigate individual limitations, leading to more balanced and nuanced personality assessments. Figure 8 illustrates a simplified process for integrating AI to complement humans in recruitment and selection.

5.2.3 Use of AVIs and General vs. Trait-Specific Questions

Asynchronous Video Interviews (AVIs) provide a multidimensional platform for personality assessment by combining both textual and visual data. Studies such as Hickman et al. (2022) and Koutsoumpis et al. (2024) have advanced the field by incorporating verbal, non-verbal, and paraverbal information alongside larger datasets, expanding beyond the text-based approach of the current study. However, validity challenges remain in effectively incorporating non-verbal and paraverbal cues in text-based personality evaluations. As a result, AVI providers in practice have shifted to relying solely on transcripts of AVIs instead of incorporating audio or visual features (Koutsoumpis et al., 2024).

This study focused on AVI questions and responses that were not designed to target specific traits, allowing for an exploration of whether non-trait-activating or general interview questions could yield meaningful personality evaluations. This contrasts with the approaches of Koutsoumpis et al. (2024) and Holtrop et al. (2022), who highlight the benefits of using trait-specific questions. Additionally, this approach deviates from studies (Hickman et al., 2022; Holtrop et al., 2022; Koutsoumpis et al., 2024; Manteli & Galanakis, 2022), which argue that trait-specific or activating questions strengthen the expression of personality traits. Nonetheless, the study achieved significant correlations and reliability evidence using this approach with more general interview questions, providing some optimism for utilising general interview or competency-based questions in the future, as explored by Liff et al. (2024). With further research, these questions could support effective personality evaluations, which may further enhance the effectiveness and efficiency of personnel selection assessments via AVIs.

5.2.3.1 Automated Transcription

This study addressed an AVI limitation: automated AVI transcription. Overcoming the limitation highlighted by Holtrop et al. (2022) regarding the challenges of manual transcription, this study adopted automated transcription using Microsoft Office. As mentioned by Holtrop et al., (2022), practically, organisations are unlikely to favour manual transcription due to its costliness and inefficiency, which can compromise data quality and reduce the validity of findings. By using automatic transcription, this study saved

considerable time and potentially enhanced the data's reliability, contributing to the observed correlations.

5.3 Practical Implications

The study highlights the strengths of AI algorithms and the need for further advancements to enhance accuracy and contextual relevance in personality evaluation. While text-based analysis from AVIs currently represents the gold standard in practice (Hickman et al., 2022; Holtrop et al., 2022; Koutsoumpis et al., 2024), the closed-dictionary keyword-counting technique may pose constraints regarding variability and the scope of analysis. On the other hand, open-vocabulary techniques and machine learning approaches could significantly enhance the ability of AI scoring to assess personality traits more comprehensively and accurately while maintaining consistency (Eichstaedt et al., 2021; Hickman et al., 2024; Holtrop et al., 2022).

For organisational applications, combining AI evaluations with human supervision and insights, alongside other self-report measures, could improve fairness and depth in recruitment and selection processes for personality evaluations. AI offers consistency and scalability, while human evaluators contribute nuanced understanding, ensuring a balanced and comprehensive assessment approach (Fan et al., 2023; Holtrop et al., 2022; Koutsoumpis et al., 2024; Stevenor et al., 2024).

The adoption of AI in selection and recruitment is also shaped by broader organisational and theoretical considerations. Given that organisations seek AI primarily for its time-saving benefits (Campion & Campion, 2023; Oostrom et al., 2024), it is noteworthy that in this study, human evaluators took at least five to 30 minutes or even longer to rate a single participant, depending on the response length. In contrast, the algorithm completed the task almost instantly, scoring 805 responses (161 participants, five responses each) in approximately 10 seconds. While this may demonstrate the potential time-saving benefits of using AI, organisations and recruiters may integrate AI-driven personality assessment tools based on factors outlined in the Technology Acceptance Model (TAM) and institutional theory (König et al., 2010; Oostrom et al., 2013). Their willingness to use AI may depend on perceived usefulness and ease of use, aligning with TAM, while institutional theory highlights the role of perceived applicant acceptance, cost considerations, and industry norms in shaping selection procedures. These factors suggest that the role of AI in personnel selection may become clearer and continue to expand as its acceptance grows and as its perceived benefits outweigh concerns about reliability, accuracy and fairness.

This study serves as an example of how to assess the scoring validity of AI-based evaluations, as recommended in the Society for Industrial and Organisational Psychology's (SIOP) "Considerations and Recommendations for the Validation and Use of AI-Based Assessments for Employee Selection" (Nye et al., 2023). By comparing human and AI scoring of HEXACO traits in AVIs, this research may contribute to the validation process of AI in personnel selection.

5.4 Limitations of the Study

A number of limitations should be acknowledged in this study. Firstly, the study included the use of a closed-dictionary AI approach, which likely constrained the analysis of nuanced personality expressions. Potential biases or limitations within the dataset, such as a restricted dictionary of words, may also have influenced the results (Holtrop et al., 2022).

Additionally, the study's reliance on non-trait-specific asynchronous video interviews (AVIs) questions, a relatively small sample size, and a USA-only sample may limit the generalisability of the findings. Since the AVIs were not explicitly designed to elicit HEXACO traits, participants' responses may not fully reflect the breadth or complexity of these dimensions, potentially leading to less precise personality assessments. The small and geographically limited sample further reduces the diversity and representativeness of the data, making it harder to identify patterns that apply to broader or international populations (Bourdage et al., 2021; Cheon et al., 2020; Hickman et al., 2022; Holtrop et al., 2022; Koutsoumpis et al., 2024). These limitations highlight a possible need for future research employing trait-specific prompts and larger, more diverse samples to improve the robustness and applicability of the findings.

Furthermore, while human raters relied on verbal and non-verbal cues, AI depended solely on textual input, which may have impacted the correlations and variance observed in AI ratings. Future research could address this by using a broader AI model incorporating open-vocabulary techniques and machine-learning capabilities to evaluate a wider range of cues, increasing alignment with human scores and providing greater variance in ratings.

Finally, studies conducted in controlled laboratory settings, such as mock interviews with participants who are not job seekers (e.g., Prolific samples), may lack fidelity and realism, as the interview performance does not impact any job outcomes (Highhouse, 2009). To address this limitation in the current study, human raters were trained, and all held Bachelor's degrees in Organisational Psychology, with all pursuing or having attained postgraduate studies. Additionally, both interviewees and raters were compensated, and their responses were checked for quality to ensure reliability.

5.5 Future Research Directions

In addition to the study's key findings, future research can explore several alternative aspects that could further enrich the understanding and application of AI and human evaluations in personality assessment, as provided below.

5.5.1 Demographics of Raters

Firstly, it could be valuable to examine the demographics of the raters, as these may influence how personality traits are assessed (De Kock et al., 2020; Yun et al., 2005). This includes considering factors such as personality, age, cultural background, education, and experience, as well as the impact of personality evaluation from AVIs. While the current study aimed to involve postgraduate students in organisational psychology, mainly Honours degree students and those with master's and doctorate level (PhD) degrees, further research exploring a more diverse sample could be valuable.

5.5.2 Sample Diversity and Cross-Cultural Research

The study relied on a USA-based interviewee sample, which may not fully represent the South African or broader global diversity. Future research should consider expanding the sample to include participants and raters from different regions and cultural contexts. This may allow researchers to assess whether biases or variations emerge based on demographic factors. A larger and more representative sample could increase the robustness and generalisability of the findings, which may ensure broader applicability.

Cross-cultural research is another potentially important area to explore. Raters' cultural values, beliefs, norms, and ethnic affiliations could influence the ratings in selection contexts (Bourdage et al., 2021; De Kock et al., 2020). Understanding how raters from different cultures interpret and evaluate personality traits could provide valuable insights into the universality or cultural specificity of certain traits, aligning with the realistic accuracy model (RAM) and the concepts of cue detection and 'good information' (Letzring et al., 2006). Refining AI algorithms to account for these cultural differences could also enhance their accuracy and applicability in diverse settings.

5.5.3 Enhancing Interviewee Responses

Future studies could consider encouraging interviewees to provide longer, more detailed responses when responding to the AVI questions. The brief nature of many responses in the current study may limit the depth of the evaluations, potentially missing nuanced aspects of personality. Longer, more thoughtful responses can provide richer data for analysis, allowing for a more precise and comprehensive evaluation of personality traits. This

approach could also contribute to more thoroughly comparing between human and AI-generated evaluations.

5.5.4 Advancements in AI Assessment Tools

The study successfully replicated human AVI scoring with algorithmic scoring at a rudimentary level. However, moving forward in research and practice, a multimodal AI approach, which assesses not only text but also verbal, non-verbal and related cues, appears to be the most promising for enabling the AI assessment method to analyse more complex personality descriptions more effectively and accurately (Hickman et al., 2022; Holtrop et al., 2022; Koutsoumpis et al., 2024).

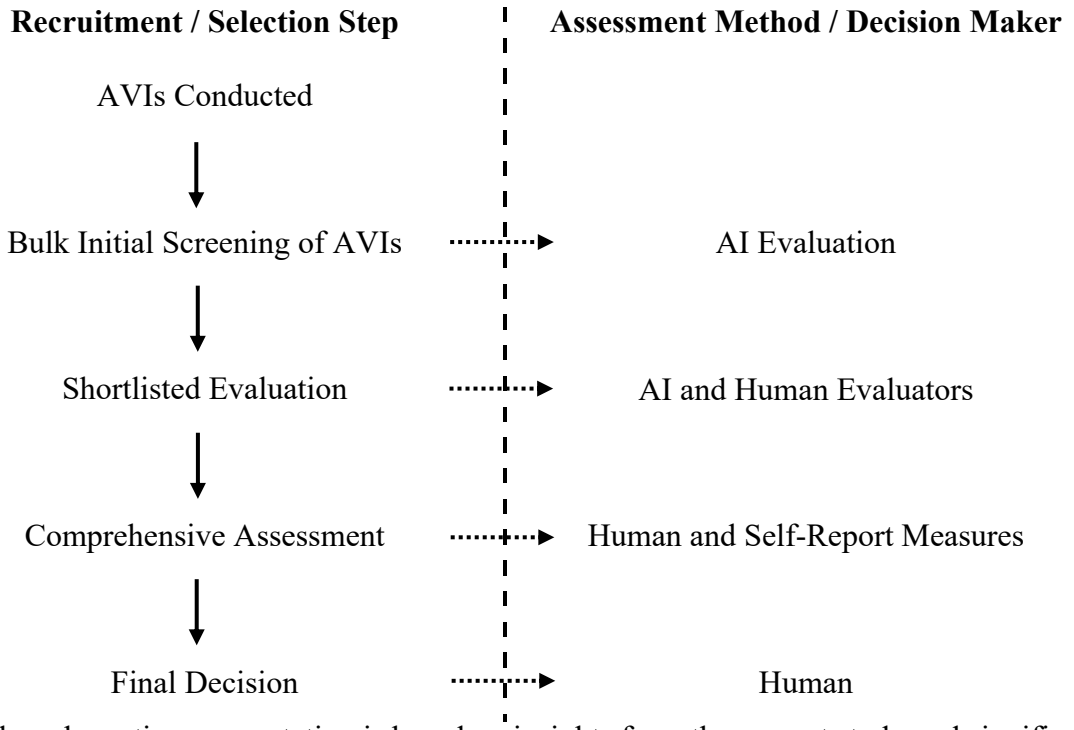
Additionally, as highlighted by Holtrop et al. (2022), automatic personality evaluation from AVIs (AVI-PA) should not replace traditional personality assessment methods but should instead serve as a supplementary approach. For now, hybrid methodologies integrating AI with human evaluators represent a promising future research avenue. While AI can process large volumes of data quickly and consistently, human evaluators contribute intuition and contextual understanding. Combining these approaches and self-reports could likely enhance the accuracy and effectiveness of personality assessment tools. Based on the current study's results, Figure 8 explain the recommended assessment tools for each selection step.

5.5.5 Triangulation of Assessments

Although it was beyond the scope of this study, it is worth noting that future analyses could incorporate self-report ratings as a potential 'true score' source to triangulate human (observer ratings) and AI assessments in alignment with the realistic accuracy model. The current methods analysed only the correlation between the two approaches, AI and human evaluators, and not necessarily their accuracy, nor can we determine which one is the most 'correct'. However, by including self-report ratings, individuals' self-perceptions can be captured, revealing insights into their behaviours and traits that human observers or AI might overlook, thus enriching the accuracy and depth of the overall personality assessment. Therefore, this approach of multiple correlation could enhance validity by comparing multiple perspectives and emphasises accuracy in personality judgements (Connolly et al., 2007; De Kock et al., 2020; Powell, 2008; Powell & Bourdage, 2016).

Figure 8

Schematic Depiction of AI's Role and Position in the AVI Recruitment and Selection Process



Note. The schematic representation is based on insights from the current study and significant sources (see Alexander III et al., 2025; Cascio & Aguinis, 2024; Fan et al., 2023; Hickman et al., 2022; Holtrop et al., 2022; Koutsoumpis et al., 2024).

Note. Solid arrows indicate the steps in the selection process, whereas the dotted arrow represents each step's applicable assessment method or decision-maker.

5.6 South African Relevance and Legislation

For AI evaluations to comply with the assessment-related legislation in South Africa, such as the Health Professions Act, the Employment Equity Act, and the Protection of Personal Information Act (Coetzee et al., 2021; Foxcroft & Roodt, 2013; HPCSA, 2008), further studies would need to be conducted, particularly focusing on score triangulation between human raters, AI evaluations, and self-report measures. This may allow for a more comprehensive assessment of AI-generated personality evaluations' accuracy, validity, and reliability. While the present study demonstrated correlations between AI and human ratings, indicating construct or convergent (mono-trait) validity (Foxcroft & Roodt, 2013; Plouffe et al., 2017), and consistency in scoring, reflecting reliability, additional research is necessary to meet legislative requirements. Given that South African regulations mandate that psychological assessments be valid, reliable, and fairly applied (Employment Equity Act, 1998, Chapter 2, Section 8, p. 15), future studies should further examine AI-based scoring methods to ensure compliance with these legal and ethical standards.

Chapter 6: Conclusion

This study contributes to understanding the complex interplay between human and AI evaluations of HEXACO personality traits, particularly in the context of asynchronous video interviews (AVIs) and personnel selection. By integrating the findings with existing literature, the study comprehensively analysed both the convergences and divergences between human and AI assessments. It also addressed the theoretical and practical implications of using AI closed-dictionary evaluation approaches in personality psychology, building on prior research (e.g., Hickman et al., 2022; Holtrop et al., 2022; Koutsoumpis et al., 2024).

The results of this study were mixed as they revealed correlation, convergence and differences in how human and AI evaluators assessed personality traits. Moderate and significantly positive overall convergence across traits was observed. Small to moderate, significant, positive trait-based convergence was also found for Honesty-Humility and Conscientiousness. However, weaker and non-significant alignment was observed for the other traits. These findings reflect both the strengths and limitations of each scoring method. Human evaluators demonstrated greater sensitivity to nuanced, context-dependent behaviours, reflected in their ratings' variability. In contrast, AI ratings, driven by a predefined algorithm and a closed vocabulary, showed greater consistency, which could aid in ensuring reliability and replicability. However, AI scoring was constrained by an inability to interpret complex, subtle cues beyond the programmed algorithm. Since these subtle variations often contribute to accurate personality assessments, the closed-dictionary keyword counting approach may fail to capture the depth and richness of certain traits, particularly those less easily expressed through predefined vocabulary (Holtrop et al., 2022). Nevertheless, AI evaluations can provide valuable insights, especially in contexts requiring large-scale data processing, but they may lack the interpretative flexibility that human evaluators bring, as observed in the current study. The findings of this study, alongside those of Hickman et al. (2022), Holtrop et al. (2022) and Koutsoumpis et al. (2024), suggest that AI algorithms would benefit from advancements, particularly in the area of open-vocabulary models and machine learning techniques. Such improvements appear to enhance the capacity of AI to interpret a broader range of linguistic and even non-verbal cues, making it a more flexible and context-sensitive evaluator. As AI continues to evolve, its integration with human evaluators could bridge the gap between consistency and contextual understanding, enabling more accurate and efficient personality assessments.

The differences observed between human and AI evaluations suggest that by combining the strengths of both methods, a hybrid approach could offer a more comprehensive and accurate personality assessment. AI can be a valuable tool for ranking candidates in large-scale selection processes, providing consistency and efficiency (Alexander III et al., 2025; Campion & Campion, 2023; Hunkenschroer & Luetge, 2022). However, human evaluations appear more suited for capturing the subtleties of interpersonal dynamics and non-verbal cues, ensuring a more holistic understanding of personality, if a closed-vocabulary approach is chosen as the AI counterpart (Holtrop et al., 2022; Koutsoumpis et al., 2024). This complementary approach may enhance the validity and fairness of personality assessments, particularly in high-stakes contexts such as recruitment.

The study's findings also open avenues for future research in several areas. First, significant and positive correlations were observed between AI and human ratings across different traits, beyond the six specific HEXACO traits expected to correlate, suggesting some overlap in how the evaluators see or assess these traits. Secondly, the demographic characteristics of both raters and participants warrant further exploration, as these factors could influence personality assessments (Bourdage et al., 2021; De Kock et al., 2020). Expanding the sample of participants and raters to include more diverse cultural, educational, and professional backgrounds could help address potential biases and improve the generalisability of the findings. Further investigation into non-trait-specific AVIs could provide valuable insights into how different interview questions influence personality expression and evaluator ratings while offering relevant data points for personality assessment. This approach could also streamline selection processes by integrating competency-related factors and personality insights into a single measurement intervention (see Liff et al., 2024). Finally, machine learning models that combine advanced AI techniques, insights from human evaluations, and other assessment tools, such as self-reports, could represent a promising direction for future research. By integrating multiple data sources, these models could offer a more accurate and nuanced understanding of personality (Booth et al., 2021; Campion & Campion, 2023; Koutsoumpis et al., 2024).

In conclusion, the findings of this study demonstrate the importance of refining both traditional and technology-driven assessment methods. This research highlights the value of leveraging human insight and AI capabilities to advance personality evaluation. By integrating the strengths of both approaches, future studies can develop more innovative, reliable, and comprehensive methods for understanding and measuring personality traits in diverse settings and applying these methods effectively across various organisations.

References

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16*(2), 270–301. https://doi.org/10.1177/1094428112470848

Alexander III, L., Song, Q. C., Hickman, L., & Shin, H. J. (2025). Sourcing algorithms: Rethinking fairness in hiring in the era of algorithmic recruitment. *International Journal of Selection and Assessment, 33*(1), Article e12499. https://doi.org/10.1111/ijsa.12499

Allik, J., Realo, A., Mõttus, R., Esko, T., Pullat, J., & Metspalu, A. (2010). Variance determines self-observer agreement on the Big Five personality traits. *Journal of Research in Personality, 44*(4), 421–426. https://doi.org/10.1016/j.jrp.2010.04.005

Anderson, N., Ones, D. S., Sinangil, H. K., & Viswesvaran, C. (Eds.). (2001). *Handbook of industrial, work & organizational psychology: Volume 1: Personnel psychology* (Vols. 1–2). SAGE Publications Ltd. https://doi.org/10.4135/9781848608320

Anglim, J., & O'Connor, P. (2019). Measurement and research using the Big Five, HEXACO, and narrow traits: A primer for researchers and practitioners. *Australian Journal of Psychology, 71*(1), 16–25. https://doi.org/10.1111/ajpy.12202

Arkkelin, D. (2014). *Using SPSS to understand research and data analysis*. Psychology Curricular Materials, 1. Valparaiso University. https://scholar.valpo.edu/psych_oer/1

Arthur, W. Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*(2), 435–448. https://doi.org/10.1037/0021-9010.93.2.435

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*(2), 150–166. https://doi.org/10.1177/1088868306294907

Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*(4), 340–345. https://doi.org/10.1080/00223890902935878

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: A review of research and theory. *Personality and Social Psychology Review, 18*(2), 139–152. https://doi.org/10.1177/1088868314523838

Ashton, M. C., & Lee, K. (2020). Objections to the HEXACO model of personality

structure and why those objections fail. *European Journal of Personality, 34*(4), 492-510. https://doi.org/10.1002/per.2242

Atkinson, R. K., Mayer, R. E., & Merrill, M. M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology, 30*(1), 117–139. https://doi.org/10.1016/j.cedpsych.2004.07.001

Auer, E. M., Mersy, G., Marin, S., Blaik, J., & Landers, R. N. (2022). Using machine learning to model trace behavioral data from a game-based assessment. *International Journal of Selection and Assessment, 30*(1), 82–102. https://doi.org/10.1111/ijsa.12363

Bae, C. L., Therriault, D. J., & Redifer, J. L. (2019). Investigating the testing effect: Retrieval as a characteristic of effective study strategies. *Learning and Instruction, 60,* 206–214. https://doi.org/10.1016/j.learninstruc.2017.12.008

Basch, J. M., Brenner, F., Melchers, K. G., Krumm, S., Dräger, L., Herzer, H., & Schuwerk, E. (2021). A good thing takes time: The role of preparation time in asynchronous video interviews. *International Journal of Selection and Assessment, 29*(3-4), 378–392. https://doi.org/10.1111/ijsa.12341

Basch, J. M., Melchers, K. G., & Büttner, J. C. (2022). Preselection in the digital age: A comparison of perceptions of asynchronous video interviews with online tests and online application documents in a simulation context. *International Journal of Selection and Assessment, 30*(4), 639–652. https://doi.org/10.1111/ijsa.12403

Baumgartner, S., Bartels, L., & Levashina, J. (2024). Improving structured interview acceptance through training. *International Journal of Selection and Assessment, 32*(6), 512–520. https://doi.org/10.1111/ijsa.12473

Bergh, Z. C., & Geldenhuys, D. (2013). *Psychology in the work context* (5th ed.). Oxford University Press.

Bergner, R. M. (2020). What is personality? Two myths and a definition. *New Ideas in Psychology, 57,* Article 100759. https://doi.org/10.1016/j.newideapsych.2019.100759

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6*(2), 205–212. https://www.jstor.org/stable/257876

Berry, K. J., & Mielke, P. W., Jr. (2000). A Monte Carlo investigation of the Fisher Z transformation for normal and nonnormal distributions. *Psychological Reports, 87*(3, Suppl), 1101–1114. https://doi.org/10.2466/pr0.2000.87.3f.1101

Bishara, A. J., & Hittner, J. B. (2015). Reducing bias and error in the correlation coefficient

due to nonnormality. *Educational and Psychological Measurement, 75*(5), 785–804. https://doi.org/10.1177/0013164414557639

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

Bourdage, J. S., Derous, E., Holtrop, D., Roulin, N., de Kock, F. S., Powell, D. M., & Dunlop, P. D. (2021). *Cross-cultural interview practices: Research and recommendations* (SIOP White papers). Society for Industrial and Organizational Psychology. https://www.siop.org/Portals/84/docs/White%20Papers/crosscultint.pdf

Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Lindenberger, U., & Hertzog, C. (2018). Precision, reliability, and effect size of slope variance in latent growth curve models: Implications for statistical power analysis. *Frontiers in Psychology, 9*, 294. https://doi.org/10.3389/fpsyg.2018.00294

Brunswik, E. (1956). *Perception and the representative design of experiments*. University of California Press.

