UNIVERSITY OF TWENTE.

DEVELOPING A VIRTUAL STEALTH ASSESSMENT OF ORDERLINESS AND EXPLORING THE EFFECT OF ENVIRONMENTAL ORDERLINESS IN VIRTUAL REALITY

Master thesis

Educational Science and Technology

Author: Chun Syuan Chu

Enschede, January 2022

Examination Committee

First supervisor: Bas Kolloffel

Second supervisor: Ilona Friso-van den Bos

University of Twente

Faculty of Behavioral, Management and Social Sciences

SERIOUS VR

Table of Contents

Table of Contents 2
Acknowledgement 4
Abbreviation List 5
Abstract 6
1. Introduction 7
2. Theoretical Framework 11
2.1. Orderliness as a Personality Facet 11
2.2. Environmental Orderliness 12
2.3. Environmental Orderliness and Orderly Behaviours 13
2.4. Explicit Measures of Personality 15
2.5. Implicit Measures of Personality 17
2.6. Stealth Assessment as an Alternative Measure 19
2.7. Potential of Stealth Assessment in VR 20
3. Research Questions & Hypotheses 23
4. Method 24
4.1. Research design 24
4.2. Participants 25
4.3. Instrumentation 25
4.3.1. Orderliness Questionnaire (IPIP-NEO) 25
4.3.2. VSA on Orderliness 26
4.3.3. Post Experiment Survey 34
4.4. Procedure 34
5. Results 38
6. Discussion 44
6.1. Implications 48
6.2. Limitations 50
6.3. Future Research & Recommendations 51
6.4. Conclusion 53
7. References 55
8. Appendices 64
8.1. Appendix A: The Use of IPIP-NEO in the Current Study 64

8.2. Appendix B: Self-reported orderliness Survey from Goldberg (1999) 65
8.3. Appendix C: Post Experiment Survey on Environmental Orderliness Features 66
8.4. Appendix D: Orderliness Categories and Corresponding Observables 66
8.5. Appendix E: Flowcharts of Tidiness Observables in the Assessment 69
8.5.1. Observable: Put trash into the trash bin at the product location 69
8.5.2. Observable: Put broken cardboard into cardboard bin at the packing station 70
8.5.3. Observable: Put the broken tape in the trash bin at the packing station 70
8.6. Appendix F: Observation Notes in the Debriefing 71

Acknowledgement

At the end of any adventurous and exciting journey, one must look back on all the support that one has received and reciprocate with gratitude and appreciation, and that is my intention with this acknowledgement. Several individuals have been indispensable in this project, and to whom I must express my gratefulness. First of all, I would like to thank my supervisor Bas Kolloffel, who has guided me on this journey with patience, kindness, and wisdom, even at times of uncertainty. Ilona Friso-van den Bos, my second supervisor, offered me words of encouragement and reassurance whenever I needed them. They both gave me integral constructive feedback which is fundamental to the development of my thesis.

I would also like to thank the CEO of Serious VR, Marjo Nieuwenhuijse, who took a leap of faith in me to work on this project. Evelien Ydo, who is my supervisor in the company, has always provided me with valuable input and emotional support at the hardest moments. My thesis would have been way less enjoyable without you. VR developer Matthijs Baalhuis managed to build very realistic and immersive environments and interactable objects for the assessment. Luca Frösler made all the interactions in the application possible. Only with their help my vision and imagination for the application could be fulfilled.

Finally, I would like to thank my parents, Tobi, and whoever listened to me attentively about the project and offered me new perspectives and ideas. You gave me opportunities to reflect on the past and move forward into the future. I am thankful that you are involved in this chapter of my life.

Abbreviation List

The following table explains the abbreviations and the corresponding meanings used in this thesis.

Abbreviation	Meaning	Page
SA	stealth assessment	8
ECD	evidence-centred design	20
VSA	virtual stealth assessment	20
KSA	knowledge, skills, or ability	26
SMV	Student model variable	27

Abstract

Self-reported measures are the predominant method to assess job applicants' orderliness in the selection process. However, the validity of the measures has been questioned due to response bias and self-knowledge constraints. An alternative is stealth assessment, a type of embedded assessment in a game that evaluates the players' latent competencies. Combined with VR, a virtual stealth assessment can enhance ecological validity and provide substantial behavioural data. Apart from orderliness, environmental factors can also influence orderly behaviours. For individuals with different orderliness levels, their orderly behaviours can vary depending on the orderliness level of the environment they are in. Therefore, this study aimed to explore the relation between orderly performance in a virtual stealth assessment and self-reported orderliness and examine how environmental orderliness affects this relationship. To achieve this, 50 participants completed a survey on their perceived level of orderliness and underwent an assessment in a tidy and messy virtual environment. The results showed a weak relationship between self-reported orderliness and orderly performance in both environments and a moderate correlation in the tidy environment for highly orderly people. The findings indicated that the current assessment is correlated to an existing measure, and highly orderly people are more likely to behave according to their personality in a tidy virtual environment than a messy one. This study proposed advantages of the current assessment over self-reported measures, such as the enhancement of ecological validity and higher resistance to social desirability bias. This study established the foundation for future developments of virtual stealth assessments.

Keywords: stealth assessment, virtual reality, orderliness, personality, Big Five.

1. Introduction

Technological advances have drastically changed the way organisations operate on a daily basis. As a critical part of human resource management, the process of personnel selection is also in need of transformation and upgrade. Employee turnover has long been cited as a critical challenge in most organisations (Boswell, 2008). Not only does it result in direct financial costs, but it also causes indirect organisational loss, such as valuable knowledge departing employees take with them (Holtom et al., 2005), which can be detrimental to the long-term organisational development. One of the factors linked to employees' intention to leave is the degree of compatibility between an employee and the organisation they work for (Kristof-Brown et al., 2005). A mismatch can occur when personal characteristics and the work environment are not corresponding, or when there is a discrepancy between an individual's attributes and job demands. A misfit between an employee and the organisation can lead to negative influence, such as lower task performance and undesirable work attitudes (Kristof-Brown et al., 2005). Therefore, to reduce the risk of low compatibility between employees and organisations, human resource management professionals need to look for candidates with characteristics that correspond well to the organisation in the selection process (Kristof-Brown et al., 2005).

The practice of screening talents based on their compatibility with the characteristics of the organisation is prevalent in every industry, and the logistics industry is no exception. One of the most integral characteristics of a well-functioning warehouse is order, and this is evident in daily operations. Materials and products come and go constantly, and if they are not overseen properly, errors and clutter occur, which bring about direct operational loss (de Koster et al., 2007). To maintain the order of the warehouse, companies uphold high standards of the environment by applying various strategies, such as regularly monitoring the level of cleanliness (de Koster, 2012). However, the attributes of the people working in an environment are as crucial as the environment itself regarding maintaining the condition of a place. This is because an efficient and well-organized environment requires a collective effort of everyone working in the warehouse to sustain. A sign of disorder can lead to further chaos and disorganisation if not eliminated quickly (Vohs et al., 2013). Therefore, one of the most desirable qualities that a warehouse candidate can have is orderliness, which can be defined as the preference of order and keeping things tidy and clean. According to the work of Goldberg (1999), a highly orderly individual is more likely to follow routines, keep track of one's belongings, and tidy up the environment if they encounter clutter compared to low orderly ones. As such, orderliness is a beneficial trait to have when applying for a job in the logistics industry. Yet, despite the specific characteristics the logistics industry demands, there is a lack of assessment methods in the personnel selection process that can objectively evaluate a candidates' orderliness level.

In practice, the most common assessment method on orderliness is self-reported personality tests where candidates answer questions about their attitudes, behaviours, or feelings (Robins et al., 2007). Although explicit measures have some apparent benefits, they are often criticized for their disadvantages. The main limitations of self-reported measures can be divided into two categories, response bias and self-knowledge constraint (Robins et al., 2007). Studies have shown that self-reported results are prone to social desirability bias (Robins et al., 2007), and respondents are more likely to provide false answers when the test result involves personal gain, such as in the hiring process (Tett & Simonet, 2011). Furthermore, explicit measures require respondents to conduct a truthful and objective reflection on one's general behavioural patterns over time, but not every individual can do so and one's introspection capability might be influenced by memory limitation and the reaction towards test questions (Robins et al., 2007). This might cause workers who have low-level reading skills and reflective abilities to be put at a disadvantage, which can lead to bias in the test results by favouring those who have better reading comprehension and introspection skills. Apart from the inherent challenges of self-report, the easy accessibility of personality tests is also another source of worry. In this digital age, candidates can easily search online for tips on how to score high on personality tests and get access to answer keys. These challenges raise serious concerns for organisations that implement personality tests as a part of the personnel selection process. Nevertheless, self-reports remain the most popular assessment method in the field of personality assessment due to the lack of suitable alternative measures (Robins et al., 2007).

In light of all the challenges in the self-reported measure mentioned above, there is a need to explore a new approach that can be used to evaluate individuals' orderliness. Stealth assessment (SA) is a developing measure that evaluates an individual's knowledge, skills, and latent attributes without revealing the target competency being assessed (Shute, 2011). The covert nature of this method reduces test anxiety and overcomes the limitations of explicit measures while maintaining the validity and reliability of the assessment. The advent of virtual reality (VR) technology provides an opportunity to utilize SA to its full potential. There are many advantages of using VR as an assessment vehicle, according to Chicchi Giglioli et al. (2017). First, it provides a more realistic perception of how people perform in a similar context. With highly realistic simulations, VR can recreate authentic experiences and elicit genuine behaviours which are expected to be similar to those in real life. This increases the ecological validity of the test, which is a challenge that traditional assessment struggles to overcome (Kourtesis et al., 2021; Tarnanas et al., 2013). Additionally, VR allows us to collect an abundance of behavioural data that other tests fail to do, offering more information to make inferences about an individual.

Another beneficial attribute of VR is that it provides a controlled environment for simulated conditions, which allows researchers to implement prompts and features that might otherwise be too difficult to conduct in real life while maintaining experimental control (Alcañiz et al., 2018; Chicchi Giglioli et al., 2017; Parsons, 2015). This is a great advantage in designing a SA on personality because one must look beyond dispositional factors and take situational factors into consideration to gain a deeper understanding of how people behave. Individual behaviours can be attributed to many reasons, and environmental orderliness is one of them. Studies have shown that the orderliness of the environment can have an influence on people's behaviours and attitudes (Mateo et al., 2013; Vohs et al., 2013). Namely, a tidy environment can promote cleaning behaviours (Ramos & Torgler, 2012) and is associated with higher task accuracy (Mateo et al., 2013). A messy one, on the contrary, can trigger littering behaviours (Ramos & Torgler, 2012). The flexibility and adaptability of VR technology allow researchers to observe and compare behaviours in different environments. By gaining more insight on the effect of environment on behaviours, VR developers can keep the condition of environments in mind when designing a new training to develop a valid assessment of individual's attributes that can withstand environmental changes.

This study aims to develop a SA on orderliness in VR, examine the relation between orderly performance and self-reported level of orderliness, and explore the effect of environmental orderliness on the relation. In order to obtain a better understanding of the relationship between the current assessment and an existing measure, this study will compare the SA results with those of the personality survey IPIP-NEO, which is the acronym for "International Personality Item Pool – Neuroticism, Extraversion & Openness" (Goldberg, 1999). By investigating how individuals' orderly performance in VR relates to their self-reported orderliness, the outcome will give insight into the possibility of establishing VR as a personality assessment method. Consequently, companies can use this information to identify candidates' qualities and improve the efficiency of the selection process as a result. In addition, by studying how the relationship between self-reported orderliness and orderly performance varies in different environments, developers can use the findings to design a test environment that allows more accurate assessments of individuals’ attributes, such as orderliness.

2. Theoretical Framework

2.1. Orderliness as a Personality Facet

Orderliness is a personality facet that has a long history in the field of psychology, dating back to 1938. In Murray's list of psychogenic needs (1938), order is identified as one of the universal needs that drive or prompt an individual's behaviour. In his view, the inclination to make things clean and tidy is an innate motivation that everyone shares to a certain degree, and the varying levels of needs are what lead to different personalities. From a psychiatric perspective, however, orderly behaviours reflect one's mental state as an excessive desire for order can be seen as a form of obsessive-compulsive disorder, whereas extreme disorderliness can be associated with compulsive hoarding (Fenske & Schwenk, 2009). Nevertheless, Murray's theory (1938) became a basis of further personality research, and order remains to be a prevalent construct in later research. For instance, orderliness appears as one of the scales in the Interpersonal Style Inventory (Lorr & Youniss, 1973) under the cluster self-control. However, it was not until the advent of the Five-Factor Model that it received substantial attention.

Since its emergence in the 1980s, the Five-Factor Model has remained the most popular and widely accepted theory on personality in the field of psychology (McCrae & Costa, 1987). It is a taxonomy that categorizes personality into five traits: extraversion, neuroticism, openness to experience, agreeableness, and conscientiousness, alternatively known as the Big Five (McCrae & Costa, 1987). Each trait consists of multiple facets, which are constructed differently in various models. In the trait conscientiousness, orderliness appears to be a salient underlying facet that exists in all the models (MacCann et al., 2009). Although it is conceptualized differently by researchers, orderliness generally refers to the preference for organisation, order, tidiness, cleanliness, and a desire to keep the environment organized (Goldberg, 1999; MacCann et al., 2009). It is also associated with rules, routine, conservatism, and tradition (Xu et al., 2020).

In the International Personality Item Pool (IPIP), which is a database that stores over 3000 items that construct 250 personality scales, orderliness is associated with the inclination for order, material organisation, tidiness, precision to detail, and time management (Goldberg, 1999). A close examination of the items shows that inclination for order is manifested in behaviours such as following schedules or routines; material organisation is demonstrated by actions such as keeping track of one's belongings; tidiness is represented by the act of tidying up and not leaving a mess; precision to detail is exhibited by the urge for everything to be just right and to be precise in one's work; lastly, time management is conveyed by completing unpleasant tasks immediately and not putting off undesired chores. All 22 items in the orderliness measure are closely related, exhibiting a high internal consistency (.78) with similar predictive power to orderliness (Goldberg, 1999). It is therefore expected that when one shows a high tendency of one of the items, the individual is also likely to demonstrate a similar preference to the others. For example, someone that likes order, in general, is more inclined to tidy up and keep track of their belongings.

The 22 items on orderliness are constructed into multiple scales with slight variations, while inclination for order, material organisation, and tidiness are the overarching themes that can be identified in most scales. In IPIP-NEO, which is an IPIP version of a proprietary personality test NEO-PI-R, orderliness is constructed by the aforementioned themes (Goldberg, 1999).

In summary, there is no conclusive categorization of the personality aspect orderliness in the literature, but it can be concluded that orderliness is closely connected to one's attitude towards the order of the environment. In the current study, the working definition of orderliness is based on the common themes that can be found in most scales, which include an inclination for order, material organization and tidiness.

2.2. Environmental Orderliness

While designing an assessment that gathers an individual's behavioural data which interpretations on personality are based on, it is important to factor in the influence of environmental factors on human behaviours. Making a distinction between environmental orderliness and orderliness as a trait is an appropriate starting point. Despite the shared connection to order and disorder, environmental orderliness concerns the physical order in the surroundings while orderliness focuses on the individual preference for physical order. The concept of order and disorder is prevalent across multiple disciplines, and it is generally believed that the environment has an impact on our lives and how we behave (Vohs et al., 2013).

