The psychometric structure of a game-based assessment

The psychometric structure of a game-based assessment

Liam Kevin Close

This thesis is submitted to Birkbeck, University of London for the degree of Doctor of Philosophy

Submitted July 2021

This is to confirm that the entire work presented in this thesis is the result of my own work.

_____

Liam Kevin Close

July 2021

Abstract

This thesis examines the psychometric properties of a game-based assessment (GBA) by utilizing generalizability theory to estimate the reliability and the effect of aggregation on GBA results. The results of this study provide insight into the theoretical standpoints of trait theory, situational theory and the interactionist perspective, and explores the extent to which the reliability of a GBA can be impacted by a chosen theoretical standpoint.

Secondary data were used within this study, compiled of three cohorts of candidates that completed a psychometric GBA as part of the selection process. The largest contribution to variance in the GBA was the Person × Level × Dimension effect (14.35% – 15.26%), with little variance accounted for by the Person × Level (0.67% – 11.04%) or Person × Dimension effects (<0.01% – 4.09%). These findings support the argument that dimensions have little effect within GBAs.

Aggregating scores to dimensions were found to result in poor reliability estimates (0.36 – 0.66) but when aggregating to an overall score, the G coefficient approached acceptable levels (0.63 – 0.86). However, aggregating to an overall score raises concerns about the interpretability of these scores. Attention within the literature has focused on the validity of dimension-related scores within a GBA, so further research is required to understand the validity of an overall score generated from a GBA.

With a negligible amount of variance associated with the Person × Dimension effect, this suggests that dimensions do not behave in a stable and consistent nature, which contradicts the theoretical underpinnings of trait theory. Little level-related variance was also noted within this study, which suggests that the situation has less overall effect than outlined by traditional situational theorists. However, the Person × Level × Dimension interaction

supports the interactionist theoretical perspective which posits that behaviour is determined by an interaction between the trait and the situation.

This study highlights the need for further understanding of the measurement structure of GBAs. Like other multifaceted measures, dimensions were found to contribute little variance, and the impact of the situation (or level within this study) on scores suggests that there is little or no evidence here that dimensions are cross-situationally stable. Aggregated overall scores offer a more reliable measure of behaviour, but the validity of these scores needs additional investigation to help understand what overall scores measure from a psychological perspective. Further implications for research and practice are discussed.

Acknowledgements

I am indebted to many for their support and guidance offered throughout my PhD. Firstly, to my secondary supervisor's, Dr Chris Dewsbury and Dr Lisbeth Drury, thank you for your wisdom over the years, and the continuous motivation to reach the finish line. To my principal supervisor, Dr Duncan Jackson, thank you for the extended support, the continuous guidance, your insightful expertise, and your limitless patience. Your supervision has developed me in ways I did not expect and I cannot thank you enough for the opportunity to and pleasure of working with you.

I am also grateful for the support and insight offered throughout the process from Professor Almuth McDowall and Dr Rebecca Whiting who have helped develop a sense of community and support in the PhD cohort at Birkbeck. I am also grateful to the Organisational Psychology department at Birkbeck, University of London, who provided me with the opportunity to pursue my passion in this area and offer me the tools to help develop my skills and experience as a researcher. I am also thankful to all the department staff and fellow PhD students who have been part of my journey and have offered insightful feedback and ideas throughout my time at Birkbeck. I am also thankful to Dr Kevin Teoh for their encouragement, advice, general banter, wisdom, and friendship throughout my PhD.

I would also like to thank my parents for their continued love and support in everything I do. Your encouragement to follow my passion and your unyielding belief has been powerful motivation. I would also like to thank my fellow 'Hallows'– Lara Montefiori and Kirsty Lauder. You have been sources of inspiration, motivation and calm throughout this process. You have both helped exponentially, and for that, I cannot thank you both enough. To Liam Dwyer, my biggest cheerleader, thank you for the motivation and support

you have provided. Your kind words and pride have helped me through this process. Finally, I would like to thank Dani Galvan for your continued patience and support throughout this thesis. You have been my steady hand and calming voice throughout this process, and I am thankful to have had the opportunity to share this process with you.