Buchanan, E., Valentine, K. D., & Pavlacic, J. (2021). Hypothesize once, plan twice. *Academia Letters*, *Article 1274*. https://doi.org/10.20935/AL1274

Camp, R. R., Schulz, E., Vielhaber, M. E., & Wagner-Marsh, F. (2011). Human resource professionals' perceptions of interviewer training. *Journal of Managerial Issues, 23*(3), 250–268. https://www.jstor.org/stable/23209116

Campion, M. C., Campion, M. A., Campion, E. D., Reider, M. H., & Chen, G. (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

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. https://doi.org/10.1111/peps.12621

Canagasuriam, D., & Lukacik, E.-R. (2025). ChatGPT, can you take my job interview? Examining artificial intelligence cheating in the asynchronous video interview. *International Journal of Selection and Assessment, 33*(1), Article e12491. https://doi.org/10.1111/ijsa.12491

Cascio, W. F., & Aguinis, H. (2024). Applied psychology in talent management (9th ed.). SAGE Publications.

Chapman, C., & Stolee, K. T. (2016, July). Exploring regular expression usage and context in

Python. In *Proceedings of the 25th International Symposium on Software Testing and Analysis* (pp. 282–293). https://doi.org/10.1145/2931037.2931073

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Cheek, J. M. (1982). Aggregation, moderator variables, and the validity of personality tests: A peer-rating study. *Journal of Personality and Social Psychology, 43*(6), 1254. https://doi.org/10.1037/0022-3514.43.6.1254

Chen, R., Rafaeli, E., Bar-Kalifa, E., Gilboa-Schechtman, E., Lutz, W., & Atzil-Slonim, D. (2018). Moderators of congruent alliance between therapists and clients: A realistic accuracy model. *Journal of Counseling Psychology, 65*(6), 703. https://doi.org/10.1037/cou0000285

Cheon, B. K., Melani, I., & Hong, Y. (2020). How USA-centric is psychology? An archival study of implicit assumptions of generalizability of findings to human nature based on origins of study samples. *Social Psychological and Personality Science, 11*(7), 928–937. https://doi.org/10.1177/1948550620927269

Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., & Wang, L. C. (2024). Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia Pacific Journal of Management, 41*(3), 745–783. https://doi.org/10.1007/s10490-023-09871-y

Christiansen, N., & Tett, R. (2013). *Handbook of personality at work*. Routledge. https://doi.org/10.13140/2.1.3105.9847

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. *Proceedings of the 2016 International Conference on Management of Data*, 2201–2206. https://doi.org/10.1145/2882903.2912574

Clark, W. A., & Avery, K. L. (1976). The effects of data aggregation in statistical analysis. *Geographical Analysis, 8*(4), 428-438.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Coetzee, M., Botha, E., & De Beer, L. (Eds.). (2021). *Personnel psychology: An applied perspective* (3rd ed.). Oxford University Press Southern Africa.

Cong, G., Fan, W., Geerts, F., Jia, X., & Ma, S. (2007). Improving data quality: Consistency and accuracy. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)* (pp. 315-326).

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

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment, 15*(1), 110–117.

Cummings, T. G., Worley, C. G., & Donovan, P. (2020). *Organizational development and change*. Cengage Learning EMEA Ltd.

Dai, Y., Jayaratne, M., & Jayatilleke, B. (2022). Explainable personality prediction using answers to open-ended interview questions. *Frontiers in Psychology, 13*, Article 865841. https://doi.org/10.3389/fpsyg.2022.865841

Das, K. R., & Imon, A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics, 5*(1), 5–12. https://doi.org/10.11648/j.ajtas.20160501.12

Department of Communications and Digital Technologies. (2024). South Africa National Artificial Intelligence Policy Framework (Towards the Development of South Africa National Artificial Intelligence Policy). https://fwblaw.co.za/wp-content/uploads/2024/08/South-Africa-National-AI-Policy-Framework.pdf

Department of Labour. (1998). *Employment Equity Act 55 of 1998*. Government Gazette, 400 (19370). Retrieved from https://www.labour.gov.za/DocumentCenter/Acts/Employment%20Equity/Act%20-%20Employment%20Equity%201998.pdf

De Kock, F., Lievens, F., & Born, M. (2020). The profile of the “good judge” in HRM: A systematic review and agenda for future research. *Human Resource Management Review, 30*(1), 1-21. https://doi.org/10.1016/j.hrmr.2018.09.003

De Raad, B., & Barelds, D. P. H. (2020). Models of personality structure. In P. J. Corr & G. Matthews (Eds.), *The Cambridge handbook of personality psychology* (pp. 115–128). Cambridge University Press. https://doi.org/10.1017/9781108264822.012

Dunlop, P. D., Holtrop, D., & Wee, S. (2022). How asynchronous video interviews are used

in practice: A study of an Australian-based AVI vendor. *International Journal of Selection and Assessment*, *30*(4), 448–455. https://doi.org/10.1111/ijsa.12372

Dupré, S., & Wille, B. (2025). Personality development goals at work: A new frontier in personality assessment in organizations. *International Journal of Selection and Assessment, 33*(1). https://doi.org/10.1111/ijsa.12490

Ebrahim, F. (2022). *Designing semi-automated video interviews (SAVI): Does stimulus format (video vs. text) of instructions and interview questions affect applicant perceptions of social presence?* [Master's dissertation, University of Cape Town, School of Management Studies]. *University of Cape Town*. https://open.uct.ac.za/server/api/core/bitstreams/dc58b5ef-c6ab-43af-8f58-6d055ac5e991/content

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V. A., Smith, L. K., Buffone, A., Iwry, J., Seligman, M. E. P., & Ungar, L. H. (2021). Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods, 26*(4), 398–427. https://doi.org/10.1037/met0000349

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

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology, 94*(2), 334–346. https://doi.org/10.1037/0022-3514.94.2.334

Feher, A., & Vernon, P. A. (2021). Looking beyond the Big Five: A selective review of alternatives to the Big Five model of personality. *Personality and Individual Differences, 169*, 110002. https://doi.org/10.1016/j.paid.2020.110002

Feng, C., Wang, H., Lu, N., Tu, X. M., Lian, Q., & Zhou, X. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry, 26*(2), 105–109. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). SAGE Publications.

Foxcroft, C., & Roodt, G. (2013). *Introduction to psychological assessment in the South African context* (4th ed.). Oxford University Press.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*(4), 652–670. https://rap.ucr.edu/FunderPR1995.pdf

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21*(3), 177–182. https://doi.org/10.1177/0963721412445309

George, D., & Mallery, P. (2019). *IBM SPSS statistics 26 step by step: A simple guide and reference* (16th ed.). Routledge. https://doi.org/10.4324/9780429056765

Goretzko, D., & Israel, L. S. F. (2022). Pitfalls of machine learning-based personnel selection: Fairness, transparency, and data quality. *Journal of Personnel Psychology, 21*(1), 37-47. https://doi.org/10.1027/1866-5888/a000287

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94*(5), 1336–1344. https://doi.org/10.1037/a0016476

Grotenhuis, M., & Visscher, C. (2014). *How to use SPSS syntax*. SAGE Publications, Inc. https://doi.org/10.4135/9781483378503

Health Professions Council of South Africa (HPCSA). (2008). *Health Professions Act 56 of 1974: Regulations defining the scope of the profession of psychology*. Government Notice R993 in Government Gazette 31433, 16 September 2008, amended by GN R704 GG 34581, 2 September 2011. https://www.hpcsa.co.za/Uploads/professional_boards/psb/regulations/regulations_gnr993_2008.pdf

Heimann, A. L., Ingold, P. V., Lievens, F., Melchers, K. G., Keen, G., & Kleinmann, M. (2021). Actions define a character: Assessment centers as behavior-focused personality measures. *Personnel Psychology.* Advance online publication. https://doi.org/10.1111/peps.12478

Henry, S., Baker, W., Bratko, D., Jern, P., Kandler, C., Tybur, J. M., de Vries, R. E., Wesseldijk, L. W., Zapko-Willmes, A., Booth, T., & Mõttus, R. (2024). Nuanced HEXACO: A meta-analysis of HEXACO cross-rater agreement, heritability, and rank-order stability. *Personality and Social Psychology Bulletin*, 1-20. https://doi.org/10.1177/01461672241253637

Herde, C. N., & Lievens, F. (2023). Multiple, Speeded Assessments Under Scrutiny: Underlying Theory, Design Considerations, Reliability, and Validity. *Journal of Applied Psychology, 108*(3), 351-373. https://doi.org/10.1037/apl0000603

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/apl0000938

Hickman, L., Saef, R., Ng, V., Woo, S. E., Tay, L., & Bosch, N. (2024). Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews. *Human Resource Management Journal, 34*(2), 255–274. https://doi.org/10.1111/1748-8583.12356

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), Article 3. https://doi.org/10.25035/pad.2019.03.003

Highhouse, S. (2009). Designing experiments that generalize. *Organizational Research Methods, 12*(3), 554–566. https://doi.org/10.1177/1094428107300396

Hilgert, L., Kroh, M., & Richter, D. (2016). The effect of face-to-face interviewing on personality measurement. *Journal of Research in Personality, 63*, 133–136. https://doi.org/10.1016/j.jrp.2016.05.006

Hinds, J., & Joinson, A. N. (2024). Digital data and personality: A systematic review and meta-analysis of human perception and computer prediction. *Psychological Bulletin, 150*(6), 727–766. https://doi.org/10.1037/bul0000430

HireVue. (n.d.-a). *HireVue*. LinkedIn. Retrieved May 25, 2025, from https://www.linkedin.com/company/hirevue/

HireVue. (n.d.-b). HireVue. https://www.saasworthy.com/product/hirevue

HireVue. (2024). *The 2024 global guide to AI in hiring*. HireVue. https://www.hirevue.com/resources/report/ai-in-hiring-report

Hmoud, B. I. F., & Várallyai, L. (2019). Will artificial intelligence take over human resources recruitment and selection? *University of Debrecen*. https://dea.lib.unideb.hu/server/api/core/bitstreams/cff8e1b9-7db6-47e9-9642-6fe8e19f565d/content

Holtrop, D., Oostrom, J. K., Van Breda, W. R., Koutsoumpis, A., & De Vries, R. E. (2022). Exploring the application of a text-to-personality technique in job interviews. *European Journal of Work and Organizational Psychology, 31*(6), 799–816. https://doi.org/10.1080/1359432X.2022.2051484

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics, 178*(4), 977–1007. https://doi.org/10.1007/s10551-022-05049-6

IBM. (2020). *Confidence interval for Cronbach's Alpha in SPSS*. IBM Support. Retrieved from https://www.ibm.com/support/pages/confidence-interval-cronbachs-alpha-spss#:~:text=Resolving%20The%20Problem,available%20for%20the%20Reliability%20procedure

Jayaraman, A. K., Ananthakrishnan, G., Trueman, T. E., & Cambria, E. (2024). Chapter four – Text-based personality prediction using XLNet. *Advances in Computers, 132*, 49–65. https://doi.org/10.1016/bs.adcom.2023.08.002

Julian, A. M., Novitsky, C., Lee, K., & Ashton, M. C. (2022). Convergent validity of three brief six-factor measures of personality. *Personality and Individual Differences, 188,* 111436. https://doi.org/10.1016/j.paid.2021.111436

König, C. J., Klehe, U.-C., Berchtold, M., & Kleinmann, M. (2010). Reasons for being selective when choosing personnel selection procedures. *International Journal of Selection and Assessment, 18*(1), 17–27. https://doi.org/10.1111/j.1468-2389.2010.00485.x

Koszalka, T. A., Pavlov, Y., & Wu, Y. (2021). The informed use of pre-work activities in collaborative asynchronous online discussions: The exploration of idea exchange, content focus, and deep learning. *Computers & Education, 161*, 104067. https://doi.org/10.1016/j.compedu.2020.104067

Koutsoumpis, A., Ghassemi, S., Oostrom, J. K., Holtrop, D., Van Breda, W., Zhang, T., & De Vries, R. E. (2024). Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2023.108128

Kühl, N., Schemmer, M., Goutier, M., & Satzger, G. (2022). Artificial intelligence and machine learning. *Electronic Markets, 32*(4), 799–815. https://doi.org/10.1007/s12525-022-00598-0

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., Roulin, N., & Oostrom, J. K. (2023). Diversity and technology—Challenges for the next decade in personnel selection. *International Journal of Selection and Assessment, 31*(3-4), 355–360. https://doi.org/10.1111/ijsa.12439

Lee, K., Ogunfowora, B., & Ashton, M. C. (2005). Personality traits beyond the Big Five:

Are they within the HEXACO space? *Journal of Personality, 73*(5), 1437–1463. https://doi.org/10.1111/j.1467-6494.2005.00354.x

Letzring, T. D., Wells, S. M., & Funder, D. C. (2006). Information quantity and quality affect the realistic accuracy of personality judgment. *Journal of Personality and Social Psychology, 91*(1), 111–123. https://doi.org/10.1037/0022-3514.91.1.111

Leyland, A. H., Groenewegen, P. P., & Michalos, A. C. (2014). Intraclass correlation coefficient (ICC). In *Encyclopedia of quality of life and well-being research* (pp. 3367–3368). Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_1528

Lievens, F., & Chapman, D. (2019). Recruitment and selection. In *SAGE Handbook of Human Resource Management* (pp. 123-150). SAGE. https://ink.library.smu.edu.sg/lkcsb_research/5988

Lievens, F., & Corstjens, J. (2018). New approaches to selection system design in healthcare: The practical and theoretical relevance of a modular approach. In F. Patterson & L. Zibarras (Eds.), *Selection and recruitment in the healthcare professions*. Palgrave Macmillan. https://doi.org/10.1007/978-3-319-94971-0_7

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., & Sanchez, J. I. (2007). Can training improve the quality of inferences made by raters in competency modeling? A quasi-experiment. *Journal of Applied Psychology, 92*(3), 812–818. http://doi.org/10.1037/0021-9010.92.3.812

Liff, J., Mondragon, N., Gardner, C., Hartwell, C. J., & Bradshaw, A. (2024). Psychometric properties of automated video interview competency assessments. *Journal of Applied Psychology, 109*(6), 921–948. https://doi.org/10.1037/apl0001173

Lukacik, E. R., Bourdage, J. S., & Roulin, N. (2022). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review, 32*, 100789. https://doi.org/10.1016/j.hrmr.2020.100789

Lukacik, E.-R., & Bourdage, J. S. (2025). Does design matter? The (limited) effects of six asynchronous video interview design features on impression management, reactions, and evaluations. *International Journal of Selection and Assessment, 33*, e12511. https://doi.org/10.1111/ijsa.12511

Luthans, F., Luthans, B. C., & Luthans, K. W. (2021). *Organizational behavior: An evidence-based approach* (14th ed.). Information Age Publishing.

Macan, T. H. (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

Manteli, M., & Galanakis, M. (2022). The new foundation of organizational psychology: Trait activation theory in the workplace. *Psychology Research, 12*(12), 939–945. https://doi.org/10.17265/2159-5542/2022.12.004

Martin, M. A. (2019). *Training interviewers to spot 'faking' in employment interviews: Can frame of reference training enhance cue detection, cue utilisation, and overall profile accuracy for rating candidate deceptive impression management?* [Master's thesis, University of Cape Town]. University of Cape Town Repository. http://hdl.handle.net/11427/30932

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

Mayer, R. E. (2021). Evidence-based principles for how to design effective instructional videos. *Journal of Applied Research in Memory and Cognition, 10*(2), 229–240. https://doi.org/10.1016/j.jarmac.2021.03.007

McKinney, W. (2022). *Python for data analysis* (3rd ed.). O'Reilly Media, Inc.

Mirowska, A. (2020). AI evaluation in selection: Effects on application and pursuit intentions. *Journal of Personnel Psychology, 19*(3), 142–149. https://doi.org/10.1027/1866-5888/a000258

Moerdyk, A. (2022). *The principles and practice of psychological assessment* (Third Edition ed.). Van Schaik.

Mori, M., Sassetti, S., Cavaliere, V., & Bonti, M. (2024). A systematic literature review on artificial intelligence in recruiting and selection: A matter of ethics. *Personnel Review*. https://doi.org/10.1108/PR-03-2023-0257

Mosier, K. L., & Kirlik, A. (2004). Brunswik's lens model in human factors research: Modern applications of a classic theory. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 48, No. 3, pp. 350–354). SAGE Publications. https://doi.org/10.1177/154193120404800316

Mount, M. K., Barrick, M. R., & Strauss, J. P. (1994). Validity of observer ratings of the big five personality factors. *Journal of Applied Psychology, 79*(2), 272. https://doi.org/10.1037/0021-9010.79.2.272

Naidu, S., Saurombe, M. D., & Mogoai, D. V. (2025). The candidate experience of virtual

interviews in a South African company. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 51*(0), Article a2226. https://doi.org/10.4102/sajip.v51i0.2226

Neumann, M., Niessen, A. S. M., Hurks, P. P. M., & Meijer, R. R. (2023). Holistic and mechanical combination in psychological assessment: Why algorithms are underutilized and what is needed to increase their use. *International Journal of Selection and Assessment.* https://doi.org/10.1111/ijsa.12416

Nikolaou, I. (2021). What is the role of technology in recruitment and selection? *The Spanish Journal of Psychology, 24*, e2. https://doi.org/10.1017/SJP.2021.6

Norris, A. E., & Aroian, K. J. (2004). To transform or not to transform skewed data for psychometric analysis: That is the question. *Nursing Research, 53*(1), 67–71.

Norton, L. W., Howell, A. W., DiGirolamo, J. A., & Hayes, T. L. (2024). Using artificial intelligence in consulting psychology. *Consulting Psychology Journal, 76*(2), 137–162. https://doi.org/10.1037/cpb0000274

Nye, C., Hough, L., Jones, K., Landers, R., Locklear, T., Macey, W., et al. (2023). *Considerations and recommendations for the validation and use of AI-based assessments for employee selection*. Society for Industrial and Organizational Psychology. https://www.siop.org/wp-content/uploads/2024/06/Considerations-and-Recommendations-for-the-Validation-and-Use-of-AI-Based-Assessments-for-Employee-Selection-January-2023.pdf

Ones, D. S., Stanek, K. C., & Dilchert, S. (2025). Beyond change: Personality-environment alignment at work. *International Journal of Selection and Assessment, 33*(1), Article e12507. https://doi.org/10.1111/ijsa.12507

Oosthuizen, R. M. (2022). The Fourth Industrial Revolution – Smart technology, artificial intelligence, robotics, and algorithms: Industrial psychologists in future workplaces. *Frontiers in Artificial Intelligence, 5*, Article 913168. https://doi.org/10.3389/frai.2022.913168

Oostrom, J. K., van der Linden, D., Born, M. P., & van der Molen, H. T. (2013). New technology in personnel selection: How recruiter characteristics affect the adoption of new selection technology. *Computers in Human Behavior, 29*(6), 2404–2415. https://doi.org/10.1016/j.chb.2013.05.025

Oostrom, J. K., Holtrop, D., Koutsoumpis, A., van Breda, W., Ghassemi, S., & de Vries, R. E. (2024). Applicant reactions to algorithm-versus recruiter-based evaluations of an asynchronous video interview and a personality inventory. *Journal of Occupational*

and Organizational Psychology, 97(1), 160–189. https://doi.org/10.1111/joop.12465

Orji, K., Roulin, N., & Bangerter, A. (2025). Is anybody watching me? Effects of information about evaluators on applicants' use of impression management in asynchronous video interviews. *International Journal of Selection and Assessment, 33*, Article e12515. https://doi.org/10.1111/ijsa.12515

Ostroff, C. (1993). Comparing correlations based on individual-level and aggregated data. *Journal of Applied Psychology, 78*(4), 569. https://doi.org/10.1037/0021-9010.78.4.569

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17,* 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2014). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology.* Advance online publication. http://dx.doi.org/10.1037/pspp0000020

Patel, A. (2022). How audio-visual stimuli in automated asynchronous video interviews affect applicant reactions: Social presence, fairness, and organisational attractiveness (Master's dissertation, University of Cape Town). *University of Cape Town*. https://open.uct.ac.za/server/api/core/bitstreams/b1c535f3-c479-4cda-adb1-33e8f5066890/content

Patel, R. D., Powell, D. M., Roulin, N., & Spence, J. R. (2025). Tell me more! Examining the benefits of adding structured probing in asynchronous video interviews. *International Journal of Selection and Assessment, 33*, e12514. https://doi.org/10.1111/ijsa.12514

Pek, J., Wong, O., & Wong, A. C. M. (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Frontiers in Psychology, 9*, 2104. https://doi.org/10.3389/fpsyg.2018.02104

Petersheim, C., Lahey, J., Cherian, J., Pina, A., Alexander, G., & Hammond, T. (2022). Comparing student and recruiter evaluations of computer science résumés. *IEEE Transactions on Education, 66*(2), 130–138.

Platt, J. (2012). The history of the interview. In J. F. Gubrium, J. A. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The SAGE handbook of interview research: The complexity of the craft* (2nd ed., pp. 9–26). Sage Publications.

Plouffe, R. A., Paunonen, S. V., & Saklofske, D. H. (2017). Item properties and the convergent validity of personality assessment: A peer rating study. *Personality and*

Individual Differences, 111, 96–105. https://doi.org/10.1016/j.paid.2017.01.051

Ployhart, R. E., Schmitt, N., Tippins, N. T., & Chen, G. (2017). Solving the supreme problem: 100 years of selection and recruitment at the *Journal of Applied Psychology*. *Journal of Applied Psychology, 102*(3), 291–304. https://doi.org/10.1037/apl0000081

Potočnik, K., Anderson, N. R., Born, M., Kleinmann, M., & Nikolaou, I. (2021). Paving the way for research in recruitment and selection: Recent developments, challenges and future opportunities. *European Journal of Work and Organizational Psychology,30*(2), 159–174. https://doi.org/10.1080/1359432X.2021.1904898

Powell, D. M. (2008). *Assessing personality in the employment interview: The impact of rater training and individual differences on rating accuracy* (Doctoral dissertation, The University of Western Ontario). ProQuest Dissertations Publishing. https://www.proquest.com/docview/304319772

Powell, D. M., & Goffin, R. D. (2009). Assessing personality in the employment interview: The impact of training on rater accuracy. *Human Performance, 22*(5), 450–465. https://doi.org/10.1080/08959280903248450

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

Prinzing, M., Bounds, E., Melton, K., Glanzer, P., Fredrickson, B., & Schnitker, S. (2024). Can an algorithm tell how spiritual you are? Using generative pretrained transformers for sophisticated forms of text analysis. *Journal of Personality*, *92*(1), 1-13. https://doi.org/10.1111/jopy.13006

Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies, 1*(1), 33–36. https://doi.org/10.1002/hbe2.117

Rizi, M. S., & Roulin, N. (2024). Does media richness influence job applicants' experience in asynchronous video interviews? Examining social presence, impression management, anxiety, and performance. *International Journal of Selection and Assessment, 32*(1), 54–68. https://doi.org/10.1111/ijsa.12448

Roediger, H. L., Putnam, A. J., & Sumeracki, M. A. (2011). Ten benefits of testing and their applications to educational practice. In *Psychology of Learning and Motivation* (Vol. 55, pp. 1-36). Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Roch, S. G., & O'Sullivan, B. J. (2003). Frame of reference rater training issues: Recall, time, and behavior observation training. *International Journal of Training and Development, 7*(2), 93–107. https://doi.org/10.1111/1468-2419.00174

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology, 85*(2), 370–395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Romano, D., Costantini, G., Richetin, J., & Perugini, M. (2023). The HEXACO adjective scales and its psychometric properties. *Assessment*, 1–23. https://doi.org/10.1177/10731911231153833

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). McGraw-Hill.