An overview of the literature that investigated the operationalization of environmental orderliness shows that it can be measured by actual cleanliness and perceived cleanliness. Actual cleanliness is an objective measure that is generally monitored by predetermined factors (e.g., the amount of litter, visible dirt, and stains) whereas perceived cleanliness is evaluated by the end-users' perception of cleanliness (van Ryzin et al., 2008). Whitehead et al. (2007) investigated perceived cleanliness in a hospital setting and concluded that patients connected it to the appearance of the environment (e.g., whether it was well-maintained and well-lit), physical cleanliness (e.g., the actual cleanliness of the toilets and beds), and staff's behaviours (e.g., behaviours related to personal hygiene and the tidiness of their uniforms). In addition, other factors that affect the perception of cleanliness have been identified in other studies. Wells and Daunt (2015) discovered that the condition of the environment also has an impact on cleanliness perception. Specifically, when comparing old and new lecture rooms, the older and less attractive environments are perceived as less clean even though their levels of actual cleanliness are considered similar.

To conclude, environmental orderliness is distinctive from orderliness as a personality aspect. Its operationalisation is well-documented, and its relationship with human behaviours has been discussed in prior literature, which will be explained further in the next section.

2.3. Environmental Orderliness and Orderly Behaviours

Understanding how environments relate to orderly behaviours is an essential question for behavioural research and environmental studies. There is a sizable amount of empirical research that shows the link between environmental orderliness and orderly behaviours. Specifically, the orderliness of an environment can influence cleaning behaviours (Ramos & Torgler, 2012), which can be seen as the characterisations of orderliness, namely, tidiness and material organisation. The reason for this is that material organisation can be interpreted as a manifestation of tidiness, since someone that prefers tidiness should also take good care of personal objects so as not to cause clutter in the future. The relationship between environment and cleaning and littering behaviours has been well-discussed in literature. Ramos and Torgler (2012) explored the effect of disorder in the setting of the university, and they discovered that it is more likely for the staff to litter in a disordered room compared to an orderly room. Another environmental factor that influences human behaviours revolves around trash cans. Namely, the availability of trash bins, the distance between the subject and the trash cans, and even the design of the trash cans play a role in reducing littering behaviours. Arafat et al. (2007), Bator et al.

(2011), and Schultz et al. (2013) found that increasing the availability, decreasing the distance to the trash cans, and implementing more persuasive trash can design can lead to the improvement of the actual cleanliness. Furthermore, Kallgren et al. (2000) discussed that information strategies such as flyers, posters, signs, and banners can also contribute to reducing littering because they evoke external pressure.

Looking from a different perspective, the implication of littering behaviours goes far beyond a simple act of transgression. There is empirical evidence showing that litter was related to crime rates in neighbourhoods (Brown et al., 2004), and the presence of litter was associated with an increase in theft (Keizer et al., 2008). Hence, littering behaviours are generally strongly connected with disobeying the rules. Furthermore, an orderly environment is often associated with following the rules and adhering to conventions (Vohs et al., 2013). These concepts are strongly related to another characteristic of orderliness, which is the inclination for order. Compliance with rules and orders is a necessary component for the correct execution of skills and the decrease of mistakes (Dunham et al., 2020). The research of Mateo et al. (2013) provides great relevance to the current study as they discovered that highly conscientious individuals committed far more errors in a messy environment than in a tidy environment, while the task accuracy of low conscientious people remained similar in different environments. The errors highly conscientious individuals made even exceeded those of low conscientious people in the messy environment. This suggests that physical disorder is especially detrimental to the performance of high conscientious individuals while leaving little impact on less conscientious people. The findings can be attributed to the low person-organisation fit, which is a mismatch between employees' personal characteristics and organisation (Kristof-Brown et al., 2005). While conscientious individuals can be categorized as orderly, responsible, industrious, reliable, and having a preference for order and organisation (MacCann et al., 2009), once they are placed in an environment with clutter and signs of disorganisation, a disassociation between their personality trait (conscientiousness) and organisation (working environment) occurs, which can have a damaging effect on performance and attitudes. Their findings demonstrate that environment can be a moderator in the relationship between conscientious individuals and human accuracy.

To conclude, environmental orderliness is likely to have an effect on individuals' orderly behaviours due to the reasons mentioned. Behaviours related to orderliness are likely to be affected by both the individual's personality trait and the environment he/she is in. More importantly, based on the person-organization fit theory (Kristof-Brown et al., 2005), highly orderly people are expected to perform in a manner that matches their orderliness in a clean environment compared to a messy environment. For low orderly people, however, the environment has little impact on the relationship between their performance and their orderliness, so the relationships in the two environments stay similar.

2.4. Explicit Measures of Personality

As previously mentioned, environmental orderliness can be observed and measured with perceived and actual cleanliness. By contrast, personality facets such as orderliness are hard to measure due to their latent nature. For this reason, psychologists have been seeking different ways to assess personalities ever since the birth of personality psychology (Robins et al., 2007). The existing measures can be categorized into explicit and implicit. The former refers to the type of tests that directly ask the respondents to provide a subjective self-assessment of their attributes, while the central character of implicit measures is assessing psychological traits in a way that does not involve the introspection of the respondents (Gawronski & De Houwer, 2014). In psychological testing, self-report is the predominant assessment method on personality traits which requires individuals to answer questions about their behaviours, feelings, or attitudes. Robins et al. (2007) pointed out that there are four main persuasive advantages of using a self-reported measure to assess personality. First, self-reports usually contain a wealth of information about an individual that other people do not have access to. They reflect a part of respondents' identity and their perceptions of themselves, which constitutes a part of their personality. By understanding their self-representations, more information on their personalities can be gathered. Self-reported methods also benefited from the motivation of respondents. They are expected to spend more time and effort answering when the test relates to their own personalities, which leads to higher validity. Another advantage self-report has is the high practicality which makes it a popular choice among researchers. They are efficient and cost-effective, require few resources, and can be administered to a large group of participants simultaneously. In comparison, other data collection methods such as interviews and observations have more demands and specifications. Finally, the easy interpretability of the data is also an advantage. Self-report consists of the common language that is understood by the assessors and the respondents, which is less likely to cause ambiguity. In contrast, quantitative behavioural data are subject to different interpretations, which adds more obscurity (Robins et al., 2007).

Despite many advantages, self-report has been criticized for suffering from some inherent limitations. An overarching concern is the credibility of self-report, namely, should we believe that people are telling the truth? The validity issue boils down to two main challenges: response bias and self-knowledge constraints. The first one, response bias, is a phenomenon where respondents answer the questions falsely depending on the context and topic. There are different types of response bias that self-report is susceptible to, including extreme responding (only selecting extreme answers), acquiescent responding (agreeing with most questions), and social desirability bias (Gawronski & De Houwer, 2014; Robins et al., 2007). Social desirability, as one of the most prominent response biases, is an inclination to attribute socially acceptable traits to oneself and refuse to acknowledge any of their negative traits. This bias takes many forms, including exaggerating positive traits and under-reporting undesirable behaviours. It is worth noting that the motivation for faking answers is heightened when personal interests are at stake (Tett & Simonet, 2011). This concern is particularly alarming when self-report tests are administered in the personnel selection setting. Ziegler et al. (2011) pointed out that although it remains inconclusive that respondents would fake personality tests in a high-stake context such as in a hiring process, it is evident that individuals can do so. Previous research has resulted in rival arguments with opposing evidence on the issue of whether individuals false report their answers in a motivational setting. However, when individuals are aware of their capability to fake and perceive faking as necessary to achieve a valuable goal, then faking is more likely to occur (Ziegler et al., 2011). The severity of this issue has only enhanced in this digital era where personality tests can be easily accessible online along with answer keys (Goldberg, 1999). When books and articles that specifically coach individuals on how to cheat on the tests are available for everyone (Whyte, 1956), it raises serious concerns on the validity of the tests not only for researchers but also organisations that filter job applicants with self-reported methods.

Although it is often assumed that honest answers are sufficient to generate an accurate presentation of one’s traits, the constraint of self-knowledge could still keep researchers from obtaining the truth of individuals even if they are candid in their answers. Self-reported measures require one to go through the process of introspection, and not everyone is capable of doing that (Gawronski & De Houwer, 2014). People might fail to accurately recall information or past experiences related to testing questions due to limitations of memory. They could find the questions too challenging, feel overwhelmed in the process, and end up providing unreliable answers (Robins et al., 2007).

To sum up, despite many perks of explicit measures on personality, this method has some problematic drawbacks that cast doubt on its validity. Even though it remains the most widely used measurement tool for personality applied in the selection process, it should be utilized with caution.

2.5. Implicit Measures of Personality

In light of the limitations of explicit measures, researchers have been exploring an alternative method that does not require the introspective ability of the respondents and is less susceptible to response bias. In contrast to explicit measures, the main characteristic of implicit measures is that it assesses psychological attributes, such as attitudes, feelings, or personalities, without demanding self-knowledge or self-representation of the individuals, which prevents validity issues that explicit measures suffer from (Nosek et al., 2007). A wide variety of implicit measures have been developed, which includes implicit association test (IAT), evaluative priming task, affect misattribution procedure, and implicit relational assessment procedure (Gawronski & De Houwer, 2014). Among them, IAT by Greenwald et al. (1998) is one of the best-known implicit measure tools and has been used to measure subconscious including stereotypes, self-esteem, and attitudes (Greenwald & Banaji, 1995), which can be challenging to assess with direct measures. The test involves a series of classification tasks, and it measures the strength of the association with response time in classifying stimuli into different categories (Grumm & von Collani, 2007). The underlying assumption of the test is that it takes less response time when reacting to groups of concepts that are more strongly associated than the ones that have a weaker association. For example, Grumm and von Collani (2007) conducted an IAT on the personality traits in the Five-Factor Model. The test included three categories, self, others, and attribute. The former one consisted of 6 stimuli with the respondents' personal characteristics, such as family name and gender, whereas the "other" category involved the same stimuli but with characteristics of another person (different family name and gender). The attribute category included adjectives items corresponding to the five personality dimensions. Respondents were asked to respond to multiple combinations of the stimuli (e.g., Extraversion + self, introversion + other) in different categories by pressing two keys on the screen, and their response times were recorded to make inferences on the strength of the associations between stimuli.

Despite the popularity of implicit measures, there seems to be a common disassociation between explicit and implicit measures. Some evidence suggests that explicit and implicit measures of the same construct would correlate, as studies have found significant correlations between implicit association tests and validated questionnaires (Friedman et al., 2021; Hofmann et al., 2005). On the other hand, other studies produced mixed results or failed to find a meaningful relationship between the two measures (Anderson, 2009; Karpinski, 2004). This poses a challenge on establishing convergent validity of the instrument and begs the question, “What exactly are implicit measures measuring?”. Some researchers assumed that explicit measures assess the conscious part of self-presentation while implicit measures tap into the unconscious reflection of oneself (Greenwald & Banaji, 1995), yet this perspective is far from conclusive. Interestingly, in the study of Back et al. (2009), both explicit and implicit measures on personality traits partly predicted actual behaviours. This might suggest that both measures are independent of each other, and they are linked to a part of human behaviours in different ways.

Although implicit measures were developed to overcome the limitations of response factors and self-knowledge constraints in external measures, the main problem in traditional psychological measurements remains. Over the years, researchers have discussed the issue of ecological validity as the results of psychological measurements often failed to be generalized to a real-life setting (Parsons, 2015). However, methods with high ecological validity often lack experimental control and internal validity (Parsons, 2015). This brought about debates over the importance of naturalistic and controlled assessment settings and which one demands more emphasis. Therefore, striking a balance between the two has been the goal of many researchers over the years. One of the methods proposed in the literature that can offer a natural situation while maintaining experimental control is SA, which has the potential to increase the ecological validity of a test with the support of VR technology (Chicchi Giglioli et al., 2017).

2.6. Stealth Assessment as an Alternative Measure

In addition to explicit and implicit measures, SA has gathered considerable academic attention in the past decade, and a growing number of studies supporting its validity and reliability has emerged (Shute, 2011). The term was coined by researcher Shute to describe a type of embedded assessment that is seamlessly woven into a digital game environment to evaluate learners' competencies, knowledge, and skills (Shute, 2011). The defining feature of SA is to examine in-game behaviours while the players are unaware of the competency being assessed. Players' performance data and behavioural information are continuously collected by the algorithm in the game and updated in real-time so that interpretations on relevant competencies can be made. In addition, the assessment is meticulously designed to maintain the flow of learning, where players are so focused on the goal of the task that they are motivated not by extrinsic awards, but inner motivation. This is what is called the "optimal learning state" where self-consciousness disappears, and players are completely immersed in the game, even to the extent where they lose track of time.

SA provides an environment that induces behaviours related to implicit attributes which can be difficult to assess in traditional standardized settings. It has been proposed to be one of the most auspicious methods to measure complicated soft skills, and various video games have been developed to examine various competencies to date. For example, Use Your Brainz was designed to measure middle school students' problem-solving skills (Shute et al., 2016), and systems thinking in Taiga Park (Shute, 2011). In Physicals playground, the personality facet persistence was evaluated (Shute et al., 2013; Ventura et al., 2013), along with cognitive skills such as creativity (Shute & Rahimi, 2021) and qualitative physics knowledge (Shute et al., 2013).

SA poses several strengths over explicit and implicit measures regarding assessing personalities. Similar to the advantages of implicit measures, SA is able to tackle the issues of response bias and self-knowledge constraints because it is not based on the introspection ability of the test-takers. The inconspicuous nature of the assessment allows for lower susceptibility to social desirability bias since the test-takers are not assessed by their response but their actions (Moore & Shute, 2017). The other strength is its higher resistance to cheating or fake answers due to the lack of information on the target competency and low accessibility. Considering the blurred distinction between learning and assessment in SA, it is assumed to be more difficult for test-takers to fake their actions or cheat since they are not certain if they are being assessed at all. Also, because SA is often developed specifically for a certain competency and is not easily replicated (Moore & Shute, 2017), it is less accessible to the general public. This prevents test-takers from acquiring tips and tricks for better scores beforehand, which might otherwise result in misconceptions and misjudgements. Finally, active engagement in the game is expected to reduce test anxiety while maintaining the validity and reliability of the assessment (Shute, 2011).

Despite its advantages, it is equally important to address the limitations of SA. One of the biggest potential challenges when using this method is that it is not suitable to evaluate all non-cognitive competencies (Moore & Shute, 2017). Some latent constructs are not appropriate or almost impossible to measure in a single game session (e.g., dependability) and researchers have to make sure the testing environment is incongruous with the target competency (Moore & Shute, 2017). Other limitations include the time-consuming process that involves experts from different disciplines. It requires educational scientists to review the literature and develop an assessment based on the evidence-centred design (ECD; Mislevy et al., 2003), psychometricians to refine the measurement instrument, and game developers to design the digital game. Therefore, it is recommended to determine the amount of time and resources that can be employed to spend on the project before commencing (Moore & Shute, 2017). Another potential obstacle lies in the technical aspect of the development. Namely, as continuously collecting data during the game constitutes a defining feature of SA, it is crucial to ensure data collection can be embedded in the game. If this cannot be achieved, then SA might not be an appropriate instrument (Moore & Shute, 2017).