Table of Contents

Abstract **3**

Acknowledgements **5**

Table of Contents **7**

List of Tables **11**

List of Figures **12**

Chapter 1 : Introduction ***13***

1.1 The History of Psychometric Testing **13**

1.2 Personality and Job Performance **15**

1.3 The Evolution of Measurement **17**

1.4 Gamification in Selection **20**

1.5 Game-based Assessment (GBA) **23**

1.6 Measuring Personality in GBA **27**

1.7 Thesis Structure **31**

Chapter 2: The Reliability of Measurement ***33***

2.1 The Operationalisation of Reliability **35**

2.2 Test-retest Reliability (TRT) **37**

2.3 Estimating Reliability Using G Theory **39**

2.4 Reliability in GBAs **42**

2.5 The Application of G Theory to Other Multifaceted Measures **49**

2.6 Aggregation of Scores **56**

Chapter 3: The Theoretical Perspective of Trait Measurement *59*

3.1 Trait Theory **59**

3.2 Cross-situational Consistency of Traits **61**

3.3 The Situationist Perspective **63**

3.4 The Interactionist Perspective **68**

3.5 Theoretical Implications of Aggregation **71**

Chapter 4: Research Aims *75*

4.1 The Reliability of a GBA **75**

4.2 The impact of Aggregation in a GBA **80**

4.3 The Psychometric Structure of a GBA **81**

4.4 The impact of the situation in a GBA **82**

4.5 The Theoretical Understanding of Behaviour Within a GBA **84**

Chapter 5: Methodology *88*

5.1 Secondary Data Analysis and Ethical Considerations **88**

5.3 Participants **92**

5.4 Materials **95**

5.4.1 Constructs measured in the GBA 102

5.5 Procedure **105**

5.6 Data Analysis **107**

5.6.1 Data cleaning and transformation. 107

5.6.2 Descriptive and correlational analysis 109

5.6.3 Psychometric structure and reliability 110

5.6.4 Comparison of different theoretical perspectives 110

5.6.5 Levels of aggregation 111

Chapter 6: Results 113

6.1 Descriptive Statistics 113

6.2 Correlational Analysis 114

6.2.1 Monotrait heteromethod correlations 118

6.2.2 Heterotrait monomethod correlations 119

6.2.3 Replicability of findings across samples 120

6.3 Generalizability Study 123

6.3.1 Replicability of G-Study findings across samples 124

6.4 Alternative Theoretical Perspectives Results 129

Chapter 7: Discussion 133

7.1.1 Cross-Sample Consistency 139

7.2 Variance Components Analysis of the GBA 141

7.2.1 Person × Dimension Effect 143

7.2.2 Person × Level Effect 144

7.2.3 Participant Main Effect 145

7.2.4 Person × Dimension × Level Effect 146

7.3 GBA Reliability 148

7.3.1 Aggregation of scores within a GBA 149

7.4 Theoretical Perspectives to Reliability 153

7.4.1 The trait perspective 154

7.4.2 The situationist perspective 156

Chapter 8: Limitations and Future Research Directions 161

8.1 Practical Implications **161**

8.2 Theoretical Implications **165**

8.3 Limitations **167**

8.4 Future Research Directions **170**

8.5 Final Summary **176**

References ***179***

Appendices ***203***

Appendix I: GBA Invitation Email 203

Appendix II: GBA Levels 205

Appendix III: Candidate Terms and Conditions 207

Appendix VI: Collaboration Agreement 215

List of Tables

Table 2.1 Decomposition of variance in a GBA *49*

Table 5.1 Demographic characteristics of each sample *93*

Table 5.2 The classification of effects across each perspective *111*

Table 6.1 Descriptive Statistics of Item Scores Across Each Sample *115*

Table 6.2 Correlations Between the Composite Dimension Scores Measured in Sample 2 ..... *115*

Table 6.3 Correlations Between the Composite Level Scores *117*

Table 6.4 Correlations of Composite Dimension Scores Measured Across All Levels Within the GBA *122*

Table 6.5 Variance Decomposition of Game-based Assessment Scores *126*

Table 6.6 Composition of Variance and Generalization for Dimension and Situationist based perspectives *132*

List of Figures

Figure 1.1: Image of Gamified Numerical Reasoning Assessment (Arctic Shores, 2017). ...22

Figure 1.2: Image of Game-based Assessment; Wasabi Waiter (Knack, 2016).....24

Figure 5.1: Original image of the Balloon Analogue Risk Task (Lejuez et al., 2002). 97

Figure 5.2: Image of the Balloon Burst Level from the Skyrise City GBA (Arctic Shores, 2018). 98

Figure 5.3: Image of the Focus Group level from the Skyrise City GBA (Arctic Shores, 2018).. 99

Figure 5.4: Image of the Power Source level from the Skyrise City GBA (Arctic Shores, 2018).. 101

Figure 5.5: Instruction screen image taken from BB level in Skyrise City.. 106

Chapter 1 : Introduction

In this chapter, I will introduce psychometric testing and how it has evolved over time. I will then discuss how gamification has been applied to psychometric testing. I will outline where the gaps in knowledge are, and how some of these gaps will be addressed within this thesis. The chapter finishes with an overview of how the thesis is organised.