Roulin, N., Langer, M., & Bourdage, J. S. (2021). "I" feel (s) left out: The importance of information and communication technology in personnel selection research. *Industrial and Organizational Psychology, 14*(3), 423-427. https://doi.org/10.1017/iop.2021.79

Salgado, J. F. (2017). Personnel selection. In *Oxford encyclopedia of research in psychology*. https://doi.org/10.1093/acrefore/9780190236557.013.8

Schmid Mast, M., Bangerter, A., Bulliard, C., & Aerni, G. (2011). How accurate are recruiters' first impressions of applicants in employment interviews? *International Journal of Selection and Assessment, 19*(2), 198–208. https://doi.org/10.1111/j.1468-2389.2011.00547.x

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609-612. https://doi.org/10.1016/j.jrp.2013.05.009

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023). Personality traits in large language models. https://doi.org/10.48550/arXiv.2307.00184

Soleimani, M., Intezari, A., & Pauleen, D. J. (2022). Mitigating cognitive biases in developing AI-assisted recruitment systems: A knowledge-sharing approach. *International Journal of Knowledge Management, 18*(1). https://doi.org/10.4018/IJKM.290022

Soliev, B. N., Odilov, A., & Sh, A. (2023). Leveraging Python for enhanced Excel functionality: A practical exploration. *Al-Farg'oniy avlodlari, 1*(4), 267–271. https://cyberleninka.ru/article/n/leveraging-python-for-enhanced-excel-functionality-a-practical-exploration/pdf

Speer, A. B., Wegmeyer, L. J., & Delacruz, A. Y. (2022). Factors leading to interview question decisions: Introducing the model of interviewer question preferences.

International Journal of Selection and Assessment, 30(5), 392–410. https://doi.org/10.1111/ijsa.12383

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., & Bühner, M. (2020). Personality research and assessment in the era of machine learning. *European Journal of Personality, 34*, 613–631. https://doi.org/10.1002/per.2257

Stenegren, F. (2023). *An analysis of data cleaning tools: A comparative analysis of the performance and effectiveness of data cleaning tools*. Mittuniversitetet. https://www.diva-portal.org/smash/get/diva2:1776404/FULLTEXT01.pdf

Stevenor, B. A., Hickman, L., Zickar, M. J., Wimbush, F., & Beck, W. (2024). Validity evidence for personality scores from algorithms trained on low-stakes verbal data and applied to high-stakes interviews. *International Journal of Selection and Assessment, 32*(6), 544-560. https://doi.org/10.1111/ijsa.12480

Suen, H. Y., Chen, M. Y., & Lu, S. H. (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? *Computers in Human Behavior*, *98*, 93–101. https://doi.org/10.1016/j.chb.2019.04.012

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Trull, T. J., Widiger, T. A., Useda, J. D., Holcomb, J., Doan, B. T., Axelrod, S. R., Stern, B. L., & Gershuny, B. S. (1998). A structured interview for the assessment of the Five-Factor Model of Personality. *Psychological Assessment, 10*(3), 229–240. https://doi.org/10.1037/1040-3590.10.3.229

University of Queensland. (n.d.). Dictation and transcription using Microsoft 365. *UQ Systems Training*. https://systems-training.its.uq.edu.au/systems/collaboration-tools/microsoft-365/blog/dictation-and-transcription-using-microsoft-365

van Aarde, N., Meiring, D., & Wiernik, B. M. (2017). The validity of the Big Five personality traits for job performance: Meta-analyses of South African studies. *International Journal of Selection and Assessment, 25*(3), 223–239. https://doi.org/10.1111/ijsa.12175

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

van Kempen, F. (2024). Predicting personality from trait-relevant vs trait-irrelevant questions

using trait activation theory in asynchronous video interviews (Master's thesis). Tilburg University.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*(5), 372–376. https://doi.org/10.1037/h0026244

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

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

Wortman, J., Lucas, R. E., & Donnellan, M. B. (2012). Stability and change in the Big Five personality domains: Evidence from a longitudinal study of Australians. *Psychology and Aging, 27*(4), 867–874. https://doi.org/10.1037/a0029322

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Yun, G. J., Donahue, L. M., Dudley, N. M., & McFarland, L. A. (2005). Rater personality, rating format, and social context: Implications for performance appraisal ratings. *International Journal of Selection and Assessment, 13*(2), 97–107.

Annexure A

Predictor Method Factors in Modular Framework (Lievens & Sackett, 2017)

Table A1

Modular Framework.

Predictor Method Factor	Description
Stimulus Format	Stimulus format refers to the modality used to present test stimuli, such as text, pictures, audio, and video. Some formats, like audiovisual presentations, tend to reduce cognitive load and subgroup differences compared to text alone.
Contextualisation	The extent the stimuli are embedded in realistic contexts. More contextualised stimuli tend to have higher validity than decontextualised stimuli like traditional personality tests.
Stimulus Presentation Consistency	The standardisation in how stimuli are presented. Higher consistency such as structured interviews reduces measurement error and increases validity.
Response Format	Response format refers to the modality that test-takers use to respond, such as multiple choice, written responses, audio or video response. Constructed response formats tend to reduce cognitive load and subgroup differences compared to multiple choice.
Response Evaluation Consistency	This refers to the standardisation in how responses are scored. Higher consistency through rubrics, training, or automated scoring increases reliability and validity.
Information Source	Combining assessment sources, such as self-reports and behavioural observations, tends to increase validity.
Instructions	Instructions refer to how explicit the directions are about the perspective candidates should take. Overly specific instructions can reduce validity by making the situation too strong.

Note. Modular framework obtained and summarised from Lievens and Sackett (2017).

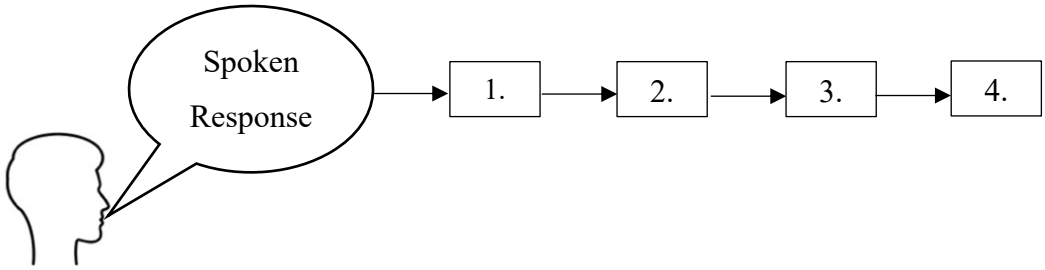
Annexure B

AI Scoring Process

As depicted in Figure B1, the spoken words of the AVI participant are transcribed using MS Office 365. The algorithm then searches for matching words based on the loadings for each HEXACO trait in the HEXACO dictionary, tallying the final matches per trait to provide a score. The converted score can be compared to scale descriptors in Table B3, for a possible interpretation of the individual's personality according to the HEXACO model.

Figure B1

Target 1. Question 5: How do you manage your time and prioritise tasks?



Step 1: Transcribed Answer

The spoken words in the video of the AVI participant are transcribed using MS Office 365, as shown in the example from an actual participant used for training purposes below.

*" In the past I have tended to do the **easy** tasks first thing in the morning. Just to get them marked off the To Do List and make myself feel like I was making progress toward my task list. However, I've recently taken some online courses on time management and **productivity** and **learn** that most people are more **productive** in the morning. I certainly am… And… So, to instead attack the more difficult tasks, the tasks that may take input from other people so that I would need to make phone calls or contact people via text or e-mail for their input, get that ball rolling first thing in the morning so that I can successfully finish that test throughout the day and then perhaps later in the afternoon. During… Maybe less productive… Maybe a little bit sleepier time in the late afternoon, I can do less brain taxing tasks, such as responding to emails, **organising** lists and so forth. "*

Step 2: Algorithm and Dictionary Matching

The algorithm analyses transcribed text by identifying words that correspond to HEXACO trait loadings in the HEXACO dictionary, as per Table B1, ultimately calculating a score based on the total matches per trait.

Table B1

HEXACO Dictionary and Algorithm Word Matching

	H	E	X	A	C	O
easy	0,104	-0,224	0,162	0,319	-0,123	-0,283
productive	0,121	-0,029	-0,024	-0,079	0,270	-0,005
difficult	-0,328	0,310	-0,205	-0,342	-0,010	0,273
organise	0,113	0,042	-0,008	0,017	0,396	-0,246
learn	0,068	-0,110	0,068	0,005	0,174	0,183
Total	0,078	-0,011	-0,007	-0,079	0,707	-0,078
Conversion	3,156	2,978	2,986	2,842	4,414	2,844

Step 3: Conversion

The raw score from the algorithm is converted to align with the human rating scale as per the formula below:

Converted Score = ((Raw Score + 1) x 2) + 1

Table B2

AI Score Conversion

	H	E	X	A	C	O
Total	0,078	-0,011	-0,007	-0,079	0,707	-0,078
Conversion	3,156	2,978	2,986	2,842	4,414	2,844

Step 4: Interpretation

Table B3

HEXACO Descriptors

Rating	H	E	X	A	C	O
1: Very Low	Very high self-importance; strongly materialistic; willing to break rules for gain.	Very unemotional; detached; fearless.	Very reserved; socially indifferent; awkward in the spotlight.	Very resentful; holds grudges; highly critical and stubborn.	Very disorganised; avoids challenges; impulsive.	Uninterested in art and sciences; avoids creativity.
2: Low	High self-importance; often materialistic; may manipulate for success.	Unemotional; detached; low anxiety.	Reserved; somewhat indifferent to socialising.	Often resentful; critical; somewhat stubborn.	Often disorganised; avoids hard tasks; somewhat impulsive.	Limited interest in art/science; avoids creativity.
3: Average	Average self-importance; rarely breaks rules; little interest in luxury.	Moderate emotions; some anxiety under stress.	Balanced; neither overly social nor reserved.	Moderate patience and forgiveness; balanced in criticism.	Moderately organised; balanced impulse control.	Moderate interest; balanced approach to creativity.
4: High	Avoids manipulation; little temptation to break rules or seek luxury.	Fearful of danger; seeks support; feels anxiety.	Confident; enjoys social interactions; positive energy.	Cooperative; forgiving; controls temper well.	Organised; disciplined; strives for accuracy.	Enjoys art and nature; curious; imaginative.
5: Very High	Rejects manipulation; uninterested in luxury or social status.	Very fearful; high need for support and empathy.	Very confident; loves socialising; high enthusiasm.	Very cooperative and forgiving; extremely patient.	Very organised; highly disciplined; perfectionistic.	Deeply absorbed in art and science; highly curious and imaginative.

Note. Obtained and Adapted from Koutsoumpis et al. (2024)

Annexure C

Recruitment Invitation for Human Evaluators

Curious about how humans and AI compare in evaluating personality from verbal expression, such as interview responses?

Invitation to Participate in the Study:

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI / ML Scoring.

What We Need:

Approximately ten (10) UCT Honours students in Industrial/Organisational Psychology.

What You Will Be Doing:

No prior experience with interviews or personality scoring? No problem! You will receive appropriate training as part of the study. After training, you will score and evaluate personality based on the HEXACO model using provided interview responses.

What's in It for You?

- Receive payment and a certificate for your participation*
- Receive training on the universally used and respected HEXACO personality model.
- Gain hands-on experience with personality assessment and interview scoring.
- Enhance your CV with valuable research experience.
- Contribute to cutting-edge research in Industrial/Organisational Psychology.

Interested?

If interested, please email crnjac009@myuct.ac.za before Friday, 26 July 2024, at 17:00. In your email, include a brief motivation (one or two sentences) explaining why you would like to participate. Applications will not be considered after this deadline. If successful, you will be contacted with details on the next steps by no later than 9 August 2024.

About the Researcher:

Jaco Cronje is currently completing his Master's degree in Organisational Psychology at UCT and holds professional registration with the HPCSA as a Psychometrist (Independent Practice) and Student Psychologist.

Technical Information:

- The study has obtained ethical clearance: **COM/00889/2024.**
- The study has received DSA100 clearance to involve UCT students.
- *Note: Payment is only provided at the end of the study upon meeting all objectives to the required standard.

Annexure D

Letter of Informed Consent for Research Participants (Human Evaluators)

Dear Student Participant.

Research Study Overview

The Department of Organisational Psychology at UCT is conducting a research study as part of the requirements for the Master's degree programme. The study aims to explore the convergence between human evaluators and an Artificial Intelligence (AI) text-to-personality algorithm in assessing personality from Asynchronous Video Interviews (AVIs). This research has received ethical approval (COM/00889/2024) from both the Faculty of Commerce Ethics in Research Committee and the Department of Student Affairs.

Participation Details

Your participation in this study is voluntary. After completing the training, you will score and evaluate personality traits based on the HEXACO model using provided interview responses. Your anonymised evaluations/ratings will be used in this study. If you choose to participate, rest assured that your ratings and responses will be treated with confidentiality and anonymity. The findings will be used solely for academic purposes within UCT.

Incentive and Quality Assurance

Upon completion of the study, you will be eligible for a cash incentive of R149 (Honours' students) or R179 (Master's students) per hour of participation. Please note that the incentive will be provided at the end of the year or study, not immediately after participation, and only if the quality standards of the evaluations are met. Quality assurance will be conducted.

Voluntary Participation and Withdrawal

You are free to withdraw from the study at any point without facing any negative consequences. However, withdrawing early will mean that you will not receive the incentive. Your participation poses no harm to yourself or others.

Time Commitment

The training will last approximately 120 minutes. Your total participation in the study is estimated to be a maximum of 5 hours, including training time. However, this may vary depending on the final number of evaluators.

Contact Information

Please do not hesitate to contact the researcher for further clarification: Jaco Cronje (crnjac009@myuct.ac.za) or the research supervisor: Francois De Kock.

Annexure E

Preparatory Training of Human Evaluators

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring

Human Evaluator Training: Prework

Jaco (JF) Cronje
CRNJAC009

Faculty of Commerce
School of management studies
Masters in Organisational Psychology

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Ethical Approval: COM/00889/2024

Prework Overview & Objectives

- The main objective of the prework is to familiarise yourself with the HEXACO personality model, the evaluation criteria, and the concept of Asynchronous Video Interviewing (AVI). This prework is designed to take approximately 30–60 minutes.
- The prework is important as it sets the foundation for the training session, which will build on the assumption of this knowledge.
- To track progress and assess the starting level of understanding at the beginning of the training session, you will be required to write a short knowledge test (± 10 minutes).
- Therefore, it is required that you read through the information before the training session.
- Note any uncertainties or areas where you need further clarification, and bring the questions to the training session.
- This prework includes three sections: Section 1: Asynchronous Video Interviewing, Section 2: HEXACO Model, and Section 3: Scoring Criteria.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Research Study Overview

This study aims to examine the convergence between human and artificial intelligence (AI) scoring of HEXACO personality traits from asynchronous video interviews (AVIs). Participants in this study will engage in scoring exercises and training sessions designed to enhance their understanding of the HEXACO model, AVIs, and the nuances of personality assessment from video and text-based (transcribed) responses. The findings may contribute to a broader understanding of how AI can complement human judgement in psychological evaluations, particularly in personality assessments during recruitment and selection processes.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

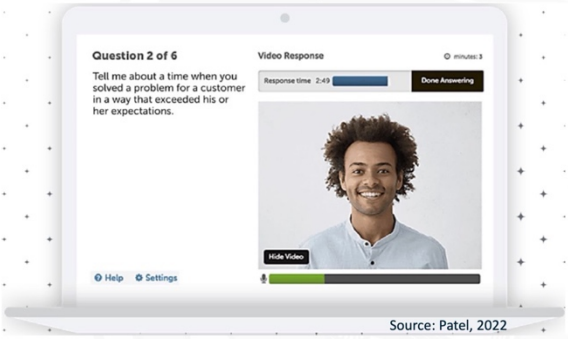
Section 1: Asynchronous Video Interviewing (AVI)

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Asynchronous Video Interviewing (AVI)

- **The study uses data from asynchronous video interviewing (AVI).**
- **AVI is a modern way to do job interviews without needing to be online at the same time as the interviewer.**
- **Instead of a live in-person or virtual discussion, candidates record themselves answering questions on a website. These recorded responses are then reviewed later by either a person or computer software.**

AVI Example



Source: Patel, 2022

You will be watching these videos and reading the transcripts to evaluate the participant's personality based on the HEXACO model.

(Lukacik et al., 2022)

Section 2: HEXACO Personality Model

The HEXACO personality model is a six-dimensional framework for understanding human personality traits, including: **Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience**. It was developed by Ashton and Lee in the early 2000s as an expansion of the traditional Big Five personality traits, with the addition of Honesty-Humility.

(Lee & Ashton, n.d.)

HEXACO Inventory: Summary

Trait	Low Scores	High Scores
Honesty-Humility	Flatter others to get what they want, inclined to break rules for personal profit, motivated by material gain, strong sense of self-importance.	Avoid manipulating others for personal gain, little temptation to break rules, uninterested in wealth and luxuries, feel no special entitlement to elevated social status.
Emotionality	Not deterred by the prospect of physical harm, little worry even in stressful situations, little need to share concerns with others, emotionally detached from others.	Experience fear of physical dangers, anxiety in response to life's stresses, need emotional support from others, empathy and sentimental attachments with others.
eXtraversion	Consider themselves unpopular, feel awkward when the centre of social attention, indifferent to social activities, feel less lively and optimistic.	Feel positively about themselves, confident when leading or addressing groups, enjoy social gatherings and interactions, experience positive feelings of enthusiasm and energy.
Agreeableness (vs. Anger)	Hold grudges against those who have harmed them, critical of others' shortcomings, stubborn in defending their point of view, feel anger readily in response to mistreatment.	Forgive wrongs suffered, lenient in judging others, willing to compromise and cooperate, easily control temper.
Conscientiousness	Unconcerned with orderly surroundings or schedules, avoid difficult tasks or challenging goals, satisfied with work containing some errors, make decisions on impulse or with little reflection.	Organise time and physical surroundings, work in a disciplined way toward goals, strive for accuracy and perfection, deliberate carefully when making decisions.
Openness to Experience	Unimpressed by most works of art, little intellectual curiosity, avoid creative pursuits, little attraction toward ideas that may seem radical or unconventional.	Absorbed in the beauty of art and nature, inquisitive about various domains of knowledge, use imagination freely in everyday life, take an interest in unusual ideas or people.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

(Lee & Ashton, n.d.)

Section 3: Scoring Criteria

- In this section, you will learn about the scoring criteria used to evaluate the AVI responses.
- Each response and HEXACO trait will be rated on a 5-point scale.
  - 1 indicates "very low indication of the trait"
  - 5 indicates "very high indication of the trait"
- This scale will help ensure consistency and objectivity in evaluating candidates' AVI responses
- Please review the scoring guidelines carefully, as they form the basis for accurate and fair assessments.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring

Scoring Rubric / Scale

Score	1	2	3	4	5
HEXACO Trait Presence (Rating)	Very low	Low	Average	High	Very high

You need to decide, based on your knowledge and the training, by watching the AVI and reading the transcript:

1. Which HEXACO traits are present (hopefully all 6).
2. How strongly each trait is expressed by allocating a score per interview response and HEXACO trait.
3. There are 5 interview questions (and responses) per participant.
4. In these responses, they portray their personalities to us.
5. The next few pages will provide you with more indicators per trait

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Koutsoumpis et al. (2024)

Trait 1: Honesty-Humility

Scale	1 = Very low	2 = Low	3 = Average	4 = High	5 = Very high
Honesty-Humility	• Very high sense of self-importance • Strongly motivated by material gain • Tempted to "bend" laws for personal profit, • Highly likely to flatter others for personal success	• High sense of self-importance, they • Motivated by material gain, • Tempted to "bend" laws for personal profit, • May flatter others for success.	• Feel an average sense of self-importance, • Not strongly motivated by luxury and status, • Seldomly break rules and manipulate or flatter others for personal gain.	• Avoid manipulating others for personal gain, • Feel little temptation to break rules, • Less interested in wealth and luxuries, • Feel little need for social status or privilege.	• Avoid manipulating others for personal gain, • Feel no temptation to break rules, • Uninterested in wealth and luxuries, • Do not feel any entitlement to elevated social status or privilege.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Koutsoumpis et al. (2024)

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring

Trait 2: Emotionality

Scale	1 = Very low	2 = Low	3 = Average	4 = High	5 = Very high
Emotionality	• Very unemotional, • Detached, • Independent, • Feel no anxiety or fear even under stressful or frightening circumstances.	• Unemotional, • Detached, • Independent • Feel little anxiety or fear even under stressful or frightening circumstances.	• Neither emotional nor unemotional • Feel some - but not too much - anxiety and fear when faced with stressors.	• Fear physical dangers • Anxiety in response to stressors. • Need emotional support, attachments and empathy	• High fear of physical dangers • Anxiety in response to stressors. • Feel a very strong need for emotional support from others, attachments and empathy.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Koutsoumpis et al. (2024)

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring

11

Trait 3: eXtraversion

Scale	1 = Very low	2 = Low	3 = Average	4 = High	5 = Very high
eXtraversion	• Very reserved and awkward when at the centre of social attention. • Consider themselves very unpopular and much less lively than others. • Highly indifferent to social activities.	• More reserved and awkward than others when at the centre of social attention. • Consider themselves somewhat unpopular and less lively than others. • Mostly indifferent to social activities.	• Neither reserved nor confident when leading a group or at the centre of attention. • Feel neither popular nor unpopular • Average levels of enthusiasm and energy.	• Feel more confident than others when leading or addressing groups of people. • Enjoy social gatherings and interactions. • Feel mostly positive about themselves. • Experience mostly positive feelings of enthusiasm and energy.	• Very confident when leading or addressing groups of people. • Very much enjoy social gatherings and interactions. • Feel very positive about themselves. • Very often experience positive feelings of enthusiasm and energy.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Koutsoumpis et al. (2024)

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring

12

Trait 4: Agreeableness

Scale	1 = Very low	2 = Low	3 = Average	4 = High	5 = Very high
Agreeableness	• Feel lots of anger in response to mistreatment. • Bear strong grudges against those who have insulted or deceived them. • Very critical of others' shortcomings. • Very stubborn in defending their point of view.	• Feel anger in response to mistreatment. • Tend to bear grudges against those who have insulted or deceived them. • More critical than others of people's shortcomings • Tend to be stubborn in defending their point of view.	• Average levels of patience and anger in response to mistreatment. • Neither very forgiving nor bearing strong grudges against those who have insulted or deceived them. • Neither very lenient nor highly critical of others' shortcomings. • Sometimes stubborn and sometimes compromising and cooperative.	• Tend to compromise and cooperate with others. • Tend to be - more than others - lenient in judging others. • Tend to remain patient. • Tend to control their temper and forgive the wrongs that they have suffered.	• Always compromise and cooperate with others. • Very lenient in judging others. • Always remain patient. • Very easily control their temper. • Always forgive the wrongs that they have suffered.