2.7. Potential of Stealth Assessment in VR

To date, the application of SA is mainly confined to computer-based games. However, the advancement of technology in recent years makes it possible to use this assessment approach in other digital environments that can optimize its potential. VR for instance, can be a suitable vehicle for conducting SA. Chicchi Giglioli et al. (2017) introduced the term “virtual stealth assessment” (VSA), and they proposed several reasons why the integration of SA into VR can be advantageous. The first one is that VR can enhance ecological validity by providing a sense of presence and immersion and allowing participants to be fully immersed in the simulations. The two important concepts of immersion and presence are explained in the paper of Slater and Wilbur (1997). Immersion is associated with the objective technical ability of VR to deliver an environment that closely resembles reality. It is closely connected to visual and auditory senses, and what is heard and seen are crucial for building immersion. Presence, on the other hand, refers to the psychological state the user experiences while being in the VR. In Slater and Wilbur’s viewpoint (1997), it is a subjective response to the simulation, and it is established when the user becomes conscious of the environment they are in, even to the point where they forget they are in a virtual setting. They believe that for individuals who experience a high level of presence, their behaviours in VR are expected to be consistent with those which occur in a similar real-life circumstance (Slater & Wilbur, 1997). By establishing immersion and presence, VR is able to overcome one of the main shortcomings of traditional assessment, which is the lack of ecological validity. Ecological validity is the extent to which an assessment can be generalized to real-life contexts (Parsons, 2015), and it is often achieved with the verisimilitude approach. Verisimilitude refers to the similarity between the cognitive demands of the task in the assessment with those in a task in real-life, and studies have employed VR simulations to increase verisimilitude in the hopes of enhancing the ecological validity of assessments on cognitive functions (Kourtesis et al., 2021; Tarnanas et al., 2013). This is because even though traditional assessments show great validity, the test performance often does not have sufficient predictive power for real-life behaviours due to the great inherent difference between the measures and realistic situations (Kourtesis et al., 2021). In VR, however, the highly realistic environment and sense of being are sufficient to simulate real experiences and elicit authentic behaviours, providing a realistic perception of how an individual performs in a similar context (Pizzi et al., 2019; Xu et al., 2021). Hence, using SA in VR is expected to generate more authentic behavioural data due to the increase of ecological validity.

Another advantage discussed in Chicchi Giglioli et al. (2017) is that VR can collect a wide range of direct behaviours when traditional assessments fail to do so. Behavioural data (e.g., motion tracking, eye-tracking) and physiological data (e.g., heart rate, skin conductance level) can be tracked to offer a powerful source of information (Weibel et al., 2018). A few studies on consumer behaviour research have shown that behaviours in VR are comparable and similar to real life (Pizzi et al., 2019; Xu et al., 2021). The large quantity of potential data can be used to make more inferences in SA, which presents more opportunities in the assessment field.

The third benefit is that VR provides a controlled environment that allows naturalistic observations for simulated real-life events (Parsons, 2015). Observations of human reactions in a real-world setting have been proven to be difficult without the ability to control the stimuli involved, and they are usually labour intensive and costly. In contrast, VR can achieve high realism without sacrificing experimental control, which allows researchers to embed stimuli that are difficult to implement in real life to elicit different responses (Chicchi Giglioli et al., 2017). For instance, variables such as environments of different degrees of orderliness and stimuli of orderly behaviours can easily be manipulated in VR in comparison to reality.

In conclusion, there are many opportunities the combination of VR and SA presents. Higher ecological validity, a larger amount of behavioural data, and an experimentally controlled environment can lay the foundation of an innovative collaboration of two prospective tools. However, despite the potential benefits that have been listed, the application of VR in assessments is severely understudied, and there is no empirical evidence of SA in the context of VR to my knowledge. Therefore, more investigation is needed to gain more insights into the limitations and challenges of VSAs, and to confirm the presumed advantages and opportunities it leads to.

3. Research Questions & Hypotheses

The goal of this study is to explore how orderly performance in VSA correlate to self-reported orderliness while examining the potential effect of environmental orderliness on the relationship between orderly performance in VSA and self-reported orderliness. The following research questions guide this study:

R1: How does orderly performance in VSA relate to self-reported orderliness?

R2: How does environmental orderliness affect the relation between orderly performance in VSA and self-reported orderliness?

Three hypotheses were made based on the research questions. The first hypothesis is related to the first question.

H1: Orderly performance in VSA is related to self-reported orderliness in both tidy and messy environments.

In addition, environmental orderliness is expected to have different influences on the relation between orderly performance in VSA and self-reported orderliness for different orderly groups. This projection was made based on the research of Mateo et al. (2013) and the person-organization theory by Kristof-Brown et al. (2005). The rationale is that when highly orderly individuals are in a tidy environment that fits their personality, it allows them to perform in a way that matches their inherent orderliness, hence the relation between their performance and self-reported orderliness would be stronger than in a messy environment. For low orderly individuals, however, as was mentioned in the work of Mateo et al. (2013), their performance was less susceptible to the tidiness of the environment. As such, the relation between how they performed and their orderliness scores should remain unaffected. Therefore, an interaction effect is expected between environmental orderliness and self-reported orderliness on orderly performance.

Based on the above arguments, the following hypotheses were formulated:

H2: The relation between orderly performance in VSA and self-reported orderliness for the highly orderly group is stronger in the tidy environment than in the messy one.

H3: The relation between orderly performance in VSA and self-reported orderliness for the low orderly group is similar in the tidy environment and the messy environment.

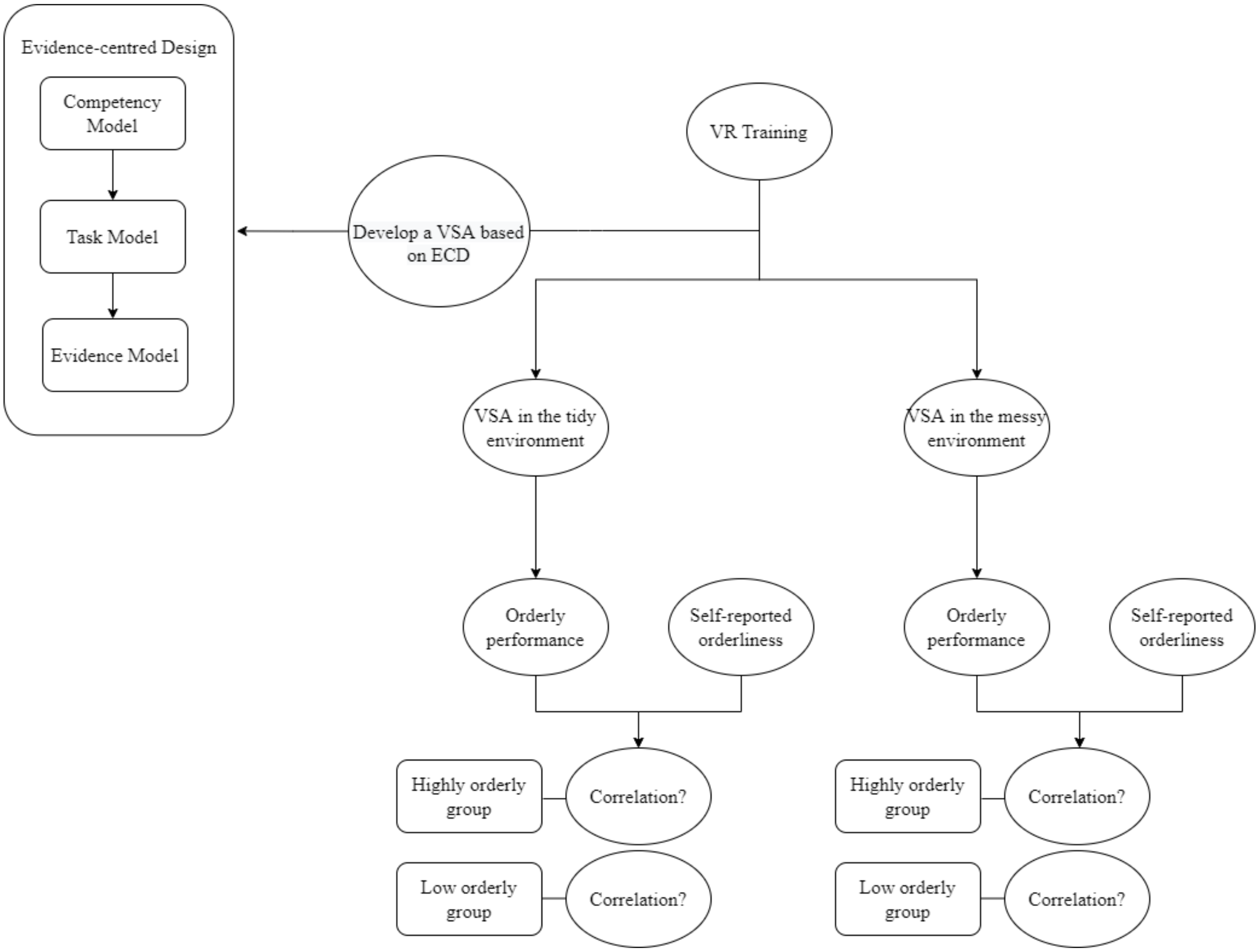
4. Method

4.1. Research design

The research design entailed a repeated measures design with self-reported orderliness as the independent variable and orderly performance in the VSA as the dependent variable (see Figure 1). Environmental orderliness functioned as an interaction variable that might influence the relationship between orderly performance and self-reported orderliness for different orderly groups.

Figure 1

Overview of the Research Design



Self-reported orderliness was evaluated by the questionnaire IPIP-NEO which was developed by Goldberg (1999). Orderly performance was assessed by the observable variables in the assessment. There were two levels of environmental orderliness, one was the tidy condition, and the other was the messy condition, which were developed accordingly in the VR environment. Specifically, a tidy warehouse with visible organization and a messy warehouse with physical clutter. In the VR training, participants went through the instruction and scenario phases in a moderately tidy environment and then conducted the assessment in both tidy and messy conditions. To control order effects, a counterbalancing design was implemented. Half of the participants underwent the assessment in the tidy condition first and then the messy condition. The other half of the participants went through the messy condition first followed by the tidy condition.

4.2. Participants

50 participants were recruited based on convenience sampling. The sample included 33 females, 16 males, and one non-binary person. Participants came from diverse cultural backgrounds with a variety of nationalities. The average age of the participants was 21.58 (*SD* = 3.08), with the lowest being 21 and the highest being 29. The group included 37 bachelor’s students, 12 master’s students, and two teachers at the University of Twente. Among them, 33 people have a high school degree, 14 have a bachelor’s degree or equivalent, one has a master’s degree, and one has a doctorate. The majority of the participants did not have experience in VR or the order-picking task. The experiments were conducted in the BMS lab at the University of Twente. For all experiments, participants with cold symptoms or a history of epilepsy and seizures were not permitted to participate.

4.3. Instrumentation

4.3.1. Orderliness Questionnaire (IPIP-NEO)

Self-reported orderliness was measured by ten items on the facet orderliness under the trait Conscientiousness from IPIP-NEO (Goldberg, 1999). The complete survey is shown in Appendix B. Each item is a description of behavioural tendency, and the central question was “How accurately does each statement describe you?” The responses are given on a 5-point Likert scale, where "Inaccurate" is assigned a value of 1, "Moderately Inaccurate" a 2, "Neither Inaccurate nor Accurate" a 3, "Moderately Accurate" a 4, and "Accurate" a 5. Three out of ten items are reverse-coded. Examples of the items on orderliness are: Like to tidy up. Leave a mess in my room (reverse-coded). The survey also included eight statements about other aspects of conscientiousness as filler items, including five items on self-efficacy, two items on achievement-striving, and one item on self-discipline. Incorporating statements related to different facets was assumed to prevent participants from detecting that orderliness was the target competency in this study, which had the potential to affect their performance in the virtual assessment.

The self-reported orderliness questionnaire showed good internal consistency (α = .81), and the structure of the instrument was confirmed by factor analysis. The total score of self-reported orderliness was the sum of all item values. The rationale of the use of IPIP-NEO survey is in Appendix A.

4.3.2. VSA on Orderliness

The virtual assessment on orderliness was developed in collaboration with the company Serious VR. The assessment was embedded in the order-picking training which was built in the development platform Unity. It entailed a room-scale experience where participants could move freely within the play area (1.5*2 square metres) while their movements were reflected in VR in real-time. Motion-tracked and hand-held controllers allowed participants to use their natural movements when completing the task, such as grabbing and moving objects in VR. Participants could teleport to the next station by clicking on the handles of the tool cart. The order-picking training application involved three phases, instruction, scenario, and assessment. The purpose of the first two phases was for trainees to learn and practice the procedure in the task, and the VSA only lied in the assessment phase.

4.3.2.1. Evidence-centred design.

SA can be developed based on the ECD model by Mislevy et al. (2003). The underlying premise of this framework is that assessment is a chain of evidentiary reasoning that connects test items with specific claims about one’s performance. In other words, evidence collected from the assessments can be analysed to make inferences on individuals’ knowledge, skills, and ability. An integral part of ECD is the conceptual assessment framework which outlines the operational details of assessment development. It is comprised of 5 models: student model, evidence model, task model, assembly model, and presentation model (see Figure 2). The first three form the core of this framework. First, the student model specifies the knowledge, skills, or ability (KSA) that the assessment aims to evaluate. In the context of ECD, KSA is also known as the student model variable (SMV). SMV are generally the target latent competencies that are not directly observable, and they function as the blueprint for this conceptual framework. There could be a single variable that represents the overall performance or multiple variables reflecting different aspects of SMV.

After defining the main SMV in the assessment, the task model is needed to design the assessment environment in a way that it can instigate evidence mirroring the target competency. Two types of information are generated in this model. First, it defines the form of the final product or the environment the assessment takes place in. This could be on a website, in a VR environment, or a videogame, etc. Second, it entails a list of key features and indicators in the assessment, such as instruction, levels, help functions, tools, and feedback system. Each feature is a task model variable, and each variable plays a part in the assessment to provide opportunities for students to demonstrate their competency that can be measured at a later stage. They can also be used to contribute to the assessment design in several ways: controlling task difficulty, concentrating evidence on a certain aspect of SMV, supporting task designing, and guiding test assembly (Mislevy et al., 2003).

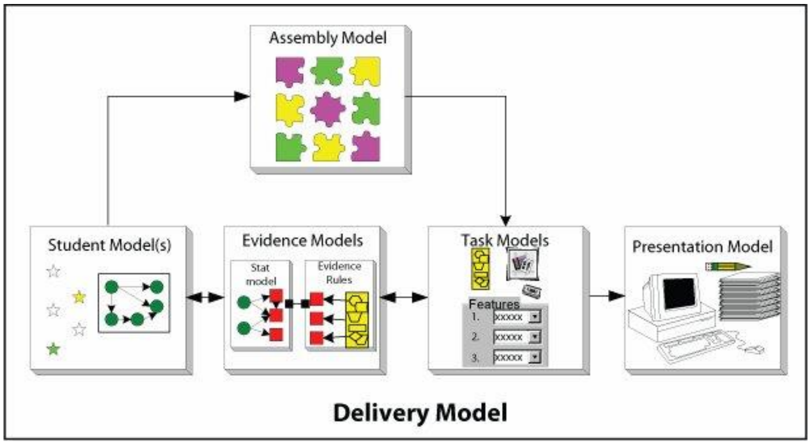
Next, the evidence model explicates the evidentiary argument behind the assessment and defines what constitutes the evidence of the SMV. It acts as the bridge between the student model and the task model by connecting student actions and their competency. There are two parts to this model. The first one is the evaluation component, which can also be called task-level scoring. It defines the rules on evaluating student actions and converting them into observable variables. More specifically, the rules examine student actions in the assessment and decide which value should be assigned to each observable variable concerning its importance in the SMV. Task scoring can be done with an automated algorithm, manual judgement, or a combination of both. The second part of this model is the measurement component, which can be compared to test scoring. It involves the psychometric model that transforms observations into inferences, or from a statistical perspective, from observable variables to SMVs (Mislevy et al., 2003). This component explains the relation between observable variables and SMV in probabilistic terms. The values of observable variables are used to update the inference of SMV.

The other two important models in ECD are the assembly model and the presentation model. The assembly model entails how the student model, task model, and evidence model work together in the assessment. It describes the strategies for choosing the mix of tasks that are used to reflect the student's competency. This is where test designers evaluate the accumulation of evidence and determine which tasks are selected for collecting evidence to constitute the SMV. This model also determines the structure of the test and establishes the conditions necessary for building the validity and reliability of the assessment.