1.1 The History of Psychometric Testing

Organisations are concerned with hiring the best talent to improve their overall effectiveness (Crook et al., 2011) and to increase productivity and financial outcomes (Brown, 2011; Gatewood et al., 2008; Guest, 1997). Psychometric concern psychological characteristics that can be systematically measured (Anunciação , 2018). Psychometric assessments can be used to select candidates that are more likely to perform better in the role and measure work-related attributes that cannot be assessed through other methods, such as interviews (Van der Merwe, 2002).

However, psychometric testing has a varied history, with candidates being assessed on psychometric characteristics for selection purposes in China dating back over 2,000 years (Bowman, 1989). Overtime, psychometrics has evolved considerably into what we see today. Early personality theorists hypothesised that personality can be dictated by bodily functions like the production of phlegm or blood (Whissell, 2010); participants' interpretation of inkblots (Vernon, 1933); or the shape of a person's skull (Simpson, 2005). Although approaches to measuring psychometric constructs have appeared throughout history, the first mention of psychometrics was established in a thesis written by Cattel in 1886 (Anunciação , 2018).

In regards to personality assessment, Galton (1884) developed the lexical hypothesis, which started to unify and evolve into a psychometric sound and academically robust research area. The lexical hypothesis proposes that personality attributes are encoded into natural language and to measure these attributes, one needs to ensure that the personality trait is sufficiently represented in the personality lexicon (Ashton & Lee, 2005). This means that it is possible to measure intangible aspects of personality with scale items based on how language is used to describe these characteristics. The lexical hypothesis was further investigated in the 20th Century, most notably by Allport and Odbert (1936) who found 17,953 words that could be used to describe behaviour, 4,504 of which they labelled 'trait-names'.

Shortly after this seminal research, Cattell (1947) reduced this list down to 35 variables and used factor analysis to identify 12 distinct personality factors and Fiske (1949) used the same factor analysis methodology to identify four distinct factors of personality. Tupes and Christal (1958) found that individual items loaded onto five distinct factors. Their research design was then replicated by several others who found five broad emergent factors using independent samples (Borgatta, 1964; Digman & Takemoto-Chock, 1981; Norman, 1963).

This was later defined as the five-factor model (FFM; Goldberg, 1990). Different models of personality have been proposed with some authors finding evidence of supplementary traits in addition to the FFM (Ashton et al., 2004; de Vries et al., 2009; Paunonen & Jackson, 2000). Others have identified additional models of personality such as the dark triad which are not encompassed in the FFM (Paulhus & Williams, 2002), but the FFM is the most widely accepted model of personality (Funder, 2006).

Personality research is based on the theoretical understanding of trait theory, which dictates that personality as an inherent disposition is stable across different situations (Costa & Mccrae, 1991). However, although personality has been found, in certain instances, to predict performance, this correlation tends to be quite low, and personality often adds only a small amount of incremental validity to other measures, such as cognitive ability (Salgado, 1998; Schmidt et al., 2016; Schmidt & Hunter, 1998). Nevertheless, personality assessments remain one of the popular methods used in selection based on perceived validity by practitioners (Furnham & Fudge, 2008).

The FFM has come under intense scrutiny since its inception. One of the most prominent arguments being that it lacks theoretical justification, as the premise of the lexical hypothesis is data-driven and does not rely on any theoretical perspective (DeYoung et al., 2007; Eysenck, 1992). Other criticisms include the lack of agreed definition surrounding the trait descriptions (Block, 1995; Eysenck, 1991)and that it does not encompass the entirety of human personality characteristics (Pervin, 1994). However, the model is not meant to be an exhaustive description of personality (Mccrae & John, 1992). It is a higher-order structure that encompasses the majority of personality traits and one of the most widely used and accepted personality frameworks (Bowler et al., 2009; Gosling et al., 2003). The traits in the FFM taxonomy can be labelled differently depending on the test developer, but generally, they represent five constructs: openness to experience, conscientiousness, extraversion, agreeableness and neuroticism.

1.2 Personality and Job Performance

The Underlying Traits of **Conscientiousness** is generally found to be the strongest overall predictor of job performance in the FFM model across all jobs, regardless of

occupational type or job complexity (Barrick & Mount, 1991; Hurtz & Donovan, 2000; Salgado, 1997; Seddigh et al., 2016). It is the extent to which an individual is highly organised, hardworking and dependable as opposed to unorganised, distractible and unmotivated (Salgado, 1997). Those extremely high on conscientiousness are fastidious with perfectionist qualities that can be perceived as negative in certain environments (Rothmann & Coetzer, 2003).