Koutsoumpis et al. (2024)

Trait 5: Conscientiousness

Scale	1 = Very low	2 = Low	3 = Average	4 = High	5 = Very high
Conscientiousness	• Very unconcerned with orderly surroundings or schedules. • Strongly avoid difficult tasks or challenging goals. • Are - much more than others - satisfied with work that contains some errors. • Often decide on impulse with no reflection.	• Tend to be - more than others - unconcerned with orderly surroundings or schedules. • Avoid difficult tasks or challenging goals. • Tend to be satisfied with work that contains some errors. • Decide on impulse or with little reflection.	• Have an average concern for the orderliness of their surroundings and schedules. • Neither very accurate nor do they make a lot of errors in their work. • Sometimes deliberate carefully and sometimes decide on impulse. • Average discipline when working toward their goals.	• Tend to organise things (both time and physical surroundings) • Work in a disciplined way toward their goals. • Strive for accuracy and perfection in their tasks. • Deliberate - more than others - careful when deciding.	• Always organise things (both time and physical surroundings) • Work in a highly disciplined way toward their goals. • Very strongly strive for accuracy and perfection in their tasks. • Deliberate very carefully when deciding.

Koutsoumpis et al. (2024)

Trait 6: Openness

Scale	1 = Very low	2 = Low	3 = Average	4 = High	5 = Very high
Openness	• Very unimpressed by most works of art. • Have no interest in the natural or social sciences. • Strongly avoid creative pursuits. • Feel no attraction toward ideas that may seem radical or unconventional	• Unimpressed by most works of art. • Feel little interest in the natural or social sciences. • Tend to avoid - more than others - creative pursuits. • Feel little attraction toward radical or unconventional ideas.	• Average interest in art and the natural or social sciences. • Do not take a strong interest, but neither do they avoid creative pursuits. • Feel relatively neutral toward radical or unconventional ideas.	• Tend to become absorbed in the beauty of art and nature. • Tend to feel intellectual curiosity in various domains of knowledge. • They - more than others - use their imagination freely in everyday life. • Take interest in unusual ideas or people.	• Very easily absorbed in art and nature. • Feel strong intellectual curiosity in various domains of knowledge. • Always use their imagination freely in everyday life. • Take a strong interest in unusual ideas or people.

Koutsoumpis et al. (2024)

References

Koutsoumpis, A., Ghassemi, S., Oostrom, J. K., Holtrop, D., Van Breda, W., Zhang, T., & De Vries, R. E. (2024). Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning. Computers in Human Behavior. https://doi.org/10.1016/j.chb.2023.108128

Lee, K., & Ashton, M. C. (n.d.). *The HEXACO personality inventory - Revised: A measure of the six major dimensions of personality*. HEXACO. https://hexaco.org/scaledescriptions

Lukacik, E. R., Bourdage, J. S., & Roulin, N. (2022). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review, 32*. https://doi.org/10.1016/j.hrmr.2020.100789

Patel, A. (2022). *How audio-visual stimuli in automated asynchronous video interviews affect applicant reactions: Social presence, fairness, and organisational attractiveness* (Master's dissertation, University of Cape Town)

Annexure F

Training Material: Human Evaluators

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring

Human Evaluator Training

Jaco (JF) Cronje
CRNJAC009

Faculty of Commerce
School of management studies
Masters in Organisational Psychology

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Ethical Approval: COM/00889/2024

Welcome and Introductions

shutterstock.com · 2036834303

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Training Session Overview

		Duration
Step 1	Pre-reading (completed before session)	± 1 Hour
Step 2	Training Overview and HEXACO Test	20 Minutes
Step 3	Asynchronous Video Interviewing (AVI) Overview	5 Minutes
Step 4	HEXACO Overview and Scoring Guidelines	20 Minutes
Step 5	Scoring Demonstration	15 Minutes
Step 6	Break	5 Minutes
Step 7	Practice and Feedback (FORT)	50 Minutes
Step 8	Closing and Next Steps	5 Minutes

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Research Study Overview

This study aims to examine the convergence between human and artificial intelligence (AI) scoring of HEXACO personality traits from asynchronous video interviews (AVIs). Participants in this study will engage in scoring exercises and training sessions designed to enhance their understanding of the HEXACO model, AVIs, and the nuances of personality assessment from video and text-based (transcribed) responses. The findings may contribute to a broader understanding of how AI can complement human judgement in psychological evaluations, particularly in personality assessments during recruitment and selection processes.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Ethical approval (COM/00889/2024)

Ethical Considerations

- **Confidentiality & Avoiding Misuse of Data**
  - We work with real people. Ensure that all personal information and results are kept safe and only used for this study.
- **Respect for Participants; Non-Discrimination; Objectivity and Impartiality; Transparency and Accountability**
  - **Awareness of Biases:**
    - Confirmation Bias: Avoid seeking information that confirms pre-existing beliefs about a participant.
    - Halo Effect: Be cautious of allowing a positive impression in one area to influence overall ratings.
    - Horn Effect: Avoid letting a negative impression in one area affect the overall evaluation.
    - Cultural Bias: Recognise and adjust for cultural differences in communication styles and behaviours.
    - Stereotyping: Avoid applying generalisations based on gender, age, race, or other group characteristics.
    - Recency Effect: Ensure that recent responses or behaviours do not disproportionately influence the overall rating.
    - Anchoring: Be mindful not to let the first few ratings set a "baseline" that skews subsequent ratings.
    - Contrast Effect: Avoid comparing participants against each other instead of evaluating them independently.
    - Leniency/Severity Bias: Be aware of tendencies to rate more leniently or harshly based on personal tendencies or moods.
    - Similar/Different to Me: Be aware of giving more (or less) favourable ratings because someone is similar to (or less like) you.
- **Competence, Consistency, Awareness of the Evaluation Environment**
  - Distraction-Free Environment: Work in a quiet, focused environment to avoid distractions that could influence judgements.
  - Emotional State: Be aware of how your current emotional state may impact their ratings and strive for neutrality.
  - Standardisation: Use the same criteria for each evaluation to maintain consistency across all ratings.
  - Clear Definitions: All evaluators have a shared understanding of the traits being rated, with clear definitions and examples.
  - Quality checks (per consent form) will be performed.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

(Cascio & Aguinis, 2019; Coetzee, 2021; Foxcroft & Roodt, 2013)

5

HEXACO Test

Instructions

- You will be provided with six descriptions. Each description corresponds to one of the six HEXACO personality traits.
- Your objective is to correctly identify which HEXACO trait each description represents. The six HEXACO traits are:
  - Honesty-Humility (H)
  - Emotionality (E)
  - eXtraversion (X)
  - Agreeableness (A)
  - Conscientiousness (C)
  - Openness to Experience (O)
- Carefully read each description provided.
- For each description, select the trait that you believe best corresponds to the description.
- Ensure that you only assign one trait to each description.
- Next to each description on your answer sheet, write the name of the trait you have selected.
- Your answers will be scored based on the accuracy of the trait identification. The test is an indication for our training process and a way to see where everyone is at.
- You have 10 minutes to complete this task. Please manage your time accordingly.
- Once you have completed the test, submit your answer sheet to the researcher.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring

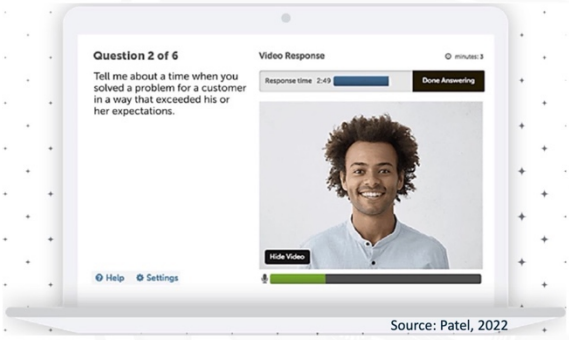
6

Test QR

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Asynchronous Video Interviewing (AVI)

- Asynchronous Video Interviewing (AVI) presents an alternative to conventional real-time interviews, offering a solution that is consistently available and easily accessible (Lukacik et al., 2022).
- Unlike traditional interview processes, AVIs entail accessing an online platform to record video responses to interview questions, without direct interaction with interviewers (Dunlop et al., 2022; Lukacik et al., 2022).
- We have a USA sample from USA, obtained from Prolific and a previous UCT IOPM study.



Source: Patel, 2022

Therefore, AVI is a modern way to do job interviews without needing to be online at the same time as the interviewer. Instead of a live in-person or virtual discussion, candidates record themselves answering questions on a website. These recorded responses are then reviewed later by either a person or a computer.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Asynchronous Video Interviewing (AVI)

Welcoming Screen

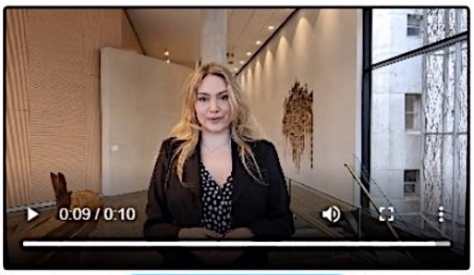
Next

UNIVERSITY
IYUNIVESITHI YASEKAPA

Asynchronous Video Interviewing (AVI)

Question Instruction

After you click **Start Recording**, you will be given the question and 60 seconds to prepare your answer. Once ready, click **Start Now** to record. For assistance, click the red button or contact your recruiter.



Start Recording

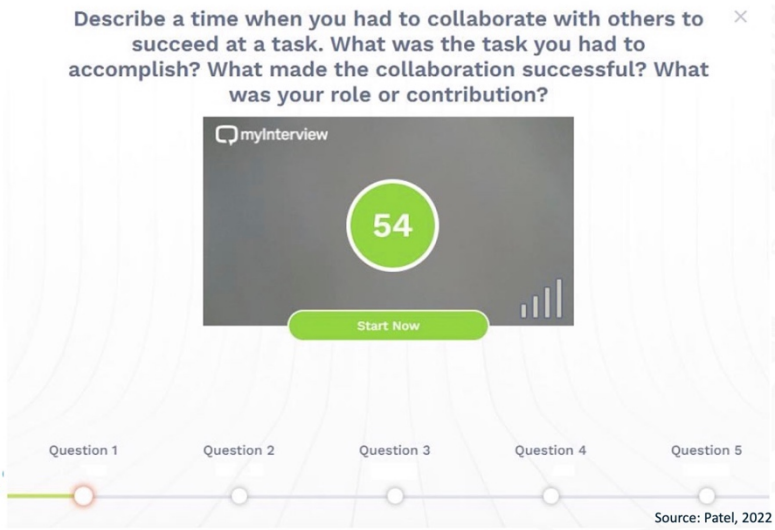
Source: Patel, 2022

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring

Asynchronous Video Interviewing (AVI)

Example of Question 1



Source: Patel, 2022

AVI Questions

Number	Question
1	Describe a time when you had to collaborate with others to succeed at a task. What was the task you had to accomplish? What made the collaboration successful? What was your role or contribution?
2	Describe a situation where you had to evaluate the risks, benefits, and potential outcomes of a decision. For example, buying something important, investing in something, starting a new project, etc. How did you handle it? And what was the outcome?
3	Describe a time when you took the lead on a group project. What was the project, how did you behave as a leader, and what was the outcome?
4	How would you handle a situation where your work colleagues ignore your ideas and input?
5	How do you manage your time and prioritise tasks?

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring

HEXACO Inventory

Trait	High Scores	Low Scores
Honesty-Humility	Avoid manipulating others for personal gain, little temptation to break rules, uninterested in lavish wealth and luxuries, feel no special entitlement to elevated social status.	Flatter others to get what they want, inclined to break rules for personal profit, motivated by material gain, strong sense of self-importance.
Emotionality	Experience fear of physical dangers, anxiety in response to life's stresses, need emotional support from others, empathy and sentimental attachments with others.	Not deterred by the prospect of physical harm, little worry even in stressful situations, little need to share concerns with others, emotionally detached from others.
eXtraversion	Feel positively about themselves, confident when leading or addressing groups, enjoy social gatherings and interactions, experience positive feelings of enthusiasm and energy.	Consider themselves unpopular, feel awkward when the center of social attention, indifferent to social activities, feel less lively and optimistic.
Agreeableness (vs. Anger)	Forgive wrongs suffered, lenient in judging others, willing to compromise and cooperate, easily control temper.	Hold grudges against those who have harmed them, critical of others' shortcomings, stubborn in defending their point of view, feel anger readily in response to mistreatment.
Conscientiousness	Organise time and physical surroundings, work in a disciplined way toward goals, strive for accuracy and perfection, deliberate carefully when making decisions.	Unconcerned with orderly surroundings or schedules, avoid difficult tasks or challenging goals, satisfied with work containing some errors, make decisions on impulse or with little reflection.
Openness to Experience	Absorbed in the beauty of art and nature, inquisitive about various domains of knowledge, use imagination freely in everyday life, take an interest in unusual ideas or people.	Unimpressed by most works of art, little intellectual curiosity, avoid creative pursuits, little attraction toward ideas that may seem radical or unconventional.

(Lee & Ashton, n.d.)

Scoring Rubric / Scale

Score	1	2	3	4	5
HEXACO Trait Presence (Rating)	Very low	Low	Average	High	Very high

You need to decide, based on your knowledge and the training, by watching the AVI and reading the transcript:

1. Which HEXACO traits are present (hopefully all 6).
2. How strongly each trait is expressed by allocating a score per interview response and HEXACO trait.
3. There are 5 interview questions (and responses) per participant.
4. In these responses, they portray their personalities to us.
5. The next few pages will provide you with more indicators per trait

Koutsoumpis et al. (2024)

Trait 1: Honesty-Humility

Scale	1 = Very low	2 = Low	3 = Average	4 = High	5 = Very high
Honesty-Humility	• Very high sense of self-importance • Strongly motivated by material gain • Tempted to "bend" laws for personal profit, • Highly likely to flatter others for personal success	• High sense of self-importance, they • Motivated by material gain, • Tempted to "bend" laws for personal profit, • May flatter others for success.	• Feel an average sense of self-importance, • Not strongly motivated by luxury and status, • Seldomly break rules and manipulate or flatter others for personal gain.	• Avoid manipulating others for personal gain, • Feel little temptation to break rules, • Less interested in wealth and luxuries, • Feel little need for social status or privilege.	• Avoid manipulating others for personal gain, • Feel no temptation to break rules, • Uninterested in wealth and luxuries, • Do not feel any entitlement to elevated social status or privilege.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Note: Middle score. True? Rater understanding?

Koutsoumpis et al. (2024)

Trait 2: Emotionality

Scale	1 = Very low	2 = Low	3 = Average	4 = High	5 = Very high
Emotionality	• Very unemotional, • Detached, • Independent, • Feel no anxiety or fear even under stressful or frightening circumstances.	• Unemotional, • Detached, • Independent • Feel little anxiety or fear even under stressful or frightening circumstances.	• Neither emotional nor unemotional • Feel some - but not too much - anxiety and fear when faced with stressors.	• Fear physical dangers • Anxiety in response to stressors. • Need emotional support, attachments and empathy	• High fear of physical dangers • Anxiety in response to stressors. • Feel a very strong need for emotional support from others, attachments and empathy.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Koutsoumpis et al. (2024)

Trait 3: eXtraversion

Scale	1 = Very low	2 = Low	3 = Average	4 = High	5 = Very high
eXtraversion	• Very reserved and awkward when at the centre of social attention. • Consider themselves very unpopular and much less lively than others. • Highly indifferent to social activities.	• More reserved and awkward than others when at the centre of social attention. • Consider themselves somewhat unpopular and less lively than others. • Mostly indifferent to social activities.	• Neither reserved nor confident when leading a group or at the centre of attention. • Feel neither popular nor unpopular • Average levels of enthusiasm and energy.	• Feel more confident than others when leading or addressing groups of people. • Enjoy social gatherings and interactions. • Feel mostly positive about themselves. • Experience mostly positive feelings of enthusiasm and energy.	• Very confident when leading or addressing groups of people. • Very much enjoy social gatherings and interactions. • Feel very positive about themselves. • Very often experience positive feelings of enthusiasm and energy.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Koutsoumpis et al. (2024)

Trait 4: Agreeableness

Scale	1 = Very low	2 = Low	3 = Average	4 = High	5 = Very high
Agreeableness	• Feel lots of anger in response to mistreatment. • Bear strong grudges against those who have insulted or deceived them. • Very critical of others' shortcomings. • Very stubborn in defending their point of view.	• Feel anger in response to mistreatment. • Tend to bear grudges against those who have insulted or deceived them. • More critical than others of people's shortcomings • Tend to be stubborn in defending their point of view.	• Average levels of patience and anger in response to mistreatment. • Neither very forgiving nor bearing strong grudges against those who have insulted or deceived them. • Neither very lenient nor highly critical of others' shortcomings. • Sometimes stubborn and sometimes compromising and cooperative.	• Tend to compromise and cooperate with others. • Tend to be - more than others - lenient in judging others. • Tend to remain patient. • Tend to control their temper and forgive the wrongs that they have suffered.	• Always compromise and cooperate with others. • Very lenient in judging others. • Always remain patient. • Very easily control their temper. • Always forgive the wrongs that they have suffered.

Koutsoumpis et al. (2024)

Trait 5: Conscientiousness

Scale	1 = Very low	2 = Low	3 = Average	4 = High	5 = Very high
Conscientiousness	• Very unconcerned with orderly surroundings or schedules. • Strongly avoid difficult tasks or challenging goals. • Are - much more than others - satisfied with work that contains some errors. • Often decide on impulse with no reflection.	• Tend to be - more than others - unconcerned with orderly surroundings or schedules. • Avoid difficult tasks or challenging goals. • Tend to be satisfied with work that contains some errors. • Decide on impulse or with little reflection.	• Have an average concern for the orderliness of their surroundings and schedules. • Neither very accurate nor do they make a lot of errors in their work. • Sometimes deliberate carefully and sometimes decide on impulse. • Average discipline when working toward their goals.	• Tend to organise things (both time and physical surroundings) • Work in a disciplined way toward their goals. • Strive for accuracy and perfection in their tasks. • Deliberate - more than others - careful when deciding.	• Always organise things (both time and physical surroundings) • Work in a highly disciplined way toward their goals. • Very strongly strive for accuracy and perfection in their tasks. • Deliberate very carefully when deciding.

Koutsoumpis et al. (2024)

Trait 6: Openness

Scale	1 = Very low	2 = Low	3 = Average	4 = High	5 = Very high
Openness	• Very unimpressed by most works of art. • Have no interest in the natural or social sciences. • Strongly avoid creative pursuits. • Feel no attraction toward ideas that may seem radical or unconventional	• Unimpressed by most works of art. • Feel little interest in the natural or social sciences. • Tend to avoid - more than others - creative pursuits. • Feel little attraction toward radical or unconventional ideas.	• Average interest in art and the natural or social sciences. • Do not take a strong interest, but neither do they avoid creative pursuits. • Feel relatively neutral toward radical or unconventional ideas.	• Tend to become absorbed in the beauty of art and nature. • Tend to feel intellectual curiosity in various domains of knowledge. • They - more than others - use their imagination freely in everyday life. • Take interest in unusual ideas or people.	• Very easily absorbed in art and nature. • Feel strong intellectual curiosity in various domains of knowledge. • Always use their imagination freely in everyday life. • Take a strong interest in unusual ideas or people.

Koutsoumpis et al. (2024)

Example 1: Let's Discuss Together

Koutsoumpis et al. (2024)

Q: Could you describe how you make decisions, do you do this rather spontaneously or very deliberately?

A: So, with this one I think sort of depends on the situation. So I guess there is advantages of both having spontaneous decisions or more deliberate ones. So if you need to adapt to change or if you are in quite a demanding environment (especially at work). So, if you have customers coming up to you or new things happening all the time, you've got to be able to make spontaneous decisions. So I try and make those decisions when necessary, but I do prefer planning things out and being more deliberate. Sort of, looking at the end goal and knowing what is expected. Well, yeah. I also work in a tutoring school and often we have kids for an hour and half so we need to plan out what we are going to do. So, when I am deciding what work we are going to do first I try and talk with the student and be more deliberate in my choices, so take into account what they have already done that day, what they need to learn and what is the highest priority.

	Score				
HEXACO Trait	**1**	**2**	**3**	**4**	**5**
Honesty-Humility	Very low	Low	Average	High	Very high
Emotionality	Very low	Low	Average	High	Very high
eXtraversion	Very low	Low	Average	High	Very high
Agreeableness	Very low	Low	Average	High	Very high
Openness to Experience	Very low	Low	Average	High	Very high

Example 1: Researcher's Thoughts

Honesty-Humility (3):
The response doesn't provide clear indicators of this trait. No evidence of humility or manipulation, so this trait might be rated as Average.

Emotionality (3):
The response reflects a balanced approach to decision-making without evident anxiety or detachment, suggesting Average.

Extraversion (3):
The respondent mentions interacting with customers and students, showing some level of social engagement. However, the focus is more on task management than social energy, so Average.

Agreeableness (4):
The response is cooperative, especially when planning with students, showing a willingness to compromise and consider others' needs. This suggests High.

Conscientiousness (5):
The respondent clearly prefers planning and being deliberate, showing a strong focus on organisation and goal-oriented behaviour. This would rate as Very High.

Openness to Experience (4):
The respondent shows some openness to adapting to change and being flexible, which suggests a moderate level of openness. This might be rated as High.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Example 2: Let's Discuss Together

Koutsoumpis et al. (2024)

Q: How do you think that others perceive you in a social setting?

