Universiteit Leiden

# Master Media Technology

## Assessing Personality Creatively

Name: Dewi Becu
Student ID: 4057953
Date: 22 / 09 / 2025

Specialisation: Creative Intelligence and Technology

1st supervisor: Max van Duijn
2nd supervisor: Bernhard Hilpert

Master’s Thesis in Creative Intelligence and Technology

Leiden Institute of Advanced Computer Science
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

# Assessing Personality Creatively

Master's Thesis in Creative Intelligence and Technology

Name: Dewi Becu
Student ID: 4057953

Date: 22 / 09 / 2025

First supervisor: Max van Duijn
Second supervisor: Bernhard Hilpert

Leiden Institute of Advanced Computer Science
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

**ARTICLE TYPE**

# Assessing Personality Creatively

Dewi Becu

Leiden University, LIACs, Creative Intelligence and Technology Master Programme, Master Student
**Author for correspondence:** Dewi Becu, Email: s4057953@vuw.leidenuniv.nl.

**Abstract**

Minigames and writing have been shown to hold potential for personality assessment. Currently, personality is primarily assessed through questionnaires, but these assessments have limitations, such as participant fatigue and self-reporting bias. Therefore, this research explored an alternative assessment approach that utilised writing and minigames. Focusing on the trait conscientiousness (from the Big Five personality model), open-ended questions and minigames were formulated and later analysed for participant behaviour. The collected data showed correlations between personality characteristics, such as the way an individual sorts books can predict how disciplined they are. The research suggests that, although the tool developed in this study is too rudimentary to be reliably used, there is significant potential in utilising LLM-analysed writing and minigames for personality assessment.

**Keywords:** personality, assessment, methodology, minigames, serious game, game based assessment, open-ended questions

## 1. Introduction

Personality refers to the patterns of thought, feelings, and behaviour an individual exhibits. Modern research shows that the Big Five model of personality is the most widely and up-to-date model that identifies an individual's personality (Roberts et al. 2014). The Big Five model explains that personality comprises five traits that work independently of each other (DeYoung, Quilty, and Peterson 2007), although there are overlaps, and they do influence each other at times. Later, a hierarchical model of the Big Five was further researched such that each trait encompassed several facets, and each facet has nuances and specific items. Modern and recent research continues to validate these split and specified facets (Soto and John 2017). For example, the trait 'conscientiousness' has 'orderliness' as a facet, which includes 'likes to tidy up' as a nuance and item. Specifying and identifying these facets, nuances, and items helps make personality definitions less vague and more accurate for testing and identification.

The most well-known and extensively researched assessment of an individual's personality is a 5-point Likert questionnaire called the NEO-PI-R (NEO Personality Inventory Revised) (McCrae, Costa, and Martin 2005), also considered the 'gold standard' for personality assessments (Kabigting 2021). It identifies individuals' personalities according to the Big Five model. There are now multiple revised versions of this inventory.

Many methodologies have been created for research. However, questionnaires have stayed as the principal methodology for personality assessment. This is for a good reason, as questionnaires have the ability to be easily be handed out and assessed on the spot (Frew, Whynes, and Wolstenholme 2003). Furthermore, questionnaires are also standardised, reliable and validated across many applications.

However, questionnaires have limitations. In terms of personality, questionnaires focus only on the conscious thoughts and feelings of the participants themselves. This means that the unconscious section of personality, such as behaviour, is omitted. Behaviour traits and specific experiments have been tested to see if they relate to personality. For example, specific experiments have shown correlations between personality (specifically the Big Five traits) and recorded behaviour through reaction times in an online video game (Tekofsky et al. 2013).

Questionnaires are specifically methodologically limited as participants are restricted to the point scale and, therefore, cannot give answers outside of that. They are limited to what the researcher predicts or assumes about an item of personality (Mõttus, Kandler, and Bleidorn 2017). They also rely on self-reports, which are less reliable in personality assessments because participants can only answer what they believe is true, rather than who they explicitly are[a] (Kim, Di Domenico, and Connelly 2019). Other limitations include how participants interpret the Likert scale[b] (Ogden and Lo 2012) and how the questionnaire may often not be taken seriously. This is sometimes attributed to questionnaire fatigue (Welz and Alfons 2025) (Kim, Di Domenico, and Connelly 2019).

This research aims to fill the existing research gap in assessing personality through individuals' behaviour using methods not typically employed, namely gamified tools such as utilising minigames and open-ended questions

The trait 'conscientiousness' was explicitly chosen so that the research can be applied more rigorously in a specific domain, and a more definitive answer can be concluded. It is hypothesised that if there are conclusive results about conscientiousness, this research can be replicated for the other four personality traits, as each trait can be measured and researched independently.

This research presents a novel approach to measuring con-

[a] Such as people appearing more extroverted and agreeable in job application and dating applications (Rau, Schömann, and Grosz 2025)

[b] For example, agreeing strongly to "I am someone who can be somewhat careless at times" is ambiguous to: (1) very careless at time or (2) somewhat careless many times or (3) certain they are careless at times (Zhang, Wing-Yee Tse, and Savalei 2019)

scientiousness, thereby opening the door for future research on this type of measurement. Furthermore, due to this research's exploratory nature, more insights can be found about personality, specifically the trait of conscientiousness. Namely, this research examines how conscientious behaviour contributes to personality and looks into assessing said behaviour and its characteristics.

The contribution of this study includes the following:

- Developing a gamified assessment for personality and specifically conscientiousness.
- Evaluating the validity of minigames for assessing conscientiousness.
- Evaluating the validity of open-ended questions for assessing conscientiousness.
- Evaluating the validity of LLMs for assessing conscientiousness.
- Developing a more fun and attractive assessment such that participation motivation and attention improves.
- Finding behavioural data otherwise remained unknown if tested only through questionnaires.

## 2. Related Work

### *2.1 Personality*

Personality is a set of characteristics that an individual possesses, determining how they behave, think, and feel (Roberts et al. 2014). Much research has been conducted in order to define personality, such as where it originates and how to categorise it.

With the continued research on personality, Gordon William Allport brought up the trait theory. The trait theory is the approach in which an individual's personality is able to be categorised and measured by multiple classifications or 'traits'. Allport and his partner, Odbert, listed 1800 adjectives in the English language to describe someone (Nicholson 1998), or notably their personality. However, 1800 adjectives were too many to measure, which is why this list was later reduced to 16 adjectives by Raymond Catell (1947). This was still criticised because these adjectives were unsuitable, as they were too statistically intercorrelated, partly due to their number. Therefore, it was later further reduced to five adjectives, in which the five traits were statistically distinct enough, as determined by Principal Component Analysis. These five adjectives were officially titled the "Big Five", which is now the most widely employed and reliable personality model.

In the Big Five model, and subsequent models, each trait is measured on a scale of 0% to 100%. This is often misunderstood, as many believe a trait to be completely categorial (therefore, you are only one trait or another) rather than on a spectrum.

While the Big Five came to be in the 1950s, HEXACO emerged in the 2000s (Hassan et al. 2023), which adds the Honesty/Humility trait on top of the Big Five traits. Much research points to this six-trait model being more thorough than the Big Five, as it removes intercorrelations between the traits and defines each trait more thoroughly. Nevertheless, the Big Five model is used in this study due to its extensive research.

#### *2.1.1 Big Five*

The Five Factor Model of Personality (FFM) was the next biggest milestone in personality trait theory (McCrae, Costa, and Martin 2005), relating directly to the Big Five. The FFM is considered the more reliable model of personality, and while it has limitations, it has been deemed stable, predictive, and cross-culturally valid (Johnson 2014). It has been developed in a way that proves that while each trait shows intercorrelation, they work independently from each other (DeYoung, Quilty, and Peterson 2007).

The Five-Factor Model of Personality Theory (FFM) from (Kabigting 2021) is demonstrated in Figure 1.

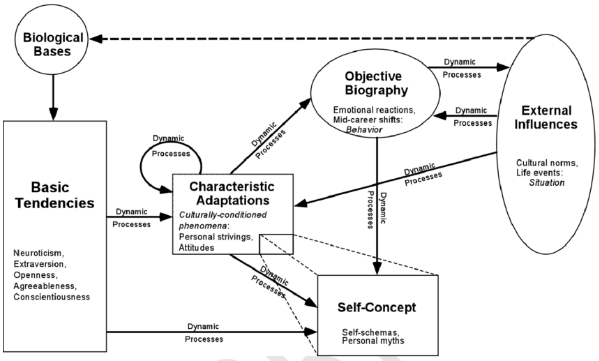

Biological Bases
Basic Tendencies
Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousness
Characteristic Adaptations
Culturally-conditioned phenomena: Personal strivings, Attitudes
Objective Biography
Emotional reactions, Mid-career shifts: Behavior
External Influences
Cultural norms, Life events: Situation
Self-Concept
Self-schemas, Personal myths
Dynamic Processes
Dynamic Processes
Dynamic Processes
Dynamic Processes
Dynamic Processes
Dynamic Processes
Dynamic Processes
Dynamic Processes
Dynamic Processes

Figure 1. Five-Factor Model of Personality Theory (FFM) (Kabigting 2021)

The five traits of the Big Five are as follows, as explained in (Kabigting 2021)'s review:

- Openness to Experience: the tolerance and appreciation of new ideas and experiences.
- Conscientiousness: the impulse of control, discipline, persistence and the natural inclination to organisation.
- Extraversion: the preferred personal interactions and involvement with others, and mood states in which a person is when they are with others.
- Agreeableness: the interpersonal orientation of trust and modesty.
- Neuroticism: the level of personal adjustment emotional instability, and hostility.

As mentioned earlier, each of these traits is measured on a scale of 0% to 100%, thereby signifying a movement from having a lack of a trait to having it altogether. For example, many believe that concerning extraversion, a person is either an extrovert or an introvert. However, according to the FFM, the definition of an introvert is a person who lacks the traits of an extroverted person.

#### *2.1.2 Hierarchy*

The Big Five model came through the effort of much personality research. Hundreds of words used to define and characterise personality were reduced to five, which had been statistically tested to be valid. There was less emphasis on deeply searching the different traits in past research, and instead, further emphasis on finding traits that worked best as personality. Therefore, the challenge now shifted to understanding more in depth these traits, which is why the five traits now require more precise definitions and sub-level terms, hence getting deeper in the hierarchical model.

Nowadays, researchers use a more hierarchical framework to define each trait more systematically, such as to find its correlations to disorders (KA et al. 2021). Soto (Soto and John 2017) revised McCrae's work on the Big Five personality traits model in his terms, which is the model used for this research.

Research has shown that there is a 25% increase in the predictive value of an individual's personality (Stewart et al. 2022) when evaluating from a facet-trait perspective (or, more effectively, a nuanced perspective). Hence, for a more precise representation of an individual's personality, it is essential to examine their traits (e.g., Conscientiousness), then delve into the respective facets (e.g., Orderliness), and further explore unique nuances (e.g., "Does not leave a mess behind") (Nielsen and Kajonius 2024). In addition, it is advisable to avoid domain-level (trait-level) questions when assessing, as they bring forth uncertainty and may often be interpreted differently by each individual (Kim, Di Domenico, and Connelly 2019). This hierarchical approach is more specific and, therefore, more accurate, enabling participants to relate to and respond to questions more effectively.

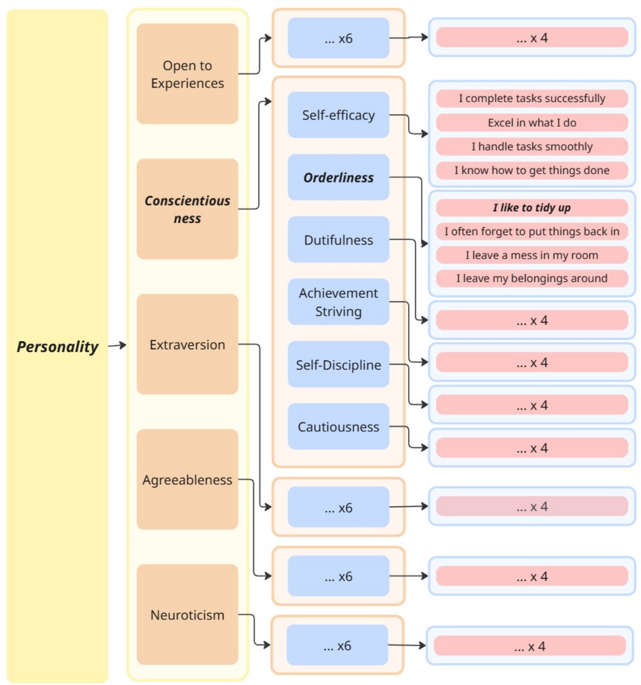


**Figure 2.** Personality hierarchy

Figure 2 shows the hierarchical approach that shows a detailed assessment of individual personality traits. It shows how personality is composed of five traits, one of which is conscientiousness. Each trait is divided into six facets. In the case of conscientiousness, one of the facets is orderliness. To go further, assessment methods differ in how they represent each facet of time items/nuances. In the case of the IPIP-NEO-120, each facet has four nuances/items. In the case of orderliness, the nuance/item "I like to tidy up" is one of them.