Extraversion is the extent to which people are outgoing, sociable and gregarious, as opposed to timid and reserved (Salgado, 1997). People who score higher in extraversion tend to be more energetic and optimistic, while introverts tend to be more reserved and independent (Rothmann & Coetzer, 2003). Extraversion has been found to predict job performance in sales, managerial roles and customer service roles (Hurtz & Donovan, 2000) and to predict task performance and creativity in the workplace (Rothmann & Coetzer, 2003). Training performance and managerial performance have also been found to correlate with extraversion (Barrick et al., 2001).

Agreeableness is the extent to which a person is more supportive and nurturing as opposed to cold, indifferent and antagonistic (Salgado, 1997). Agreeable people tend to often be altruistic and cooperative, as opposed to less agreeable people who tend to be competitive and sceptical of others (Rothmann & Coetzer, 2003). Agreeableness has been found to correlate with occupational measures such as teamwork (Barrick et al. 2001), professional efficiency (Seddigh et al., 2016) and to predict performance in roles such as customer service (Hurtz & Donovan, 2000).

Neuroticism is a lack of emotional stability and psychological adjustment (Judge et al., 1999). It is linked to a candidate's instability and self-confidence, with highly neurotic people showing less impulse control and an inability to deal with stress productively,

compared to more emotionally stable individuals who are relaxed and even-tempered (Rothmann & Coetzer, 2003). Neuroticism is negatively correlated with job satisfaction and professional efficiency (Seddigh et al., 2016) and with training efficiency (Salgado, 1997).

Openness to experience is defined by the participant's propensity to think creatively, adapt well to change and be tolerant of uncertainty (Mccrae & John, 1992). A person who scores high on openness to experience is likely to be more imaginative and intellectually curious (Rothmann & Coetzer, 2003). In an organisation, openness to experience is linked to innovation, creative problem solving and creating new and unconventional ways of doing things (George & Zhou, 2001). Those lower in this trait are likely to feel more comfortable with convention, practicality and tried and tested means of solving a problem and are less likely to feel comfortable with change (Mccrae & John, 1992; Salgado, 1997). It has the lowest correlation with overall job performance of the other traits in the FFM (Barrick & Mount, 1991) however, a more recent meta-analysis shows that this trait is the second-best predictor of overall job performance (Schmidt et al., 2016).

1.3 The Evolution of Measurement

Since the inception of assessments, computerised technology has advanced, as a result, assessments have moved away from being presented on pen-and-paper, and are now commonly conducted online via a computer. When assessments were adapted from pen-and-paper to computerised versions, little regard was paid to the equivalency of both forms of assessment (Noyes & Garland, 2008). Initial research highlighted a lack of equivalence between pen-and-paper and computerized versions of assessments. This was attributed to a number of reasons including; differing reading speed and accuracy between paper and screen;

and general fatigue experienced between methods and preference frequency of using screens (Dillon, 1992). However, evidence of these findings were collected at a time in which users were more naive in their experience of using computers, and pen-and-paper based assessments were the norm. More recent research has found that that although equivalence is unlikely to be found consistently, as time proceeds, equivalency is better (De Beuckelaer & Lievens, 2009; Noyes & Garland, 2008). This has been attributed to the shift from preference over pen-and-paper to more computerised systems due to participants increased experience with using computers (Vispoel et al., 2001).

One of the benefits of moving to computers to conduct assessments, was the ease at which computer adaptive testing could be applied (CAT). Whereas traditional assessments aligned with classical test theory formats (CTT), CAT could utilize an item-response theory (IRT) approach to assessment. Both CTT and IRT aim to measure a participants standing on a latent, unobservable construct (Rdz-Navarro, 2019), however, CTT treats scores on an assessment as an indicator of a candidates ability level, normally by aggregating or combining scores across multiple items (Embretson & Reise, 2000). In contrast, IRT treats items as individual indicators, and responses to items are determined by a candidates ability level (Reeve & Fayers, 2005). For example, if a candidate has a high cognitive ability, they are more likely to respond correctly to an item that is difficult than someone lower in cognitive ability. Therefore, with the development of computerized assessment, it is possible to automatically present items based on a participants ability level, based on how they responded to previous items, which has led to a decreased assessment length in comparison to traditional testing formats (Embretson & Reise, 2000).