A: I'm very chatty, if you haven't realized already, so most people do find me as an extrovert but if I do, if I talk, if I see someone and I have something to say, I would say it to them. For example, completing my lab report yesterday in one day I've just been like blowing it out to everyone. Anyone I could see, I would tell everyone oh my gosh I achieved something like impossible, so I would tell them because I just needed to release that energy I guess. People to find me very extrovert but I'm actually like I do my personality qualities are very extrovert but in terms of habits and hobbies they're quite introvert but they have to be with other people, that's the thing. I get really lazy going to parties that's, so I guess that's two types of extroverts or more than that actually. I rather just sit at home and watch a movie with a group, like with a two of my best friends rather than going to a party with massive group of people. Once in a long while sounds good but the whole, especially winter I just need to like be in my sweat pants so yeap. I think other people also perceive me as very polite, very easy going, very easy going to talk to. I can easily chat to anyone, have a comfortable conversation with them. I think, a lot of people actually told me in high school that I made them feel comfortable. Made them feel comfortable at school and to a group of people they wouldn't you know start talking to if it wasn't for me. I think everyone thinks I'm very polite and very nice and I do have, I do, I value respect so that's one, I value respect and honesty so the two things I would do and I think people do perceive me as that as well.

	Score				
HEXACO Trait	**1**	**2**	**3**	**4**	**5**
Honesty-Humility	Very low	Low	Average	High	Very high
Emotionality	Very low	Low	Average	High	Very high
eXtraversion	Very low	Low	Average	High	Very high
Agreeableness	Very low	Low	Average	High	Very high
Openness to Experience	Very low	Low	Average	High	Very high

Example 2: Researcher's Thoughts

Honesty-Humility:
The respondent mentions valuing respect and honesty, which suggests a higher level of this trait. Therefore, this might be rated as **4 (High)**.

Emotionality:
The response does not provide strong indicators of emotionality, either in terms of anxiety or dependence. This trait might be rated as **3 (Average)**.

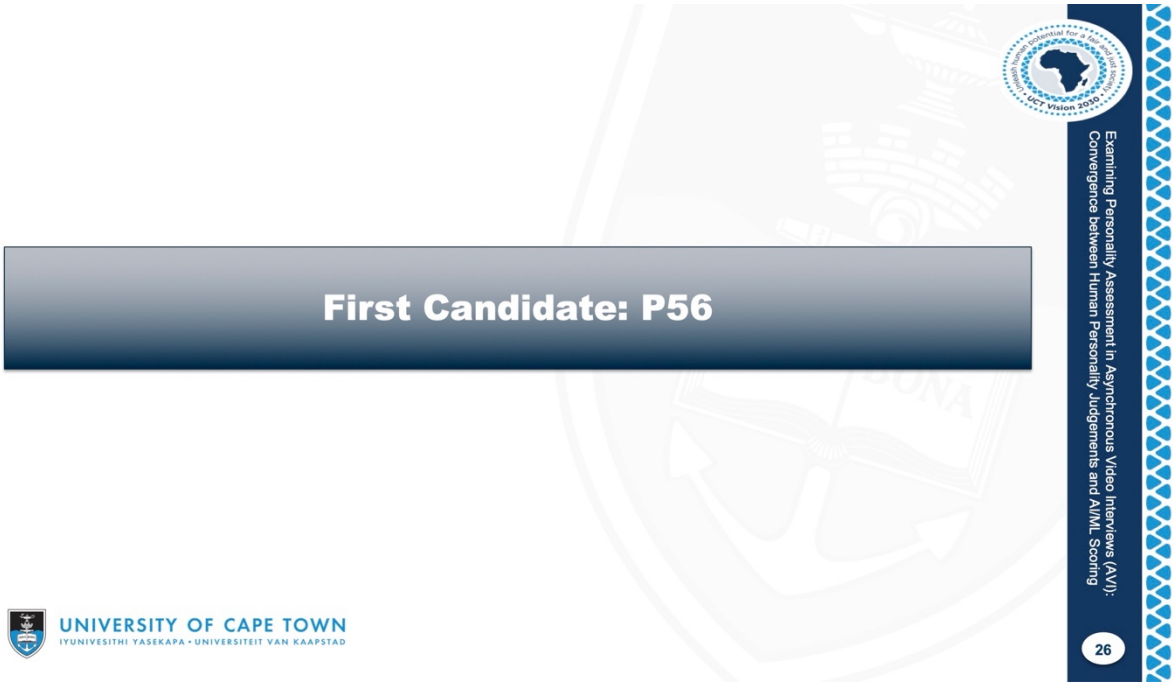
Extraversion:
The respondent identifies as chatty, enjoys socialising, and is seen by others as extroverted, even though they also enjoy some introverted activities. This would likely be rated as **4 (High)**.

Agreeableness (vs. Anger):
The respondent is described as polite, easy-going, and able to make others feel comfortable, which suggests a high level of agreeableness. This might be rated as **5 (Very High)**.

Conscientiousness:
The response doesn't provide strong indicators of this trait. No specific details on planning or organization are mentioned, so this might be rated as **3 (Average)**.

Openness to Experience:
The respondent mentions enjoying a variety of social activities and valuing respect and honesty, which could indicate a moderate level of openness. This trait might be rated as **3 (Average)**.


Unleash human potential for a fair and just society
UCT Vision 2030
Let's take a 5-minute break
UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD


First Candidate: P56
UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring
26

Videos to Be Scored Together
P56

Question 1:

Describe a time when you had to collaborate with others to succeed at a task.

What was the task you had to accomplish?

What made the collaboration successful?

What was your role or contribution?

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together
P56

Question 1:

Describe a time when you had to collaborate with others to succeed at a task. What was the task you had to accomplish? What made the collaboration successful? What was your role or contribution?

The most recent time that I had to collaborate with others to succeed at a task was doing a residential remodeling project for some clients. Their flooring that they wanted that we had installed a couple of years ago had been discontinued. And no longer carried by the local flooring distributor. I was able to track down the the main distributor online. I called and spoke to them in California. Found out that they were ordering a shipment due to customer demand from China. Actually, I take that back. From Vietnam for some back stocked item. Was able to order some online. Figure out how to get… How to meet the freight carrier. My husband and I met the freight carrier, got the client what they wanted. It was a little extra trouble but made them happy and didn't tell them about it till after the fact. They were so happy about it. That they gave us a nice gift certificate to a local restaurant.

Let's Discuss !

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring

Videos to Be Scored Together
P56

Question 1:
Describe a time when you had to collaborate with others to succeed at a task. What was the task you had to accomplish? What made the collaboration successful? What was your role or contribution?

HEXACO Trait	Rating	Reasoning
Honesty-Humility	4 (High)	Showed some humility by going the extra mile for the client
Emotionality	3 (Average)	The response does not show strong emotional reactions
Extraversion	3 (Average)	The response focuses more on task completion than social interaction.
Agreeableness	4 (High)	The respondent was highly agreeable, aiming to make the client happy even with extra effort.
Conscientiousness	5 (Very High)	The respondent demonstrated strong conscientiousness in solving the client's issue effectively.
Openness to Experience	3 (Average)	The respondent showed some creativity in problem-solving, but the task was relatively routine.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together
P56

Question 3:
Describe a time when you took the lead on a group project. What was the project, how did you behave as a leader, and what was the outcome?

Note: The candidate unfortunately did not answer question 2. The AVIs that you will evaluate will be complete.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together
P56

Question 3:
Describe a time when you took the lead on a group project. What was the project, how did you behave as a leader, and what was the outcome?

In my working life, I really have not been a position to take the lead on projects. I've always had more of a supportive role. I am not shy about sharing my opinions and trying to add value to the group with suggestions and collaborations. But really, in terms of leadership, I cannot think of a time when I was actually the lead on a group project.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Let's Discuss !

Videos to Be Scored Together
P56

Question 3:
Describe a time when you took the lead on a group project. What was the project, how did you behave as a leader, and what was the outcome?

HEXACO Trait	Rating	Reasoning
Honesty-Humility	5 (Very High)	The respondent is honest about their experience and limitations in leadership roles.
Emotionality	3 (Average)	The response is neutral in emotional expression, neither overly emotional nor detached.
Extraversion	2 (Low)	While not clear, the respondent acknowledges a more supportive role, which may indicate a slight avoidance of the centre stage.
Agreeableness	4 (High)	The respondent is collaborative and values contributing to the group's success.
Conscientiousness	3 (Average)	The respondent demonstrates a supportive role but does not indicate strong conscientiousness.
Openness to Experience	3 (Average)	The respondent is open to sharing opinions and suggestions, showing a moderate level of openness to new ideas.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together
P56

Question 4:
How would you handle a situation where your work colleagues ignore your ideas and input?

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together
P56

Question 4:
How would you handle a situation where your work colleagues ignore your ideas and input?

I have been fortunate throughout my working career that my work colleagues, male and female alike, have been actually good listeners and open to input, so I'm in the enviable position of not having had coworkers who ignored my ideas and input. I currently work with my husband. UM. Whatever issues we may have while we work, he actually is a good listener. The few times that I have felt ignored, I will try and bring something up later, or perhaps put it in writing in an e-mail. Saying Visa V, the previous project we were working on or the meeting that we had. These are some ideas I have and just put those virtual pen to paper. That way maybe somebody was distracted at the time or preoccupied with something else. They can take their time and and read through it and respond to it and maybe listen better when it's actually on paper. Or computer.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Let's Discuss !

Videos to Be Scored Together
P56

Question 4:
How would you handle a situation where your work colleagues ignore your ideas and input?

HEXACO Trait	Rating	Reasoning
Honesty-Humility	4 (High)	The respondent appears honest about their positive experiences with colleagues and their approach to communication.
Emotionality	3 (Average)	The response reflects a balanced approach to dealing with feeling ignored, showing neither high anxiety nor detachment.
Extraversion	3 (Average)	The respondent values communication and ensures their ideas are heard, indicating a moderate level of sociability.
Agreeableness	5 (Very High)	The respondent shows a high level of patience and understanding, addressing issues calmly and constructively.
Conscientiousness	4 (High)	The respondent takes proactive steps to ensure their ideas are communicated effectively, showing responsibility and organisation to make sure their heard.
Openness to Experience	4 (High)	The respondent seems open to different methods of communication and problem-solving, indicating adaptability and creativity.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together
P56

Question 5:
How do you manage your time and prioritise tasks?

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together
P56

Question 5:
How do you manage your time and prioritise tasks?

In the past I have tended to do the easy tasks first thing in the morning. Just to get them marked off the To Do List and make myself feel like I was making progress toward my task list. However, I've recently taken some online courses on time management and productivity and learn that most people are more productive in the morning. I certainly am. And. So, to instead attack the more difficult tasks, the tasks that may take input from other people so that I would need to make phone calls or contact people via text or e-mail for their input, get that ball rolling first thing in the morning so that I can successfully finish that test throughout the day and then perhaps later in the afternoon. During. Maybe less productive, maybe a little bit sleepier time in the late afternoon, I can do less brain texting tasks, such as responding to emails, organising lists and so forth.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Let's Discuss !

Videos to Be Scored Together
P56

Question 5:
How do you manage your time and prioritise tasks?

HEXACO Trait	Rating	Reasoning
Honesty-Humility	4 (High)	The respondent honestly reflects on their past behaviour and shows humility by acknowledging the need to improve their time management strategies.
Emotionality	3 (Average)	The response shows a balanced approach to handling tasks and productivity, without displaying extreme anxiety or detachment.
Extraversion	3 (Average)	The respondent acknowledges the need to communicate with others for input, but there is no strong indication of sociability or assertiveness.
Agreeableness	4 (High)	While not very clear, the respondent shows a willingness to adapt their behaviour based on new knowledge.
Conscientiousness	5 (Very High)	The respondent demonstrates strong conscientiousness by reorganising their task priorities to maximise productivity and efficiency.
Openness to Experience	4 (High)	The respondent shows openness to new ideas and learning by taking courses and applying learned strategies to improve their work habits.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together

P56

MS Excel

- Use this sheet while scoring the videos per question.
- Our example did not answer question two, but your participants will be complete sets
- The coloured columns are the averages.

Participant ID	H	Q1	Q2	Q3	Q4	Q5	E	Q1	Q2	Q3	Q4	Q5	X	Q1	Q2	Q3	Q4	Q5	A	Q1	Q2	Q3	Q4	Q5	C	Q1	Q2	Q3	Q4	Q5	O	Q1	Q2	Q3	Q4	Q5
P56	**4**	4		5	4	4	**3**	3		3	3	3	**3**	3		2	5	3	**4**	4		4	4	4	**4**	5		3	4	5	**4**	3		3	4	4

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Second Candidate: P7

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together
P7

Question 1:

Describe a time when you had to collaborate with others to succeed at a task.

What was the task you had to accomplish?

What made the collaboration successful?

What was your role or contribution?

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together
P7

Question 1:

Describe a time when you had to collaborate with others to succeed at a task. What was the task you had to accomplish? What made the collaboration successful? What was your role or contribution?

So I the time when I had to collaborate with the others was when I was, I had to prepare my annual report for the project that I've been involved in. It was a biodiversity protection. Protection of the national Parks and biodiversity in my country. So I was the I was not the literal project, but my department supervisor was the leader of the project, but she gave me that task, so I was supposed to do it and we had a lot of associates who are working with us. So because I need to take. Opinion of others and to take to to write in manual for what they know, because I was just a leader. Let's say I didn't have all information I had to call them. I had to make interviews with the five of them who were involved in the. Just. And so I would have a meeting separately with them and then after that they would give me the written report. So after that report, I had to put it on one big, let's say, annual report and to give it to my manager. So I, as I said, I had to collaborate with five different people. One of them was professor of biology. One of them was one of those people were NGO NGO organization leaders. And I would call them As for their opinion. I had a meeting with them and I would say I had very successful contribute. And I was the leader in the project, but also I was one of the active member, I would say on the same level as them. So I said I would say that was a very, very successful collaboration. And so my task was to give the report annual report of the steps that were implemented. And so my manager is, as he gave me the the timeline. What I was supposed to do, she gave me some instruction and I I made my report. My task was done in the due in the given time given time frame.

Let's Discuss !

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring

Videos to Be Scored Together
P7

Question 2:
Describe a situation where you had to evaluate the risks, benefits, and potential outcomes of a decision. For example, buying something important, investing in something, starting a new project, etc. How did you handle it? And what was the outcome?

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together
P7

Question 2:
Describe a situation where you had to evaluate the risks, benefits, and potential outcomes of a decision. For example, buying something important, investing in something, starting a new project, etc. How did you handle it? And what was the outcome?

HEXACO Trait	Rating	Reasoning
Honesty-Humility	4 (High)	The respondent shows honesty and humility by openly acknowledging the risks and challenges involved in the decision-making process.
Emotionality	4 (High)	The respondent expresses concern for their sister's well-being and shows a strong emotional investment in ensuring a secure outcome.
Extraversion	3 (Average)	The respondent takes initiative in contacting others for advice but does not display strong sociability or assertiveness beyond that.
Agreeableness	3 (Average)	The respondent is cooperative but cautious, showing a balanced approach by not rushing into the purchase without thorough investigation.
Conscientiousness	5 (Very High)	The respondent demonstrates strong conscientiousness by diligently investigating the legality of the property and making informed decisions.
Openness to Experience	4 (High)	The respondent is open to taking a significant risk by considering an overseas property purchase, showing adaptability and willingness to face uncertainty.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together
P7

Question 2:

Describe a situation where you had to evaluate the risks, benefits, and potential outcomes of a decision. For example, buying something important, investing in something, starting a new project, etc. How did you handle it? And what was the outcome?

I think one of the biggest. The risk lately that I wanted to take for me was to buy a small property from my from my sister. She lives in Europe in small country in Europe, Montenegro and she she is she. She rents apartment which is very expensive and she doesn't have enough money. I want to help her to buy something small because I'm not able to buy anything. Big for her. And in our neighbor neighborhood, there was one small apartment, just one room house that was on sale. And I was getting ready to find money to take money from my bank account here in USA and to buy a house. But I don't know. I'll try to investigate and to see because, you know, there is not everything very clean and clear in my country with The Who has a property whose land is it? The Who build the house. So I said. Can I please have a paper for the for the House so when we sign a contract I want to sign a contract actually in front of the like a lawyer. I want to be sure that I gave you the money to have A to have approved that I gave you the money and to give. To you can give me the paper. From the house. So we bought the house. It's our the it's the house is out. So that man who was selling the house, he was trying to play some games or I'll give you later. You give me money now. So and we were in the rush because my sister was not able to pay any more the rent and she didn't have a place to live. So I had to. To to think quickly, like what to do. Then I I called some people that were other people in the neighborhood who who knew the person who was selling the house. So I asked them about this potential risk. Or benefit of buying this property so they gave me the very how can I say very, very strong important information about the house. They told me that the house actually is not his and he's trying to sell the property that doesn't belong to him. So I told him again I went to this man who was saying the house and said this house is very cheap. Even though it's small, but why do you sound so cheap? He was trying to hide information from me. He didn't want to say. Anything. And then I called the National Agency for the property. That's how they call it. Like something where you can see the building or the house on which name is it? So it's it's it's not easy to get the information. But I tried it and I did and I found out that. That house didn't belong to him. Actually. It's a it's illegal building. It's a city property. So that's how I I was very happy. I didn't buy the house from him because I would lose the money and it would be it would be a very, maybe even involve the police or something so they can throw me out because that house doesn't belong to him. It doesn't belong to me, doesn't belong to anyone actually. So if the city wants to throw you out and take it away from you, you don't have any rights. So I think I my outcome of my decision not to buy it and to investigate everything, not to go just like ohh, go buy it. It's a good deal. It was good and beneficial for me and I think I. Handled it very. Well, for the first time I was entering something risky buying buying bigger property or something.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Let's Discuss !

Videos to Be Scored Together
P7

Question 3:

Describe a time when you took the lead on a group project. What was the project, how did you behave as a leader, and what was the outcome?

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and AI/ML Scoring

Videos to Be Scored Together
P7

Question 3:
Describe a time when you took the lead on a group project. What was the project, how did you behave as a leader, and what was the outcome?

I had a couple of this kind of situation when I was a leader of the group project, maybe one that I remember very. It was like. Maybe the first one that I've been taking care of is when I was a youth leader and I had to take care of the. Like a youth group. With the kids who are elementary school and we had to do some group project of preparing the performance. For the public. And uh and I had a partner who was a leader with me in that group. But she gets sick. So I had to do it all by myself. It was very, very scary for me the first time I was doing that, I. I was lacking information actually or or. Experience to do it, but. Somehow I said, OK, it's let's try it. So my my biggest like achievement was that I involve the kids like say 100%. I always talk with the kids. I had a meeting in Group with them. We we interact in different kind of levels. And then we had to, like, do the performance in front of their family, their friends, and they were very scared. So I called some actors real actors from the our theater. And the actor, who who are very famous in our country and I called them so they were giving the free lesson for the kids. And the the kids were very happy. They were very relaxed. They prepared their performance very good and I think they all enjoy, they make their own script, their own text. What they're gonna do, how they gonna look, they may. They made a costume for themselves and the. So I was very happy with the outcome of this. I think the project was very successful. They even they even put it on the TV. And they make some, uh news like report. And it was a national TV that what did we do like? I was very happy. I was very proud being the leader of the group that the kids were enjoying. So enjoy. Enjoy so much.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Let's Discuss !

Videos to Be Scored Together
P7

Question 4:
How would you handle a situation where your work colleagues ignore your ideas and input?

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together
P7

Question 4:

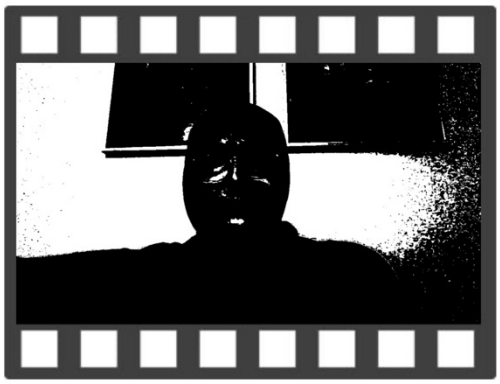
How would you handle a situation where your work colleagues ignore your ideas and input?

I well, I do have some experience with handling situation where my colleague ignore my idea because I am a a lead supervisor in in the store where I work so many times when I have some new associate they they just ignore you. Let's say they don't want to do what you're what you want from them. Or the way you tell them to prepare the store how it should be looking so like but I I try to handle it without stressing too much without you know, like not being upset. So if I I I give them advice and I tell them always like I'm not trying to be bossy. Like can you please like do like this or do like that. Or I always try to ask them. Like, what do you think about this situation? Why do you think you need to do it differently or so? Let's say I would. I would try to stay calm, like not to be like too much annoyed. If they really don't show like any kind of interest to work with me or they think that I would just, I'm very direct person, I like straight communication. Like if I think just depends what level of the project is or something that we're involved with and that my ideas need to be. Also put in that prop. That my idea needs to be valued as a part of the team work, so I would, I would ask them to talk with me to be straight. Why? What's the problem? Is some problem with me personally or something that I said or they don't just don't like my ideas. If that situation is, let's say not able to solve, I would. I will try to talk with my supervisor with my manager and to tell them that I tried to collaborate, that I tried to give my ideas and input, but they were ignored. So if my supervisor thinks that the manager, whoever is there, things that it's, it's it shouldn't be working like that. They can call the other person. They could talk with them. But I always try to do this. Take communication and to tell them what I think. How I think, why my ideas are ignored. So is there any problem? That we can solve. So I think that's it. Like I'm just trying to handle it very well and not to be not to feel like down or sad or it's OK people are different so. That's.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Let's Discuss !

Videos to Be Scored Together
P7



Question 5:

How do you manage your time and prioritise tasks?

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Videos to Be Scored Together

P7

Question 5: How do you manage your time and prioritise tasks?

Well, how do I manage my time? Well, that's depends like I. So in my life, my daily life, what I have learned is to always have time for myself, for my kids and my family, and for my job. So that's how I divide my time. Like in the morning. Like. I want to have time for myself wife I think make is full in the school and then I have to relax like 30 minutes 40 minutes to drink my coffee to take and so whatever and when I'm drinking coffee and relax and I always like to learn something new. So if whatever I let's say I have to do some tests for my job. I can always look for some information while I'm, let's say relaxing. I know there's something a little bit silly to say, like you're relaxing and working, but I can I can approach my test on the, let's say relax way, but I want to have time for myself. I want to have time for my kids and my family because they're they're education is very important. For me, I have. To to give my my own time to them and also when I go to my job, I always like to be on time there to finish everything step by step what needs to be done. Sometimes you really need to balance like nothing can go by the plane. Many times actually things cannot go by the way you want it to be or how you plan it. But at the end of the time, what's important is to achieve the the. Most important goals. Like, say now it's approaching the holidays and Christmas time, so I have prioritized list what to do in the store, how to finish everything, how to prepare the store, how to my? How my department should be looking? I already already have ideas and then I prioritize. Like test, what is the most important step? So is it to make space in my store, let's say for the new stuff that's gonna come like a lot of gift present for Christmas or thing. So let's just spend the night thought to consult people. What we're going to do, how we're going to manage. We can then just I want to be sure that I my task. Is done. Also my free time, I just have my free time like I have a let's say a little deal with my husband that I need to take time for myself. We need to go sometimes out, take coffee or sit in the park just to relax. And then I have a time to talk with my friends and family. Also. Mostly it's gonna be a viral because I live in USA. They live in other countries, most of them and now with the COVID it's very hard to see anyone these. Just it's not like. Specific specific timeline that I have. Sometimes it's a more general idea what I'm going to do, but that's it. I try to do my. Time like it's. My family, me, my job. So desert tree. Fill the time trying to handle the best I can.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Let's Discuss !