### 2.1.3 Conscientiousness

To narrow the scope of this research and enhance focus, the trait of "conscientiousness" was examined. It was assumed that using the conscientiousness trait would minimise participation bias during the experiment. When it came time to recruit participants through the researcher's socials, it was believed that individuals who tend to be more extroverted or agreeable (traits within personality) are more likely to agree to participate, hence the reason of avoiding those traits. Secondly, there is extensive literature on conscientiousness as a trait and its possible characteristics, making the study more thorough due to the availability of literature. Thirdly, conscientiousness as a trait is able to be explicitly gamified through various experimental designs, making it a suitable facet to measure through gamification.

The IPIP-NEO-120, the personality inventory from Soto (Soto and John 2017), is structured into six facets. Each facet has four items, each providing nuance and further understanding of the personality traits. Below is an explanation of each facet with its accompanying items (Johnson 2014).

- Self-efficacy: This facet examines an individual's confidence and belief in their ability to accomplish specific tasks and goals. It includes items such as completing tasks successfully and smoothly, excelling in their work, and knowing how to perform tasks.
- Orderliness: This facet examines an individual's tidiness and how organised they are, which also includes having habits of keeping a clean environment. It contains items such as enjoying tidying up, remembering to put things back into their proper place, and not leaving belongings or messes behind.
- Dutifulness: This facet examines an individual's sense of responsibility and obligation to tasks and commitments. It is also correlated to their morals and ethics. It includes items such as keeping promises, telling the truth, and obeying rules.
- Achievement Striving: This facet examines an individual's drive for success and work ethic. However, this does not have to be specifically related to jobs, such as setting high expectations for themselves. It includes items such as exceeding expectations, dedicating time and effort to work, and working diligently.
- Self-Discipline: This facet examines the individual's ability to control impulses, maintain focus, and persevere. It includes items such as being prepared, carrying out plans, being time-efficient, and having ease with starting tasks.
- Cautiousness: This facet looks at whether an individual thinks carefully through decisions, as well as having strategies for problem-solving and assessing risk. It includes

items such as not making rash decisions and not jumping to conclusions.

However, there are additional characteristics recognised in the literature concerning conscientiousness, namely self-control, perfectionism, tidiness, the ability to refrain from procrastination, task planning, and perseverance. Furthermore, related dimensions include delay of gratification, ego control, effortful control, impulsivity, constraint, and grit (Roberts et al. 2014), as well as risk-taking behaviour (Joseph and Zhang 2021). Grit and conscientiousness are notably interconnected, with grit associated with proactive conscientiousness and perseverance, akin to self-discipline (Schmidt et al. 2018). This relationship indicates that individuals higher in grit are more likely to persist on a task, especially when a reward is perceived as attainable (Dale et al. 2018).

Understanding these characteristics is especially important because they point to behaviours and patterns that individuals exhibit in direct correlation to conscientiousness. They allow us to understand the effects of conscientiousness better (as well as the other personality traits).

## 2.2 Questionnaires

There are multiple ways of measuring personality. The most common methods are through questionnaires, though there are different measurement types too.

Questionnaires, especially with the five-point Likert scale, are the most often used assessment tools to assess personality. Questionnaires and personality inventories like the NEO-PI-R and the IPIP-NEO-120 are well-adjusted and conform to the theory of personality (Big Five) very well. This is because it is structured in such a manner that participants go through every single item within a facet of a trait, which is precisely how the hierarchical personality model is based. Secondly, it is very easy to hand out this type of tool to participants (especially through online means nowadays), as it is structured, and determining your personality type is simply a matter of adding specific numbers together.

The NEO-PI-R (and its later version counterparts such as the NEO-PI-3) is the most current and widely used method for assessing an individual's personality based on FFM. It is known as the 'golden standard' (Kabigting 2021) of personality tests. It is a 5-point Likert scale questionnaire with approximately 300 questions.

IPIP-NEO-120 Most of the questions of the NEO-PI-R questionnaire are domain-level rather than facet-specific, making them vague and relatively weak, often lacking specificity (Seeboth and Mõttus 2018), though later revisions implemented these improvements. The IPIP-NEO-120 is an improved and publicly available instrument version of the NEO-PI-R (Johnson 2014). The IPIP-NEO-120 adopts the Big Five framework, where each trait comprises six facets, and each facet is represented by five items that capture nuances.

This research uses the model that the questionnaire IPIP-NEO-120 is based on, in which the facets and nuances were detailed earlier in the "Conscientiousness" subsection.

### 2.2.1 Critiques on questionnaires

Loss of nuance There are limitations to questionnaire-type assessment tools, despite this format being most widely used. For example, many psychological constructs cannot be measured directly in these linear ways because they tend to be more complex and nuanced. Furthermore, in psychology, it is often best to use indirect measurements to gain a broader perspective (Wagner et al., n.d.). Nevertheless, it is possible to capture the bigger picture by assessing more specific items, just as is done with the IPIP-NEO-120.

Furthermore, the use of closed-ended questions and Likert scales may oversimplify concepts and fail to capture specific nuances, as participants are limited to selecting preset options. Due to these presets, responses may unintentionally lead individuals to conform to specific norms, whereas they may have thought of something entirely differently otherwise (Hansen and Świderska 2024).

Quiz fatigue In lengthy questionnaires, participants often lose interest and become careless in their responses (Welz and Alfons 2025). Their carelessness translates to sloppiness in how they answer, resulting in datasets that create unclear conclusions and overall less reliable conclusions.

Differences in context & interpretation Other limitations include how different participants interpret statements differently in Likert scale questionnaires. Therefore, how a participant interprets a question may be different from how another participant interprets it. This brings forth uncertainty in answers, and analysing results becomes more vague too.

Furthermore, people are susceptible to self-reporting, which includes an attempt to interpret themself. For example, if a participant is tasked to rank themselves on how 'careless' they are, this could be interpreted in different manners, such as: they are 'very careless sometimes' or 'somewhat careless all the time' (Zhang, Wing-Yee Tse, and Savalei 2019). These differences in interpretation cloud the data.

This is further highlighted in the contradictions participants unknowingly adhere to. For example, in a separate study, there was an experiment where homeless individuals rated their level of tiredness on the Likert scales compared to their qualitative responses. It showed that the Likert scale and qualitative response were disproportionate. This can be attributed to their interpretations of the scale in the Likert scales, as they can be misleading, and context is often lost in Likert scales. Additionally, habits normalise situations, which in turn alter how specific individuals perceive situations compared to others (Ogden and Lo 2012).

### 2.2.2 Self-reporting

It is also important to note that all of these inventory items reflect how participants view themselves (Soto and John 2017). Self-reporting examines how individuals view themselves rather than how they perform in activities. It is observed that self-assessments correlate weakly with actual performances (also personality behaviours) across various domains and that many individuals overestimate their abilities. Oftentimes, these biases are unconscious mechanisms, and having perfect self-knowledge is not achievable (Karpen, n.d.).

There is little difference between self-report and stranger analysis (where an external observer assesses an individual's personality), except for a slight difference in openness and intellect (Kim, Di Domenico, and Connelly 2019). In (Kreitchmann et al. 2019), they tested social desirability response (where a participant answers in a socially acceptable manner, despite what they genuinely think) and acquiescence (where they agree with statements regardless of the contents). They observe a high correlation between the social desirability bias and its limited predictive value for real-world outcomes, particularly with Likert-scale questionnaires. However, despite this, self-reports and stranger-based reports do not show much difference, as also mentioned by Kim (Kim, Di Domenico, and Connelly 2019). Nevertheless, Likert questionnaires are still not the perfect control method to eliminate self-reporting.

Furthermore, emotional states alter how participants rank on the IPIP-NEO-120, with notable changes occurring when a participant is induced to feel sad. In such cases, neuroticism increases, while extroversion and agreeableness decrease moderately (Querengässer and Schindler 2014). That is to say, a participant's self-assessment is not always be as accurate as it could be.

One of the primary limitations of questionnaires that this research focuses on is the participant's self-reflection. Participants' self-interpretation is not accurate as to who they truly are. They answer what they believe themselves to be, while in actuality, that might not be accurate to what they are.

## 2.3 Alternative measurements

### 2.3.1 Open-ended questions

Open-ended questions are well-suited for spontaneous reasoning and yield rich and nuanced responses when there are numerous answers (Connor Desai and Reimers 2019). They prove to be beneficial in allowing participants to express and write their opinions freely (Baburajan, Abreu e Silva, and Pereira 2022) (Hansen and Świderska 2024). In contrast, with questionnaires or Likert scale assessments, the participant is already confined to the examiner's pre-set answers and is therefore limited to expressing themselves through simple means. With open-ended answers, participants are allowed to share their reasoning behind their statements.

### 2.3.2 LLMs to assess personality through text

Multiple research studies have shown that LLMs accurately estimate personality traits based on written texts (Bubaš, n.d.), despite their confidence not being high (Piastra and Catellani 2025). When AI-generated analysis was compared to experts' analysis, a high correlation was found between them, pointing out how AI-generated analysis is capable of replacing experts' analysis(Oeljeklaus, Höft, and Danner 2025). Furthermore, it is observed that ChatGPT's performance is similar to that of humans', although with typically more positivity biases and greater sensitivity to the actual prompts given (Derner et al. 2024). While LLMs help assess personality, it is noted that using personas, providing better prompts, and correctly classifying questions result in a more accurate analysis of personality traits (Olea et al., n.d.).

Quiz fatigue While quiz fatigue appears during questionnaires, it appears more extensively in open-ended questionnaires. Participants become more tired when answering open-ended questions, as they are more time-consuming and require more thought (Baburajan, Abreu e Silva, and Pereira 2022). This is also seen with a higher amount of drop-outs in open-ended question sections rather than in close-ended sections (Connor Desai and Reimers 2019). Furthermore, open-ended questions are more challenging for the examiner to analyse. With both the time for the participants to answer and for the examiner to analyse, open-ended questions take much longer compared to closed-ended questions.

### 2.3.3 Game-Based Assessment

There has been extensive research on serious games and their positive effects. Serious gaming refers to games not intended for entertainment purposes, and are therefore used as educational tools or research aids. Their positive impact includes enhancing specific learning outcomes and allowing certain topics to be learned more easily (Facchino et al. 2025). They also demonstrate greater immersion in understanding particular issues, which makes learning a less mentally demanding task (Westera, n.d.).

A part of serious games includes Game-Based Assessments (GBA), which examines the research-oriented aspect of serious games. These types of assessment tools examine how individuals react within the actual serious games, as well as the scores they achieve in these games that specifically relate to their research topic. It has been observed that game assessments are efficient in engaging participants, reducing specific anxieties, and providing more authentic assessments, as they tend to be more immersed in this type of assessment (Su and Zou 2024) (Noroozi, Dehghanzadeh, and Talaee 2020). This also notably reduces self-reporting bias, as participants link GBA less to actual assessments and more to a fun experience, allowing their authentic behaviour to shine through with further ease (Santos et al., n.d.). Some researchers say that GBA has the potential to become an alternative to personality questionnaires (Ramos-Villagrasa et al. 2024).

According to current research on personality, specific characteristics and behaviours are found to relate to certain facets of conscientiousness. Regarding GBA, it was seen that personality facets can be indirectly measured through these behaviours in games. For example, the facets of achievement striving, dutifulness, and self-discipline relate closely to the characteristic of attempting to succeed in a specific task (Roberts et al. 2014), which can be translated to succeeding in games. Similarly, quick reaction time is hypothesised to relate to the subfacet of self-efficacy and achievement striving, as those individuals understand how to handle a situation more effectively. Then, on the contrary, slower reaction times may reflect individuals who think for a longer amount of time and consider risks, as well as being more patient. This correlates to constraint, which is a sign of conscientiousness, specifically in the the self-discipline and cautiousness facets (Roberts et al. 2014). Furthermore, the facet self-discipline shows characteristics of patience. Therefore, a more patient person has more self-discipline, which may translate into them writing a larger quantity of words.

## 3. Research Statement

This research aims to address the knowledge gap in how personality assessments are conducted by exploring methods beyond traditional questionnaires and evaluating the validity of these alternative methods in comparison.

To analyse this, multiple research questions and hypotheses were formulated in order to quantify and test how open-ended questions and mini-games perform in assessing personality traits compared to questionnaires.

The main research question is as follows:

> **Main RQ:** Would a combination of mini-games and open-ended questions be a valid and attractive alternative to the IPIP-NEO-120 for assessing conscientiousness?

To further explore this central question, multiple sub-questions were developed, each accompanied by several related hypotheses.

> **Sub-RQ 1:** How do measures obtained for playing mini-games compare to the conscientiousness level assessed using traditional testing?

- H1.1: Higher mini-game success rates (sorting, maze, spin wheel) will correlate positively with conscientiousness, especially on achievement striving, dutifulness, and self-discipline.
- H1.2: Faster reaction times will correlate positively with high self-efficacy and achievement striving
- H1.3: Slower reaction times will correlate positively with self-discipline and cautiousness

> **Sub-RQ 2:** How do writing style and content from open-ended question answers compare to the conscientiousness level as assessed using traditional testing?

- H2.1: Richer and thoughtful content correlates with conscientiousness, especially with dutifulness.
- H2.2: Structured and formatted writing style correlates with conscientiousness, especially with orderliness.
- H2.3: Participants with lower conscientiousness (especially low self-discipline) will write less (word count and quality) with each open-ended question due to frustration and fatigue.

> **Sub-RQ 3:** How do participants evaluate taking the alternative compared to traditional measurements?