With the steady evolution of technology, alternative forms of assessments are developing such as video interviewing. Originally, this method allowed interview questions to be asked and answered via computer, either in real time with then interviewer present via

online video calling, or via recorded questions (Lukacik et al., 2022). However, now it is possible to apply text analytics and machine learning techniques to gain insights from a candidates response to interview questions, their facial expressions, and the tone of their voice (Ihsan & Furnham, 2018). However, the evidence of this approach is currently lacking, with mixed reaction from a candidate perspective and fear of AI-based solutions (Kim & Heo, 2021). Furthermore, changes in the use of facial expression data within video interview scoring has also been removed from specific AI video interview offerings due to legal actions that have been raised by candidates applying for roles using this method of assessment (Harwell, 2019; Maurer, 2021).

Web and social media scraping is another form of technology enhanced assessment process that makes use of text analytics and machine learning techniques (Chamorro-Premuzic et al., 2016). Previously, it was common practice for employers to screen candidates' social media to ensure that they did not portray themselves in a way that would negatively impact the organizations reputation, however this approach has been found to add bias into the selection process (Acquisti et al., 2020). With more advanced analytics and easier access to data, it is possible to "scrape" this information and use this to infer a persons individual differences (Chamorro-Premuzic et al., 2016). For example, researchers have used data gathered from social media to predict personal information, such as gender, race, religion, drug use, relationship status, political view, and sexual orientation with a high accuracy (65-95%; Kosinski et al., 2013). Within the same study, authors found correlations with psychometric constructs of personality (r= .29 - .43) that has also been replicated in other studies (Schwartz et al., 2013). Furthermore, authors have been able to use data collected from Facebook profile pictures to predict personality with more accuracy than human ratings (Segalin et al., 2017).

This process has also received negative backlash through issues with data privacy. The most contentious being that surrounding that of Cambridge Analytica. This British consulting company gained access to social media users data, and data of their online friends via Facebook, when asking them to complete an online personality survey (Schneble et al., 2018). This data was then used to generate "psychographic" identities of the users, and this information was then used to target personalised advertisements to up to 87 million users on Facebook, which may have played a part in influencing the 2016 US election and UK referendum (Richterich, 2018) . This has led to stricter monitoring of data privacy and data usage, to ensure users are made aware of how their data will be used, and to ensure informed consent is given for the data to be used in this manner (Houser & Voss, 2018).

1.4 Gamification in Selection

Gamification is another form of technology enhanced assessment, and the focus of this thesis. The definition of gamification provided above is a very broad umbrella term. It is the addition of game-like qualities to improve the effectiveness of the process (Armstrong et al., 2016). This gives the ability to harness the fun elements associated with games to improve motivation, attention, persistence and achievement (Richter et al., 2015; Ryan et al., 2006).

For the context of this thesis, there are two levels of gamification that can be applied to psychometric assessments used for selection. The first is gamified assessments, which is a traditional assessment with added gaming elements (Landers & Armstrong, 2017). This form of assessment can be very similar to traditional assessments, but have added game-like

qualities. However, constructs and items will remain the same, changing very little of the psychometric structure of the assessment. The added game elements can include leader boards (Landers & Landers, 2014), the use of avatars (Downes-Le Guin et al., 2012), animation (Fetzer et al., 2017), or more game-like response options (Geimer et al., 2015).

A number of researchers have developed a taxonomy of game elements (Bedwell et al., 2012; Carvalho et al., 2015; Deterding et al., 2011; Fetzer et al., 2017; Jia et al., 2016). However, these can apply generally to gamification and all these elements may not be appropriate or applicable to gamify a selection tool. It is through the addition of game elements that test developers seek to increase engagement by and sense of enjoyment in participants (Bharathi et al., 2016). It has also been suggested that if participants are more engaged and less able to fake their results, this may result in better criterion-related validity of assessments (Geimer et al., 2015). However, due to the lack of research into the area of game-based psychometric testing, this assumption has yet to be supported.

An example of how game elements are applied to a traditional assessment can be found in Figure 1.1. The image is of a numerical reasoning gamified assessment (Arctic Shores, 2018). The assessment measures numerical reasoning in a traditional format where participants are required to calculate the missing number in the sequence. The assessment is timed, and questions are scored as either correct or incorrect. However, game elements have been applied to this assessment in the form of sensory stimuli. The candidate will see animation that will outline the narrative of the assessment; participants are playing a pirate and you sail across an island with a volcano that is about to erupt. Participants encounter a group of ancient statues that contain numbers, some of which are obscured by ash from the

volcano. The participants are required to identify and record the missing numbers in a log before the volcano erupts and the statues are buried. This game-like storyline and

animation are used to increase the engagement and motivation of This game-like storyline and animation are used to increase the engagement and motivation of the participant; however, no changes are made to the constructs being measured, nor the scoring mechanism. The assessment could be presented in a traditional format without the game-like features. This differs in comparison to other levels of gamification which move further away from traditional assessment formats.