List of Steps

Step	Task	Details
1	**Access Your Participants**	Receive access to a OneDrive folder with your specific participants. This will be sent to your university email by Friday, 30 August. These are your specific participants to rate.
2	**Watch the Video**	Begin by watching your first participant's AVI video. Read through the provided transcript and modify it if necessary.
3	**Refer to the Rating Scale**	Use the rating scale provided in the Excel sheet. Decide on a rating for each HEXACO trait per question.
4	**Input Ratings in Excel**	Enter your ratings into the Excel sheet. The sheet will automatically calculate the averages for you.
5	**Complete All Ratings**	Ensure the accuracy and completion of all your ratings in the excel sheet. Feel free to add any relevant comments in the last column that comes to mind.
6	**Consult Training Materials**	Regularly refer to your pre-training and training materials. While researchers should not intervene with specific ratings, I am available for general support if needed, though not specific to your participant.
7	**Trust Your Training**	Avoid spending too much time on each participant. You are trained and are working toward advanced psychology degrees. Trust your judgement, avoid overthinking, and adhere to the ethical guidelines discussed.
8	**Submit Your Work & Timesheet**	Complete your evaluations within two weeks, by 15 September. Upload your Excel sheets to your specific OneDrive university folder, along with any comments you wish to share. You may submit your **timesheet** (template to be provided) in the same folder.

References

Cascio, W. F., & Aguinis, H. (2019). Applied Psychology in Talent Management (8th Edition ed.).

Coetzee, M. (2021). Personnel Psychology: An Applied Perspective (3rd Edition ed.). Cape Town: Oxford University Press Southern Africa.

Foxcroft, C., & Roodt, G. (2013). Introduction to Psychological Assessment in the South African Context (4th Edition ed.). Oxford University Press.

Koutsoumpis, A., Ghassemi, S., Oostrom, J. K., Holtrop, D., Van Breda, W., Zhang, T., & De Vries, R. E. (2024). Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning. Computers in Human Behavior. https://doi.org/10.1016/j.chb.2023.108128

Lee, K., & Ashton, M. C. (n.d.). *The HEXACO personality inventory - Revised: A measure of the six major dimensions of personality.* HEXACO. https://hexaco.org/scaledescriptions

Lukacik, E. R., Bourdage, J. S., & Roulin, N. (2022). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review, 32.* https://doi.org/10.1016/j.hrmr.2020.100789

Patel, A. (2022). *How audio-visual stimuli in automated asynchronous video interviews affect applicant reactions: Social presence, fairness, and organisational attractiveness* (Master's dissertation, University of Cape Town)

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Annexure G

Ethical Clearance

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA · UNIVERSITEIT VAN KAAPSTAD

2024/06/06

COM/00889/2024

RE: Research Ethics Committee Project Approval Letter

Dear Jacobus Cronje,

Your application for ethics review of your project titled

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and Artificial Intelligence (AI) / Machine Learning (ML) Scoring

has been reviewed and evaluated by the

Commerce Research Ethics Committee.

You may proceed with your research project titled:

Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgements and Artificial Intelligence (AI) / Machine Learning (ML) Scoring

Expiration date of approval: 2025/02/28

Please note that should:

(i) any serious or adverse effects to participants occur and/or,

(ii) aspect(s) of your current project change and/or

(iii) any unforeseen events that might affect continued ethical acceptability of the project occur then you should immediately report this to the approving REC. You may be required to submit an amendment to this application, in order to determine whether the changed aspects increase the ethical risks of your project.

Based on the information supplied your application has been successful and is approved.

Please note the following additional conditions associated with this approval:

(i) * Ethics approval granted through 28 february 2025
* Permission must be sought from DSA to use students as participants: further details available at https://uct.ac.za/research-support-hub/integrity/accessing-uct-staff-or-students-research-population

Regards,

Commerce Research Ethics Committee.

Annexure H

DSA100 Approval

Department of Student Affairs
Liberating the Soul for Well-being & Flourishing

RESEARCH ACCESS TO STUDENTS ACCESS APPLICATION DSA100a

NOTES

1. This form must be **FULLY** completed by all applicants who want to access UCT students for the purpose of research or surveys.
2. Return the fully completed **(a) DSA100** application forms **by email**, in **the same word format**, together with **your: (b) full research proposal inclusive of your research methodology process, i.e. questionnaire/interview document, tests, etc, (c) copy of your ethics approval letter / proof, (d) informed consent letter, (e) information and invitation, (f) home institution ethics approval if non UCT applicant,** to: Nadierah.Pienaar@uct.ac.za. You application will be attended to by the Executive Director, Department of Student Affairs (DSA), UCT.
3. The turnaround time for a reply is **approximately 2 to 4 weeks**.
4. NB: It is the responsibility of the researcher/s to apply for and to obtain **ethics approval and to comply with amendments that may be requested;** as well as **to obtain** approval to access UCT staff and/or UCT students, from the following, at UCT, respectively:
   (a) **Ethics**: Chairperson, Faculty Research Ethics Committee' (FREC) for ethics approval, (b) **Staff access:** Executive Director: HR for approval to access UCT staff, and (c) Student **access:** Executive Director: Student Affairs for approval to access UCT students.
5. **Note:** UCT Senate Research Protocols requires compliance to the above, **even if prior approval has been obtained from any other institution/agency. UCT's research protocol requirements applies to *all* persons, institutions and agencies from UCT and external to UCT who want to conduct research on human subjects for academic, marketing or service related reasons at UCT.**
6. **Should approval be granted to access UCT students for this research study, such approval is effective for a period of one year from the date of approval (as stated in Section D of this form), and the approval expires automatically on the last day.**
7. **The approving authority reserves the right to revoke an approval based on reasonable grounds and/or new information.**

SECTION A: RESEARCH APPLICANT/S DETAILS

Position	Staff / Student No	Title and Full Name	Contact Details (Email & Cell & Land line)
A.1 Student Number	CRNJAC009	Mr Jacobus Fouché Cronje	crnjac009@myuct.ac.za / jaco.cronje13@gmail.com / +27 74 326 6945
A.2 Academic / PASS Staff No.			
A.3 Visitor/ Researcher ID No.			
A.4 University at which a student or employee	University of Cape Town (UCT)	Address if *not* UCT:	
A.5 Faculty & Department/School	Faculty of Commerce; School of Management Studies; Section of Organisational Psychology		
A.6 APPLICANTS DETAILS If different from above	Title and Name	Tel.	Email

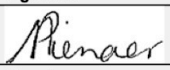
SECTION B: RESEARCHER/S SUPERVISOR/S DETAILS

Position	Title and Name	Tel.	Email
B.1 Supervisor	Prof Francois de Kock	+27 21 650 2181	francois.dekock@uct.ac.za
B.2 Co-Supervisor/s			

SECTION C: APPLICANT'S RESEARCH STUDY FIELD AND APPROVAL STATUS

C.1 Degree – if applicable	Master of Industrial and Organisational Psychology [CM037BUS028]
C.2 Research Project Title	Examining Personality Assessment in Asynchronous Video Interviews (AVI): Convergence between Human Personality Judgments and AI/ML Scoring
C.3 Research Proposal	Attached: Yes ☐ No ☐
C.4 Target population	Honours degree students in Organisational Psychology at the University of Cape Town
C.5 Lead Researcher details	If different from applicant:
C6. Will use research assistant/s	Yes ☐ No ☒ If yes, provide a list of names, staff/student no., e-mail and contact details:
C.7 Research Methodology and Informed consent	**Research methodology:** This study will use a quantitative approach and a within-subjects post-test only experimental design. AVIs will be evaluated by human raters (honours degree students) and an AI text-to-personality algorithm. Comparisons will be made between human and AI ratings. **Is there Informed consent?** Yes, please refer to Annexure E of the research proposal.
C.8 Ethics clearance status from UCT's Faculty Ethics in Research Committee /Chair (EiRC)	Approved by the UCT EiRC: Yes ☒ With amendments: Yes ☐ No ☒ (a) Attach copy of your UCT ethics approval. Attached: Yes ☒ No ☐ (b) State date / Ref. No / Faculty of your UCT ethics approval: 6/06/2024 Ref. / Faculty: COM/00889/2024

SECTION D: APPLICANT/S APPROVAL STATUS FOR ACCESS TO STUDENTS FOR RESEARCH PURPOSE
(To be completed by the ED, DSA or NOMINEE)

D.1 APPROVAL STATUS	**Approved / With Terms / Not**	*** Conditional approval with terms**		**Applicant/s Ref. No.:**
	(i) Approved [redacted] **~~(ii) With terms~~** **~~(iii) Not approved~~**	**a) Access to students for this research study must only be undertaken after written ethics approval has been obtained.** ***b)*** **In event any ethics conditions are attached, these must be complied with before access to students.**		**CRNJAC009 / Mr Jacobus Fouché Cronje**
D.2 PREPARED BY:	**Designation**	**Name**	**Signature**	**Date of Approval**
	Personal Assistant	***Nadierah Pienaar***		**1/07/2024**
D.3 APPROVED BY:	**Designation**	**Name**	**Signature**	**Date of Approval**
	Executive Director / Nominee Department of Student Affairs	***Mr Loki Manise***		***03/07/2024***

Annexure I

AsPredicted Registration

Personality Assessment (HEXACO) in AVIs: Convergence between Human and AI Scores (#193715)

Author(s)
Jaco Cronje (University of Cape Town (UCT)) - CRNJAC009@myuct.ac.za

Pre-registered on: 10/12/2024 04:48 AM (PT)

1) Have any data been collected for this study already?
It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What's the main question being asked or hypothesis being tested in this study?
Advancements in artificial intelligence (AI) have enabled the scoring of asynchronous video interviews (AVI) and personality assessments using algorithms to. Despite emerging research, a gap remains in understanding the convergence between AI and human scoring of personality in AVIs, which this study specifically examines using the HEXACO model. Research Questions "Q" & Hypotheses "H":
Q1: What is the relationship between human evaluator ratings of personality and Artificial Intelligence (AI)-generated personality scores in asynchronous video interviews? Q2: How does the inter-rater reliability of AI-generated personality scores compare to that of human evaluator ratings, and what implications does this have for the reliability and consistency of personality assessment in asynchronous video interviews? Q3: Will human rater judgements of HEXACO (Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience) personality traits correlate positively with the AI-scored personality trait scores? Q4: Among the HEXACO personality traits (Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience), which trait exhibits the lowest correlation between human and AI scorers in personality assessments? Null Hypothesis (H0): There is no significant relationship between human evaluator ratings and AI algorithm ratings in evaluating each HEXACO (Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience) personality trait from asynchronous video interviews. Alternative Hypothesis (H1a): There is a positive relationship between human evaluator ratings and AI algorithm ratings in evaluating each HEXACO (Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience) personality trait from asynchronous video interviews. Alternative Hypothesis (H1b): Convergence between AI and observer ratings of personality will be strongest for extraversion, compared to other personality traits.

3) Describe the key dependent variable(s) specifying how they will be measured.
Evaluations provided by human scorers on each HEXACO trait, following their training in personality assessment, will serve as the independent variables. The scores for each HEXACO trait generated by the AI HEXACO text-to-personality (HTTP) algorithm will serve as the dependent (criterion) variables.

4) How many and which conditions will participants be assigned to?
This study will employ a quantitative approach with a within-subjects, post-test-only experimental research design. All participants' AVI transcripts (±160) will be evaluated for personality by both human scorers and the AI text-to-personality algorithm.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Since data analysis has not yet been conducted and the study focuses on the convergence between AI and human scores, the general approach will be to compare candidate personality traits as assessed by both the AI algorithm and human evaluators. Correlation and regression analyses will be used to compare the scores generated by the AI algorithm with those from human evaluators. Additionally, guidance from prominent researchers and relevant studies in the field will be followed to ensure appropriate data analysis, such as Holtrop et al. (2022) and Koutsoumpis et al. (2024).

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
Only a U.S. sample will be used. Approximately 198 AVIs are checked for factors such as whether all five questions have been answered and whether the audio is clear. Participants will be excluded if the audio is unclear, if one or more questions were not recorded, or if responses are not at least 10 seconds (approximately) in duration to allow for a meaningful response. Currently, only 162 videos are considered suitable for proceeding to the scoring phase (both human and AI). During the data analysis phase, data cleaning, screening, and checking will be conducted as per the guidance from prominent sources like Tabachnick and Fidell (2019): Tabachnick, B. G., & Fidell, L. S. (2019). Using multivariate statistics (7th ed.). Pearson.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
Although the U.S. sample included 198 AVI participants, only 162 were considered useful. These participants provided all five videos, with clear audio, and each video was at least 10 seconds (approximately) in duration. A total of 34 participants were excluded, and 2 were used for training the human evaluators. Thus, 162 AVI participants (each with five videos) proceeded to the scoring phase.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
No form of data analysis has been conducted to date in this study (no analysis of convergence between human and AI ratings: regression/correlation). While some data have been collected, and the human evaluators have been trained and provided their ratings, these ratings have yet to be checked and analysed. The AI algorithm is ready for use; however, it must undergo additional checks before being utilised for data collection and final scoring.

Annexure J

SPPS Additional Outputs Supporting Chapter 4

Table J1

Full Variable Descriptive Statistics

Variable	N	Mean	Std. Deviation	Missing Count	Percent Missing	Low Extremes	High Extremes
H_Human	161	3.74	.58	0	.00	0	0
E_Human	161	2.95	.52	0	.00	0	0
X_Human	161	3.45	.46	0	.00	4	3
A_Human	161	3.65	.49	0	.00	0	0
C_Human	161	3.94	.55	0	.00	1	0
O_Human	161	3.42	.50	0	.00	11	14
H_AI	161	3.08	.10	0	.00	5	2
E_ AI	161	3.00	.08	0	.00	2	2
X_ AI	161	3.07	.07	0	.00	2	1
A_ AI	161	3.01	.09	0	.00	2	2
C_ AI	161	3.15	.08	0	.00	3	1
O_ AI	161	2.96	.04	0	.00	0	2

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Table J2

Extreme Values Data

Variable	Extreme	Rank	Case Number	Value
H_Human	Highest	1	100	5.00
		2	17	4.80
		3	41	4.80
		4	82	4.80
		5	26	4.60a
	Lowest	1	77	2.20
		2	60	2.40
		3	115	2.60
		4	75	2.60
		5	64	2.60b

E_Human	Highest	1	156	4.40
		2	16	4.00
		3	19	4.00
		4	155	4.00
		5	81	3.80c
	Lowest	1	83	1.60
		2	101	1.80
		3	32	1.80
		4	37	2.00
		5	33	2.00
X_Human	Highest	1	13	4.80
		2	100	4.80
		3	17	4.60
		4	32	4.40
		5	33	4.40d
	Lowest	1	93	2.00
		2	160	2.20
		3	154	2.40
		4	136	2.40
		5	112	2.60b
A_Human	Highest	1	150	4.80
		2	105	4.60
		3	151	4.60
		4	153	4.60
		5	26	4.40d
	Lowest	1	15	2.20
		2	87	2.60
		3	35	2.60
		4	160	2.80
		5	143	2.80e
C_Human	Highest	1	13	5.00
		2	100	5.00
		3	106	5.00

		4	145	5.00
		5	147	5.00
	Lowest	1	69	2.00
		2	136	2.40
		3	95	2.80
		4	127	3.00
		5	85	3.00f
O_Human	Highest	1	12	4.60
		2	33	4.60
		3	100	4.60
		4	16	4.40
		5	104	4.40d
	Lowest	1	84	1.60
		2	85	1.80
		3	94	2.20
		4	86	2.20
		5	81	2.20g
H_AI	Highest	1	110	3.45
		2	30	3.32
		3	138	3.29
		4	149	3.27
		5	21	3.25
	Lowest	1	86	2.69
		2	20	2.77
		3	115	2.79
		4	161	2.83
		5	65	2.86
E_ AI	Highest	1	11	3.22
		2	108	3.19
		3	122	3.19
		4	23	3.18
		5	69	3.15
	Lowest	1	103	2.80

		2	96	2.81
		3	151	2.82
		4	56	2.83
		5	70	2.83
X_ AI	Highest	1	144	3.30
		2	135	3.23
		3	132	3.22
		4	128	3.22
		5	35	3.20
	Lowest	1	34	2.86
		2	160	2.89
		3	62	2.91
		4	80	2.92
		5	48	2.92
A_ AI	Highest	1	160	3.35
		2	110	3.31
		3	80	3.21
		4	132	3.20
		5	29	3.20
	Lowest	1	154	2.71
		2	86	2.77
		3	155	2.82
		4	40	2.85
		5	61	2.86
C_ AI	Highest	1	137	3.35
		2	64	3.35
		3	158	3.34
		4	91	3.33
		5	23	3.30
	Lowest	1	107	2.83
		2	84	2.85
		3	83	2.87
		4	76	2.88

		5	42	2.89
O_ AI	Highest	1	107	3.35
		2	109	3.32
		3	61	3.27
		4	94	3.24
		5	119	3.21
	Lowest	1	25	2.84
		2	59	2.85
		3	41	2.86
		4	45	2.86
		5	105	2.86

Table J3

Skewness of Variables

Variable	Skewness
Human	
H	-.35
E	.04
X	.12
A	-.05
C	-.29
O	-.44
AI	
H	-.48
E	-.02
X	-.02
A	.18
C	-.30
O	.07