- H3.1: They prefer the mini-games.
- H3.2: They do not prefer the open-ended questions
- H3.3: They were more immersed in the alternative
- H3.4: The alternative will produce less self-reporting

> **Sub-RQ 4:** Can the alternative measures further our understanding of the construct of conscientiousness? (+ In what ways are mini-games and open-ended questions measurable?)

- H4.1: Each mini-game correlates with its corresponding conscientiousness facet
- H4.2: Certain writing style aspects correlate with conscientiousness facets
- H4.3: Certain content aspects correlate with conscientiousness facets
- H4.4: Certain words (terminology) will correlate with conscientiousness facets

## 4. Tool

This section goes over the tool that was constructed to evaluate and measure the hypotheses and research questions. This was done to determine if the tool has the potential to be a possible alternative to questionnaires (the traditional method for assessing personality).

To tackle the assessment of gamified assessment, two different methods were designed: open-ended questions and minigames. How these two parts were chosen, designed and analysed is explained here.

### *4.1 Open-ended questions*

#### *4.1.1 Designing the questions*

Three open-ended questions were constructed for this section, where each targeted one of the facets of conscientiousness. The facets that are being specifically targeted are self-efficacy, dutifulness, and achievement striving. These facets were specifically chosen to compensate for the designed minigames (explained in the next section), where they were designed with the three other conscientiousness facets in mind.

Instead of asking the participant what they believe themselves to be in those specific facets, more situational questions were asked to initiate a more role-playing atmosphere, as this gamified aspect has been shown to have a better engagement and immersion factor (Santos et al., n.d.). The participant is then asked to consider how they would respond and write that down.

The open-ended questions in question are:

- (Self-efficacy) You notice that money was taken from your bank account an hour ago from an unknown address. You are writing an email to your bank, but your computer goes black and shuts off. What is your reaction, how do you feel, and what will you do next?
- (Dutifulness) On your way home, a stranger on the street gives you a pamphlet to donate money to animals (pets) in need. You are in a hurry to get home, so you tell them you will, and they thank you. You arrive home, and you are holding onto this pamphlet. What will you do?

- (Achievement striving) You tell yourself that you should go for a walk during the afternoon. You step outside, and you notice it is cloudy. The weather app says it will drizzle lightly in an hour. What do you do? How do you spend your afternoon?

Each question was designed and written with the facet of conscientiousness it referred to in mind. Therefore, the researcher brainstormed various questions, and the ones included in the experiment are those below (Roberts et al. 2014).

- Self-efficacy relates to being able to handle situations on your own, so the open question focused on a problem where the participant had to react in a certain manner.
- Dutifulness relates to more ethical or moral implications, such as people keeping promises. Therefore, this situation looked at how the participant responds when promising to do a small task, and what they would actually do when alone.
- The achievement striving facet looks at how likely someone is to continue with a task they set for themselves despite possible inconveniences. That is why the goal of going for a walk, despite the potential for rain, was chosen. It represents a task that many individuals could set for themselves, but might also postpone or delay.

### 4.1.2 LLM Analysis of answers

The open-ended question answers were analysed through an LLM, notably Claude.AI, due to its transparency in how it trains its system and the possibility to remove user inputs. The prompt given to Claude.AI, also shown in Appendix 1, consists of certain criteria that the LLM ranks the answer on. There were three different sections in which the LLM had to analyse: the answer's contents, writing style, and terminology. According to what the LLM was ranking, it assigns a specific characteristic a score out of ten to see how closely the participant's answer encompassed that specific, defined characteristic. Characteristics of conscientiousness were defined and added to the prompt to feed into the LLM, such that the LLM better understands what to look for and rank the answers. Characteristics include the use of more work-related words and individuals who have a work ethic and are more achievement-oriented, as those inclined to be more conscientious (Hirsh and Peterson 2009).

For the contents, it was asked to analyse the answer's level of goals, responsibilities, emotional experiences, reflection, objective description and technical language, as well as the six facets of conscientiousness. The characteristics of the answer's contents were explicitly chosen by literature (Ragsdale et al., n.d.) (Troiano, Oberländer, and Klinger 2022) that listed which characteristics were within a piece of writing's contents.

Similarly, the same was done for the writing style, such that the complexity of words chosen, the sentimental analysis, grammar and punctuation, style and tone (Argamon et al. 2007) (Verma and Srinivasan 2019), as well as structural complexity, were analysed and ranked.

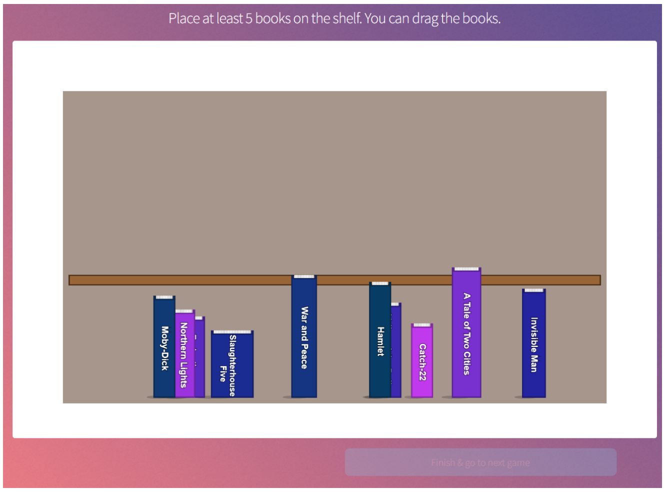
Lastly, for the terminology section, only the six facets of conscientiousness were listed, along with words related to each facet. For each of these characteristics (content, writing style, and terminology), the LLM ranked from 0 to 10 on how much a specific characteristic seemed apparent in the text.

For example, in the writing style 'grammar and punctuation', if there were no grammar mistakes and the writing was done very well, the LLM is tasked to rate that characteristic as a '10'. With these numbers, it was possible to analyse whether an LLM is able to analyse a piece of text and if that has the potential to be used to assess personality.

## 4.2 Minigames

In this section, three different minigames were designed and developed, where they also represented one of the facets of conscientiousness. For a complete list of all data points collected in all minigames, along with their specifics, please refer to the codebook in the OSF directory (Appendix 1).

### 4.2.1 Book minigame



Place at least 5 books on the shelf. You can drag the books.

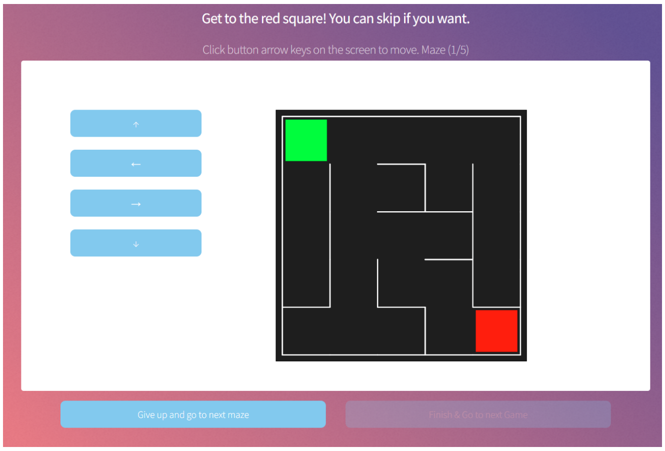
**Figure 3.** Book Minigame

The first game was a sorting books minigame (see figure 3), where the intended conscientiousness facet was orderliness. Orderliness is defined as the tendency to be organised and logical (McCrae, Costa, and Martin 2005), as well as a preference for a structured environment (Roberts et al. 2014). In this specific minigame, the participant has a set of books in the lower compartment that they are asked to set atop a shelf, but only for a few of the books that exist. For example, if six books existed, the participant is asked to place at least three of them. It is completely up to them which books they wish to set on top, and if they want to follow a specific order.

The data collected from the book minigame include the total time to complete the task, timestamps of mouse clicks, the order of books picked up, the final order of books set, and a list of all times the books were ordered.

### 4.2.2 Maze minigame

The maze minigame's intended conscientiousness facet was achievement striving. The idea and design of this maze (see



Get to the red square! You can skip if you want.

Click button arrow keys on the screen to move. Maze (1/5)

**Figure 4.** Maze Minigame setup

figure 4) came from an experiment where the participant was asked to play a game to guess a specific word from a set of other words. However, the twist was that they will always get the answer wrong. This experiment was designed to test perseverance (Määttänen et al. 2021), especially in situations where a solution is impossible to find. Since achievement striving is related to persistence in complex tasks, setting high standards (McCrae, Costa, and Martin 2005), and perseverance in the face of difficulty is a key behavioural trait of conscientious individuals (Duckworth et al. 2007), a minigame was designed with that in mind.

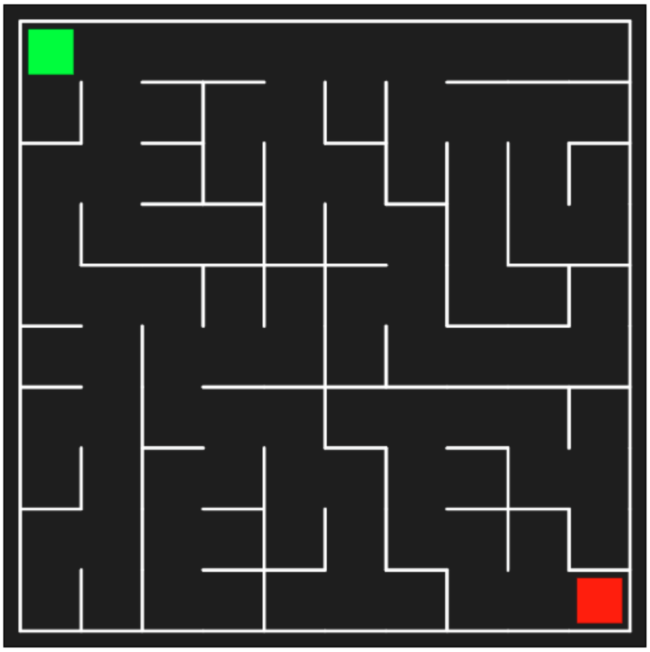
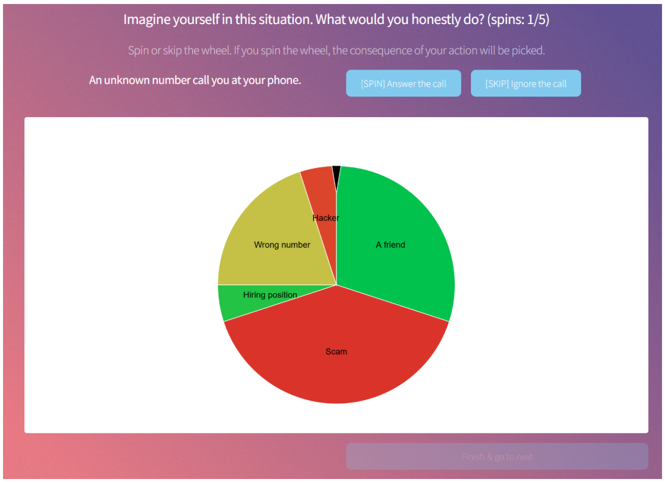


**Figure 5.** Impossible Maze

To align with this literature, the possible and impossible maze was designed. The participant was presented with a maze to complete, starting simply at first. There are five mazes to complete, which become more complex as they increase in size. The third and fifth mazes were intentionally designed to be impossible to solve, which tests the participant's perseverance (see figure 5). The complexity of the maze itself is used to assess whether the participant is willing to undertake harder and more challenging puzzles.

The maze minigame collected the following information: time to completion, the path the participant took to complete the maze, all key press timestamps, and the timestamps when the participant clicked the exit and enter buttons.

### 4.2.3 Spin-the-wheel minigame



Imagine yourself in this situation. What would you honestly do? (spins: 1/5)

Spin or skip the wheel. If you spin the wheel, the consequence of your action will be picked.

An unknown number call you at your phone.

**Figure 6.** Spin the chance wheel minigame

The last minigame was spinning the chance wheel (see figure 6), where the intended facet was cautiousness. There are specific characteristics associated with cautiousness, and risk-taking is one of them, as it influences decision-making. Cautiousness is the facet most associated with careful decision-making (McCrae, Costa, and Martin 2005). Highly conscientious individuals tend to exhibit greater caution and consider consequences more thoroughly (Johnston et al. 2008). Therefore, the spin-the-wheel minigame presented a specific scenario in which a wheel displayed various possible outcomes. Any participant decides to spin the wheel and take a chance on something happening, or skip a scenario entirely. Unknown to them, however, is that the answer is already predetermined, making it easier to find any correlations in a more controlled environment.

The spin-the-wheel minigame collected the following information: total completion time and the click, hover, and enter times for both the spin and skip buttons.c

## 5. Method

The personality facets and terms used in this research are based on (Johnson 2014)'s research

### 5.1 Instrument

An online experiment was developed for this research, which took place on Qualtrics and a website built (the tool) by the researcher. The website is hosted online, and participant data is also stored in the associated database, in which the data is not be linked back to individuals.

### 5.2 Procedure

This 'Procedure' section gives an overview of what the participant goes through during the experiment's entirety and therefore does not go deeper into the details. A more thorough explanation of the experiment is detailed in the following subsection: 'Experiment Design'.