Finally, the presentation model specifies the delivery environment where the assessment is presented, along with organisational details such as how tasks are scheduled and how students interact with the assessment.

Figure 2

The Conceptual Assessment Framework



Note. From "A Brief Introduction to Evidence-Centered Design", by R. Mislevy, R. R. G. Almond, and J. F. J. Lukas, 2003, *CSE Report 632*, p. 6 (https://doi.org/10.1002/j.2333-8504.2003.tb01908.x). Copyright 2003 by The Regents of the University of California.

4.3.2.2. Developing a VSA with ECD.

The VSA in this current study was designed based on ECD (Mislevy et al., 2003) and the development was guided by the core five models in the framework.

First, orderliness was confirmed to be the SMV. The conceptualization of orderliness from IPIP-NEO was selected as the basis of the student model. Based on the items in IPIP-NEO, orderliness was categorized into an inclination for order, material organization, and tidiness, and each reflected a particular aspect of the facet. Then, key features and indicators in the assessment environments were decided. They entailed a list of environmental orderliness features based on the literature on perceived cleanliness (Wells & Daunt, 2015; Whitehead et al., 2007). Important functions in the assessment were also determined in this phase. It included a dashboard that gathered behavioural data in real-time and provided an overview of orderliness performance, correct actions, and mistakes immediately after the assessment. Moreover, features such as feedback and a support system were also established. When an incorrect box was selected, a dialogue box with a warning sign would appear and suggest that a different box size should be chosen. In addition, a warning icon would appear on the scanner when the wrong product location was scanned, indicating a mistake was made in the scanning process. The help system was put in place so that when the participants got stuck on a certain step, they could press a button on the controller and request help. As a result, highlights on the objects would then show up to signal the correct next step.

Next, in the evidence model, observables were designed by drawing inspiration from the behavioural statements in the IPIP-NEO questionnaire on orderliness. Each observable was a potential mistake an individual could make that was closely linked to a certain category of orderliness (see Appendix D). Concerning one’s inclination for order, the corresponding observables consisted of mistakes of doing the steps in the wrong order and failing to complete the steps. The underlying assumption is that when an individual is capable of completing a task perfectly, then the task performance is considered one’s willingness to follow instructions and it can be seen as the manifestation of one’s inclination for order. As such, the mistakes one makes symbolize the preference of disorder and disobedience to the instructions.

The observables for material organization contained mistakes related to keeping track of the objects in the training, such as not placing the tools back to the original location and not bringing necessary objects to the next station. For example, to measure if someone put the tools back to the original location, four available spots on the tool cart were set up, and only one was the correct location. Lastly, one’s tidiness was represented by the reactions towards trash at different stations.

In total, three pieces of interactable trash were embedded at two stations in the assessment phase. A paper ball was put on the correct product at the product station (see Figure 3), and there was a piece of broken cardboard and a broken tape on the conveyor at the packing station (see Figure 4).

Figure 3

A Paper Ball on the Product at the Product Station

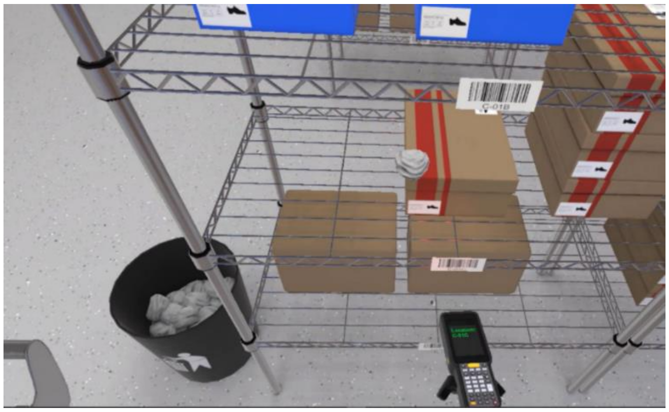
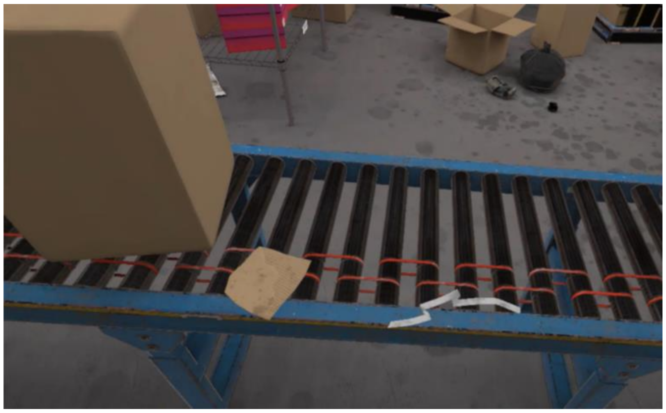


Figure 4

A Broken Cardboard and A Broken Tape on the Conveyor at the Packing Station



Participants could interact with the trash in the same way as they would in real life. They could pick it up, throw it on the floor, throw it in the bin, pick it up from the floor, or place it on any object. The interactions with the trash were then categorized into “left the trash untouched” or “grabbed the trash but did not put it in the bin”. Although the result of these two reactions was the same, which was that the participant did not put the trash into the bin, the differentiation between these two behaviours could potentially lead to more insight. As an example, if someone did not touch the trash, perhaps it was because he did not notice it at all. In contrast, if someone grabbed the trash, then it meant he did notice the trash but chose not to put it in the bin. One could argue the latter reaction could be a more revealing demonstration of one's untidiness[1].

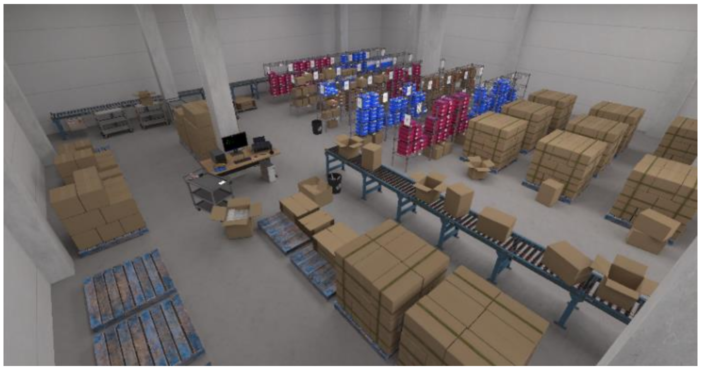
In addition, all trash was deliberately placed in the participant's working area without obstructing the task. In other words, even if participants did not interact with the trash, they could continue with the task and their performance would not be affected in any way. The choice of reacting to the trash entirely depended on each participant without any additional incentive or deterrent.

4.3.2.3. Virtual assessment environment.

There were three environments in the training in total. The environment in the first two phases, the instruction and scenario, had a moderate level of orderliness (see Figure 5). This was an intentional decision because the similarity between the environment where the instruction is learnt and where the assessment is conducted is assumed to have an influence on the task performance. There is a chance that learning in a clean environment results in different performances in a clean warehouse compared to a messy one. Therefore, to avoid biases against either the tidy or messy environment, the first two phases in the training were conducted in a moderately tidy warehouse. There were some boxes and cardboard lying on the floor and the conveyor, but they did not look crowded. Products on the shelf were not perfectly aligned.

Figure 5

The Moderately Messy Warehouse

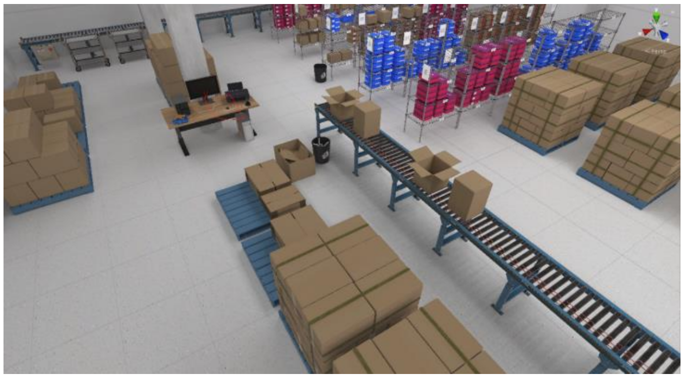


[1] Mann-Whitney U tests were conducted to see if people who grabbed the trash and did not put it in the bin scored significantly differently on self-reported orderliness compared to people who did not commit this mistake, and the results were non-significant.

To explore the influence of environmental orderliness, a tidy and messy warehouse environment were developed for the assessment phase. The design features incorporated to distinguish the different levels of orderliness were based on previous studies on perceived cleanliness (Wells & Daunt, 2015; Whitehead et al., 2007). Specifically, in the tidy environment (see Figure 6), the warehouse was constructed in an orderly manner. There was no visible clutter, and the environment looked well-organised. The walls, pillars, and floor looked shiny, clean, and new. No excess objects were on the ground. In the product location, the products were well-aligned on the shelf.

Figure 6

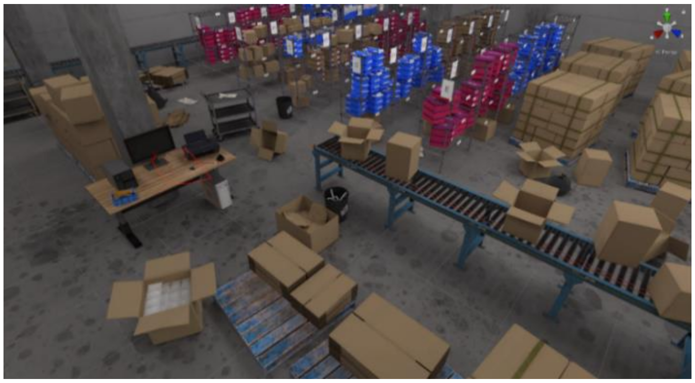
The Clean Warehouse



On the contrary, the warehouse in the messy environment had many excessive cardboards, boxes, and trash lying on the floor (see Figure 7). Walls, pillars, and the floor looked dirty, dull, and old. Products on the shelves were not well-aligned. It is worth noting that the orderliness features in the environment should not affect the task performance of the participants, and they did not hinder or facilitate the task that was required to complete.

Figure 7

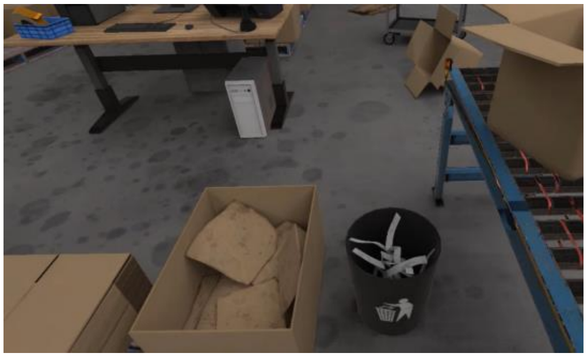
The Messy Warehouse



The common features in all environments included a trash bin filled with paper balls at the product station (see Figure 3), a cardboard box filled with cardboards and a bin filled with broken tapes in the packing station (see Figure 8).

Figure 8

Cardboard Bin and Tape Bin at the Packing Station



It was important to include the bins in all environments because participants might be alerted when they noticed the inconsistency of the objects. The sudden appearance of the bins in the assessment phase could generate suspicion of the target competency of the experiment and consequently trigger their cleaning behaviours. As such, in all environments, the bins were placed in the same locations. In addition, all the bins were placed within close proximity to the corresponding trash and the participants. This was deliberate as the accessibility of the trash cans could be an influencing factor in cleaning behaviours according to previous literature (Arafat et al., 2007; Bator et al., 2011; Schultz et al., 2013). Providing easy access to the trash cans could potentially prevent participants from being too lazy to throw the trash away.

Additionally, all the trash bins were filled with corresponding trash which looked almost identical to the interactable trash the participants could react to. For instance, the cardboard bin was full of broken cardboards and the tape bin was filled with broken tapes. This was intended to demonstrate that different types of trash belonged to different bins, while implicitly assuring the participants that the interactable trash was something that could be thrown away. This feature could decrease the chance of someone wanting to throw away the trash but did not because he was not sure if it was something that should be discarded.

4.3.3. Post Experiment Survey

To investigate participants' perceptions on the manipulation of environmental orderliness, a post-experiment survey was designed. One question was included, which was "What did you notice in the last two sessions in the scenario phase?" Eight environmental features implemented in the tidy and messy environments were listed as options. For example, the floor was clean. The survey also provided an option to write down additional environmental features that were not mentioned in either environment. The complete survey is shown in Appendix C.

4.4. Procedure

This study was approved by the BMS Ethics committee before the experiments, and the application number is 210842. The procedure was as followed. First, participants were asked to read two documents, which were two pages of information on the study and one page of a consent form. The consent form included information on the physical risks of participating in the experiment. It was also elaborated that if they experienced any nausea or motion sickness during the experiment, they should let the researcher know immediately because if one did not remove from the VR environment quickly, the discomfort might prolong for a few hours. It was made clear that their involvement was voluntary, and they could discontinue the experiment at any time. Then, participants were asked to sign the consent form that represented their permission to commence.

Afterwards, they received an overview of the experiment and what can be expected in the three phases of the VR training game. It is important to note that although the assessment phase was developed to test participants' orderly performance, participants should go through the assessment without knowing that their orderliness was being evaluated. Therefore, they were told that the goal of the training game was to perform an order-picking task in VR, and their performance was assessed by their actions. There were three phases in the game in total. In the instruction phase, they learned the procedure for the order-picking task through step-by-step guidance. After that, they entered the practice phase, where they could practice doing the task without guidance. Finally, they carried out the task two more times in the assessment phase. It is worth mentioning that the training was described as an interesting game, and the three phases were named differently in the information of the study. Instead of the instruction, scenario, and assessment phase, they were introduced as instruction, practice, and scenario phases. By framing the training as a game and replacing the name “assessment” with “scenario”, test anxiety could be avoided so that participants could demonstrate their most genuine behaviours. Similarly, to relieve performance stress and to ensure they had enough time to notice and interact with the trash, it was highlighted that the task completion time was not taken into account, so they could spend as much as they needed on the task. In the information on the study document, it was mentioned that if they noticed clutter in the environment, they were free to clean it up, and it would not affect their task performance in any way.

Afterwards, participants were asked to scan a QR code on a piece of paper to complete an 18-item survey on their mobile device, which entailed rating the items to the extent to which they correspond to the description of themselves. The survey started with a paragraph emphasizing the importance of answering honestly and truthfully about themselves. Otherwise, the experiment would be compromised. They were also assured that their data and responses would remain confidential and only the researcher had access to them.

Then, they were shown an instructional video on a laptop showing the order-picking task procedure with three types of guidance, including voiceover, highlight, and ghosting. They could only watch the video once. Then, they were instructed on how to use the VR head-mounted device (HMD) and the controllers.

Afterwards, participants proceeded with the VR training game, which consisted of three phases, instruction, scenario, and assessment. The first two phases were carried out in a moderately tidy condition, and the assessment phase included a tidy and messy warehouse. In the instruction phase, participants learned the sequence of actions for performing the order-picking task in a warehouse. Several types of guidance were used in this section, including visual highlights on objects involved in the actions that should be carried out and ghosting, which showed a white outline of the movements that should be conducted. If the participant performed the correct actions, the guidance for the next step would occur. It was impossible to make mistakes in the instruction phase. If one made an error, the instruction simply did not continue until the right action was performed.

The scenario phase was followed by the instruction. This was the stage where participants had opportunities to practice the procedure they had learned in the instruction without automatic guidance. The tasks in this phase remained the same with a few variations that required them to make certain choices to perform the correct actions. They could make mistakes in this phase and continue to the next step, such as forgetting to put certain objects on the tool cart. Immediate feedback was implemented for some of the wrong actions. For instance, choosing the wrong box for the product would prompt a textbox with a warning message on the screen. If participants committed mistakes that did not trigger immediate feedback, they would be verbally informed of their errors and the correct actions.