Adding different types of gamified elements to an assessment can also allow researchers to measure traits and dimensions that are more difficult to measure in a traditional assessment. For example, measuring visuomotor and visuospatial skills is difficult using traditional assessment methods, but allowing participants to interact with stimuli presented on a screen may allow it and be more accurate than traditional assessment formats allow (Delgado et al., 2014). Adding game-like elements can also reduce anxiety which could result in improved candidate reactions to the assessment process (McPherson & Burns, 2007; McPherson & Burns, 2008; Porter, 1995; Washburn, 2003) .

Figure 1.1: Image of Gamified Numerical Reasoning Assessment (Arctic Shores, 2017). Reprinted with permission.

1.5 Game-based Assessment (GBA)

GBA is a different form of gamified assessment in that it is more advanced and represents a form of enhanced technological assessment that can include more gamified elements such as awarding points for completing tasks, narrative storylines built into the assessment, or adding animation or 3D graphics to an assessment (Fetzer et al., 2017). Games are well suited to measure behaviour. Recording both player choices and the game's para-data (data about *how* the player arrived at their choice; Stieger & Reips, 2010) allows the assessment of information that often cannot be captured by traditional or gamified psychometric tests (Landers, 2015; Shute & Ventura, 2013). The ability to extract this data and connect it to the traits and constructs being assessed is known as stealth assessment and allows the direct measurement of a candidate's behaviour (Shute, 2011). However, depending on how the game is configured, it is not always obvious what information should be extracted from the large amount of data logged by the system (DiCerbo, 2014; Hao et al., 2015). Because user actions can be assessed at multiple levels of granularity, this also provides a large pool of variables to measure traits such as processing speed and learning agility and can develop an in-depth personality profile of the user (Fu et al., 2014; Mandryk & Birk, 2019).

An example of a GBA is shown in Figure 1.2. The GBA is set in an interactive restaurant. The object of the game is to serve as many customers as possible in a given time. There are several competing tasks a candidate has to complete within the level. First, they must serve the customer as quickly as possible, or they earn fewer points (reflected in the star rating next to the avatar). Second, they must choose between serving a customer quickly and clearing dishes to allow more customers to sit down. Third, they must try and serve the highest number of customers possible in the time frame. Depending on the emotion reflected

in the customers' facial expressions, they must also serve them a dish that matches their mood as reflected in the labels that appear next to the dishes.

The technology enabled game elements in this level include video and animation which are used to frame the task. A sense of control is also given to the candidate as they can control how they complete the task, while still adhering to the rules. The whole premise of the game involves interactive problem-solving and the candidate is required to make lots of decisions based on the information that is presented that can be recorded and used as variables to infer personality traits.

Figure 1.2: Image of Game-based Assessment; Wasabi Waiter (Knack, 2016).

With the addition of game-elements to the assessment process, there have been a number of concerns raised about the psychometric properties of this method of assessment. For example, authors have found that when adding game-elements to an assessment, this can change the constructs being measured. McPherson and Burns (2008) designed a cognitive ability GBA, and through the purposes of adding game-elements to the assessment, found that the GBA measured a different cognitive ability (working memory) to what they had anticipated (visual-spatial reasoning). However, little research has investigated the impact of this issue to other forms of GBA.

There are also a number of game elements that can be added to any form of assessment and the effect of these additions has not been fully researched. For example, evidence suggests that specific game elements such as awarding points to participants who complete a task and displaying their results on a leader board can positively motivate some participants whilst demotivating others (Christy & Fox, 2014). When this form of motivation is unrelated to the construct of interest, it becomes unfair to make comparisons between participants as scores from in the GBA may not accurately and fairly represent all participants. Messick (1984) noted that construct-irrelevant variance or added construct-irrelevant difficulty could affect the validity of an assessment. With the addition of game elements into the assessment process, construct-irrelevant variance or difficulty could reduce the effectiveness of the psychometric properties of the assessment such as reliability and validity in comparison to traditional assessments; however, this has yet to be investigated in the literature. The difference in how participants respond to different types of game elements has also been highlighted as an issue in the gamified learning literature (Bedwell et al., 2012).

Although there is little research into the psychometric properties of GBA, there has been a number of papers that have investigated the validity of GBAs. Several studies have been conducted using GBAs which measure key performance constructs. These include measures

of fluid intelligence (Buford & O'Leary, 2015), persistence (DiCerbo, 2014), problem-solving skills (Shute et al., 2016), GMA (Peters et al., 2021) and sustained attention (Godwin-Jones, 2016). Buford and O'Leary (2015), DiCerbo (2014), Shute et al. (2016), and Peters et al. (2021) created GBAs by adapting commercially available games for their purposes: *Portal 2*, *Poptropica*,*Plants vs. Zombies*, and *Minecraft*, respectively), while Godwin et al. (2015) developed their own games to measure specific constructs.