It is divided into three parts, including a pre-experiment and a post-experiment. As a whole, the participant go through the following: pre-experiment, the open-ended questions, the minigames, the traditional assessment, and lastly the post-experiment.

**Pre-experiment** Before the experiment, potential participants are approached through different online social media of the researcher. If an individual accepts the invitation, they are brought into the first Qualtrics webpage. This is where the experiment officially begins. Here, the participant is informed about the experiment and its concerning research. They are also asked to do the experiment alone, as outside distractions influences the experiment. They are asked for their consent to the experiment, such that the data collected is used for analysis. Demographic information is requested, which includes their age, English proficiency, computer science skills, and whether they have auto-correct enabled. Once they have entered all the necessary information, they are directed to the first part of the experiment.

**Part 1: Open-ended questions** The first part of the experiment consists of answering three open-ended questions. The participant is presented with each question at a time, where each question pertains to everyday life situations (e.g., do they clean their kitchen when they come home at 6 PM?) that they may encounter, to which they are asked how they would react and respond. They then write their answer in the provided text box, and once finished, can click the button to move to the next question or section.

**Part 2: Minigames** The participant is instructed to go to the separate website. There, they are given an ID code to enter into Qualtrics, which links their data from the minigames to Qualtrics. Once they have inputted the ID Code, they can continue with the experiment.

The contents of each minigame section are first explained, and then the structure of each of them is detailed.

- Sorting books. The participant will have a number of books to put on a shelf. They have a minimum number of books to put in. They can freely choose if they wish to sort the books.
- Completing a maze. The participant will have to be presented with a maze to complete. They have 5 of them to complete, each becoming slightly more complex. A chosen number of them will be impossible to solve, to test their perseverance. They may skip a maze.
- Spinning a chance wheel. The wheel will show how much money they could win or lose. They will have five wheels to spin, each becoming riskier. This answer is already set (unknown to them). They may skip.

The structure of the minigames is as follows: sorting books (Level 1), going through the maze (Level 1 through 5), sorting books (Level 2), spinning the wheel (Level 1 through 5), and sorting books (Level 3). Each level increases in difficulty. Therefore, for the 'sorting the books' minigame, the first level has a minimum of three books to sort, while the third level has a minimum of eight. The maze minigame also increases in difficulty, where the maze becomes larger with each level. For example, the level of the maze is a 5-by-5 grid, while the last maze is a 25-by-25 grid. Furthermore, the third and fifth levels of the maze minigame are impossible to solve. Lastly, spinning the chance wheel minigame does not necessarily increase in difficulty, but it does conceptually propose riskier situations in which the participant must decide.

After completing all minigames, they are redirected back to Qualtrics. They will continue and finish the experiment there.

**Part 3: Traditional assessment** This last part of the experiment is the traditional assessment. In this section, the participant is given 32 statements about conscientiousness, in which they need to respond using a 5-point Likert scale to indicate how much they agree or disagree with each statement. With those 32 statements, two additional statements are asked to test the participant's focus. For example, this could be to check the right-most option. Once the participant completes all the questions, they are directed to the post-experiment.

**Post-experiment** With the use of Qualtric's services[c], it was possible to see if the participant has failed the one or both of the focus questions in the previous section. If they failed the focus questions (meaning they did not select the correct box), then they were asked to rate the amount of attention they had given to each section of the experiment, on a scale of zero to ten. After this, they continued onto the next part of this section.

The participant evaluation comes next. The participant is asked, on a scale from zero to ten, if they enjoyed the experiment, were immersed in it, and believed they self-reported. They asked all of this on a scale from zero to ten. After this is all completed, they have officially finished the experiment.

### 5.3 Experiment design

This section, experimental design, looks further into the design of the experiment, and includes the reasoning and literature that helped conceptualise it. It also provides references to the

[c] Qualtrics, the online survey platform used in this study provided by Leiden University, has multiple functionalities.

appendix (Appendix 1) and specific examples of how certain sections were formulated.

#### 5.3.1 *Pre-experiment*

The participant is informed before the experiment whether they consent to participate and are given the minimum necessary information about the experiment itself. Nothing is revealed about how it is assessed explicitly for personality, nor what type of data is being recorded regarding the minigames and how the open-ended questions are utilised. The reason for not telling them is that letting them know will make them aware that this data is specifically collected, and, therefore, it may change their behaviour and results. This deception is therefore necessary. They are completely disclosed in the post-experiment, however.

This section also receives the participant's demographic information, such as their age, English proficiency, their skills in using a computer, and whether they have a grammar check enabled. The reason this is used is to potentially exclude participants who are under 21, as their personality is still developing and more volatile at that stage (Atherton et al. 2022). English proficiency is required to understand some of the questions that are asked. Since minigames on a computer are a part of the experiment, participants need to have at least a minimal amount of computer skills to use the mouse and keyboard. The reason the grammar check is requested is to provide a control check on the open-ended questions, as grammar is one of the characteristics of the analysis check.

#### 5.3.2 *Part 1: Open-ended questions*

The three open-ended questions presented here aim to encapsulate the facets of self-efficacy, dutifulness, and achievement striving. They were placed one after another, each with a text box underneath for the participant to write into. It was decided that the open-ended questions are grouped in sequence rather than intertwined with the minigames, to increase the overall ease of use. This way, the participant does not need to switch between window tabs.

#### 5.3.3 *Part 2: Minigames*

With the three minigames designed, they had to be organised and structured accordingly. The book minigame had three levels, and the maze and spin-the-wheel minigames had five. The reason for the difference in levels is that the book minigame could technically take a very long time to complete (and also take the shortest amount of time). To balance this (and to reduce the possibility of the participant dropping out), it was reduced to three levels.

To keep the entire minigame section structure less exhaustive for the participant, it was structured in such a manner that the first level of the book minigame happens, then the five levels of the maze, then the second level of the books, then the five levels of the spin minigame, and only then the last level of the books. Switching between the minigames has shown that it makes for a more enjoyable and less exhaustive experience. Lastly, another reason for keeping the maze and spin levels together is that having each level follow the next relates to the actual facets, unlike the book ordering minigame. The maze minigame relies on the sequence of possible and impossible mazes, which are presented one after the other, to bring the participant off guard when they encounter the impossible maze, to push them into thinking whether they have indeed interpreted correctly that the maze was impossible or not. Similarly, the spin-the-wheel minigame involves risk-taking and comparing risks from one spin to another. Having each level occur subsequently helps the participant compare each level more easily.

#### 5.3.4 *Part 3: Traditional Assessment*

The third part of the experiment involved a traditional assessment, specifically the IPIP-NEO 120 survey for conscientiousness (Johnson 2014). Here, the participant is asked 32 questions concerning conscientiousness, which uses a 5-point Likert scale. After the participant has finished, the results are accumulated to show scores for each facet of conscientiousness, as well as the overall score for conscientiousness. This is the result that is compared to the open-ended questions and the minigames results.

#### 5.3.5 *Post-experiment*

Here, the participant is asked if they enjoyed the experiment, how their immersion was, and if they experienced any self-reporting. This is particularly important because it is essential to understand the participant's perspective on the experiment, as one of the motivations behind this study is to determine if there is a more enjoyable and fulfilling way to assess personality. Additionally, a few focus questions were asked after the participant evaluation to determine if they were paying attention or not. Some specific individuals did not pay attention, and later, when they confirmed they had not paid attention, they were excluded from the results so as not to bring further outliers to the analysis.

### 5.4 *Variables & Measurements*

The data collected from the open-ended questions came directly from the answers participants provided. However, the data was then transformed using the LLM, such that their answers (strings of words) are translated into graded characteristics interpreted by the model (numbers). With the data now in numerical form, it becomes much easier to find correlations between the open-ended question answers and the grades from the traditional assessments.

The data collected in the minigames consists of information that participants are not aware that is being recorded, and they were debriefed about this after the experiment. This data includes the number of times they hovered over or clicked any button shown in the minigames, as well as the timestamps of these actions. If they used a mouse, it also includes the number of mouse clicks. In addition, minigame-specific datasets were collected, such as which books participants clicked on and released in the book minigame, and the path they took in the maze minigame. All recorded data is found in the Appendix 1.

The gathered raw data of the minigames was transformed and manipulated to ensure its usability. For example, the data on the number of mouse clicks in a minigame was stored in a long string of timestamps (e.g., 100, 128, 193, 329), and this data is run through code to calculate the precise number of clicks that have occurred. Furthermore, some of these results were used to create scores for the minigames. For example, in the sorting book minigames, 'optimal orders' were compared against the actual order of books the participant made, therefore creating a score.

Lastly, raw data were also collected from the traditional assessment, namely the IPIP-NEO item inventory. Using the Likert scale, the questions that pertained to certain facets were either added to or subtracted from the total sum of that facet's score (as some items are phrased in negation). This then provided a total score for each individual's standing on each conscientiousness facet, as well as on conscientiousness as a whole.

Excel and Python were used to handle and transform the raw data, such that it is used as a variable to understand if there was any correlation between these numbers and the traditional assessment.

### *5.5 Data analysis*

Data was first prepared and formatted properly in Excel so that the sheets are easily accessed within a self-built local Python program (Appendix 1). After the the raw data was properly formatted, statistical analysis began.

Before examining any correlations, a pre-analysis was conducted. Two different tests were conducted for the pre-analysis: the Shapiro-Wilk test for assessing normality and examining z-scores for outliers. The Shapiro-Wilk test examined the p-values and was compared with the standard significance level of 0.05. Then, the z-score test was conducted to identify normal outliers and extreme outliers, using thresholds of two standard deviations and three standard deviations, respectively. These variables were all tested using the data gathered from the open-ended questions and the minigames.

Then there was the choice of selecting between two different correlation tests to measure the strength of the relationship between the variables and the facets of conscientiousness. This is where the variables from the open-ended questions and minigames were tested against the scores retrieved from the traditional assessment (IPIP-NEO-120 survey). Since Spearman does not assume a normal distribution and is less susceptible to outliers, Spearman correlation was used when the pre-analysis revealed that the variable was not normally distributed or had a few outliers (namely, more than five moderate outliers and more than two extreme outliers). If the variable was normally distributed and had few outliers, then Pearson's method was used.

The correlation tests (Pearson or Spearman) examined the p-values and were compared against two different significance levels: 0.05 for any value tested against simply conscientiousness, and 0.083 for all the other sub-facets. This is because, since there are six different sub-facets, the standard 0.05 p-value was divided by 6 for the Bonferroni correction. If the p-values of these correlation tests were below these levels, then it meant that the correlation was significant. For those values specifically, the coefficient of the correlation test was examined to determine its magnitude and direction. This was to identify a strong or weak correlation and to determine whether it was positive or negative.

Lastly, to analyse the participant evaluations, paired sample t-tests were used with the Bonferroni correction. This was done to determine whether the calculated p-values indicated any significant correlations rather than random noise.

The following section shows the main results and the analysis that was conducted. To see all of the results, please see the appendix (Appendix 1).

## 6. Results

### *6.1 Demographics*

The research aimed to recruit 50 participants, which is the minimum required to conduct a conclusive evaluation using statistical analysis on the research questions and hypotheses. This number was calculated using G*Power. During the experiment's development, a pilot version was tested on four different participants to refine the experiment. Their data was still utilised. Over 70 participants took part in the experiment; however, data from only 51 of them were used and analysed for this research. This was because, for more than 10 participants, the data was not used, either because they had not completed the experiment (the minigames or the IPIP-NEO questionnaire at the end) or because they indicated that they had not paid attention during the focus checks towards the end of the experiment.

Participants were recruited at random. The online experiment was distributed through the researcher's personal connections and social media platforms, which were promoted by the researcher, including WhatsApp, Instagram, and LinkedIn. No participant was selected based on a specific characteristic. However, there was an inclination towards individuals who spoke English and were aged 21 or older[d]. This was done by reaching out to individuals in that age bracket. Furthermore, their age was requested during the pre-experiment, to potentially remove their data from the analysis. In the end, data from 51 participants was used after taking out invalid participants due to age or incomplete data.

### *6.2 Measurement details*

Conscientiousness was measured through minigames and open-ended questions. For facets (e.g., cautiousness), the significance of the correlation was calculated with a p-value of below 0.0083 (0.05/6) because of the six facets of conscientiousness, while the trait of conscientiousness was checked for a p-value of below 0.05.

The different sections of the experiment were analysed separately. Therefore, the variables of the different minigames

[d] research shows that younger individuals change their personality more easily before the age of 21 (Atherton et al. 2022)

were first analysed, then the open-ended questions were analysed, and finally, the values of what the participants thought about the experiment were examined.

### *6.3 Minigames*

Conscientiousness was measured through the minigames: book sorting, impossible and possible mazes, and spinning a chance wheel. Specifically, the minigames were measure through scores, reaction times, and overall behaviours.

#### *6.3.1 Book sorting*

In the minigame 'book sorting', the variables score, order, amount ratio, total time, clicks, average drag duration, and number of books picked up were analysed. All the significant correlations for this minigame is found in Table 1, where the non-correlations were omitted. The omitted and insignificant correlations include the other levels and facets not mentioned in Table 1, as well as the variables total time, number of clicks, aberration duration, and number of books picked up.