In one of the characterisations of orderliness, the inclination for order was operationalized by task performance. That is to say if one committed a mistake in the task, it was assumed that he made the errors on purpose. However, inadequate performance could be attributed to many reasons, and it did not necessarily result from reluctance to follow the routine. It might result from the incapability to perform the desired actions. However, only instances where participants were unwilling to work according to instruction should be considered the correct representation of one's preference of orders or the lack thereof. Therefore, to minimize the possibility that participants made errors due to inability instead of unwillingness, they were given sufficient chances to practice the task in the scenario phase. Consequently, only until they could perform the procedure perfectly without committing mistakes in the scenario phase could they enter the assessment phase.

The last phase in the training was assessment. Participants completed the same assessment in two different environments to explore the effect of environmental orderliness on orderly performance. Half of the participants underwent the assessment in the tidy condition first and then the messy condition. The other half of the participants went through the messy condition and followed by the tidy condition. Indicators and features that were used to assess orderly performance were implemented in this stage. The features and indicators in the assessment were identical in both environments and the only different factor was the tidiness level of the warehouse. Participants should complete this phase on their own without automatic guidance or feedback. However, if they were stuck on a certain sequence of actions, they had the option to ask for help by pressing the help button, where visual highlights would appear as in the instruction phase. Participants' behaviours in the VR training were recorded on a screen recorder, but only the performance in the assessment phase was used to judge orderliness.

Participants then proceeded to scan another QR code and answered one question on the environmental changes they noticed in the assessment phase. The question was included to understand their perceptions on the manipulation of environmental orderliness and to obtain confirmation on how many environmental orderliness features were perceived (see Appendix C).

Because this experiment involved deception, participants received a debriefing about the true intention of the study and how their orderliness was assessed in the SA. They were then allowed to withdraw their consent given before the experiment. Finally, comments from the participants were registered to understand more about the intention behind their behaviours in VR and how it related to their cleaning behaviours and orderliness in real life. This concluded the experiment.

5. Results

The goal of this study was to explore the relationship between orderly performance in VSA and self-reported orderliness and to investigate how the relationship differs in tidy and messy environments.

Descriptive statistics (see Table 1) showed the total score of orderliness ranged from 16 to 47 (*M* = 35.62, *SD* = 7.05). The maximum score is 50 while the lowest is 0 in theory. On average, the total mistakes made in the tidy environment was 2.16 (*SD* = 1.82), whereas the mean of the total mistakes in the messy environment was 1.82 (*SD* = 1.64). The mistakes in the tidy environment ranged from 0 to 8, while the ones in the messy environment varied from 0 to 7.

Table 1

Correlations and Descriptive Statistics of Total Score of Orderliness, Total Mistakes in the Tidy Environment, and Total Mistakes in the Messy Environment

	M	*SD*	1.	2.	3.
1. Total score of orderliness	35.62	7.05			
2. Total mistakes in the tidy environment	2.16	1.82	-.28*		
3. Total mistakes in the messy environment	1.82	1.64	-.30*	.66**	

*p < .05

**p < .01

Regarding the first research question, to investigate the relationship between orderliness performance in the VSA and self-reported orderliness, scatterplots were made to obtain the first impression of the relationship and correlation tests were performed. Results from scatterplots showed a negative relationship between the total mistakes in the tidy environment and the total score of orderliness (see Figure 9) as well as between total mistakes in the messy environment and the total score of orderliness (see Figure 10).

Figure 9

Scatterplot of Total Mistakes in the Tidy Environment and Total Score of Orderliness

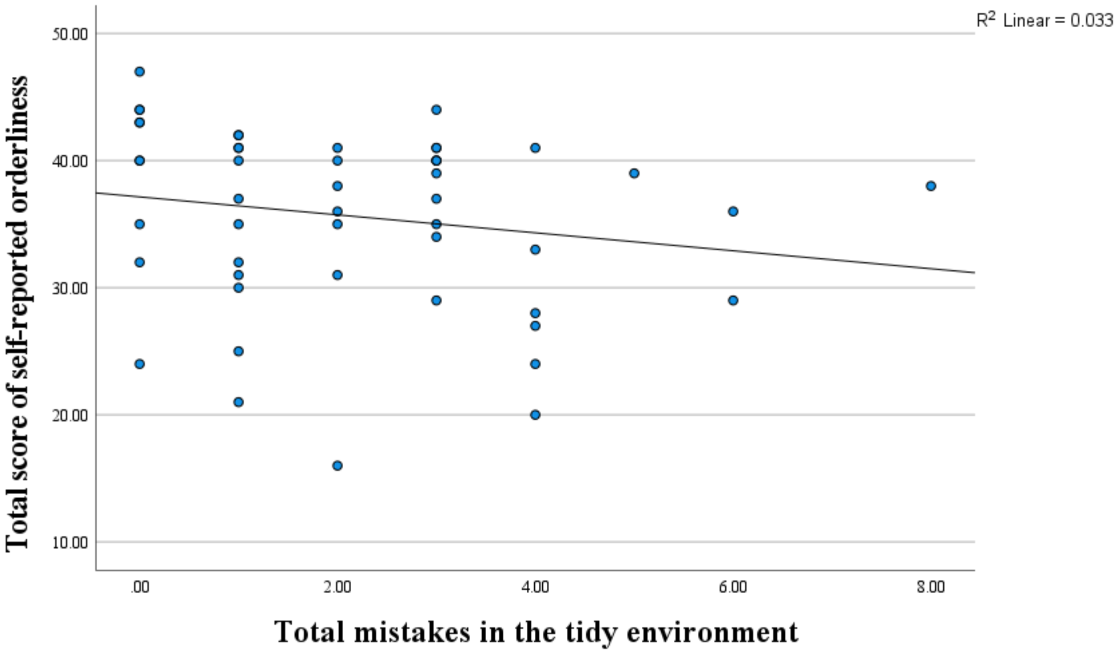
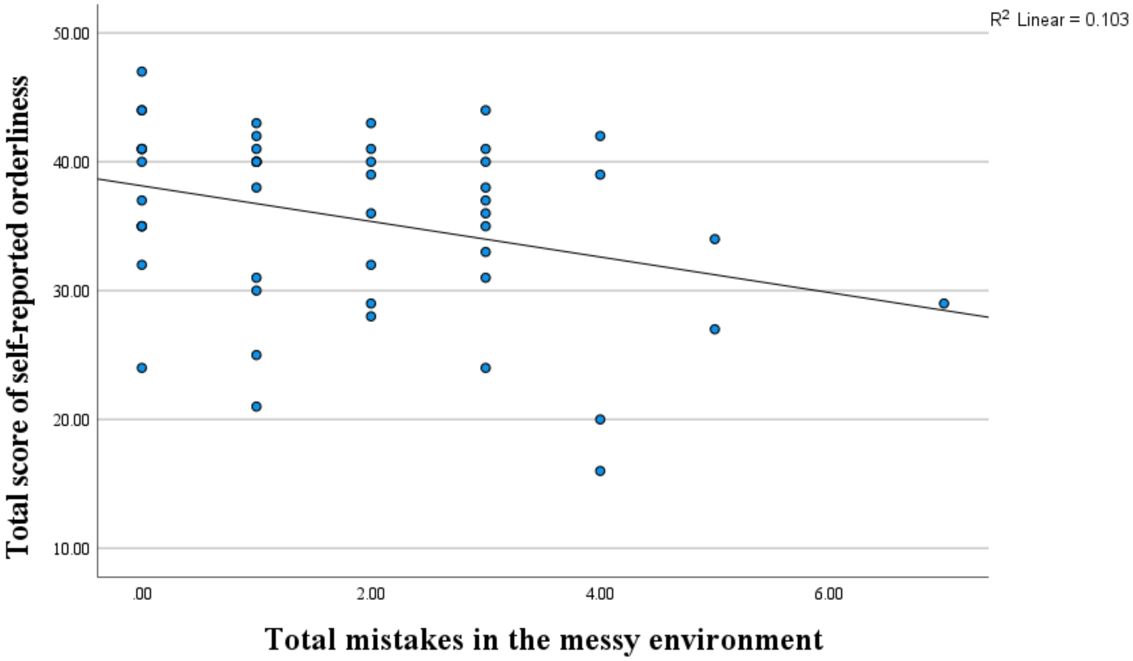


Figure 10

Scatterplot of Total Mistakes in the Messy Environment and Total Score of Orderliness



The results of Spearman’s rho showed that there was a significant negative association between mistakes in the tidy environment and total score of orderliness, although the strength of this relationship was weak (*r* = -.28, *p* = .046). Similarly, there was also a statistically significant correlation between the mistakes in the messy environment and the total score of orderliness with a weak correlation coefficient (*r* = -.30, *p* = .036). The result also showed that total mistakes in the tidy environment and the total mistakes in the messy environment were significantly correlated (*r* = .66, *p* < .001).

To answer the second research question, the effect of environmental orderliness on the relation between orderly performance in VSA and self-reported orderliness was examined. The corresponding hypotheses were the following:

H2: The relation between orderly performance in VSA and self-reported orderliness for the highly orderly group is stronger in the tidy environment than in the messy one.
H3: The relation between orderly performance in VSA and self-reported orderliness for the low orderly group is similar in the tidy environment and the messy environment.

To investigate the hypotheses, participants were first split into two groups, low orderly and highly orderly by the median of the total score of orderliness (*M* = 37.5). The average number of mistakes in the tidy environment for highly orderly people was 1.96 (*SD* = 1.93), which was lower than the average of 2.36 (*SD* = 1.73) for low orderly people. Similarly, the mean number of mistakes in the messy environment for highly orderly people was 1.44 (*SD* = 1.29), also fewer than low orderly people (*M* = 2.20, *SD* = 1.87). However, Mann-Whitney U tests indicated the difference in the performance between highly orderly and low orderly people in the tidy environment was not significant (*U* = 259.5, *p* = .295). There was also no significant difference between the performance of different orderly groups in the messy environment (*U* = 243, *p* = .168). Wilcoxon signed-ranks tests showed that for low orderly people, there was no significant difference between their performance in the tidy and messy environment (*Z* = -.45, *p* = .651). No significant difference was found for highly orderly people’s performance in different environments either (*Z* = 1.6, *p* = .109).

Correlation tests indicated how the correlation between orderly performance and self-reported orderliness differed in tidy and messy environments for different orderly groups (see Table 2). The results of Spearman’s rho showed that for highly orderly people, the total mistakes they committed in the tidy environment were significantly correlated to the total score of orderliness with a moderate association (*r* = -.54, *p* = .005). However, there was no statistically significant association between the total mistakes in the messy environment and the total score of orderliness (*r* = -.30, *p* = .144). As for the low orderly group, no significant association was found between total mistakes and total score of orderliness in the tidy environment (*r* = -.07, *p* = .758) or messy environment (*r* = -.24, *p* = .251). Total mistakes in both environments were significantly associated for both highly orderly group (*r* = .58, *p* = .003) and low orderly group (*r* = .73, *p* < .001).

Regarding environmental orderliness features, the maximum number of the total features one could notice in theory was four per environment. Spearman's rho showed a significant negative relationship between total mistakes in the messy environment and total environmental features noticed in the same environment for high orderly people (*r* = -.53, *p* = .007). No significant relationship was found between total mistakes and total environmental features noticed for low orderly people.

Table 2

Correlations and Descriptive Statistics of Total Score of Orderliness, Total Mistakes in the Tidy Environment, Total Mistakes in the Messy Environment, Total Environmental Features Noticed in the Tidy Environment, and Total Environmental Features Noticed in the Messy Environment for Different Orderliness Levels

Orderliness level	Highly orderly							Low orderly						
	M	*SD*	1.	2.	3.	4.	5.	*M*	*SD*	1.	2.	3.	4.	5.
1. Total score of orderliness	41.16	2.10						30.08	5.75					
2. Total mistakes in the tidy environment	1.96	1.93	-.54**					2.36	1.73	-.07				
3. Total mistakes in the messy environment	1.44	1.29	-.30	.58**				1.87	1.87	-.24	.73**			
4. Total environmental features noticed in the tidy environment	.80	.91	.21	-.01	-.14			1.28	1.37	-.07	-.06	.06		
5. Total environmental features noticed in the messy environment	1.76	1.05	.09	-.26	-.53**	-.30		1.64	1.19	-.30	.29	.28	-.05	

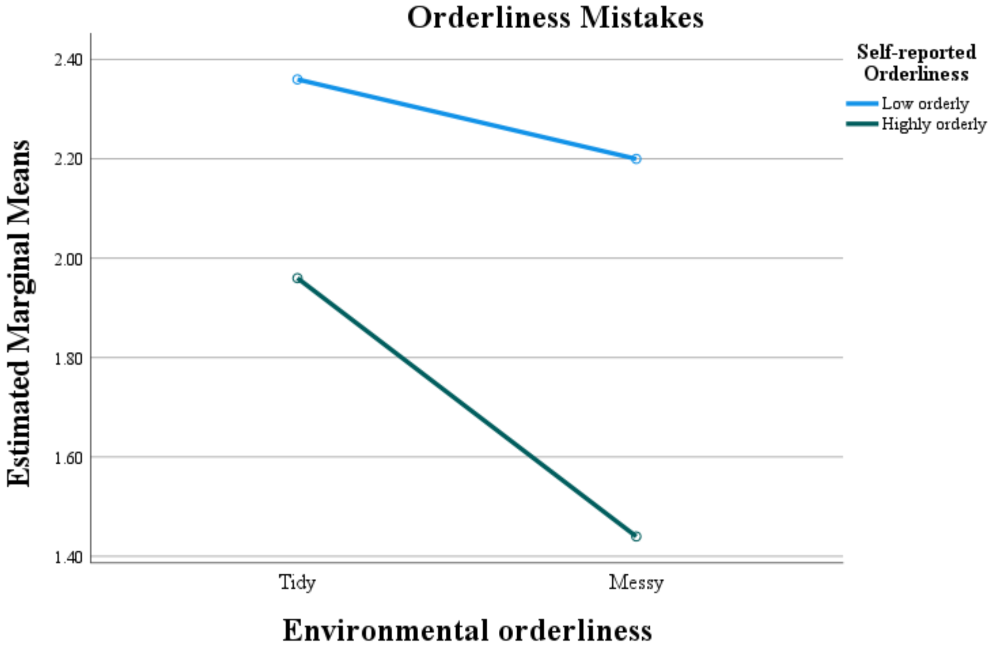
*$p < .05$

**$p < .01$

In addition, a two-way ANOVA test was conducted to examine the interaction effect of environmental orderliness and self-reported orderliness on orderly performance. The plot suggested the effect of environmental orderliness was different for highly orderly people and low orderly people on their mistakes (see Figure 11). Specifically, it indicated that for highly orderly people, there was a larger difference in their performance in different environments than low orderly people. However, the interaction effect between environmental orderliness and orderly groups was not statistically significant, *F* (1, 48) = .75, *p* = .391. The effect of environmental orderliness on orderly mistakes was also not significant, *F* (1, 48) = .2.68, *p* = .11.

Figure 11

Interaction Plot of Environmental Orderliness and Self-reported Orderliness Groups on Orderliness Mistakes



6. Discussion

This study aimed to explore how individuals' orderly performance in the VSA and their self-reported orderliness were related, as well as how the relationship might be affected by environmental orderliness for different orderly groups. To summarize the results, the findings showed that there was a significant but weak relationship between the orderly performance in both environments and self-reported orderliness. Moreover, for highly orderly people, there was a significant relationship between their orderly performance in the tidy environment and their self-reported orderliness. No significant difference was found in the correlations between self-reported orderliness and total mistakes in the two environments for the low orderly group.