Figure J1

Boxplot: Human and AI ratings: Honesty-Humility Trait

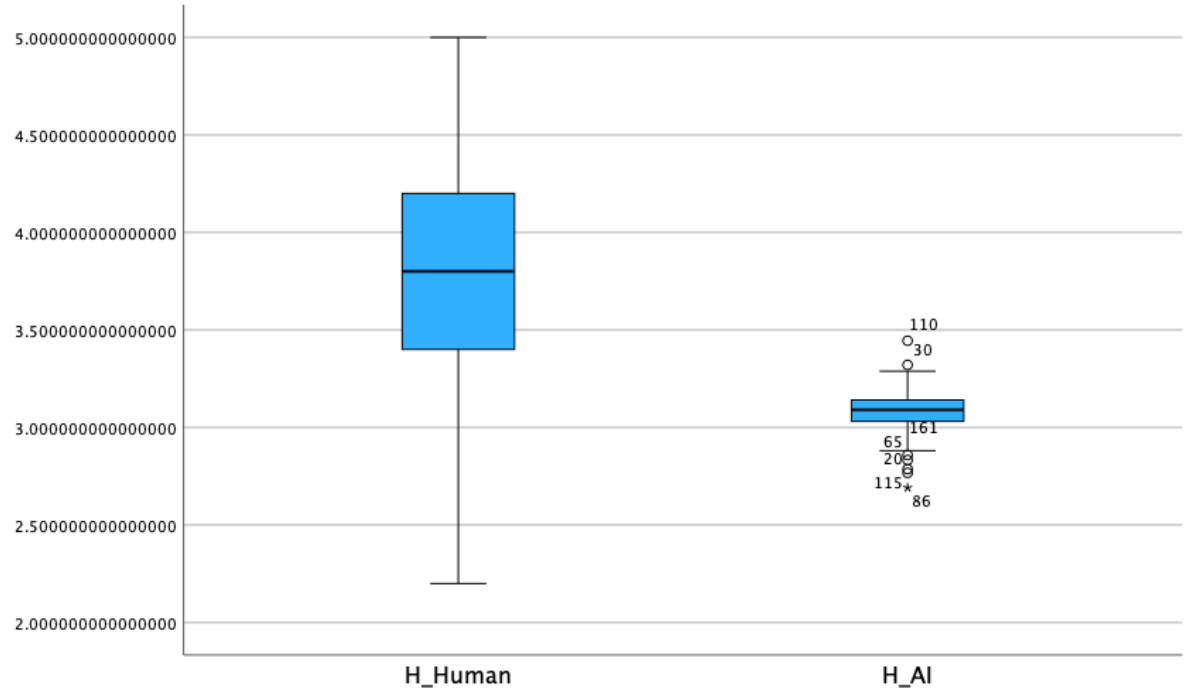


Figure J2

Boxplot: Human and AI ratings: Emotionality Trait

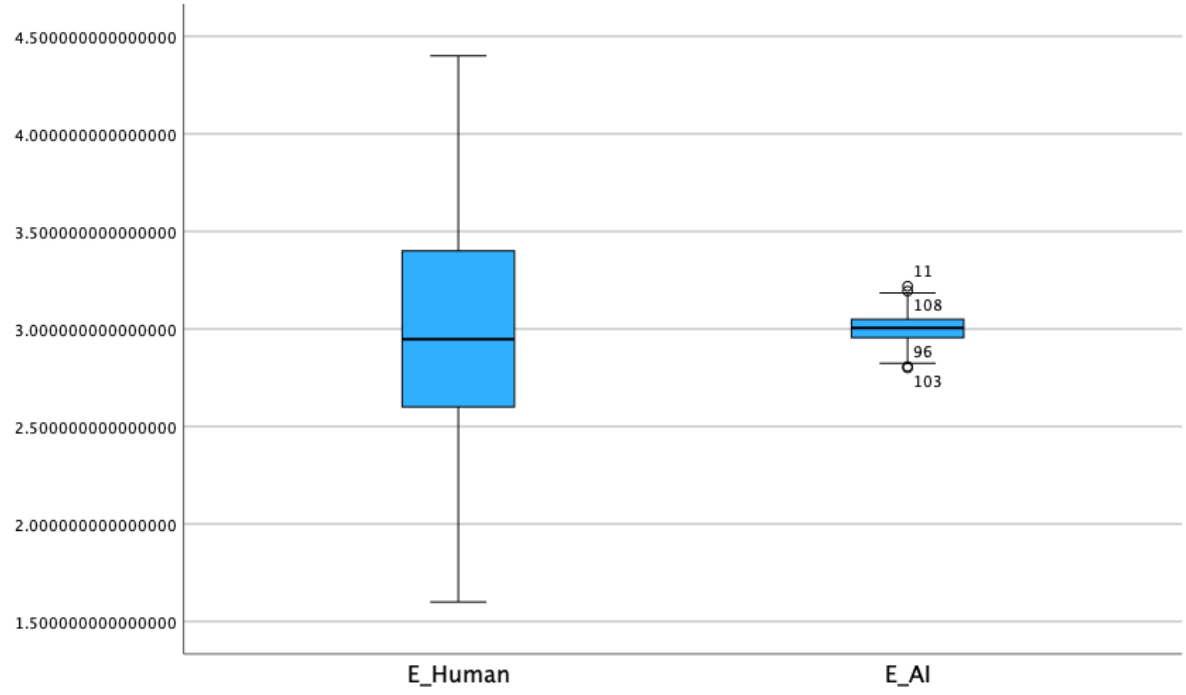


Figure J3

Boxplot: Human and AI ratings: Extraversion Trait

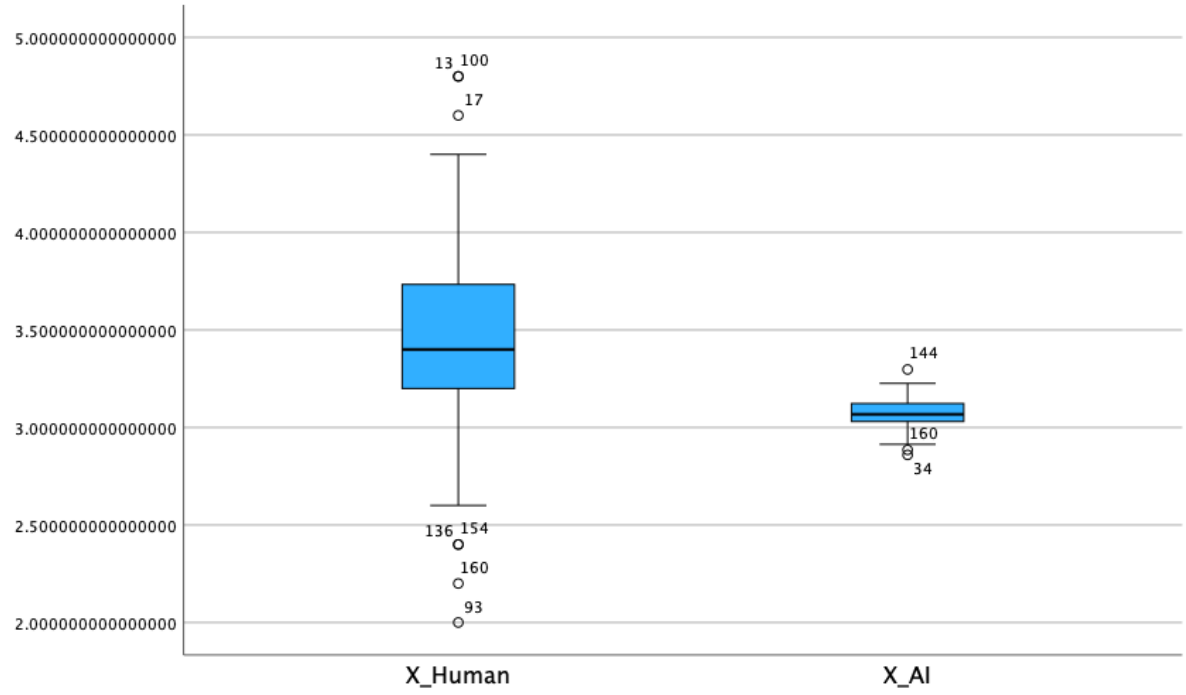


Figure J4

Boxplot: Human and AI ratings: Agreeableness Trait

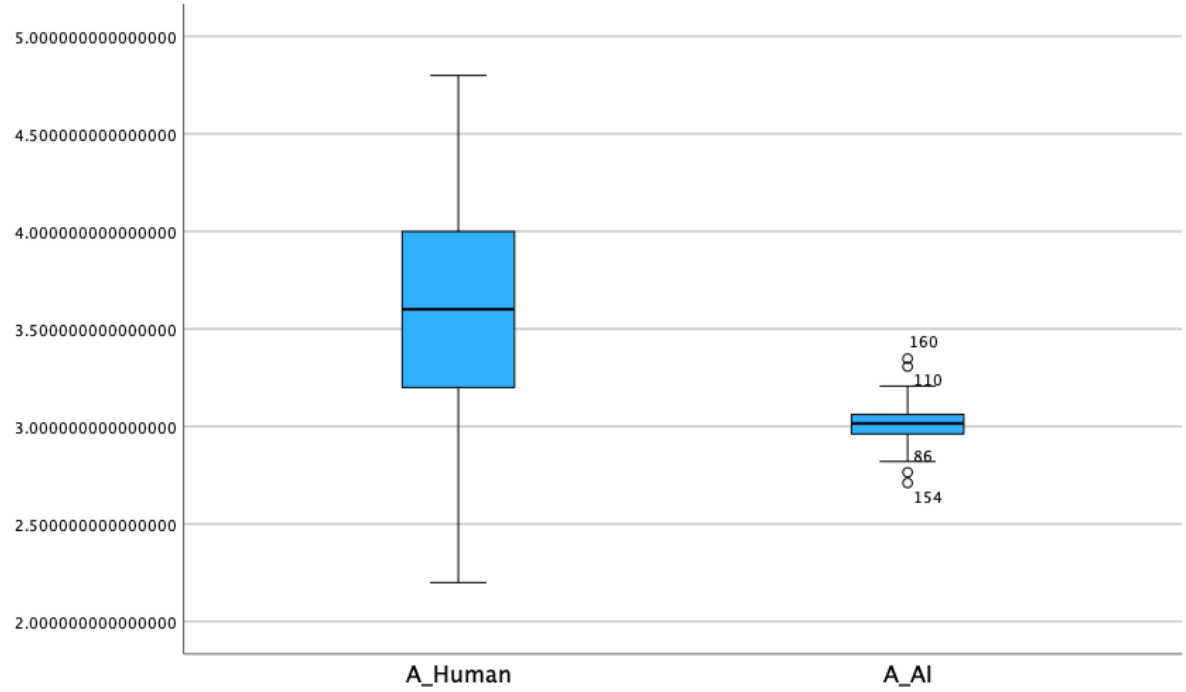


Figure J5

Boxplot: Human and AI ratings: Conscientiousness Trait

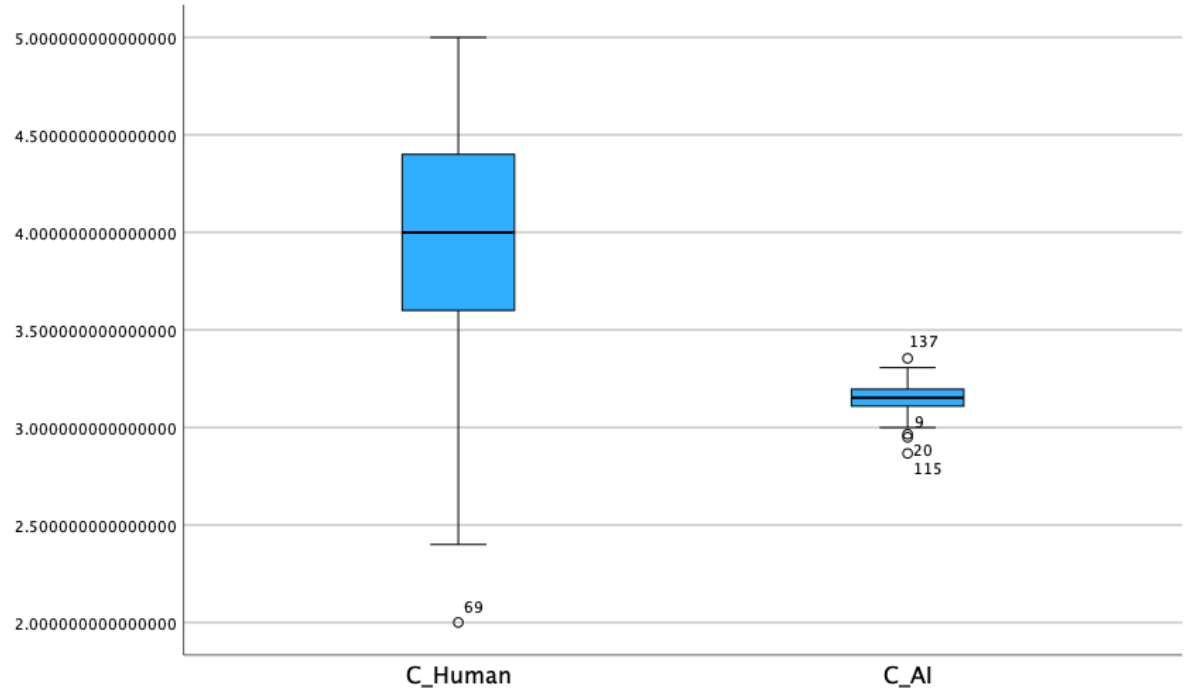
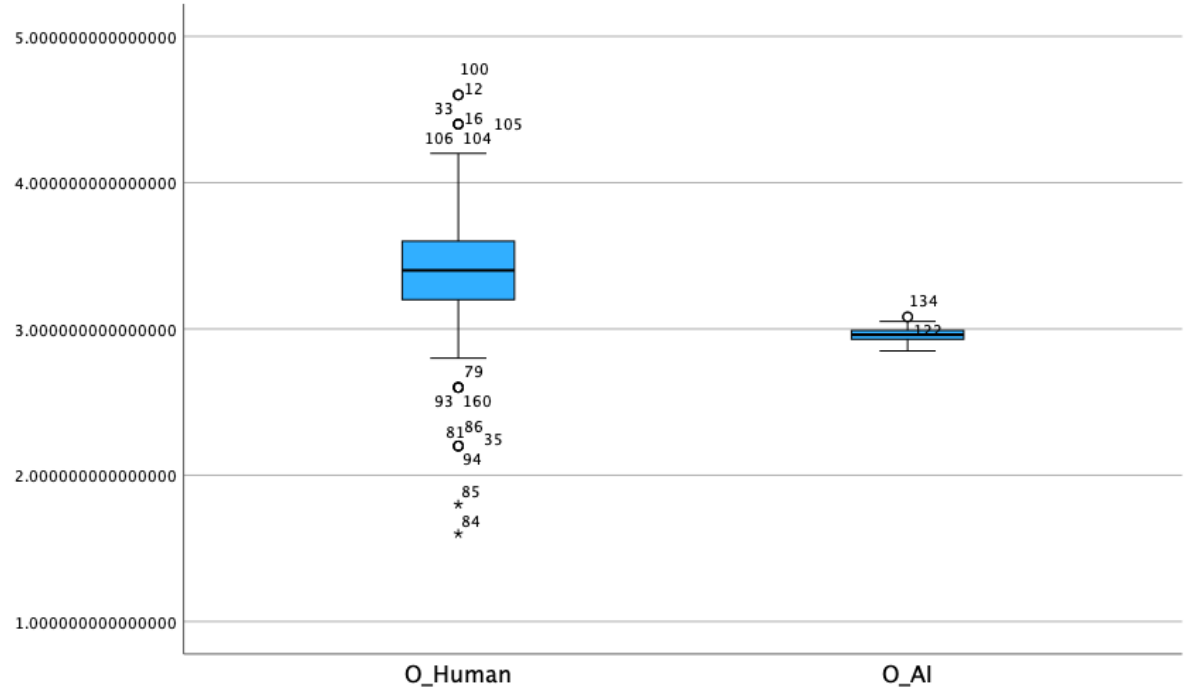


Figure J6

Boxplot: Human and AI ratings: Openness to Experience Trait



Annexure K

Managing Outliers and Further Analysis

In support of Chapters 4 and 5, which present the analyses and discussions, this section (Annexure K) provides further elaboration. It is included as an annexure, but remains aligned with the main section. Here, outlier management is explored under two alternative conditions that manage outliers, additional to the primary analyses, which proceeded with the full dataset, including outliers.

Outliers are data points that significantly differ from the rest of the dataset and can have a disproportionate impact on the conclusions drawn about the relationships between variables (Aguinis et al., 2013). As these extreme values are potential sources of bias that could skew the results (Field, 2018), reasonable steps were taken to identify and address them in the dataset. Following the guidance of Field (2018) and Tabachnick and Fidell (2019), casewise diagnostics and visual techniques were used alongside three statistical methods. The three approaches employed to identify and manage outliers were: range-based detection to confirm scoring validity (used in the main analysis), z-score analysis to flag extreme deviations, and log transformations to address skewness, as outlined below.

1.1 Range-Based Outlier Detection

The first method of outlier detection was a range-based approach, where extreme values were examined, as displayed in Table J2. Given that the scoring range is from 1 to 5, it was confirmed that no values fell outside this range, as shown in Table J2. Although some values were identified as extreme compared to the dataset, no erroneous values exceeded the scoring limits. Therefore, while all values remained within the acceptable scoring range, some extreme values were identified. Nevertheless, since the data remains within the scoring range, it was decided to proceed with a correlation study using the full dataset, as evident in the main analysis. In light of the identification of some outliers during the data cleaning process, two additional approaches were implemented. The main correlation analysis was therefore carried out on the complete dataset. Additionally, in this section, analyses were conducted on datasets after applying the z-score method and performing transformations.

1.2 Z-Score Method

The second approach involved Z-Score Analysis. Z-scores represent standardised residuals in terms of standard deviations (Field, 2018) and were calculated for each variable. Using the ±3 threshold, values exceeding this range were flagged as extreme outliers for further investigation, in line with the guidelines by Field (2018) and Tabachnick & Fidell

(2019). This method allowed for the identification of data points with significant deviations from the mean, ensuring that extreme values were addressed appropriately.

For each HEXACO trait, human and AI evaluations, z-scores were calculated using SPSS. Once calculated, a condition was applied in SPSS to identify cases where the absolute value of the z-scores exceeded ±3. For example, the condition ABS(ZH_Human) > 3 was used to flag participants with extreme outlier values for the Honesty-Humility trait provided by the human raters. Only data points within the accepted range of standard deviations were retained for further analysis.

1.3 Transformation for Skewed Data

The third and additional method addressed skewed data through log and square root transformations, as guided by Field (2018). Tests of normality, such as the Shapiro-Wilk and Kolmogorov-Smirnov tests, revealed that several variables, both human and AI ratings for the HEXACO traits, did not follow a normal distribution (see Table 4). Specifically, the human ratings for traits such as H_Human, E_Human, X_Human, A_Human, and O_Human failed to meet the assumption of normality, with significant p-values ($p < .05$) (Field, 2018). While some AI-generated ratings, including E_AI, X_AI, C_AI and O_AI, did not significantly deviate from normality ($p > .05$), others, such as A_AI and H_AI, displayed notable deviations.

According to Norris and Aroian (2004) and Tabachnick and Fidell (2019), the criteria for problematic skewness warranting transformation vary. However, the degree of skewness is considered a good indicator. Moderate skew is indicated by a skewness value of 1 or 1.25 standard deviations from the normal distribution, while high skew is indicated by values above 2.25 (Norris & Aroian, 2004). Significance tests are also often used to guide decisions regarding transformation (Norris & Aroian, 2004). Nonetheless, unless there are specific reasons to avoid transformation, it is generally recommended to proceed with it (Tabachnick & Fidell, 2019).

Although the skewness values in the data are below 1, as per Table J3, suggesting that transformation may not be strictly necessary, square root transformation, and thereafter log transformation were applied as supplementary steps to address non-normality in the dataset, as guided by Tabachnick and Fidell (2019). Square root transformation is particularly suited for addressing mild skewness. While initially the skewness observed was not substantial enough to warrant more aggressive transformations, such as log transformation, the square root transformation was applied to improve the normality of the data and reduce the influence

of extreme values, ensuring the assumptions of normality and homogeneity of variance are met (Feng, 2014; Tabachnick & Fidell, 2019).

Given that the post-square root transformed data still exhibited some skewness and kurtosis, as per Table K1, a subsequent log transformation was applied to the original data. The log transformation is often recommended as a next step to address remaining skewness. By compressing extreme values further, it can help reduce skewness and improve the overall normality of the distribution (Tabachnick & Fidell, 2019). However, considering the moderate skewness values, it is important to note that the log transformation might have been too extreme if applied initially. Therefore, it was used as a follow-up procedure after the square root transformation.

To maintain consistency, the same transformation technique was applied across all groups, even though it might not have been the most optimal for each group (or trait) individually, which is the recommended approach (Tabachnick & Fidell, 2019). However, as Feng et al. (2014) noted, transformations should be used with caution, as they can alter the original relationships in the data. Therefore, the transformations were used as additional or supplementary analyses, as described below.

1.3.1 Detailed Examination of Transformation Analyses

Transformations were applied to address skewed distributions and improve normality. The comparative analysis of transformations applied to the data (square root and logarithmic) reveals significant differences in the distribution characteristics of the HEXACO traits.

1.3.1.1 Human Evaluator Scores

For human evaluator scores (e.g., H_Human, E_Human), original data often exhibited non-normal distributions, as evidenced by skewness, kurtosis, and significant p-values in the Shapiro-Wilk test. For example, H_Human had a Shapiro-Wilk p-value of .001, indicating a significant deviation from normality. Transformations improved normality to varying degrees, with the square root transformation resulting in a skewness of -.52 and kurtosis of -.36 but a Shapiro-Wilk p-value remaining significant at <.001. Similarly, logarithmic transformations for H_Human reduced kurtosis to -.09, but the Shapiro-Wilk p-value also remained significant (<.001). Compared with the original data used in the main analysis, these results are detailed in Table K1.

1.3.1.2 AI Scores

For AI scores, some traits demonstrated closer adherence to normality in their original form. For instance, E_AI exhibited a near-normal distribution with a Shapiro-Wilk p-value of .370. Transformations further improved the normality, with the logarithmic transformation

resulting in skewness of -.106 and kurtosis of .253, alongside a Shapiro-Wilk p-value of .320. Additionally, traits such as H_AI retained significant deviations from normality across all transformations, with skewness ranging from -.477 (original) to -.674 (logarithmic) and a Shapiro-Wilk p-value consistently below <.001.

1.3.1.3 Comparative Impact of Transformations

The transformations had varying impacts across traits. For instance, traits like O_Human remained highly non-normal among human evaluator scores, with logarithmic transformations yielding a skewness of -1.363 and kurtosis of 4.443 and a significant Shapiro-Wilk p-value of <0.001. Among AI scores, some traits, such as O_AI, were naturally closer to normality and retained non-significant Shapiro-Wilk p-values after transformations.

1.3.1.4 Summary of Findings

In summary, while square root and logarithmic transformations improved normality for many variables, some traits, particularly those evaluated by human raters, remained non-normal even after transformation. The square root transformation proved more effective in stabilising variance and improving distribution symmetry, particularly for traits with moderate skewness and kurtosis. However, logarithmic transformations were more beneficial for traits requiring more significant reductions in skewness. These findings highlight the necessity of selecting transformations tailored to the specific distributional properties of the data to ensure accurate statistical analyses while also carefully considering whether such transformations are, in fact, necessary for further analyses. Data transformation is not always required or advisable when calculating correlations on skewed data, as suggested by Norris and Aroian (2004).

Table K1

Comparative Analysis of Normality

Variable	Transformation	Skewness	Kurtosis	Shapiro-Wilk Sig.
H_Human	Original	-.35	-.53	.001
	Square Root	-.52	-.36	<.001
	Logarithmic	-.69	-.09	<.001
E_Human	Original	.04	-.53	.005
	Square Root	-.16	-.42	.004
	Logarithmic	-.39	-.10	<.001
X_Human	Original	.12	.90	.003

	Square Root	-.18	1.12	.002
	Logarithmic	-.51	1.68	<.001
A_Human	Original	-.05	-.47	.013
	Square Root	-.22	-.27	.008
	Logarithmic	-.40	.08	.002
C_Human	Original	-.29	.09	.007
	Square Root	-.55	.77	<.001
	Logarithmic	-.86	1.92	<.001
O_Human	Original	-.44	1.59	<.001
	Square Root	-.87	2.57	<.001
	Logarithmic	-1.36	4.44	<.001
H_AI	Original	-.48	2.24	<.001
	Square Root	-.58	2.35	<.001
	Logarithmic	-.67	2.50	<.001
E_AI	Original	-.02	.26	.370
	Square Root	-.06	.25	.353
	Logarithmic	-.11	.25	.320
X_AI	Original	-.02	.43	.647
	Square Root	-.06	.45	.627
	Logarithmic	-.10	.46	.586
A_AI	Original	.18	1.83	.017
	Square Root	.09	1.78	.021
	Logarithmic	.01	1.76	.024
C_AI	Original	-.30	.63	.268
	Square Root	-.35	.72	.182
	Logarithmic	-.39	.82	.115
O_AI	Original	.07	.10	.928
	Square Root	.05	.09	.939
	Logarithmic	.03	.09	.945

Note. H: Honesty-Humility, E: Emotionality, X: Extraversion, A: Agreeableness, C: Conscientiousness, O: Openness to Experience.

1.4 Correlation Studies on Alternative Approaches

This section presents the correlation studies for the two additional data-handling approaches or conditions. Condition 1, as used in the primary analysis, included all data points, while Condition 2 excluded outliers using the Z-Score method, and Condition 3 applied transformations (logarithmic and square root).

1.4.1 Alternative Condition 2: Excluding Outliers (Z-Score Method)

This section examines the analysis conducted after excluding extreme values to investigate correlations between human and AI ratings of HEXACO traits. Excluding outliers provided an alternative perspective on the data, enabling comparisons between Condition 1 (including outliers) and Condition 2 (excluding outliers). The findings in both conditions align with previous research and highlight some alignment between human and AI ratings for specific HEXACO traits. Table K2 presents the results. Outliers were identified based on Z-scores, with values exceeding ±3 flagged as extreme, consistent with widely accepted statistical thresholds.

Table K2

Spearman's rho Correlation Coefficients: Condition Two

Variable	H_AI	E_AI	X_AI	A_AI	C_AI	O_AI
H_Human	**.181**	-.040	.006	.042	.134	-.019
p-value	.027	.626	.947	.608	.103	.818
E_Human	-.014	**.100**	.114	.138	-.130	.060
p-value	.866	.225	.165	.094	.115	.467
X_Human	.330	-.002	**.067**	.255	.040	-.092
p-value	<.001	.997	.420	.002	.632	.265
A_Human	.118	-.022	.108	**.010**	.128	-.035
p-value	.152	.788	.191	.904	.119	.669
C_Human	.146	-.106	.148	-.011	**.168**	-.007
p-value	.076	.198	.072	.895	.041	.936
O_Human	.195	-.004	.155	.156	.088	**.050**
p-value	.017	.964	.060	.057	.284	.545

N = 149

Note. Values are rounded to three decimal places due to lower values for some traits.

Note. H: Honesty-Humility, E: Emotionality, X: Extraversion, A: Agreeableness, C: Conscientiousness, O: Openness to Experience.

1.4.1.1. Identifying and Filtering Outliers

As outlined in Section 1.2. of Annexure K, values exceeding ±3 standard deviations were flagged as extreme. These cases were identified as outliers for potential exclusion to study and prevent undue influence on the results. Outliers were excluded using SPSS's Select Cases function, with the condition abs(Z_Human) < 3 applied to each human-rated HEXACO trait. This ensured consistent exclusion across all variables.