**Table 1.** Book sorting minigame significant correlations results

| Variable | Levels | Conscien.[a] | p-value | coef. |
|---|---|---|---|---|
| Score ha[b] | all | Conscien. | 0.0354 | 0.1954 |
| Score ha | all | cautious. | 0.0035 | 0.2646 |
| Score hd[c] | all | cautious. | 0.0021 | -0.2782 |
| Amount ratio | all | discipline | 0.0006 | -0.3085 |
| Amount ratio | 1 | discipline | 0.0045 | -0.4397 |
| Most picked book[d] | all | Conscien. | 0.0197 | -0.3674 |
| Unchanged order[e] | all | Conscien. | 0.0471 | -0.3197 |
| Amount times last order | 1 | self-efficacy | 0.0052 | -0.4332 |

[a] Conscientiousness
[b] Height ascending
[c] Height descending
[d] Amount most picked up book
[e] Amount unchanging order

The score of the minigame was a ratio comparing how the participants sorted the books compared to six different optimal manners in which the books could have been sorted. The books had three different characteristics: height, colour, and title, which means that they could have been sorted according to these three characteristics in either an ascending or descending manner. Therefore, six different scores were calculated according to these six different ways that the books could have been sorted.

The analysis shows that height ascending exhibits a significant positive correlation with the overall trait of conscientiousness (coef = 0.2), with a greater magnitude for the facet of cautiousness (coef = 0.26). In the same vein, the books sorted in a height-descending manner showed a significant negative correlation with cautiousness (coef = -0.28). No other score order had any significant correlation, nor did any other level in the book sorting minigame, nor any other facets.

There was a significant correlation between the number of books set, calculated by ratio of the maximum number of books set and the number the participant had set. Namely, throughout the entire book minigame, there was a significant negative correlation with self-discipline (coef = -0.31), especially in the first level (coef = -0.44). In contrast, the other two levels did not show any significant correlations.

No significant correlations were found for the book minigame's completion time, as well as for the number of mouse clicks and the average drag duration.

Concerning behaviours related to the number of times specific books were picked up, there are a few significant correlations. The maximum number of times the same book was picked up shows a significant negative correlation (coef = -0.37) with the overall conscientiousness trait across all levels. However, there was no significant correlation at the individual level, only for the minigame as a whole. Similarly, the number of times the same order of books was set also shows a significant negative correlation (coef = -0.32) with the overall conscientiousness trait over the whole minigame. This refers to the participant picking up a book to put it in the same place in terms of order, but not necessarily in the same distance. Furthermore, the number of times the last order of books that has been sorted in the same way shows a significant negative correlation (coef = -0.43) specifically with the self-efficacy facet at the first level, but not overall. Finally, no significant correlations were found between the overall number of books picked up and the other variables.

#### *6.3.2 Maze*

All the significant correlations for this minigame can be found in Table 2, where the non-correlations were omitted. The omitted and insignificant correlations include the other levels and facets not mentioned in Table 2, as well as the score, length ratio, total time, average hover time and last hover duration variables.

**Table 2.** Maze minigame significant correlations results

| Variable | Levels | Conscien.[a] | p-value | coef. |
|---|---|---|---|---|
| Amount key press | all imp[b] | Conscien. | 0.0487 | -0.2171 |
| Amount key press | 5 | Conscien. | 0.0341 | -0.3278 |
| Last press time | 5 | Conscien. | 0.0403 | -0.3177 |
| Amount hovers | 5 | discipline | 0.0073 | -0.4079 |

[a] Conscientiousness
[b] All impossible mazes

Conscientiousness was measured in the minigame 'maze' through score, length ratio, total time, number of key presses, and hovers.

No significant correlations were found in the maze score, which relates to how closely the participant's path aligned with the optimal path towards the final square, the ratio of steps taken to the minimum possible towards the maze exit, and the total completion time. Therefore, no correlations were found in any individual levels, nor the minigame as a whole.

Significant correlations were found in the number of key presses the participants took. Interestingly, no significant correlations were found when examining the entire maze minigame as a whole. However, significant correlations were specifically identified within the impossible mazes. Thus, significant negative correlations were found in the overall conscientiousness trait (coef = -0.22) for all impossible mazes, especially in the last (impossible) maze (coef = -0.33).

Secondly, the time it took participants to press the last key to move in the maze and then click the button to finish the level showed a significant correlation. Specifically, there was a significant negative correlation (coef = -0.32) with the last impossible maze to the conscientiousness trait, but not with any other individual levels or levels as a whole.

There were significant correlations found with the number of hovers above the 'go to next minigame' button. Specifically, on the last impossible maze, there was a significant negative correlation (coef = -0.41) to the facet discipline.

Lastly, no significant correlations were found between the average hover durations and the amount of time the last hover on the button took.

#### 6.3.3 Spin the wheel

All the significant correlations for this minigame is found in Table 3, where the non-correlations were omitted. The omitted and insignificant correlations include the other levels and facets not mentioned in Table 3, as well as the amount of spins in the overall minigame, the timestamp for click, total time, amount of times skip button was hovered and overall amount of times both buttons were hovered variables.

**Table 3.** Spin-the-wheel minigame significant correlations results

| Variable | Levels | Conscien.[a] | p-value | coef. |
|---|---|---|---|---|
| Avg hover spin[b] | all | Conscien. | 0.0315 | -0.3407 |
| Avg hover skip[c] | all | achieve.[d] | 0.0046 | -0.4387 |
| Amount hover spin[e] | 5 | self-efficacy | 0.002 | -0.4748 |

a Conscientiousness
b Average hover on spin button
c Average hover on skip button
d Achievement-striving
e Amount of times spin button was hovered

For the spin-the-wheel minigame, conscientiousness was also measured and examined for significant correlations. The specific variables measured were the number of spins taken by the participants, as well as specific time durations and the frequency of their mouse hovering above a button.

No significant correlations were found between the number of spins taken by the participant and the five levels of this minigame. Similarly, no significant correlations were found for the amount of time it took participants to click either the spin or skip button or to proceed to the next level.

Significant correlations were found for the average number of hover times the participant had over the spin and skip buttons. Specifically, a significant negative correlation was found in the overall conscientiousness trait (coef = -0.34) for hovering above the spin button, although no significant correlation was found at the individual levels. Significant negative correlations were found in level 2 (coef = -0.44) for hovering above the skip button. However, no significant correlation was found on other levels or for the entire minigame.

A few significant correlations were found for the amount of time the participant hovered above a button. Notably, at the spin button level 5, there was a negative correlation (coef = -0.48) with facet self-efficacy. However, no significant correlation was found on other individual levels or for the entire minigame, nor for the skip button.

### 6.4 Open-ended Questions

Conscientiousness was measured through open-ended questions, specifically through the contents and writing style of the answers, which were analysed using the LLM's analysis of conscientiousness and other writing elements. Thus, first, the answers were analysed through the LLM such that it have a ranking to each conscientiousness facets from 0 to 10. Later, it analysed the answers based on specific elements of either the answer's content or writing style.

All significant correlations for the open-ended questions are presented in the Table 4, where non-significant correlations have been omitted. The omitted and insignificant correlations include the other levels and facets not mentioned in Table 4, as well as the content's conscientiousness analysis of question 1, 2 and all, the writing style question 1, 2, 3, and all, content's characteristics on question 1, 3 and all, writing style's characteristic on question 2, 3, and all terminology.

**Table 4.** Open-ended answers significant correlations results

| Variable | Levels | Conscien.[a] | p-value | coef. |
|---|---|---|---|---|
| Content oe-achieve.[b] | 3 | Conscien. | 0.0185 | 0.3288 |
| Word count | 1 | achieve.[c] | 0.0046 | -0.4387 |
| Content objective.[d] | 2 | cautious. | 0.0062 | 0.3785 |
| WS stylistic[e] | 1 | Conscien. | 0.0224 | 0.3194 |
| WS structural[f] | all | cautious. | 0.0024 | 0.3194 |

a Conscientiousness
b LLM ranking of achievement-striving for the answer's content
c Achievement-striving
d Objective writing
e Writing style: stylistic
f Writing style: structural

Significant correlations were found in the contents of the answer and how the LLM ranked the answers according to the conscientiousness facets. It was observed that what the LLM defined as achievement striving was significantly and positively correlated with the overall conscientiousness trait, specifically in the third open-ended question (coef = 0.33). No other significant correlations were found in the other open-ended questions or the open-ended questions section as a whole. Furthermore, no significant correlations were found at all for the writing style in the LLM's analysis of the conscientiousness facets.

Significant correlations were also found for the word count of the answers and conditions. In the second open-ended question, specifically, the facet cautiousness significantly positively correlated (coef = 0.37) with word count.

According to the characteristics of the answer (therefore not the LLM's ranking of the conscientiousness facets), there appeared to be significant correlations. What the LLM ranked as reflective writing showed a significant positive correlation (coef = 0.38) with the facet cautiousness. However, no other significant correlations were found in the other questions, in the open-ended question section as a whole, or in the other characteristics the answers' contents could have.

Following the content, significant correlations were also found in the writing style characteristics of the answers. Stylistic writing exhibits a significant negative correlation (coef = -0.32) with the facet of cautiousness, and structured writing displays a significant negative correlation (coef = -0.32) with the overall trait of conscientiousness. However, no significant correlations were found in the other questions or the open-ended section as a whole.

Lastly, no significant correlations were found for terminology.

### *6.5 Participant evaluation*

Lastly, at the end of the experiment, participants were asked to reflect on their overall experience and rank each of the three sections (open-ended questions, minigames, and the state-of-the-art survey). These were all compared to one another. By subtracting the average of the ranking of the surveys from the minigames and open-ended questions, we have the following:

- Mean of -0.79 when comparing the minigames and survey enjoyment, therefore skewed toward minigames.
- Mean of -0.09 when comparing the open-ended questions and survey enjoyment, therefore slightly skewed toward open-ended questions.
- Mean of -0.5 when comparing the minigames and survey immersion, therefore skewed toward minigames.
- Mean of -0.21 when comparing the open-ended questions and survey immersion, therefore skewed toward open-ended questions.
- Mean of 0.13 when comparing the minigames and survey self-report, therefore skewed toward surveys.

The results of the participation evaluation were analysed through paired sample t-tests, which were then corrected using the Bonferroni method (dividing the 0.05 alpha by 5) to account for the five different comparisons. The analysis showed that the p-values calculated did not reveal any significant correlations. Therefore, across these five comparisons, the results are not statistically significant enough to draw any conclusions.

## 7. Discussion

With the results in place, the sub-research questions defined earlier were analysed and determined whether the hypotheses have been accepted or rejected. In the hypotheses and discussions below, some hypotheses are rejected because correlations were not found in that specific area. However, this does not mean that correlations were absent for the overall minigame. It is possible that correlations were not found in the success rates of the minigames but were instead present in the reaction times, hence the need to reject or accept specific hypotheses.

### *7.1 Sub-RQ 1*

Sub-research question: How do measures obtained for playing minigames compare to the conscientiousness level assessed using traditional testing?

#### *7.1.1 H1.1*

Hypothesis: Higher mini-game success rates will correlate positively with conscientiousness, especially on achievement striving, dutifulness, and self-discipline.

The hypothesis specifically examines the defined success measures of each minigame to see if they relate to conscientiousness. These success measures include whether participants attempted to sort the books, how close they came to solving the maze without drifting, their ratio of safe versus risky decisions in the spinner wheel game, and the speed with which they completed these tasks. Therefore, this hypothesis does not focus on smaller behaviours or reaction times, but rather measures more conscientious behaviours and the speed with which participants carried them out.

Book minigame: The hypothesis stands partly true. There is a significant correlation between minigame scores and conscientiousness, as well as with self-discipline and cautiousness, but not with any other facets.

Conscientiousness can be derived specifically by a height-sorted book ordering, often through an ascending pattern (although a descending sort can be used). How well the books are sorted indicates the conscientiousness level, more specifically in relation to the facet of cautiousness. Furthermore, the number of books sorted is also an indicator of conscientiousness, although it is mostly related to the facet of self-discipline.

This means that caution can be measured by how well the books were sorted in a height-ascending manner, while self-discipline can be measured by how many books were set on the shelf. Therefore, overall, conscientiousness can be measured by the book minigame.

Additionally, the results also show that any other method of sorting the books (therefore, colour or alphabetically) shows no significant correlation with conscientiousness as a whole or with any of its facets. This is particularly interesting because it suggests that, when individuals think about sorting or organising, the height of the books is their main priority. It may indicate that participants perceive height as the most defining characteristic of the books, making it more logical for them to focus on height rather than colour or the titles of the books.

Maze minigame: The hypothesis is rejected. In contrast to the prior minigame, no significant correlations were found for any type of scoring. This means that there were no correlations between conscientiousness and the similarity between participants' and the optimal path through a maze. However, other correlations were found, but they do not relate to success rates and are discussed under another hypothesis.

Spin-the-wheel minigame: The hypothesis is also rejected. The number of spins the participants had chosen showed no correlations to any of the conscientiousness facets or its overall trait. However, other correlations were found, but they do not relate to success rates and are discussed under another hypothesis.

Overall: The only indication that the success rate serves as a viable measure was within the book minigame, and not in any of the other two minigames. Furthermore, there were no correlations with achievement striving or dutifulness either. While success rates can correlate, it seems that minigame scores do not measure the facets of achievement striving, dutifulness, and self-discipline. As a whole, however, the success rates of minigames can measure conscientiousness in general.