Tests of convergent validity in these studies are promising. Buford and O'Leary's (2015) measure of fluid intelligence had a correlation of $r=$ 0.46 with Ravens Progressive Matrices (RPM) and $r=$ 0.49 with Shipley-2 Block Patterns test. Shute et al. (2016) found that their problem-solving GBA had a correlation of around $r=$ 0.4 with both RPM and another problem-solving assessment, while Goodwin et al. (2015) found that their custom-built GBA *Monster Mischief* had a correlation of r=0.62 with a previous measure of selective sustained attention and a correlation of r=0.52 with the memory components of the task. Furthermore, Peters and colleagues (2021) found a large correlation between their latent g factor measured in their GBA and a traditional measure of GMA (r=0.72).

In an attempt to understand how GBAs can be used to measure work-relevant behaviour, Landers et al. (2017) found that a GBA assessing cognitive ability demonstrated greater incremental validity than Spearman's g, when both were used to predict GPA, which is often used as a proxy for job performance. Furthermore, initial evidence shows correlations between constructs measured in gamified SJTs (Georgiou et al., 2019), and a GBA designed to measured honesty (Barends et al., 2021) shows small correlations with traditional personality measures used in selection (r=.29-.43). Further studies have found correlations between GBA performance and real-world ratings of behaviour. For example, performance in a GBA called *Zoo U* was found to correlate (r=0.48) with external ratings of emotional

skills (DeRosier & Thomas, 2018), and honesty scores generated on a GBA were found to share variance with cheating behaviours (*r*=-.19; Barends et al., 2021).

Several authors have suggested that GBAs, when used correctly, have the potential to predict job performance better than current means of psychometric assessment (Fetzer et al., 2017). However, these tools are under-researched and little evidence is available to support these assertions (Winsborough & Chamorro-Premuzic, 2016). This lack of research, observed by several authors, highlights a need to further develop our understanding of gamified assessments and to actively research in this area (Armstrong et al., 2016; Chamorro-Premuzic et al., 2016; Church & Silzer, 2016; Horn et al., 2016; Lievens & Van Iddekinge, 2016).

1.6 Measuring Personality in GBA

The majority of evidence noted in the previous sections focuses on psychometric constructs related to intelligence, this thesis will investigate GBAs that measure personality constructs, which has received less investigation in comparison to intelligence measures. However, measuring dimensions across different tasks or levels within a GBA, is similar to how dimensions are measured in other multifaceted measures such as assessment centres (ACs) which measure dimensions across a number of different exercises, and situational judgement tests (SJTs) which measures dimensions across different situations (Gatewood et al., 2008). Research has consistently found that within ACs, dimensions account for little variance in scores, with the majority of variance being related to the task (Jackson et al., 2016; Putka & Hoffman, 2013). SJTs also have been found to show little variance associated with dimensions, leading researchers to conclude that these methods of assessment are not

reliable measures of dimensions (Jackson et al., 2017). This issue, and how this relates to GBA will be explored in more detail in the next chapter.

The personality dimensions measured within this thesis are *Creativity*, *Novelty-seeking* , and *Sensitivity to Punishment.* In the workplace, creativity has been found to correlate strongly with job-related constructs, such as organisational citizenship behaviours, task performance and negatively with counterproductive work behaviours (Harari et al., 2016). However, the measurement of creativity is widely discussed within the literature, and proves slightly problematic. Researchers have highlighted concerns associated with trying to measure creativity, which include how to agree on a definition of the construct (Batey & Furnham, 2006; Said-Metwaly et al., 2017). Without a clear definition, this has also led researchers to measure creativity in many different ways, often with findings that show weak validity with different types of creativity assessments (Simonton, 2003). This has also resulted in poor findings of predictive validity (Lemons, 2011) which is also related to poor criterion measures (Plucker & Makel, 2012). Due to the lack of agreement and consistency in findings related to creativity has led some authors to suggest that creativity is too complex to measure using psychometric testing, and at best, it is possible to tap into a small aspect of creativity using this method of assessment (Piffer, 2012).

Although there is much disagreement about how to define creativity, it is believed to be linked to the generation of original and novel ideas (Runco & Jaeger, 2012). Within the GBA used within this thesis, Creativity was found to show a moderate relationship with self-reported measures of creativity (r=.50) indicating a large portion of variance is shared between the traditional self-report measure and the GBA (Arctic Shores, 2018). However, these findings have not been subjected to peer review, and the generalizability of these findings is not clear. Further evidence validity have been found in the tasks used within the GBA which have been found to relate to found to correlate with external measures of self-

report creativity (Agarwal & Kumari, 1982; Fong, 2006). This will be discussed in more detail in the methods chapter.