The first hypothesis was that orderly performance in both environments and self-reported orderliness would be correlated. This hypothesis was confirmed. Results showed that the orderliness mistakes committed in both environments were negatively correlated to self-reported orderliness scores. This indicated that the more orderly an individual was, the fewer mistakes he made in both environments. Despite the statistically significant association, the strength of the relationship was quite weak, which might be explained by a few potential reasons. First, this outcome could result from employing different measurements in the current study. The VSA can be categorized as a type of performance-based assessment, whereas the self-reported instrument is considered an explicit measure. The lack of alignment between the results from different assessment methods has been well-documented in the literature (Anderson, 2009; Karpinski, 2004; Ventura et al., 2013). For example, Ventura et al. (2013) measured persistence with a performance-based measure (a riddle task in a video game) and a survey. While a significant relationship was found between time spent on unsolved trials in the game and the score of self-reported persistence, the correlation was considerably low. It is possible that the construct of the two assessments measured was not entirely the same but somewhat overlapped. In the context of this study, the questionnaire merely questioned one's general orderliness over time, while the VSA observed how an individual would behave in a specific environment and context. It could be argued that the virtual assessment possessed higher ecological and face validity as the observables were directly linked to one’s orderly behaviours.

In addition, VSA might be less susceptible to social desirability bias than self-reported measures and this could give rise to the weak association between the results (Ventura et al., 2013). In the assessment, the participants were not aware that their orderliness was being assessed, and the immersive experience in VR created an illusion that they were not being observed. According to a few participants, they responded to the interactable trash as if no one was watching, in a way that was different from how they would behave in public. In comparison, self-reported measures pose a high risk of social desirability bias because it is challenging to prevent respondents from answering questions in a way that makes them appear more favourable (Robins et al., 2007).

Comments from the participants in the debriefing also provided important insights to interpret the result (see Appendix F). It was evident that there was often a gap between how orderly an individual thinks he or she was, and the cleaning behaviours displayed in the VSA. Two types of inconsistencies were identified. One was highly orderly people who cleaned up less trash than expected and the other was low orderly people who cleaned up more trash than expected. The most common explanation for the first discrepancy was that they did not notice the different types of trash or the corresponding bins because they were too focused on the task. This outcome could be attributed to selective attention, which refers to the ability to direct one’s attention to one input while ignoring external and irrelevant information (Stevens & Bavelier, 2012). Prior research has found that training in VR could induce more extraneous load compared to real-life training (Frederiksen et al., 2020), therefore, it is possible that the cognitive load was so substantial that participants did not have enough processing resources left to complete the task while simultaneously paying attention to the objects around them. For some people, however, they noticed the trash but deliberately chose not to tidy up. A few reasons were revealed. For example, some participants did not think it was their responsibility to clean up the clutter that was not caused by themselves, and since cleaning was not related to their task performance, they did not feel the need to react to the trash. Other participants thought if they had paused their action to clean up, they would have forgotten about the next steps, so they chose not to. Some people did not know they could interact with the trash (even though this point was made clear in the study information document), so they simply continued with the task without any attempt to throw it away.

Interestingly, a participant who was an experienced VR user thought that they did not behave in the way they would in real life. Specifically, they identify themselves as a frequent cleaner, but they did not pick up any trash in the assessment despite noticing them. The explanation they gave was that they did not take the assessment very seriously because they were aware that it was merely a simulation game. Consequently, there was an inconsistency between how they behaved in VR and how they would have behaved in the same context according to them. Understanding the importance of presence in VR can lead to more clarification of this disparity. Research has shown that natural behaviours and similar emotional responses are a sign of high presence, which could lead to better performance and learning (Slater & Wilbur, 1997). If presence is not achieved in the simulation, the assessment would likely fail to incite authentic reactions, which results in the gap between behaviours in VR and real life.

In contrast, even though some people characterised themselves as untidy in general, they still cleaned up more trash than anticipated. Some expressed they made more effort to keep their working areas clean in daily life because they got distracted easily. Others said that they were accustomed to cleaning the public spaces out of politeness and consideration of other people’s feelings. As a result, compared to their personal spaces, they were more concerned about maintaining the cleanliness of shared areas to the extent that they would also clean up other people’s mess just to keep the environment tidy. This mindset seems to be contrary to the highly orderly group who did not feel obligated to tidy up others’ mess. The different mentalities perhaps can be seen as a representation of the cultural syndromes of collectivism and individualism, which can be a contributing factor for personal characteristics and behaviours (Triandis, 2002).

Regarding the second and third hypotheses, it was expected that for highly orderly people the relationship between self-reported orderliness and orderly performance in the tidy environment would be stronger than in the messy environment. Moreover, for low orderly people, the relationship between self-reported orderliness and orderly performance would remain similar in both environments. The two hypotheses were also confirmed. The results showed a significantly negative correlation between orderly mistakes in the tidy environment and orderliness scores for highly orderly people, while no significant relationship was found for the messy environment. This implies that for highly orderly people, the more orderly the person was, the more orderly the performance in the clean environment was. Additionally, there was no significant association between self-reported orderliness and orderly performance in either environment for low orderly people, which indicates that environmental orderliness did not have a strong impact on the relationship. Similarly, the interaction plot also indicated that compared to highly orderly people, the orderly performance of low orderly people varied less with the environments. On the other hand, highly orderly people's performance seemed to be more affected as the difference between performance in different environments was larger than low orderly people. Considering the interaction effect was not significant, a more substantial sample size might be needed to conclude that there is an effect in the population.

The findings are in line with the study of Mateo et al. (2013) where high conscientious people performed with the accuracy level that corresponded to their personality trait in a physical tidy environment, while the accuracy for low conscientious people remained similar in tidy and messy environments. The results in this study provided evidence that a compatible virtual environment, a clean warehouse, in this case, was able to elicit behaviours that are more consistent with highly orderly people's orderliness than an incompatible one, while there is no significant difference for low orderly people.

As suggested by the person-organisation theory, human behaviours can be explained as an outcome of the interactions between an individual and the environment (Kristof-Brown et al., 2005). From this perspective, an explanation of the results could be that highly orderly individuals felt more comfortable in the tidy environment and therefore reacted more in accordance with their nature. For low orderly people, however, the discomfort resulting from the misfit with the tidy environment was not as strong, so no significant difference between the actual behaviours and perceived orderliness in different conditions was found.

Although only a weak relationship was found between performance in the messy environment and orderliness, it is interesting to mention that there was a significant negative relationship between the number of environmental orderliness features an individual noticed (e.g., dirty walls, dirty floors, unaligned products, garbage on the floor) and the total mistakes committed in the messy environment for highly orderly people. In contrast, a positive correlation was found between the same variables for low orderly people. This indicated that the more signs of disorder a highly orderly person perceived in the messy environment, the more orderliness behaviours one demonstrated. Perhaps the visible clutter and dirt functioned as triggers that prompted their desire to behave in a more orderly manner. For low orderly people, however, the result suggested that the more mess they noticed, the less orderly they behaved. Since signs of disorder could promote further disorganization (Vohs et al., 2013), it could be that highly orderly people were more resistant to the messy environment than low orderly people.

In conclusion, every individual is orderly in varying degrees and different ways. It is therefore challenging to make inferences on one’s behaviours with a self-reported measure because it is often too general and ambiguous. VSA, on the other hand, demonstrates the opportunity to assess individuals by placing them in an exact situation and context, which is a more representative assessment environment. Furthermore, VSA is assumed to be more resistant to social desirability issues because participants felt less observed as they were immersed in the simulation, and the high level of presence compelled them to behave in similar ways as in real life.

6.1. Implications

This study contributed to the scientific field in a number of ways. First, this research is the first study that provided empirical data on developing a SA in the VR environment. Over the years, SAs have been applied in different educational contexts and have shown reliability and validity in measuring competencies in many studies (Shute, 2011). Despite papers that advocated VR as a suitable environment to evaluate different qualities (Alcañiz et al., 2018; Chicchi Giglioli et al., 2017; De-Juan-Ripoll et al., 2018), there is scarce empirical evidence of SAs in the context of VR (Chicchi Giglioli et al., 2017). Therefore, this study laid the groundwork for future research in the area of VSAs by establishing the correlation between performance in VR and an explicit measure.

Second, in the field of personality psychology, the predominant assessment approach on personality traits are explicit measures such as surveys, which are susceptible to cheating and social desirability issues (Robins et al., 2007). By developing a new measure of orderliness that is impervious to the limitations of traditional instruments, this research enriched the personality literature by exploring an additional approach to complement existing measures and deepen the understanding of orderliness.

Third, although the effect of environmental orderliness on individuals' attitudes and behaviours has been established in the context of physical places (Mateo et al., 2013; Vohs et al., 2013), its transferability to virtual environments has not been studied. Therefore, this research is one of the first attempts to explore the influence of environmental orderliness on behaviours in VR.

Regarding practical relevance, VSAs showed several advantages in assessing orderliness over explicit measures. First, assessments conducted in VR can be considered to have higher ecological validity than self-reported measures (Kourtesis et al., 2021; Tarnanas et al., 2013). Assigning a specific environment and task to the candidates allows organisations to understand how they behave in a certain work context, and the results can be more insightful than self-reported behavioural patterns which are more abstract and broader in nature. After all, orderliness behaviours vary with different contexts, environments, and other situational factors. Since organisations are mostly concerned with behaviours at a workplace, a more representative assessment is assumed to be more indicative than a general one.

Second, VSA can be less prone to social desirability limitations. Due to the high degree of immersion in VSA, participants can feel less observed and therefore act more authentically. In a real warehouse, workers might feel more obligated to clean up when there are colleagues and supervisors around, but it is uncertain if the same behaviours would persist without other people present. For organisations, a candidate who cleans up out of inner motivation might be more valuable than one who only cleans up due to expectations of others, so it is important to understand to what extent one cleans up due to external factors or the natural inclination. Yet, this is quite difficult to achieve in real life as it requires stealth observation. VSA overcomes this obstacle by offering a controlled environment to see how an individual behaves without external social pressure, thereby making a more authentic representation of one's orderly behaviours.

The findings also highlighted the importance of taking individuals' personal characteristics and environments into account when developing a virtual application that aims to induce authentic reactions and behaviours. As demonstrated, a tidy virtual environment promotes highly orderly people to act in a way that reflects their orderly nature whereas a messy one does not. It can be concluded that a tidy environment is not harmful to anyone, but a messy environment might restrain highly orderly individuals from behaving according to their disposition. Additionally, as the comments from participants suggested, personal characteristics such as cultural background and prior experience with VR can potentially have an influence on orderly behaviours in VR. It is also possible that presence plays a role in behaviours in VR, as the connection between presence and emotional reactions has been reported in the literature (Diemer et al., 2015). It is therefore important to take them into consideration when interpreting the performance results.

6.2. Limitations

There are three limitations to this research. First, although experiments were conducted with the target population (MBO students studying logistics), the data was not used in the analysis due to the sample size being too small. Because of limited time and resources, it was unlikely to recruit more participants from the target group. As such, more experiments were done with students from the University of Twente, which represents a population that is quite different from prospective logistic workers. However, the focus of the study is on the relationship between the variables instead of the variables themselves, namely the association between self-reports and VR performance. Since there is no indication in the literature as to how the relationship varies with different populations, it can be argued that the use of a population with different characteristics than the target group is not detrimental to the results of this study. This limitation presents an opportunity to replicate the experiments in the current study with a larger group of participants with a professional logistics background to confirm the above assumption.

Second, due to technical issues and the way the application was designed, there were instances where participants could not react to the trash in the way that they intended. Specifically, objects get snapped to an unintended location because the participants accidentally release the trigger on the controller. For example, participants might want to put the broken tape in the trash bin, but instead, the tape got snapped to the cardboard bin. This technical issue can be considered random errors, and although they have the potential of affecting the precision of the results, this issue only occurred a few times and the errors in different directions are likely to cancel each other out without compromising the validity of this study.

Finally, another limitation lies in the categories of orderliness mistakes. The number of mistakes related to the inclination for order (see Appendix D) in both environments is considerably low. This might be explained by the requirement to perform the procedure perfectly in the scenario phase before moving on to the assessment. Consequently, most participants learned the procedure so well that it was unlikely for them to make mistakes. Future research can explore new settings and situations in VSAs where one’s preference for order is triggered and therefore measured more easily.

6.3. Future Research & Recommendations

Based on the findings of this study, there are several directions for further research to follow. First, researchers could investigate further the potential competencies VSAs can measure. 21st-century skills such as problem-solving skills and creativity are difficult to assess in traditional assessments (Hu Au & Lee, 2017), and SAs that measure these two skills have been developed and validated (Shute & Rahimi, 2021; Shute et al., 2016). Researchers could consider building on the work of prior SAs and design VSA in different contexts to measure the same skills to enhance ecological validity while exploring other 21st skills to evaluate. It is worth noting that SA is not suitable for every target competency, and not every aspect within the competency is measurable. In addition, it is important to ensure the alignment of the external measure and SA, as the discrepancy can be detrimental for validating the results. It is therefore suggested that more consistent measures should be selected such as a performance-based assessment.

Nevertheless, as demonstrated in this research, there is great potential in the development of more SA in VR, such as evaluating individuals' ability to perform under pressure. With VR, emotional responses can be recorded in VR headsets that track the heartbeat (Floris et al., 2020) and eye movements of the user (Hickson et al., 2019) when conducting a task under stress, while the behavioural responses can be measured with the SA in the application. This could provide more insight as to how emotional responses and behavioural reactions are correlated when triggered by stress (Clifford et al., 2019). Spatial ability is another competency that is suitable to measure in VR, as the simulation offers space and flexibility for individuals to exhibit their skills in terms of manipulating physical objects (González, 2018). It can be interesting to see the development in combination with SA.

Furthermore, this study demonstrated that there is potential in assessing self-reported behavioural patterns in VSA. This finding could lead to investigations on how behaviours in VR transfer to a similar real-life context. Understanding human behaviours has always been the overarching theme in psychology, and the use of VR technology offers great opportunities to measure and analyse human behaviours, as well as how they relate to actual behaviours. By assessing behaviours in VR, it can illustrate the possibility of predicting real-life behaviours in the future. The next step could be conducting a real-life observation on orderly behaviours and interpreting the relationship between the results with the VSAs.

Another direction could be to combine eye-tracking technology with VR to better comprehend users' focus and attention in the VR environment. In the experiment, seemingly obvious objects and important environmental features sometimes get ignored when conducting the task. This could potentially be explained by the increase of inattentional blindness and change blindness when individuals are in a virtual environment rather than in reality. Their perception of the objects could change in the simulation, and as a result, their ability to notice stimuli in plain sight is also different. Additionally, further research can be done on the influence of VR environments and performance. As the finding suggests, environmental orderliness did not make a strong impact on individuals’ performance, which is contrary to empirical evidence in the literature on physical environmental orderliness (Mateo et al., 2013; Vohs et al., 2013). Further research is required to understand if and what potential elements in the environments can affect individuals’ actions and performance. By using eye-tracking to understand where individuals pay attention, information on how VR applications can be developed most efficiently and effectively can be obtained. From a learning perspective, developers can also use this knowledge to better direct user attention to important information they would like learners to focus on in VR training.

Regarding practical recommendations, it is crucial for practitioners to keep in mind that the performance in VR only tells half the story. Although this study showed a correlation between behaviours in VR and an existing explicit measure, future investigation is needed to examine if the results are transferable to a real-life context. Practitioners are also advised to consider the personal characteristics and the compatibility between users and the virtual assessment environments when interpreting the assessment results, as they can influence their performance and behaviours. It is therefore beneficial to gather further qualitative information along with behavioural data to fully understand the reasons why people behaved the way they did in the virtual assessment. This could be done in several ways, such as using video-stimulated recall to understand their behaviours. The information can contribute considerably to understanding the relationship between the motives and the behaviours in the simulation and obtaining more information on their perception and opinion on using VR technology**.**

6.4. Conclusion

This research extended prior studies by providing empirical evidence of using VSA as a measurement of one’s orderliness and an exploratory perspective on how the relationship between self-reported orderliness and orderly performance varied with different environments.