1.4.1.2. Condition 2 Spearman's Correlation and Hypotheses Testing

After removing outliers, Spearman's rank correlation was used to assess the relationship between human and AI ratings, as per Table K2. Spearman's rho was chosen for consistency with Condition 1 and its applicability to non-normal data. While the correlations were generally small, they indicated notable convergence between human and AI ratings for some traits. Holtrop et al. (2022) and Koutsoumpis et al. (2024), which are similar studies, do not specify whether or how outliers were handled. Nonetheless, the correlations observed in Condition 1 of the present study (inclusive of outliers) and those in Condition 2 (excluding outliers) appear consistent with the findings of these past studies. Specifically, Holtrop et al., (2022) reported small to moderate correlations between human and AI ratings in a similar context, particularly for Honesty-Humility and Conscientiousness, which showed significant and positive correlations.

Honesty-Humility (H): The correlation between human and AI scores for Honesty-Humility is small, yet positive and significant ($r = .18$, $p < .05$), closely aligning with findings from Holtrop et al. (2022), who reported a similar correlation ($r = .20$, $p < .05$). This suggests some alignment in how this trait is assessed across the two methods and studies, supporting H1a. Therefore, this result leads to the rejection of H0 for Honesty-Humility, indicating a positive relationship between human and AI ratings of Honesty-Humility.

Emotionality (E): A weak positive, yet non-significant correlation ($r = .10$, $p = .23$) does not support H1a, suggesting no meaningful convergence between human and AI ratings of Emotionality. Consequently, this result fails to reject H0 for Emotionality, as no significant relationship was observed.

Extraversion (X): A weak positive correlation ($r = .07$, $p = .42$) fails to provide evidence for H1a, demonstrating no significant convergence between human and AI ratings of Extraversion. This result fails to reject H0 for Extraversion, and H1b is not supported, as Extraversion did not show the strongest convergence between human and AI ratings as hypothesised. Instead, Honesty-Humility exhibited the strongest convergence.

Agreeableness (A): A weak positive correlation ($r = .01$, $p = .90$) fails to support H1a, suggesting no notable alignment between human and AI ratings of Agreeableness. Thus, this result fails to reject H0 for Agreeableness.

Conscientiousness (C): A weak positive, yet significant, correlation ($r = .17$, $p = .04$) supports H1a, indicating some alignment between human and AI ratings of Conscientiousness, although the correlation is small. This finding leads to the rejection of H0 for Conscientiousness, suggesting some significant convergence between the two rating methods.

Openness to Experience (O): A weak positive correlation ($r = .05$, $p = .55$) fails to support H1a, indicating no significant convergence between human and AI ratings of Openness to Experience. Therefore, this result fails to reject H0 for Openness to Experience.

1.4.1.3. Further Observations

Several notable correlations were observed between AI and human ratings across different traits. For example, AI-rated Honesty-Humility exhibited a significant positive correlation with human-rated Extraversion ($r = .33$, $p < .001$), which was stronger than under Condition 1 ($r = .30$, $p < .001$). This unexpected finding was thus reinforced, suggesting a need for further investigation. As described under Condition 1, this result suggests potential overlap in how these traits are assessed by the two evaluators, highlighting the importance of exploring these connections in future research. Additionally, AI-rated Honesty-Humility was positively correlated with human-rated Openness to Experience ($r = .20$, $p = .02$), a stronger and more significant correlation than under Condition 1 ($r = .18$, $p = .02$), revealing further possible connections between the evaluation of these traits. Further significant correlations include AI-rated Agreeableness versus human-rated Extraversion ($r = .26$, $p < .05$), which was more substantial than under Condition 1 ($r = .21$, $p = .01$).

Therefore, excluding outliers strengthened some traits' correlations, making them more significant than under Condition 1. As detailed in Annexure L, significant correlations were also observed within raters and traits. However, since this study focuses on the convergence between methods, further discussion of intra-rater correlations is beyond the scope of this analysis, as stated under Condition 1 in the primary analysis.

Overall, the results under Condition 2, provide partial support for H_{1a}, with significant positive correlations found for Honesty-Humility and Conscientiousness, indicating some alignment between human and AI ratings. However, H_{1b} is not fully supported, as Extraversion did not demonstrate the most substantial convergence between human and AI

ratings. The relatively weak strength of these correlations and the unexpected correlations observed suggest the need for further research in this area.

1.4.2 Alternative Condition 3: Transforming Data

The third method addressed skewed data through transformations, as guided by Tabachnick and Fidell (2019). Certain variables in the dataset exhibited significant skewness, which could impact the normality assumptions necessary for statistical analyses. Transformations were applied to these skewed variables to normalise their distributions and potentially mitigate the influence of extreme values, attempting to ensure that the data met the assumptions of normality and homogeneity of variance required for subsequent analyses (Field, 2018). As discussed in the normality tests (Shapiro-Wilk and Kolmogorov-Smirnov tests), several variables, both human and AI ratings for the HEXACO traits, violated the assumption of normality (see Table J3).

1.4.2.1. Data Transformation Methods

Two transformations were applied to the dataset: logarithmic (LOG) and square root (SQRT) transformations. These transformations were selected for their effectiveness in addressing distribution issues such as skewness and heteroscedasticity, as described under Section 1.3 of the current Annexure (K). Table K3 and K4 present the Spearman correlation outputs under the two transformation conditions.

Square Root Transformation: Square root transformations are effective at reducing moderate skewness.

Table K3

Spearman's Correlation Coefficients: Condition Three: Square Root Transformation

Variable	H_AI	E_AI	X_AI	A_AI	C_AI	O_AI
H_Human	**.207**	-.034	.007	.059	.169	-.013
p-value	.008	.672	.930	.456	.032	.872
E_Human	-.004	**.118**	.142	.133	-.140	.068
p-value	.963	.136	.073	.094	.077	.392
X_Human	.295	.043	**.089**	.209	.036	-.039
p-value	<.001	.584	.260	.008	.650	.624
A_Human	.179	-.025	.133	**.053**	.186	-.028
p-value	.023	.752	.091	.503	.018	.725
C_Human	.142	-.060	.162	-.017	**.156**	.019
p-value	.073	.453	.039	.827	.048	.816

O_Human	.182	.039	.136	.141	.106	**.046**
p-value	.021	.621	.085	.075	.179	.565

N = 161

Note. Values are rounded to three decimal places due to lower values for some traits.

Note. H: Honesty-Humility, E: Emotionality, X: Extraversion, A: Agreeableness, C: Conscientiousness, O: Openness to Experience.

Logarithmic Transformation: This method is particularly useful for correcting right-skewed data (positive skewness), as it compresses larger values while preserving the overall shape of the data distribution.

Table K4

Spearman's Correlation Coefficients: Condition Three: Logarithmic Transformation

Variable	H_AI	E_AI	X_AI	A_AI	C_AI	O_AI
H_Human	**.207**	-.034	.007	.059	.169	-.013
p-value	.008	.672	.930	.456	.032	.872
E_Human	-.004	**.118**	.142	.133	-.140	.068
p-value	.963	.136	.073	.094	.077	.392
X_Human	.295	.043	**.089**	.209	.036	-.039
p-value	<.001	.584	.260	.008	.650	.624
A_Human	.179	-.025	.133	**.053**	.186	-.028
p-value	.023	.752	.091	.503	.018	.725
C_Human	.142	-.060	.162	-.017	**.156**	.019
p-value	.073	.453	.039	.827	.048	.816
O_Human	.182	.039	.136	.141	.106	**.046**
p-value	.021	.621	.085	.075	.179	.565

N = 161

Note. Values are rounded to three decimal places due to lower values for some traits.

Note. H: Honesty-Humility, E: Emotionality, X: Extraversion, A: Agreeableness, C: Conscientiousness, O: Openness to Experience.

Although the correlation values from Tables K3 and K4 closely align with the main study correlation outputs (Table 5), suggesting that the transformations had little to no impact on correlation strength, statistical comparisons were used instead of manual inspection to assess agreement between the various conditions, as outlined in the next section.

1.4.3 Fisher Z-Test Analysis of Correlations Across Conditions

The Fisher z-test was applied to compare the correlations between human and AI ratings for each HEXACO trait across three conditions: Condition 1 (Original Data), Condition 2 (Excluding Outliers), and Condition 3 (Transformed Data using Square Root and Log transformations), as per Table K5. This analysis aimed to determine whether the correlation coefficients differed significantly across conditions, assessing the robustness of the relationship between human and AI evaluations and supporting the decision to use Condition 1 as the primary evaluation dataset for the hypotheses. While Fisher's Z-method is typically applied to normally distributed data, which is rare in psychological research, it is also used cautiously and successfully with non-normal data (Berry & Mielke, 2000).

Therefore, the focus of this analysis is on testing the statistical significance of the relationships between human evaluator ratings and AI algorithm ratings, rather than merely identifying patterns and trends in the data manually. Spearman's correlation was used as the basis for Fisher's z-test, which was applied to compare the correlation coefficients across conditions, with significance assessed at the $p = .05$ level. As guided by Field (2018), a critical value of $z = 1.96$ was used to determine significant differences between correlations, providing an inferential basis for understanding the differences in the strength of these relationships. Table K5 presents Fisher's z-test results to offer a comparative overview of the relationships across the HEXACO traits and conditions. The findings are interpreted inferentially, addressing the hypotheses and providing meaningful insights into the alignment between human and AI evaluations of personality.

Table K5

Fisher z-Test Results for HEXACO Correlations Across Conditions

Trait	z-Value (Condition 1 vs. Condition 2)	p-Value (Condition 1 vs. Condition 2)	z-Value (Condition 1 vs. Condition 3)	p-Value (Condition 1 vs. Condition 3)
H	.24	.81	.00	1.00
E	.16	.87	.00	1.00
X	.18	.85	-.01	.99
A	.38	.71	.00	1.00
C	-.11	.91	.00	1.00
O	-.03	.97	.00	1.00

Trait H (Honesty-Humility): No significant differences were found across conditions (z-values ranged from .00 to .24, p-values ranged from .81 to 1.00).

Trait E (Emotionality): No significant differences were observed (z-values ranged from .00 to .16, p-values ranged from .87 to 1.00).

Trait X (Extraversion): No significant differences emerged (z-values ranged from -.01 to .18, p-values ranged from .85 to .99).

Trait A (Agreeableness): No significant differences were detected (z-values ranged from .00 to .38, p-values ranged from .71 to 1.00).

Trait C (Conscientiousness): No significant differences were found (z-values ranged from .00 to -.11, p-values ranged from .91 to 1.00).

Trait O (Openness to Experience): No significant differences were observed (z-values ranged from .00 to -.03, p-values ranged from .97 to 1.00).

As no significant differences were found across conditions for any HEXACO trait, the researcher concluded that the original data (Condition 1) provided a reliable representation of the relationships. Therefore, for hypothesis testing, the study proceeded with Condition 1 for the primary analysis, as it includes all data and no substantial changes were observed in the correlations when outliers were excluded or transformations applied.

Annexure L

Statistical Outputs Supporting HEXACO Trait Comparisons Between Humans and AI

1.1 Additional Analyses

Additional analyses were conducted to further explore the correlation between human and AI evaluators and HEXACO traits, further assess the alignment and convergence between the two methods, and support the primary analysis, as provided in Chapter 4. Given that this study aims to compare HEXACO trait ratings assigned by human raters and an AI-based scoring method to determine the extent of their convergence or divergence, this section presents a more detailed examination of scoring patterns across evaluators than briefly provided in Chapter 4. The primary objective of this additional section is to assess the strength of the association between human and AI ratings and the presence of systematic differences in trait assessments. Spearman correlations provided a substantial measure of agreement or convergence, as per the main content of the study. At the same time, Analysis of Variance (ANOVA/MANOVA) offer insights into significant differences across evaluator types and traits. The following subsections outline the statistical methodologies employed and the findings derived from these analyses.

1.1.1 Multivariate and Trait-Specific Analysis

This study included an analysis of Variance (ANOVA/MANOVA) to complement the Spearman correlation analysis by assessing further potential differences in HEXACO trait scores between human evaluators and the AI algorithm. Guided by Field (2018) and Tabachnick and Fidell (2019), while Spearman correlations examine the strength and direction of relationships between traits, ANOVA/MANOVA provided insight into whether significant differences exist in how human and AI evaluators score these traits. This additional analysis was necessary to evaluate the consistency and potential variability in the evaluation methods, aiming to ensure an even more comprehensive comparison of human and AI scoring approaches. Therefore, building on the correlation analysis, the next phase focuses on assessing multivariate differences and trait-specific effects in human and AI raters' evaluations of HEXACO traits. Through MANOVA, overarching patterns in evaluator differences were explored, followed by ANOVA, which examined how specific traits differ between evaluators. The findings from these analyses may offer a more detailed understanding of how human and AI evaluators align or diverge in their assessments of the HEXACO traits.

1.1.1.1 Comparison of Human and AI Evaluators per HEXACO Trait

This study aimed to compare HEXACO personality trait ratings derived from asynchronous video interviews (AVIs), as assessed by human evaluators and AI. Repeated Measures ANOVA, also known as Within-Subjects ANOVA, was used to investigate rating differences between the two evaluator types across the six HEXACO traits. The primary goal was to determine the extent to which human and AI evaluations aligned or diverged for each trait, in addition to the correlation studies.

The within-subjects factor in the analysis was the evaluation source (Human vs. AI), with each trait analysed separately for every participant. Mauchly's test assessed sphericity, a key assumption that requires consistent relationships between scores across conditions (i.e., ratings by Human and AI evaluators). When sphericity was violated, corrections such as Greenhouse-Geisser, Huynh-Feldt, and Lower-bound were applied, as Field (2018) recommended and provided in section 1.1.3. of the current Annexure.

This approach allowed for an even more detailed investigation of the convergence between human and AI-based evaluations of HEXACO personality across the six traits. The analysis included descriptive and inferential statistical methods to explore differences attributable to evaluator type (Human vs AI), trait, and their interaction (evaluator and trait). Each stage of the process is detailed below.

Firstly, descriptive statistics were calculated to summarise the central tendency, variability, and distribution of the HEXACO trait ratings across human and AI evaluations. The means and standard deviations for both evaluation methods were examined to identify any obvious or notable differences in scoring patterns. As shown in Tables 3 and L1, the results indicate that, consistent with the main analysis, mean scores for human evaluations tend to vary more widely than those for AI evaluations. As shown in Tables 3 and L1, the results indicate that, consistent with the main analysis, mean scores for human evaluations tend to vary more widely than those for AI evaluations. Building on this, this section examines the relationship between evaluator type and trait scores, specifically whether the two evaluator types rated the HEXACO traits significantly differently.

Table L1

Descriptive Statistics for ANOVA and Study of Convergence

Evaluator	Trait	Mean	Std. Deviation	N
Human	Honesty-Humility (H)	3.74	.58	161
Human	Emotionality (E)	2.95	.52	161

Human	Extraversion (X)	3.45	.46	161
Human	Agreeableness (A)	3.65	.49	161
Human	Conscientiousness (C)	3.94	.55	161
Human	Openness to Experience (O)	3.42	.50	161
AI	Honesty-Humility (H)	3.08	.10	161
AI	Emotionality (E)	3.00	.08	161
AI	Extraversion (X)	3.07	.07	161
AI	Agreeableness (A)	3.01	.09	161
AI	Conscientiousness (C)	3.15	.08	161
AI	Openness to Experience (O)	2.96	.04	161

N = 161

1.1.2 Multivariate Tests (MANOVA)

A Multivariate Analysis of Variance (MANOVA) assessed significant differences between human and AI evaluations across all six HEXACO traits. Following Field's (2018) guidelines, this analysis provided an overview of the relationship between evaluator type and traits before focusing on trait-specific patterns.

1.1.2.1 Multivariate Effects

Pillai's Trace was selected as the primary test statistic due to its reliability, even when the assumption of homogeneity of variance-covariance matrices is not fully met (Field, 2018). As shown in Table L2, significant multivariate effects were found for evaluator type, trait, and their interaction (Evaluator × Trait).

1.1.2.2 Evaluator Effect

Wilks' Lambda was also reported for the Evaluator effect, as it provides a more conservative estimate of multivariate differences when testing for group differences (Field, 2018; Tabachnick & Fidell, 2019). The Wilks' Lambda value of .33 ($p < .001$) indicates that human and AI evaluations diverge significantly across all six HEXACO traits combined, meaning there is a substantial difference in how the two evaluators assess these traits overall. This is further supported by a sizeable partial eta squared ($\eta^2 = .67$), which signifies a large effect size, suggesting that the evaluator type (human vs. AI) plays a significant role in determining the ratings.

1.1.2.3 Trait Effect

The Trait effect, indicated by Pillai's Trace = .71 ($p < .001$, partial $\eta^2 = .71$), shows that the six HEXACO traits are rated differently in general, meaning there is inherent

variability in how the traits are assessed. This confirms that some traits are likely to show more variation in scores than others, which could be important for understanding the nature of these traits in evaluations.

1.1.2.4 Evaluator x Trait Interaction

Additionally, the significant Evaluator x Trait interaction (Pillai's Trace = .60, $p < .001$, partial $\eta^2 = .60$) reveals that the difference between human and AI evaluations varies depending on the specific trait being assessed. For example, the discrepancy between evaluators might be larger for some traits and smaller for others. This interaction, supporting the main correlation matrix (Table 5), suggests that the relationship between evaluator type and trait is not uniform. While an overall significant and moderately positive correlation between the two raters (human and AI) of the HEXACO traits has been shown, further investigation is needed to understand why certain traits show the greatest convergence or divergence between human and AI raters.

Table L2

MANOVA Output

Effect	Test	Value	F	Sig.	Partial η^2
Evaluator	Pillai's Trace	.67	329.06	<.001	.67
	Wilks' Lambda	.33	329.06	<.001	.67
Trait	Pillai's Trace	.71	77.62	<.001	.71
Evaluator x Trait	Pillai's Trace	.60	46.98	<.001	.60

1.1.3 Repeated Measures / Within-Subjects ANOVA

1.1.3.1 Mauchly's Test of Sphericity

Before conducting the Repeated Measures ANOVA, Mauchly's Test of Sphericity was performed to check the assumption of equal variances. In the context of the current study, which examines repeated measures of human and AI evaluations across the six HEXACO traits, sphericity refers to the assumption that the variances of the differences between all combinations of evaluator conditions (i.e., human vs. AI) and trait ratings are equal (Field, 2018). This assumption is important for the validity of statistical tests like the Repeated Measures ANOVA, as violations of sphericity can distort results, leading to inaccurate conclusions (Field, 2018).

Mauchly's Test of Sphericity was conducted to assess whether this assumption holds. A significant p-value (typically $p < .05$) indicates a violation of sphericity, meaning the

assumption does not hold, and adjustments must be made to correct for this (Field, 2018). In this study, the results of Mauchly's test revealed violations of sphericity for both the Trait and Evaluator x Trait effects, with significant p-values ($x^2 = 83.18$, $p < .001$ and $x^2 = 63.14$, $p < .001$, respectively), as shown in Table L3. This means the variances of the differences between conditions were unequal, which could lead to inaccurate statistical inferences (Field, 2018). Therefore, corrections were necessary to ensure the validity of the analysis.

1.1.3.1.1 Correction for sphericity violations. To address this issue, Greenhouse-Geisser corrections were applied. This correction adjusts the degrees of freedom for the within-subjects effects, providing more accurate estimates when the sphericity assumption is violated (Field, 2018; Tabachnick & Fidell, 2019). The epsilon values for these corrections, .806 for Trait and .851 for the Evaluator × Trait interaction, indicate how much the degrees of freedom were adjusted to account for the violations of sphericity. By applying these Greenhouse-Geisser adjustments, it was ensured that the subsequent statistical tests remain reliable, despite the violations of sphericity (Field, 2018). This correction is particularly important in repeated measures designs, such as the current study, where the same participants are measured across multiple conditions (e.g., human and AI evaluations), as it prevents misleading results and ensures the accuracy of the conclusions.

Table L3

Mauchly's Test of Sphericity

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser Epsilon
Trait	.59	83.18	14	<.001	.81
Evaluator × Trait	.67	63.14	14	<.001	.85

Thus, the violation of sphericity for the Trait and Evaluator × Trait effects, as indicated by Mauchly's test, was addressed using the Greenhouse-Geisser correction. This adjustment ensures that the F-tests for these within-subjects effects are valid, even though the assumption of sphericity was not met (Field, 2018).

1.1.3.1.2 Within-Subjects Effects

Following the sphericity tests and necessary corrections, the within-subjects effects of evaluator type (human vs. AI), trait, and their interaction were analysed to assess significant differences in the evaluation of HEXACO traits across both evaluators, as guided by Field (2018). The table below, Table L4, summarises the significant main and interaction effects for evaluator type, trait, and the evaluator x trait interaction:

Table L4

Within-Subjects Effects Output

Source	df	F	Sig.	Partial η²
Evaluator	1	329.057	<.001	0.673
Trait	4.028	128.372	<.001	0.445
Evaluator × Trait	4.255	77.064	<.001	0.325

N = 161

Evaluator. The significant main effect of evaluator ($F(1,160) = 329.06$, $p < .001$, $\eta^2 = .67$) indicates that there are systematic differences in the personality trait scores provided by human and AI evaluators. These results suggest that the two evaluation methods are not equivalent in assessing the HEXACO traits.

Trait. The main effect of trait ($F(4.03, 160) = 128.37$, $p < .001$, $\eta^2 = .45$) reveals significant differences across the six HEXACO traits in how they were rated, with specific traits being rated more distinctly or more consistently than others. This shows that the way both evaluators rate the traits is not uniform, and there are marked differences in trait assessments.

Evaluator and trait interaction. The evaluator × trait interaction ($F(4.255, 160) = 77.06$, $p < .001$, $\eta^2 = .33$) indicates that the evaluator-related differences in ratings are not consistent across all HEXACO traits. It appears that AI and human evaluations may differ substantially on specific traits, as well as the degree of discrepancy varies depending on the trait being assessed. This interaction highlights the need to explore trait-specific differences in future analyses to better understand how each evaluator (human vs. AI) interacts with specific HEXACO traits.

The significant main effects in Table L4 indicate systematic differences in personality trait scores provided by humans and AI. The interaction suggests that these differences are inconsistent across traits, as outlined in the following section.

1.1.4 ANOVA/MANOVA and Correlation Outcomes

The analysis provides strong evidence of significant systematic differences between human and AI evaluations of HEXACO traits and notable trait-specific variations and interaction effects. While correlations between the two rater types (human and AI) exist, as discussed in Chapters 4 and 5, the results from Annexure L confirm significant differences in how these methods approach their evaluations. This is expected, given that the algorithm was designed as a text-analysis tool. In contrast, human evaluators had a broader range of interpretation techniques and sources to draw from when evaluating (e.g., cue availability and utilisation).

Given these findings, further examination of univariate effects may be necessary to explore trait-specific patterns of convergence or divergence, mainly since correlations between unexpected traits were also observed in Chapters 4 and 5. Therefore, while correlations were evident, the discrepancies between evaluators and the systematic differences assessing HEXACO traits highlight the complexity of personality assessment. This suggests that further research is needed to explore the sources of divergence and how each evaluation method may uniquely contribute to the overall assessment process.