However, this could also be due to the fact that these games were not designed to be competitive or to have a high score in the first place. This may contribute to the lack of correlations with these specific facets. Future research and work is encouraged to examine how success rates correlate with conscientiousness, especially when a minigame is designed to be competitive and score-based, and reexamine this hypothesis.

### *7.1.2 H1.2*

Hypothesis: Faster reaction times will correlate positively with high self-efficacy and achievement striving.

This hypothesis analysed variables such as the amount of time it took to complete the minigame, as well as the number of clicks taken throughout the entire time and the average drag durations. The amount of time and the quantity of these actions relate to the speed of the reaction times.

Book minigame: The hypothesis is rejected as no significant correlations were found. No correlations were found between this minigame and the facets of self-efficacy or achievement-striving. Furthermore, no other correlations to faster reaction times were found.

Maze minigame: The hypothesis is partially accepted. Various significant correlations were found concerning speed and the frequency of participant reactions. The time it took for the participant to press the key for the maze to the button to advance to the next level correlated negatively with the conscientiousness trait. Therefore, the longer it took for the participant to click to the next minigame after finishing the maze, the lower their conscientiousness level was. Vice versa, the quicker they reacted, the higher their conscientiousness level. This was the only variable concerning quicker reaction times in the maze minigame.

This could be interpreted as participants quickly deciding to move on to the next minigame or level once they completed the maze, due to them being more conscientious (and therefore more generally organised and confident in their assessment). This could imply that participants who complete the minigame and decide they no longer want to continue have a stronger tendency toward conscientiousness. In other words, these individuals are possibly more decisive in determining when they are finished interacting with the minigame.

Spin-the-wheel minigame: The hypothesis is also partially accepted. The average amount of time the participant hovered above the spin and skip button showed a negative correlation with conscientiousness. Therefore, a shorter amount of time spent hovering above the spin button correlated with higher conscientiousness for all five levels of the spin-the-wheel minigame.

Interestingly, this is not the case for the skip button. For the skip button, specifically on the second level, there is a negative correlation with achievement striving. Therefore, if the participant hovers for a shorter time above the skip button during level 2, they correlate with higher achievement striving. This is interpreted as a quick decision-making process on the participants' part, especially since the second level involved petting a dog, which could be seen as a low-stakes, quick-thought decision. Furthermore, as this correlation only applies on the second level, it signifies that this is not a general effect. Therefore, it shows that it is concerned with the contents of the level rather than the nature of the minigame.

These correlations specifically relate to the hypothesis, which signified a faster reaction time. However, there are no other significant correlations relating to faster reaction times.

Overall: This hypothesis shows that there are limited correlations between conscientiousness and fast speeds in minigames. It was hypothesised that individuals who want to complete a task as quickly as possible also demonstrate self-efficacy, which involves confidence in taking on tasks quickly and effectively (Baburajan, Abreu e Silva, and Pereira 2022), which was not the case here. Nevertheless, there are correlations between conscientiousness and fast reactions.

### *7.1.3 H1.3*

Hypothesis: Slower reaction times will correlate positively with self-discipline and cautiousness.

This hypothesis also analyses time-based variables and the frequency of actions within the minigame.

Book minigame: Similarly to the previous hypothesis, no significant correlations were found with any of the slower behvaiours nor the facets of self-discipline or cautiousness. Therefore, the hypothesis is rejected for the book minigame.

Maze minigame: The hypothesis is partially accepted. The number of times participants pressed a key showed a negative correlation with their level of conscientiousness. This was only found in the impossible mazes, especially the last one. Therefore, the fewer times participants pressed a key during the impossible mazes, particularly the last one, indicated a higher degree of conscientiousness. This can be attributed to participants carefully considering whether they are capable of solving the maze or not. Similarly, there was also a correlation between fewer hovers above the button and overall conscientiousness on just the last impossible maze. This suggests that participants hovered their mouse over the button less, perhaps because they were more carefully considering their actions during the maze rather than acting quickly. No other significant correlations were found, therefore saying nothing about the individual facets. Nevertheless, the hypothesis is partially accepted due to the correlation with overall conscientiousness.

Spin-the-wheel minigame: This hypothesis is partially accepted. The quantity in which the participants hovered above the spin button correlated negatively with the self-efficacy facet, specifically the last level. The last level of this minigame focuses on deciding whether to walk past a seemingly drunk man late at night. This level was designed to evoke a sense of danger and urgency in participants, specifically to assess their ability to perceive risk, a trait hypothesised to relate to risk-taking behaviour and, consequently, the facet of cautiousness.

The analyses data reveals that the fewer times the participants hovered above the spin button indicates a high level of self-efficacy at the fifth level. This can be interpreted as participants being less hesitant and less indecisive in this specific scenario, resulting in fewer 'switches' between choices, as seen in the number of times they came on and off the spin button. This suggests that participants have more self-efficacy, as they know themselves better and therefore know how to behave in this specific scene more quickly.

Overall: This hypothesis holds for both the maze minigame and the spin-the-wheel minigame, as participants take more time to consider their decisions and are more confident about their planning before proceeding. While there are no correlations between self-discipline and cautiousness, there are correlations with self-efficacy, a facet closely related to these two.

#### *7.1.4 Sub research question 1 conclusion*

Regarding this sub-research question, it appears that while minigames show potential in measuring the trait of conscientiousness in terms of their scoring and reaction speeds, the current minigames used in this study do not reliably substitute for traditional personality assessments. Currently, reaction times (such as how fast or slow a participant is) show more promise than game scores for measuring conscientiousness. Therefore, these behavioural characteristics, such as reaction times and hovering patterns, show more promise than actual success rates.

There is, however, potential in determining individuals' contingency levels through success rates, particularly in the book minigames. An individual's level of conscientiousness can be inferred by whether they store their books in a height-ascending order. This is not reflected through the specific facets, but rather through the conscientiousness trait in general.

### *7.2 Sub-RQ 2*

Sub-research question: How do writing style and content from open-ended question answers compare to the conscientiousness level as assessed using traditional testing?

#### *7.2.1 H2.1*

Hypothesis: Richer and thoughtful content correlates with conscientiousness, especially with dutifulness.

This hypothesis is partially accepted, though with limited correlation. The only correlation found in the LLM's analysis of open-ended responses was between its achievement striving rankings and participants' overall conscientiousness. Therefore, what the LLM identified as achievement striving correlates positively with conscientiousness, but this correlation only applies to the third open-ended question, not to the other two questions or the open-ended section as a whole.

This can be interpreted in two ways: either LLMs interpret conscientiousness as primarily related to achievement striving content, or people with more achievement-oriented content are more prone to being conscientious. However, since this correlation only occurred in one of the three open-ended questions, it already points to how using LLMs prove to be difficult in determining conscientiousness.

Lastly, since this correlates only with the third open-ended question, achievement striving possibly relates directly to the question's topic, which concerned how participants reacts in case they wanted to take a walk and it started drizzling outside. This open-ended question could help measure a participant's conscientiousness by examining how the LLM ranks their achievement striving in response to this scenario.

#### *7.2.2 H2.2*

Hypothesis: Structured and formatted writing style correlates with conscientiousness, especially with orderliness.

This hypothesis is rejected. No significant correlations were found between the LLM's analysis of the writing style in the open-ended questions and the traditional survey results of conscientiousness. No correlations were found with conscientiousness or with the facet orderliness.

#### *7.2.3 H2.3*

Hypothesis: Participants with lower conscientiousness (especially low self-discipline) will write less with each open-ended question due to frustration and fatigue.

This hypothesis is partially accepted. The word count on the first open-ended question was negatively correlated with achievement striving. This means that the less the participant wrote in response to the first open-ended question, the more likely they were to have a higher achievement-striving level. Since this correlation only appears for the first open-ended question, and not for any other question or the entire open-ended section, this correlation is likely specific to this question specifically.

This could be interpreted about the question's topic or by the fact that it is the first question to appear. Considering the topic about what the participant would do in case of bank problems, the amount of detail that participants want to go into (therefore relating to word count) may be interpreted as a sign of achievement striving. Similarly, since this is the first open-ended question to appear, it could also indicate that participants put in more effort on the first question when they are still less fatigued. This attitude possibly diminishes with subsequent questions.

#### 7.2.4 Sub research question 2 conclusion

Overall, while LLMs do show a few correlations, they are very few and largely fail to predict conscientiousness facet levels. There are specific findings, such as how the LLM's ranking on achievement striving correlates with conscientiousness, and how word count correlates with achievement striving. Nevertheless, it appears that LLMs cannot accurately assess conscientiousness facets solely through content and writing style. The question-specific correlations suggests that further refinement is needed in how the questions are worded and what they are about, which could help unveil underlying personality traits more effectively. Currently, it points that traditional personality assessments are superior to LLM analysis of conscientiousness (to these current open-ended questions).

### 7.3 Sub-RQ 3

Sub-research question: How do participants evaluate taking the alternative compared to traditional measurements?

The participation evaluation consisted of asking participants how they thought of the experiment. The analysis of the paired sample t-tests showed no significant correlations within the participant evaluations. Therefore, all the hypotheses in this sub-question are inconclusive, as no significant correlations were found. Consequently, no conclusions can be drawn about whether participants enjoyed or were immersed in the experiment.

Nevertheless, the remainder of this section considers the magnitude and direction of the calculated means from the participant evaluations. The following discussion is therefore hypothetical, intended to explore what the findings could have suggested if the values were significant.

#### 7.3.1 H3.1

Hypothesis: They prefer the mini-games

This hypothesis is accepted. Comparing participants' rankings of their enjoyment of minigames and surveys showed a mean difference of -0.79, where the negative sign indicates a skew towards enjoying minigames more. Since minigames were the section that contributed most to serious gaming, with a big focus on making them as game-like as possible, this continues to prove that game-like elements are more enjoyable than simple surveys.

#### 7.3.2 H3.2

Hypothesis: They don't prefer the open-ended questions

This hypothesis is rejected. Participant evaluation showed a slight skew towards enjoyment of open-ended questions. This is quite interesting because open-ended questions are typically viewed more negatively compared to surveys. The reason for further enjoyment of open-ended questions may be because of the question content itself. Instead of being purely about scientific topics, it incorporated more of the role-playing gamification aspect, which may have encouraged participants to enjoy those specific open-ended questions more.

#### 7.3.3 H3.3

Hypothesis: They were more immersed in the alternative

This hypothesis is accepted. For both open-ended questions and minigames, participants noted that they were more immersed in these activities than in surveys. Once again, this could be attributed to the game-like aspects of both parts. It is also interesting that participants were more immersed in minigames than open-ended questions, which tracks with how much they enjoyed both sections.

#### 7.3.4 H3.4

Hypothesis: The alternative will produce less self-reporting

This hypothesis is rejected. A comparison was made between the honesty of participants in open-ended questions and surveys to determine if there was any self-reporting bias. There was a skew towards surveys, indicating that participants were more honest when responding to surveys compared to open-ended questions. This could be attributed to the fact that open-ended questions require more mental effort. Since they involve role-playing in specific scenarios, there's more leniency towards being less honest. Surveys, on the other hand, are more pointed and require less deliberation.

#### 7.3.5 Sub research question 3 conclusion

Participants preferred the more gameified assessment methods compared to traditional methods. This was shown through their ranking of enjoyment and immersion for open-ended questions, minigames, and surveys. However, it does show that the traditional method of surveys brought forth more honesty from participants, which puts self-reporting at a greater risk in open-ended questions. Therefore, for more honest and reliable responses, surveys may be a better approach. However, as minigames are only assessed through behaviour, and they have the most enjoyment and immersion out of the three sections, it could be argued that minigames are more honest and better for retrieving data.

### 7.4 Sub-RQ 4

Sub-research question: Can the alternative measures further our understanding of the construct of conscientiousness?

#### 7.4.1 H4.1

Hypothesis: Each minigame correlates with its corresponding conscientiousness facet

Examining the correlation between each minigame and its intended target facet reveals few significant correlations.

Book minigame: The book sorting minigame was designed to measure the facet orderliness, but no correlation was found between this facet and the results.

Maze minigame: The maze minigame was designed to measure self-discipline, and correlations were found for that facet on the last impossible maze. Specifically, participants tended to have a higher level of self-discipline when they hovered less over the skip button without clicking. No other correlations were found within the maze minigame for self-discipline. Nevertheless, this indicates that individuals with more self-discipline understand how to control the cells better and therefore have less tendency to hover back and forth over the button, perhaps demonstrating more certainty in the decisions they make.

Spin-the-wheel minigame: The spin-the-wheel minigame was designed for cautiousness, but it did not correlate with that facet either.

Overall: These results indicate that measuring targeted facets through minigames is more challenging than expected. However, it is possible, as self-discipline was measured with the final maze instance as originally intended. Nevertheless, this represents only one successful instance out of all attempts, showing the difficulty in actually measuring intended facets through these minigames. Still, in all cases, conscientiousness as a whole trait was measurable, suggesting that while specific facets prove challenging to assess, broader personality dimensions (such as traits, e.g. conscientiousness) is able to be captured through gameplay behaviour.

#### *7.4.2 H4.2*

Hypothesis: Each open-ended question correlates with its corresponding conscientiousness facet

This hypothesis is rejected. Each open-ended question targeted a specific facet of conscientiousness: the first open-ended question about the bank related to self-efficacy, the second open-ended question about the pamphlet related to dutifulness, and the last open-ended question about the walk in the rain related to achievement striving.