The findings demonstrated promising results in using VR as the vehicle for developing SA and that the orderliness features of the simulation environment can potentially make an impact on eliciting genuine behaviours. Practitioners are advised to regard performance results in VR with caution and be mindful that while VSA is a promising measurement method, it is an addition to traditional assessments instead of a replacement. Although the pursuit of comprehending the complex nature of real-life behaviours has been and will continue to be challenging, this study laid the groundwork for future endeavours by demonstrating that VR could be a viable way to forward our understanding of how and why people behave the way they do.

7. References

Alcañiz, M., Parra, E., & Giglioli, I. A. C. (2018). Virtual reality as an emerging methodology for leadership assessment and training. *Frontiers in Psychology*, *9*(SEP), 1–7. https://doi.org/10.3389/fpsyg.2018.01658

Anderson, R. D. (2009). *The Implicit Association Test for Conscientiousness: An indirect method of measuring personality*. https://etd.ohiolink.edu/ap:10:0::NO:10:P10_ACCESSION_NUM:bgsu1237835643

Arafat, H. A., Al-Khatib, I. A., Daoud, R., & Shwahneh, H. (2007). Influence of socio-economic factors on street litter generation in the Middle East: Effects of education level, age, and type of residence. *Waste Management and Research*, *25*(4), 363–370. https://doi.org/10.1177/0734242X07076942

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting Actual Behavior From the Explicit and Implicit Self-Concept of Personality. *Journal of Personality and Social Psychology*, *97*(3), 533–548. https://doi.org/10.1037/a0016229

Bator, R. J., Bryan, A. D., & Schultz, P. W. (2011). Who gives a hoot?: Intercept surveys of litterers and disposers. *Environment and Behavior*, *43*(3), 295–315. https://doi.org/10.1177/0013916509356884

Boswell, W. R. (2008). Voluntary Employee Turnover: Determinants, Processes, and Future Directions. In J. Barling & C. L. Cooper (Eds.), *The SAGE Handbook of Organizational Behavior: Volume One: Micro Approaches* (1st ed., Vol. 1, pp. 196–216). SAGE Publications Ltd. https://doi.org/10.4135/9781849200448.n1.

Brown, B. B., Perkins, D. D., & Brown, G. (2004). Incivilities, place attachment and crime: Block and individual effects. *Journal of Environmental Psychology*, *24*(3), 359–371. https://doi.org/10.1016/j.jenvp.2004.01.001

Chicchi Giglioli, I. A., Parra, E., Cardenas-Lopez, G., Riva, G., & Alcañiz Raya, M. (2017). Virtual stealth assessment: A new methodological approach for assessing psychological needs. *Lecture Notes in Computer Science (Including Subseries*

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10622 LNCS(November), 1–11. https://doi.org/10.1007/978-3-319-70111-0_1

Clifford, R. M. S., Jung, S., Hoerrnann, S., Billinqhurst, M., & Lindeman, R. W. (2019). Creating a stressful decision making environment for aerial firefighter training in virtual reality. *26th IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2019 - Proceedings*, 181–189. https://doi.org/10.1109/VR.2019.8797889

De-Juan-Ripoll, C., Soler-Domínguez, J. L., Guixeres, J., Contero, M., Gutiérrez, N. Á., & Alcañiz, M. (2018). Virtual reality as a new approach for risk taking assessment. *Frontiers in Psychology*, *9*(DEC), 1–8. https://doi.org/10.3389/fpsyg.2018.02532

de Koster, M. B. M. (2012). Warehouse assessment in a single tour. In *Warehousing in the Global Supply Chain: Advanced Models, Tools and Applications for Storage Systems* (pp. 457–473). Riccardo Manzini. https://doi.org/10.1007/978-1-4471-2274-6

de Koster, R., Le-Duc, T., & Roodbergen, K. J. (2007). Design and control of warehouse order picking: A literature review. *European Journal of Operational Research*, *182*(2), 481–501. https://doi.org/10.1016/j.ejor.2006.07.009

Diemer, J., Alpers, G. W., Peperkorn, H. M., Shiban, Y., & Mühlberger, A. (2015). The impact of perception and presence on emotional reactions: A review of research in virtual reality. *Frontiers in Psychology*, *6*(JAN), 1–9. https://doi.org/10.3389/fpsyg.2015.00026

Dunham, S., Lee, E., & Persky, A. M. (2020). The psychology of following instructions and its implications. *American Journal of Pharmaceutical Education*, *84*(8), 1052–1056. https://doi.org/10.5688/ajpe7779

Fenske, J. N., & Schwenk, T. L. (2009). *Obsessive-Compulsive Disorder: Diagnosis and Management*. *80*(3), 239–245. https://www.aafp.org/afp/2009/0801/afp20090801p239.pdf

Floris, C., Solbiati, S., Landreani, F., Damato, G., Lenzi, B., Megale, V., & Caiani, E. G. (2020). Feasibility of heart rate and respiratory rate estimation by inertial sensors embedded in a virtual reality headset. *Sensors (Switzerland)*, *20*(24), 1–21. https://doi.org/10.3390/s20247168

Frederiksen, J. G., Sørensen, S. M. D., Konge, L., Svendsen, M. B. S., Nobel-Jørgensen, M., Bjerrum, F., & Andersen, S. A. W. (2020). Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: a randomized trial. *Surgical Endoscopy*, *34*(3), 1244–1252. https://doi.org/10.1007/s00464-019-06887-8

Friedman, A., Katz, B. A., Cohen, I. H., & Yovel, I. (2021). Expanding the Scope of Implicit Personality Assessment: An Examination of the Questionnaire-Based Implicit Association Test (qIAT). *Journal of Personality Assessment*, *103*(3), 380–391. https://doi.org/10.1080/00223891.2020.1754230

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 283–310). Cambridge University Press. https://doi.org/10.1017/CBO9780511996481.016

Goldberg, L. R. (1999). A Broad-Bandwidth, Public Domain Personality Inventory Measuring the Lower-Level Facets of Several Five-Factor Models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe*, Vol. 7 (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press. https://ipip.ori.org/A%20broad-bandwidth%20inventory.pdf

González, N. A. A. (2018). Development of spatial skills with virtual reality and augmented reality. *International Journal on Interactive Design and Manufacturing*, *12*(1), 133–144. https://doi.org/10.1007/s12008-017-0388-x

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Grumm, M., & von Collani, G. (2007). Measuring Big-Five personality dimensions with the implicit association test - Implicit personality traits or self-esteem? *Personality and Individual Differences*, *43*(8), 2205–2217. https://doi.org/10.1016/j.paid.2007.06.032

Hickson, S., Kwatra, V., Dufour, N., Sud, A., & Essa, I. (2019). Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, 1626–1635. https://doi.org/10.1109/WACV.2019.00178

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin*, *31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613

Holtom, B. C., Mitchell, T. R., Lee, T. W., & Inderrieden, E. J. (2005). Shocks as causes of turnover: What they are and how organizations can manage them. *Human Resource Management*, *44*(3), 337–352. https://doi.org/10.1002/hrm.20074

Hu Au, E., & Lee, J. J. (2017). Virtual reality in education: a tool for learning in the experience age. *International Journal of Innovation in Education*, *4*(4), 215. https://doi.org/10.1504/ijiie.2017.10012691

Kallgren, C. A., Reno, R. R., & Cialdini, R. B. (2000). A focus theory of normative conduct: When norms do and do not affect behavior. *Personality and Social Psychology Bulletin*, *26*(8), 1002–1012. https://doi.org/10.1177/01461672002610009

Karpinski, A. (2004). Measuring Self-Esteem Using the Implicit Association Test: The Role of the Other. *Personality and Social Psychology Bulletin*, *30*(1), 22–34. https://doi.org/10.1177/0146167203258835

Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science*, *322*(5908), 1681–1685. https://doi.org/10.1126/science.1161405

Kourtesis, P., Collina, S., Doumas, L. A. A., & MacPherson, S. E. (2021). Validation of the Virtual Reality Everyday Assessment Lab (VR-EAL): An Immersive Virtual Reality Neuropsychological Battery with Enhanced Ecological Validity. *Journal of the International Neuropsychological Society*, *27*(2), 181–196. https://doi.org/10.1017/S1355617720000764

Kristof-Brown, A. L., Zimmerman, R. D., & Johnson, E. C. (2005). Consequences of Individuals ' Fit At Work : a Meta-Analysis of Person-Jo. *Personnel Psychology*, *58*, 281–342. https://doi.org/10.1111/j.1744-6570.2005.00672.x

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality and Quantity*, *47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Lim, B. C., & Ployhart, R. E. (2006). Assessing the convergent and discriminant validity of Goldberg's international personality item pool: A multitrait-multimethod examination. *Organizational Research Methods*, *9*(1), 29–54. https://doi.org/10.1177/1094428105283193

Lorr, M., & Youniss, R. P. (1973). An inventory of interpersonal style. *Journal of personality assessment, 37 2*, 165-73. https://doi.org/10.1080/00223891.1973.10119847

MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009). Empirical identification of the major facets of Conscientiousness. *Learning and Individual Differences*, *19*(4), 451–458. https://doi.org/10.1016/j.lindif.2009.03.007

Maples, J. L., Guan, L., Carter, N. T., & Miller, J. D. (2014). A test of the international personality item pool representation of the revised NEO personality inventory and development of a 120-item IPIP-based measure of the five-factor model. *Psychological Assessment*, *26*(4), 1070–1084. https://doi.org/10.1037/pas0000004

Mateo, R., Hernández, J. R., Jaca, C., & Blazsek, S. (2013). Effects of tidy/messy work environment on human accuracy. *Management Decision*, *51*(9), 1861–1877. https://doi.org/10.1108/MD-02-2013-0084

McCrae, R. R., & Costa, P. T. (1987). Validation of the Five-Factor Model of Personality Across Instruments and Observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90. https://doi.org/10.1037/0022-3514.52.1.81

Mislevy, R. J. R., Almond, R. R. G., & Lukas, J. F. J. (2003). A Brief Introduction to Evidence-Centered Design. CSE Report 632. *US Department of Education*, *1522*(310). https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Moore, G. R., & Shute, V. J. (2017). Handbook on Digital Learning for K-12 Schools. *Handbook on Digital Learning for K-12 Schools*, 355–368. https://doi.org/10.1007/978-3-319-33808-8

Murray, H. A. (1938). Explorations in Personality. *Oxford University Press.* https://archive.org/details/in.ernet.dli.2015.202783

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at Age 7: A Methodological and Conceptual Review. In J. A. Bargh (Ed.), *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). Psychology Press. https://faculty.washington.edu/agg/pdf/Nosek%20&%20al.IATatage7.2007.pdf

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in Human Neuroscience*, *9*(DEC), 1–19. https://doi.org/10.3389/fnhum.2015.00660

Pizzi, G., Scarpi, D., Pichierri, M., & Vannucci, V. (2019). Virtual reality, real reactions?: Comparing consumers' perceptions and shopping orientation across physical and virtual-reality retail stores. *Computers in Human Behavior*, *96*(July 2018), 1–12. https://doi.org/10.1016/j.chb.2019.02.008

Ramos, J., & Torgler, B. (2012). Are academics messy? Testing the broken windows theory with a field experiment in the work environment. *Review of Law and Economics*, *8*(3), 563–577. https://doi.org/10.1515/1555-5879.1617

Robins, R. W., Fraley, R. C., & Krueger, R. F. (Eds.). (2007). *Handbook of research methods in personality psychology.* The Guilford Press. https://doi.org/10.1017/CBO9780511996481

Schultz, P. W., Bator, R. J., Large, L. B., Bruni, C. M., & Tabanico, J. J. (2013). Littering in Context: Personal and Environmental Predictors of Littering Behavior *Environment and Behavior*, *45*(1), 35–59. https://doi.org/10.1177/0013916511412179

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age. https://myweb.fsu.edu/vshute/pdf/shute%20pres_h.pdf

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*(December 2019), 106647. https://doi.org/10.1016/j.chb.2020.106647

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's playground. *Journal of Educational Research*, *106*(6), 423–430. https://doi.org/10.1080/00220671.2013.832970

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

Slater, M., & Wilbur, S. (1997). A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments* 1997; 6 (6): 603–616. https://doi.org/10.1162/pres.1997.6.6.603

Stevens, C., & Bavelier, D. (2012). The role of selective attention on academic foundations: A cognitive neuroscience perspective. *Developmental Cognitive Neuroscience*, *2*(SUPPL. 1), S30–S48. https://doi.org/10.1016/j.dcn.2011.11.001

Tarnanas, I., Schlee, W., Tsolaki, M., Müri, R., Mosimann, U., & Nef, T. (2013). Ecological validity of virtual reality daily living activities screening for early dementia: Longitudinal study. *JMIR Serious Games*, *1*(1), 1–14. https://doi.org/10.2196/games.2778

Tett, R. P., & Simonet, D. V. (2011). Faking in personality assessment: A "Multisaturation" perspective on faking as performance. *Human Performance*, *24*(4), 302–321. https://doi.org/10.1080/08959285.2011.597472

Triandis, H. (2002). Individualism-Collectivism and Personality. *Journal of personality*. 69. 907-24. https://doi.org/10.1111/1467-6494.696169

van Ryzin, G., Immerwahr, S., & Altman, S. (2008). Measuring Street Cleanliness: A Comparison of New York City's Scorecard and Results from a Citizen Survey. *Public Administration Review*. 68. 295 - 303. https://doi.org/10.1111/j.1540-6210.2007.00863.x

Ventura, M., Shute, V., & Zhao, W. (2013). The relationship between video game use and a performance-based measure of persistence. *Computers and Education*, *60*(1), 52–58. https://doi.org/10.1016/j.compedu.2012.07.003

Vohs, K. D., Redden, J. P., & Rahinel, R. (2013). Physical Order Produces Healthy Choices, Generosity, and Conventionality, Whereas Disorder Produces Creativity. *Psychological Science*, *24*(9), 1860–1867. https://doi.org/10.1177/0956797613480186

Weibel, R. P., Grübel, J., Zhao, H., Thrash, T., Meloni, D., Hölscher, C., & Schinazi, V. R. (2018). Virtual reality experiments with physiological measures. *Journal of Visualized Experiments*, *2018*(138), 1–8. https://doi.org/10.3791/58318

Wells, V. K., & Daunt, K. L. (2015). Eduscape: The Effects of Servicescapes and Emotions in Academic Learning Environments. *Journal of Further and Higher Education*. https://doi.org/10.1080/0309877X.2014.984599

Whitehead, H., May, D., & Agahi, H. (2007). An exploratory study into the factors that influence patients' perceptions of cleanliness in an acute NHS trust hospital. *Journal of Facilities Management*, *5*(4), 275–289. https://doi.org/10.1108/14725960710822268

Whyte, W. H. (1956). The organization man. *Simon and Schuster*. https://doi.org/10.9783/9780812209266

Xu, C., Demir-Kaymaz, Y., Hartmann, C., Menozzi, M., & Siegrist, M. (2021). The comparability of consumers' behavior in virtual reality and real life: A validation study of virtual reality based on a ranking task. *Food Quality and Preference*, *87*(August 2020), 104071. https://doi.org/10.1016/j.foodqual.2020.104071

Xu, X., Karinen, A. K., Chapman, H. A., Peterson, J. B., & Plaks, J. E. (2020). An orderly personality partially explains the link between trait disgust and political conservatism. *Cognition and Emotion*, *34*(2), 302–315. https://doi.org/10.1080/02699931.2019.1627292

Ziegler, M., Maccann, C., & Roberts, R. (2011). New Perspectives on Faking in Personality Assessment. https://doi.org/10.1093/acprof:oso/9780195387476.003.0011.