As seen in hypothesis 2.1, the third open-ended question showed a correlation between the LLM's ranking of achievement striving and conscientiousness. In this case, it was the LLM that found the correlation between the answer contents and achievement striving, rather than the actual survey itself. Since this does not relate directly to the achievement striving facet according to the survey, nothing can be said about whether the answer contents is able to assess achievement striving.

No other correlations were found for the contents. While the LLM's assessment of specific characteristics within the answer contents and writing style showed correlations with specific characteristics and conscientiousness, no relation was found to the actual facets assigned to each open-ended question.

#### *7.4.3 H4.2*

Hypothesis: Certain writing style aspects correlate with conscientiousness facets

This hypothesis is partially accepted. The only writing style aspects that correlated with a conscientiousness facet were stylistic and structural writing, where both correlated negatively with conscientiousness as a trait. Stylistic writing refers to the style and tone of a specific piece of writing, examining whether it incorporates personal reflection, objective tones, or humour within the text. According to these results, participants who wrote more stylistically were less likely to be more conscientious. Similarly, structural complexity, which referred to sentence length and paragraph structure, showed a negative correlation to conscientiousness. Therefore, the less structural their writing was, the more they tended to be conscientious.

This is particularly interesting because the hypothesis was mostly based on the opposite of these results, where it was assumed that more conscientious people have a more stylistic and structurally complex writing style. These results could be interpreted as indicating that conscientious participants are more straightforward in their writing, conveying their intended meaning more directly.

Nevertheless, this also means that other writing style characteristics did not have significant correlations to conscientiousness. This means that the formality of the writing style, as well as the sentimentality, grammatical punctuation, and complexity of constructions within the wording, did not contribute to significant correlations. Instead, conscientiousness can only be measured through a lack of style and tone within the writing, as well as simple structural complexity.

#### *7.4.4 H4.3*

Hypothesis: Certain content aspects correlate with conscientiousness facets

This hypothesis is partially accepted. The only correlation found between content characteristics and conscientiousness was a positive correlation between reflective writing and the facet of cautiousness. Therefore, if the participant's content showed more reflection (signs of thoughtfulness, introspection, and self-awareness), the participant was more likely to be cautious. This correlation was specifically observed in the second open-ended question, which examined participants' reactions to what they would do if given a pamphlet by a stranger.

This is particularly interesting because it shows that participants tended to be more cautious when reacting in this specific scenario. What could be happening is that this question was framed to touch on a person's dutifulness, and therefore touched upon their morality regarding what they would do in this situation. This may have pushed people to explain their reasoning more thoroughly, hence fostering their reflection, which we now see correlated to their caution

No other correlations were found in the other open-ended questions or the other content writing styles.

### 7.4.5 Sub research question 4 conclusion

It is limited in its ability to reliably measure the targeted facets of conscientiousness in the intended minigames. Only the maze minigame measured its intended facet (self-discipline) in its impossible mazes, while the other minigames did not show any correlations to their intended facets. As a whole, the minigames can measure conscientiousness, simply not in their more specific facets. However, the correlation of self-discipline with the maze minigame does indicate potential for measuring conscientiousness facets through the use of minigames. Currently, though, it is more reliable to measure conscientiousness as a whole, but this does not rule out the possibility of measuring facets as well.

Several interesting characteristics were identified in the open-ended questions. First of all, there was no correlation with any of the intended facets for the open-ended questions, either. However, there were signs of cautiousness in the reflective writing in the pamphlet scenario question, as well as the unexpected,, simple, and straightforward writing style, which showed conscientiousness. Instead of looking at open-ended questions as a whole, there are more reliable correlations with scenario-specific questions than with open-ended questions in general.

## 8. Limitations

Several limitations affected this research, including constraints related to the chosen LLM, the exhaustive nature of the study, time limitations, methodological improvements, and inherent limitations of current personality assessments.

### 8.1 LLM Selection

Initially, the LLM chatbot Olmo was selected for analysis. However, this choice was abandoned before data analysis began because Olmo no longer provided transparent explanations of how they stored user input. To protect participant privacy, Claude.ai was chosen instead due to its ability to remove all user inputs.

After conducting data analysis and reviewing the literature (Oeljeklaus, Höft, and Danner 2025), several pitfalls became apparent. Proper prompting when using LLMs is critically important. While extensive definitions and instructions were provided (detailed in methodology and Appendix 1), the lack of correlations for open-ended questions suggests that more specific, direct prompts with examples may have yielded better correlations.

### 8.2 Time Constraints

The six-month thesis timeframe, combined with the exploratory nature of this study, created significant limitations. Since the study examined numerous variables seeking correlations with conscientiousness, non-correlations were inevitable. Nevertheless, analysing all data points (particularly minigames) required extensive time for conceptualisation, design, data gathering, and analysis.

The exhaustive nature of this thesis means time constraints was likely to remain regardless of the timeframe given. Nevertheless, this specific timeframe limited how much data was analysed compared to what was gathered. Further analysis could have examined timestamp correlations with conscientiousness at different minigame stages (beginning, middle, end), but time constraints prevented this analysis.

Additionally, data from 51 participants were used after excluding 20 participants due to incomplete experiments, language barriers, or age restrictions. More participants could have yielded more reliable correlations.

### 8.3 Methodology

If time permitted, more variables could have been analysed. While already exhaustive, the study is able to be expanded further. For example, minigames could have included more levels, potentially revealing additional correlations. This experiment examined three facets through open-ended questions and three through minigames. An expanded format could have included six open-ended questions and six minigames for each of the six facets. This was not implemented to prevent participant exhaustion. Nevertheless, with guaranteed participant completion, this longer format is expected to better determined whether gamified assessments reliably measure personality.

Similarly, this experiment utilised the IPIP-NEO-120, which assesses conscientiousness through 32 items. Using the IPIP-NEO-300 would double the conscientiousness statements the participant answers to, providing a more thorough personality assessment through traditional means.

These methodological limitations can only be addressed with a sufficient number of participants willing to complete the entire experiment.

### 8.4 Personality Assessment Limitations

The biggest limitation is the current inability to reliably assess personality, even with current state-of-the-art traditional assessments. The IPIP-NEO-120, IPIP-NEO-300, and the well-known NEO-PI-R assessments are modern descriptions of what researchers currently define as personality.

Currently, only theories of personality exist. Given the current definition of personality, it is theoretically impossible to know someone's personality concretely, as researchers are still defining what personality is. Therefore, despite traditional assessment types providing a solid foundation for understanding and assessing personality, the inherent nature of personality cannot be fully known. The entire experiment, comparing its results to state-of-the-art personality assessments, is inherently flawed because it compares against something not entirely reliable. Nevertheless, these represent our most reliable personality assessment tools and provide a good basis for understanding whether open-ended question analysis and minigames are comparable to current assessment standards.

## 9. Further Research

Personality assessment would benefit from continued exploration of this topic, though with more rigorous and realistic expectations for specific minigames and open-ended questions. Instead, further research is suggested to focus on either minigames or open-ended questions, but not simultaneously.

Concerning minigames, it is encouraged for there to be further theoretical connection between specific game mechanics and personality facets. The minigames were designed based on existing experiments that were then translated into a game format. Examining more experiments that correlate with personality assessments and understanding how they differ as minigames could be a fruitful area of research. Nevertheless, this still requires exploratory research, which is likely to reveal many non-correlations as well.

Since minigames and open-ended question analysis are not currently sufficient to replace traditional assessment types, testing a hybrid approach between traditional and gamified elements would be interesting. Utilising the reliable and quick nature of surveys while introducing gamified characteristics has potential.

Regarding specific limitations, the methodology could be improved with a larger sample size and more rigorous testing for minigames, open-ended questions, and traditional assessment types. This would allow for more comprehensive data collection and more reliable personality assessments for comparison.

## 10. Conclusion

This study investigated whether gamified methods, specifically minigames and LLM-analysed open-ended responses, have the potential to serve as an alternative to traditional survey methods for assessing personality. The motivation was to develop a more enjoyable and gamified measurement tool to reduce self-reporting bias in personality assessment while maintaining the reliability of traditional assessments. It was found that the experiment designed in this study is not reliable enough to replace traditional survey assessments. However, there is potential.

The results show that gamified approaches, as constructed in this study, are not as reliable and refined as current traditional personality assessments and are not yet ready to replace them.

The minigames showed few correlations, indicating how behaviours correlate to personality in terms of conscientiousness and its facets. Importantly, the minigames did not fully and reliably measure their intended facets of conscientiousness, although they did indicate conscientiousness as a trait overall. This suggests that measuring conscientiousness through minigames has potential but requires significant refinement, additional testing, and further research.

The LLM analysis of open-ended answers also did not provide fully reliable assessment of conscientiousness. A few correlations were found, but like the minigames, the intended facet for each open-ended question had very few correlations. These limited correlations show promise that assessing personality through the analysis of open-ended answers is possible. However, it requires more testing with improved LLM prompting.

Furthermore, the study revealed certain personality-related behaviours through minigames and open questions that are otherwise not previously known. For example, hovering above certain buttons or the time taken to click to the next minigame showed signs of personality correlation. Similarly, certain behaviours, such as the total time to complete a minigame, showed no correlation to personality at all. By understanding which behaviours do and don't correlate, we have a more comprehensive list of actions that actively contribute to personality or not.

The same applied to open-ended questions, which revealed unexpected characteristics, such as simpler writing styles correlating to conscientiousness. These unconscious behaviours suggest that behaviour is not always a conscious decision but also an unconscious reaction.

This study demonstrates that correlations exist between minigames, open-ended answers, and personality. The fact that correlations were found suggests potential in this gamified assessment style. It provides a good starting point for future research, as there are indications that personality assessment using minigames and open-ended questions is feasible, although not in its current form.

## Acknowledgments

I would like to thank the many people who have helped and guided me throughout this thesis project. First, I want to thank my supervisors, Max van Duijn and Bernhard Hilpert, for their guidance, advice, and encouragement throughout the process. Their support has been essential in bringing this project to its final form, which I am proud of.

I also want to thank Nicolas Becu, who helped me conceptualise the idea of this thesis and provided advice and encouragement from the very beginning.

Thank you as well to the critics who reviewed my research and provided valuable feedback.

Lastly, I would like to thank all of the pilot testers and participants for their patience and effort in taking part in my experiment.

## References

Argamon, Shlomo, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and L. Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology* 58, no. 6 (April): 802–822. ISSN: 15322882. https://doi.org/10.1002/asi.20553.

Atherton, Olivia E., Angelina R. Sutin, Antonio Terracciano, and Richard W. Robins. 2022. Stability and Change in the Big Five Personality Traits: Findings From a Longitudinal Study of Mexican-Origin Adults. *Journal of Personality and Social Psychology* 122 (2): 337–350. ISSN: 00223514. https://doi.org/10.1037/pspp0000385.

Baburajan, Vishnu, João de Abreu e Silva, and Francisco Camara Pereira. 2022. Open vs closed-ended questions in attitudinal surveys – Comparing, combining, and interpreting using natural language processing. *Transportation Research Part C: Emerging Technologies* 137 (April). ISSN: 0968090X. https://doi.org/10.1016/j.trc.2022.103589.

Bubaš, Goran. n.d. *The use of GPT-4o and Other Large Language Models for the Improvement and Design of Self-Assessment Scales for Measurement of Interpersonal Communication Skills.* Technical report.

Connor Desai, Saoirse, and Stian Reimers. 2019. Comparing the use of open and closed questions for Web-based measures of the continued-influence effect. *Behavior Research Methods* 51, no. 3 (June): 1426–1440. issn: 15543528. https://doi.org/10.3758/s13428-018-1066-z.

Dale, Gillian, Danielle Sampers, Stephanie Loo, and C. Shawn Green. 2018. Individual differences in exploration and persistence: Grit and beliefs about ability and reward. *PLoS ONE* 13, no. 9 (September). issn: 19326203. https://doi.org/10.1371/journal.pone.0203131.

Derner, Erik, Dalibor Kučera, Nuria Oliver, and Jan Zahálka. 2024. Can ChatGPT read who you are? *Computers in Human Behavior: Artificial Humans* 2, no. 2 (August): 100088. issn: 29498821. https://doi.org/10.1016/j.chbah.2024.100088.

DeYoung, Colin G., Lena C. Quilty, and Jordan B. Peterson. 2007. Between Facets and Domains: 10 Aspects of the Big Five. *Journal of Personality and Social Psychology* 93, no. 5 (November): 880–896. issn: 00223514. https://doi.org/10.1037/0022-3514.93.5.880.

Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. 2007. Grit: Perseverance and Passion for Long-Term Goals. *Journal of Personality and Social Psychology* 92, no. 6 (June): 1087–1101. issn: 00223514. https://doi.org/10.1037/0022-3514.92.6.1087.

Facchino, Antonio Pio, Daniela Marchetti, Marco Colasanti, Lilybeth Fontanesi, and Maria Cristina Verrocchio. 2025. *The use of serious games for psychological education and training: a systematic review.* https://doi.org/10.3389/feduc.2025.1511729.

Frew, Emma J., David K. Whynes, and Jane L. Wolstenholme. 2003. Eliciting willingness to pay: Comparing closed-ended with open-ended and payment scale formats. *Medical Decision Making* 23, no. 2 (March): 150–159. issn: 0272989X. https://doi.org/10.1177/0272989X03251245.