8. Appendices

8.1. Appendix A: The Use of IPIP-NEO in the Current Study

Despite the challenges of explicit measures mentioned in the theoretical framework, a self-report measure on orderliness, namely, the IPIP-NEO was used to compare the results of the VSA. Since both implicit measures and SA do not require the introspection ability of the respondents, the results might be more likely to be aligned; therefore, some might argue it is more appropriate to use an implicit instrument to relate to the current assessment. However, the purpose of this study was to investigate if a VSA could potentially be an alternative option to assess one’s orderliness. In order to achieve this, contrasting the results between the VSA with the most common personality test used in practice is the first step into exploring the possibility.

A few other reasons led to the belief that explicit measures are more suitable in this study. First is that in the personality psychology field, there is hardly any reliable and valid implicit measure of personality traits (Gawronski & De Houwer, 2014). Researchers have made attempts in developing various implicit tests to assess personalities but to no avail (Anderson, 2009; Karpinski, 2004). The implicit measures that have been developed often suffer from measurement error and a lack of reliability (Gawronski & De Houwer, 2014). Moreover, it is challenging to validate them because when compared with the results of explicit measures, correlation is often not significant due to the inherent difference between these methods (Anderson, 2009; Karpinski, 2004). Second, the questionnaire implemented in the current study is a widely used and well-validated personality measure. IPIP-NEO is a 300-item questionnaire that measures the Big Five traits along with six facets belonging to each one. Multiple studies have demonstrated a high degree of internal consistency, construct validity, and convergent validity to the parent scale NEO-PI-R (Maples et al., 2014; Lim & Ployhart, 2006), with the average correlation between corresponding scales being .73 (Goldberg, 1999). Although NEO-PI-R remains to be the most comprehensive personality test with the ability to predict important outcomes, unfortunately, it is a copyrighted propriety instrument that cannot be deployed freely (Maples et al., 2014). On the contrary, IPIP-NEO has an open-access nature and can be modified to accommodate researchers’ needs

(Goldberg, 1999). In view of these two arguments, employing IPIP-NEO as the external measure to investigate the relationship between the current VSA instrument and self-reported measure appeared to be an applicable solution.

Additionally, during the data collection process, certain strategies were implemented which can be used to tackle issues of bias that external measures are often dogged by. For instance, studies have indicated a private and undiscriminating atmosphere, confidentiality, and data protection assurances prior to the test can contribute to lower social desirability bias and higher response accuracy (Krumpal, 2013). Moreover, it was expected that participants were less motivated to cheat on the questionnaire because they were in a low stake setting where there was no personal gain to fake the answers (Tett & Simonet, 2011). Considering the reasons above, it was believed that the self-report result in this research would be less susceptible to bias issues compared to the ones administered in other settings.

8.2. Appendix B: Self-reported orderliness Survey from Goldberg (1999)

Orderliness (α = .78)

+keyed	Like order.
	Follow a schedule.
	Work according to a routine.
	Like to tidy up.
	Do things by the book.
	Take good care of my belongings.
	See that rules are observed.
-keyed	Leave my belongings around
	Leave a mess in my room.
	Dislike routine.

8.3. Appendix C: Post Experiment Survey on Environmental Orderliness Features

What did you notice in the last two sessions in the scenario phase? (Select all that apply)

□The floor was clean.

□The floor was dirty.

□The wall was clean.

□The wall was dirty.

□The products on the shelf were aligned.

□The products on the shelf were unaligned.

□There was no garbage on the floor.

□There was garbage on the floor.

□I didn’t notice any of the things above.

□Other:

8.4. Appendix D: Orderliness Categories and Corresponding Observables

Self-reported orderliness items	Orderliness Categories	Observables
Like order.	Inclination for order	Scanned the product location and then the product • Did not scan any product
		Scan pick ticket • Did not scan pick ticket at the product location
Follow a schedule.		Grabbed the correct box • Grabbed a box that is too small and then the correct box
		Folded the flaps in the correct order (left and right

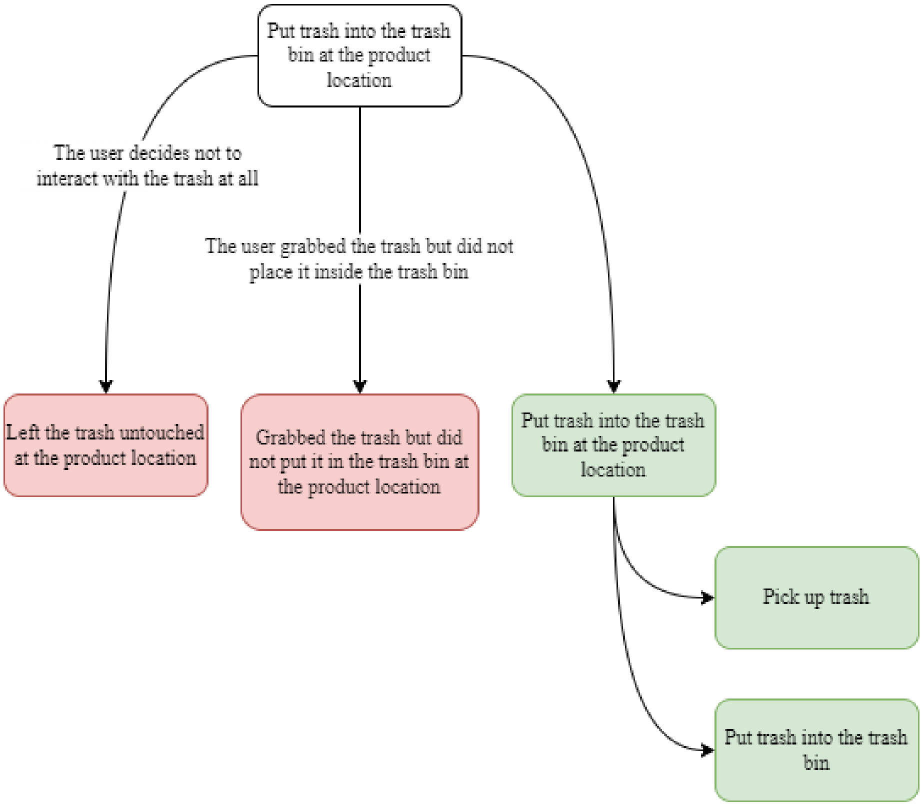
Work according to a routine.		flap first, and then front and back) • Folded the front and back flap first and then the left and right flap • Folded the left and back flap first, and then the right and front flap • Folded the front and left flap first, and then the right and back flap
Do things by the book.		Taped the right side of the box • Did not tape the right side of the box
		Taped the left side of the box • Did not tape the left side of the box
See that rules are observed.		Taped the box from left to right • Did not tape the box from left to right
Dislike routine.		Taped the box from back to front • Did not tape the box from back to front
Take good care of my belongings.	Material Organisation	Placed the scanner back to the correct location at the Check & Sending Station • Did not place the scanner back to the correct location at the Check & Sending Station
		Placed the scanner back to the correct location at Product Location • Did not place the scanner back to the correct location at the Product Location
		Placed the tape dispenser back to the correct location at the Packing Station • Did not place tape dispenser back to the correct location at Packing Station

Leave my belongings around.		Placed the pick ticket in the box • Did not place pick ticket into the box
		Take cart with the product to the check & sending station • Forgot to bring the box with the product to the check & sending station
Like to tidy up.	Tidiness	Put trash into the trash bin at the product location • Left the trash untouched at the product location • Grabbed the trash but did not put it in the trash bin at the product location
Leave a mess in my room.		Put broken cardboard into cardboard bin at the packing station • Left the broken cardboard untouched • Put broken cardboard in the trash can, instead of the cardboard bin • Grabbed the broken cardboard but did not throw it into any bin
		Put the broken tape in the trash bin at the packing station • Left the broken tape untouched • Grabbed the broken tape but did not throw it into any bin

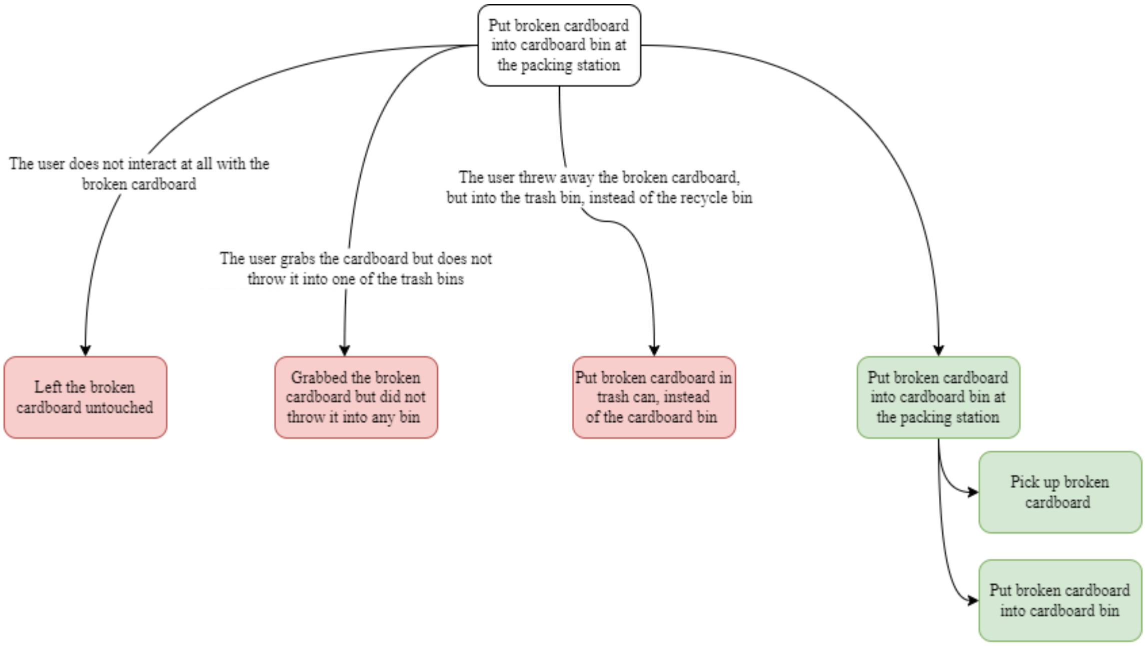
Note. Correct actions are underlined, and potential mistakes are in red.

8.5. Appendix E: Flowcharts of Tidiness Observables in the Assessment

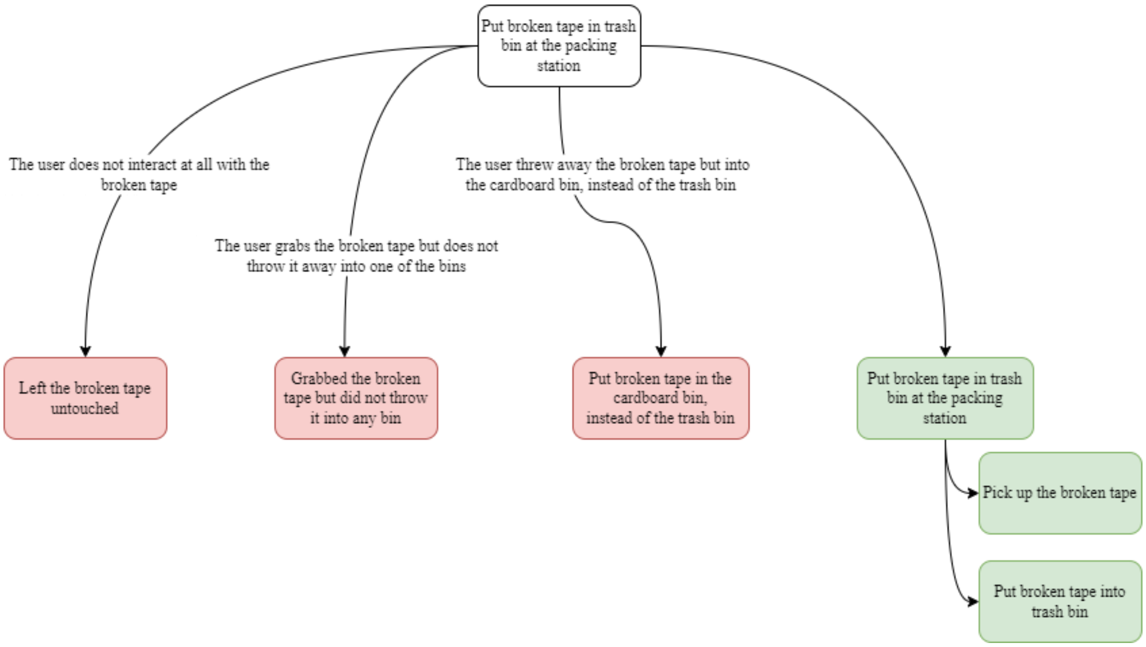
8.5.1. Observable: Put trash into the trash bin at the product location


Put trash into the trash bin at the product location
The user decides not to interact with the trash at all
The user grabbed the trash but did not place it inside the trash bin
Left the trash untouched at the product location
Grabbed the trash but did not put it in the trash bin at the product location
Put trash into the trash bin at the product location
Pick up trash
Put trash into the trash bin

8.5.2. Observable: Put broken cardboard into cardboard bin at the packing station


Put broken cardboard into cardboard bin at the packing station
The user does not interact at all with the broken cardboard
The user grabs the cardboard but does not throw it into one of the trash bins
The user threw away the broken cardboard, but into the trash bin, instead of the recycle bin
Left the broken cardboard untouched
Grabbed the broken cardboard but did not throw it into any bin
Put broken cardboard in trash can, instead of the cardboard bin
Put broken cardboard into cardboard bin at the packing station
Pick up broken cardboard
Put broken cardboard into cardboard bin

8.5.3. Observable: Put the broken tape in the trash bin at the packing station


Put broken tape in trash bin at the packing station
The user does not interact at all with the broken tape
The user grabs the broken tape but does not throw it away into one of the bins
The user threw away the broken tape but into the cardboard bin, instead of the trash bin
Left the broken tape untouched
Grabbed the broken tape but did not throw it into any bin
Put broken tape in the cardboard bin, instead of the trash bin
Put broken tape in trash bin at the packing station
Pick up the broken tape
Put broken tape into trash bin

8.6. Appendix F: Observation Notes in the Debriefing

Reasons why highly orderly people picked up less trash than expected based on participants comments and observation:

- Objects snapped to an unintended location
- Did not notice the trash bin/cardboard bin/tape bin
- Did not notice the cardboard/tape
- Did not think it was their responsibility to clean up common areas
- Did not think it was relevant to their task
- Did not know they could interact with the trash (did not notice the sentence "If you notice any clutter, you are free to clean up" in the information sheet)
- Thought if they paused to clean up, they would get distracted and forget what to do
- Have had a lot of VR experience and therefore did not behave like they would in real life (they were well aware that VR is not real, and they are only playing a game)
- Did pick up the trash but did not recycle properly→too lazy
- Cultural background

Reasons why low orderly people picked up more trash than expected according to participants:

- They may be untidy with their rooms and personal spaces, but they try to keep their working areas and public spaces clean in consideration of other people's feelings
- They get easily distracted by mess when they are working, so they tend to keep their working areas clean

Reasons why people did not notice as many environmental features as expected according to participants:

- Too focused on the task to notice their surroundings
- For experienced VR users, they might not explore the environment as much as first time users and therefore did not notice as many changes
- Participants spent too little time in the assessment so that they noticed few environmental features