Hansen, Karolina, and Aleksandra Świderska. 2024. Integrating open- and closed-ended questions on attitudes towards outgroups with different methods of text analysis. *Behavior Research Methods* 56, no. 5 (August): 4802–4822. issn: 15543528. https://doi.org/10.3758/s13428-023-02218-x.

Hassan, Saad, Muhamamd Faisal Malik, Saqlain Raza, Mulyadi Suhardi, and Wentri Merdiani. 2023. Personality and Humbleness: The Role of the HEXACO Model of Personality in Development of Humble Leaders. *SAGE Open* 13, no. 4 (October). issn: 21582440. https://doi.org/10.1177/21582440231216172.

Hirsh, Jacob B., and Jordan B. Peterson. 2009. Personality and language use in self-narratives. *Journal of Research in Personality* 43, no. 3 (June): 524–527. issn: 00926566. https://doi.org/10.1016/j.jrp.2009.01.006.

Johnson, John A. 2014. Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality* 51:78–89. issn: 10957251. https://doi.org/10.1016/j.jrp.2014.05.003.

Johnston, S., R. Taylor, M. Bailes, N. Bartel, C. Baugh, M. Bietenholz, C. Blake, et al. 2008. Science with ASKAP. The Australian square-kilometre-array pathfinder. *Experimental Astronomy* 22, no. 3 (December): 151–273. https://doi.org/10.1007/s10686-008-9124-7. arXiv: 0810.5187 [astro-ph].

Joseph, Elizabeth D., and Don C. Zhang. 2021. Personality Profile of Risk-Takers: An Examination of the Big Five Facets. *Journal of Individual Differences* 42, no. 4 (October): 194–203. issn: 21512299. https://doi.org/10.1027/1614-0001/a000346.

KA, Lyon, Elliott R, Ware K, Juhasz G, and Brown LJE. 2021. *Associations between Facets and Aspects of Big Five Personality and Affective Disorders:A Systematic Review and Best Evidence Synthesis,* June. https://doi.org/10.1016/j.jad.2021.03.061.

Kabigting, Florencio. 2021. The Discovery and Evolution of the Big Five of Personality Traits: A Historical Review. *An Interdisciplinary Journal of Human Theory and Praxis* 4, no. 3 (December). issn: 2714-2485.

Karpen, Samuel C. n.d. *The Social Psychology of Biased Self-Assessment.* Technical report.

Kim, Hyunji, Stefano I. Di Domenico, and Brian S. Connelly. 2019. Self–Other Agreement in Personality Reports: A Meta-Analytic Comparison of Self- and Informant-Report Means. *Psychological Science* 30, no. 1 (January): 129–138. issn: 14679280. https://doi.org/10.1177/0956797618810000.

Kreitchmann, Rodrigo Schames, Francisco J. Abad, Vicente Ponsoda, Maria Dolores Nieto, and Daniel Morillo. 2019. Controlling for Response Biases in Self-Report Scales: Forced-Choice vs. Psychometric Modeling of Likert Items. *Frontiers in Psychology* 10 (October). issn: 16641078. https://doi.org/10.3389/fpsyg.2019.02309.

Määttänen, Ilmari, Emilia Makkonen, Markus Jokela, Johanna Närväinen, Julius Väliaho, Vilja Seppälä, Julia Kylmälä, and Pentti Henttonen. 2021. Evidence for a behaviourally measurable perseverance trait in humans. *Behavioral Sciences* 11, no. 9 (September). issn: 2076328X. https://doi.org/10.3390/BS11090123.

McCrae, Robert R., Paul T. Costa, and Thomas A. Martin. 2005. The NEO-PI-3: A more readable Revised NEO Personality Inventory. *Journal of Personality Assessment* 84 (3): 261–270. issn: 00223891. https://doi.org/10.1207/s15327752jpa8403{_}05.

Mõttus, René, Christian Kandler, and Wiebke Bleidorn. 2017. Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology,* issn: 0022-3514. https://doi.org/10.1037/pspp0000100.supp.

Nicholson, Ian A. M. 1998. Gordon Allport, character, and the "culture of personality," 1897–1937. *History of Psychology* 1, no. 1 (February): 52–68. issn: 1939-0610. https://doi.org/10.1037/1093-4510.1.1.52.

Nielsen, Maiken Due, and Petri Kajonius. 2024. Beyond the Big Five factors: using facets and nuances for enhanced prediction in life outcomes. *Current Psychology* 43, no. 20 (May): 18621–18630. issn: 19364733. https://doi.org/10.1007/s12144-024-05662-w.

Noroozi, Omid, Hojjat Dehghanzadeh, and Ebrahim Talaee. 2020. *A systematic review on the impacts of game-based learning on argumentation skills,* August. https://doi.org/10.1016/j.entcom.2020.100369.

Oeljeklaus, Lydia, Stefan Höft, and Daniel Danner. 2025. Comparing Psychometric Properties of Expert-Developed and AI-Generated Personality Scales: A Proof-of-Concept Study. *Psychological Test Adaptation and Development* 6, no. 1 (April): 29–43. issn: 26981866. https://doi.org/10.1027/2698-1866/a000095.

Ogden, Jane, and Jessica Lo. 2012. How meaningful are data from Likert scales? An evaluation of how ratings are made and the role of the response shift in the socially disadvantaged. *Journal of Health Psychology* 17, no. 3 (April): 350–361. issn: 13591053. https://doi.org/10.1177/1359105311417192.

Olea, Carlos, Holly Tucker, Jessica Phelan, Cameron Pattison, Shen Zhang, Maxwell Lieb, Doug Schmidt, and Jules White. n.d. *Evaluating Persona Prompting for Question Answering Tasks.* Technical report.

Piastra, Marco, and Patrizia Catellani. 2025. On the emergent capabilities of ChatGPT 4 to estimate personality traits. *Frontiers in Artificial Intelligence* 8. issn: 26248212. https://doi.org/10.3389/frai.2025.1484260.

Querengässer, Jan, and Sebastian Schindler. 2014. Sad but true? - How induced emotional states differentially bias self-rated Big Five personality traits. *BMC Psychology* 2, no. 1 (December). https://doi.org/10.1186/2050-7283-2-14.

Ragsdale, Jennifer M, Neil D Christiansen, Christopher T Frost, John A Rahael, and Gary N Burns. n.d. *Content Analysis of Personality at Work.* Technical report.

Ramos-Villagrasa, Pedro J., Elena Fernández-Del-Río, Ramón Hermoso, and Jorge Cebrián. 2024. Are serious games an alternative to traditional personality questionnaires? Initial analysis of a gamified assessment. *PLoS ONE* 19, no. 5 May (May). issn: 19326203. https://doi.org/10.1371/journal.pone.0302429.

Rau, Richard, Louisa M. Schömann, and Michael P. Grosz. 2025. People "fake-good" on personality self-reports more strongly in a job context than in a dating context. *Journal of Research in Personality* 116 (June). ISSN: 10957251. https://doi.org/10.1016/j.jrp.2025.104596.

Roberts, Brent W., Carl Lejuez, Robert F. Krueger, Jessica M. Richards, and Patrick L. Hill. 2014. What is conscientiousness and how can it be assessed? *Developmental Psychology* 50 (5): 1315–1330. ISSN: 00121649. https://doi.org/10.1037/a0031109.

Santos, Joseline M, Rogevir C John Christopher Rivera, Julie R Ann Bermas, Jocelyn G Dalusung, Jeck A Richard Mendoza, Christian L Roque, and Jennylyn B Roque. n.d. The Effectiveness of Game-Based Assessment Tools (GBAT) in Teaching Technology and Livelihood Education (TLE), ISSN: 2454-6186. https://doi.org/10.47772/IJRISS. www.rsisinternational.org.

Schmidt, Fabian T.C., Gabriel Nagy, Johanna Fleckenstein, Jens Möller, and Jan Retelsdorf. 2018. Same Same, but Different? Relations Between Facets of Conscientiousness and Grit. *European Journal of Personality* 32, no. 6 (November): 705–720. ISSN: 10990984. https://doi.org/10.1002/per.2171.

Seeboth, Anne, and René Mõttus. 2018. Successful Explanations Start with Accurate Descriptions: Questionnaire Items as Personality Markers for More Accurate Predictions. *European Journal of Personality* 32, no. 3 (May): 186–201. ISSN: 10990984. https://doi.org/10.1002/per.2147.

Soto, Christopher J., and Oliver P. John. 2017. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology* 113, no. 1 (July): 117–143. ISSN: 00223514. https://doi.org/10.1037/pspp0000096.

Stewart, Ross David, René Mõttus, Anne Seeboth, Christopher John Soto, and Wendy Johnson. 2022. The finer details? The predictability of life outcomes from Big Five domains, facets, and nuances. *Journal of Personality* 90, no. 2 (April): 167–182. ISSN: 14676494. https://doi.org/10.1111/jopy.12660.

Su, Fan, and Di Zou. 2024. A systematic review of game-based assessment in education in the past decade. *Knowledge Management and E-Learning* 16, no. 3 (September): 451–476. ISSN: 20737904. https://doi.org/10.34105/j.kmel.2024.16.021.

Tekofsky, Shoshannah, Pieter Spronck, Aske Plaat, Jaap Van Den Herik, and Jan Broersen. 2013. *Class (4), Score (19).* Technical report 7. http://www.psyopsresearch.com.

Troiano, Enrica, Laura Oberländer, and Roman Klinger. 2022. Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction (October). https://doi.org/10.1162/coli{_}a{_}00461. http://arxiv.org/abs/2206.05238%20http://dx.doi.org/10.1162/coli_a_00461.

Verma, Gaurav, and Balaji Vasan Srinivasan. 2019. A Lexical, Syntactic, and Semantic Perspective for Understanding Style in Text (September). http://arxiv.org/abs/1909.08349.

Wagner, Stefan, Daniel Mendez, Michael Felderer, Daniel Graziotin, and Marcos Kalinowski. n.d. *Challenges in Survey Research.* Technical report. http://napire.org.

Welz, Max, and Andreas Alfons. 2025. When Respondents Don't Care Anymore: Identifying the Onset of Careless Responding (February). http://arxiv.org/abs/2303.07167.

Westera, Wim. n.d. *Why and how serious games can become far more effective: accommodating productive learning experiences, learner motivation and the monitoring of learning gains.* Technical report.

Zhang, Xijuan, Winnie Wing-Yee Tse, and Victoria Savalei. 2019. Improved properties of the big five inventory and the Rosenberg self-esteem scale in the expanded format relative to the likert format. *Frontiers in Psychology* 10 (JUN). ISSN: 16641078. https://doi.org/10.3389/fpsyg.2019.01286.

## Appendix 1. OSF and External Documents

All other documents and information can be found in the OSF page, found in this url: https://osf.io/u6jxk. In this page, the pre-registration can be found, as well as the documents such as the codebook, data analysis, LLM prompts, code, and the raw data.

## Appendix 2. Standards

### *Appendix 2.1 Ethical Standards*

The research meets all ethical guidelines, including adherence to the legal requirements of the study country. This study was reviewed by the Ethics Committee and approved. This means that the way data was collected from participants, how it was used and handled was reviewed and managed accordingly.

### *Appendix 2.2 Pre-registration*

This research underwent pre-registration to transparently outline its purpose, including the construction and design of the experiment, the hypotheses, and the intended raw data to be collected, all of which were defined before actual data collection. The pre-registration was conducted through OSF, and the link to this research pre-registration can be found https://osf.io/rqz4y.

The pre-registration documented the plans for the study before data collection. However, changes were later made to the experiment, research questions, hypotheses, and data analysis. The following section explains the reasoning behind these changes. In general, these adjustments were due to time constraints and refinements in the study's structure.

**Research question and hypothesis changes** The main research question was revised from: "Would a combination of minigames and open-ended questions be **more** valid and attractive than the IPIP-NEO-120 to assess conscientiousness?" to "Would a combination of minigames and open-ended questions be valid and attractive than the IPIP-NEO-120 to assess conscientiousness?". This change was made for clarity and to better focus on the alternative assessment type.

**LLM changes** The intended LLM for analysing the open-ended responses was originally OLMo. However, due to changes in their privacy policy regarding the retention of user inputs, the model was switched to Claude.ai, which allows for the removal of user inputs.

**Demographic additions** The demographic information collected was also updated. In addition to the mentioned initially demographics (age and English proficiency, asked during consent), participants were also asked whether they were completing the experiment on a computer or a smart device, and about their computer skills. This was intended to enable further research on participant demographics, though it ultimately was not conducted.

Data analysis changes The planned data analysis was reduced. Initially, scatter plots were intended to visualise patterns, but after discussion with the supervisor, it was concluded that the Shapiro–Wilk test and Z-Scores for outliers would be sufficient.

Additionally, an HLM (hierarchical linear model) analysis was planned initially to examine potential correlations between the timing of participant actions and outcomes. For example, a hypothesis involved analysing how many times participants clicked at the beginning versus the end of the minigame to assess conscientiousness. The HLM allowed analysis in 10% intervals rather than only the beginning or end. This specific analysis was not conducted due to time constraints.

Furthermore, the analysis done on the participant evaluation with paired sample t-tests was also not mentioned in the pre-registration, as this had been overlooked initially. Therefore, this analysis was conducted after the data collection.