Novelty-seeking is a less contentious personality dimension, although it has not been investigated as much in terms of it's impact within the workplace. Novelty-seeking can be defined as the tendency to wish to explore new situations and experiences (Gocłowska et al., 2018). It has been found to relate to extraversion and openness to experience constructs of the FFM (Gordon & Luo, 2011). However, openness to experience has been found to rarely show strong evidence of criterion-related validity (Barrick & Mount, 1991; Griffin & Hesketh, 2004; Salgado, 1997), however novelty-seeking has been found to hinder job performance in certain roles (Reio & Sanders-Reio, 2006).

Self-reported scores on novelty-seeking personality scales have been linked to tasks measured in the GBA (Buelow & Suhr, 2009; Lauriola et al., 2013; Suhr & Tsanadis, 2007). This indicates that variance in the tasks used in the GBA is shared with traditional measures of novelty-seeking. The GBA measure of novelty-seeking has also been found to have moderate construct validity when compared to self-reported measures of the same scale (r=.49; Arctic Shores, 2018). However, as mentioned before, the findings from the test developer have not been peer reviewed, and combining variables from different GBA tasks has not received much research.

Sensitivity to punishment is another dimension measured within the GBA that is less common within Occupational Psychology. Sensitivity to reward can be defined as an overall sensitivity, or negative reaction to adverse situations or stimuli (Sava & Sperneac, 2006). Sensitivity to punishment has been linked to neuroticism with highly neurotic individuals displaying more frustration in tasks in which reinforcement levels change abruptly and have been linked to tasks measured with the GBA used within this thesis (Ball & Zuckerman,

1990). Construct validity of the GBA measure of sensitivity to punishment with a self-reported measure of a similar construct shows moderate levels of construct validity (*r*=.46; Arctic Shores, 2018). Initial findings show a link between sensitivity to punishment to stress within the workplace (Taris et al., 2007), however, further research on the relevance of this trait to workplace settings is warranted.

The use of GBA in recruitment is increasing, and as previously mentioned, this has caused academic research to begin falling behind (Lowman, 2016). Evidence suggests that candidate respond more positively when assessed in a game-based format, compared to traditional assessments, and view the assessment as more job-related and face valid (Landers et al., 2021), Furthermore, organizations are bought into this method of assessment showing faith validity in these assessment methods to improve candidate attraction, and measure job-relevant behaviours (Ihsan & Furnham, 2018). These factors are important elements of an assessment, but do not provide insight into the psychometric properties of the assessment (Barends et al., 2021). Best practice guidelines highlight the need for assessments to valid, reliable and fair (ITC, 2001), with some authors suggesting that GBA should not to be used until there is additional evidence of the psychometric properties of GBA (Church & Silzer, 2016).

In this thesis, I aim to answer the call for research in the area of GBA. Specifically focusing on the reliability of this measure as reliability is one of the fundamental building blocks of a measure (West & Finch, 1996). Until now, there has been a focus on the validity of these methods of assessment and the findings are promising. However, there has been little study of how reliable these tools are at measuring individual differences, and this hinders the interpretation of their findings. The findings of this study will add to the knowledge base for GBA and add to our understanding of the theoretical frameworks associated with measurement. The findings will further our knowledge on the psychometric

properties of a GBA and identify how these findings can be used to inform practice regarding the use of GBAs in selection. The International Test Commission states that it is the responsibility of the assessment provider to 'ensure that current psychometric standards (test reliability, validity etc.) have been met' (Bartram, 2001, p. 20). It is also the responsibility of the test users, to ensure they have this information available for tests that they use. If this information is unavailable or under-researched, then using this type of assessment directly contradicts best practice.

1.7 Thesis Structure

Within this thesis I aim to present the first unconfounded reliability estimates associated with GBA, and present key insights into what contributes to behaviour within a GBA, and how this aligns to our theoretical understanding of behaviour. **Chapter 2** discusses the history of reliability, CTT and how we operationalise true score and error. It discusses how reliability has been applied to traditional assessments, GBAs and multifaceted measures and why traditional assessments of reliability such as internal consistency and test-retest reliability are inappropriate for multifaceted measures like GBA. It describes how generalisability theory is more appropriate and how it can be used to understand the psychometric properties of GBA, including its reliability. This chapter will outline G theory, and how I can use this methodology to answer my first 4 research questions (RQ):

RQ1: Are GBAs a reliable measures of personality dimensions?

RQ2: Does aggregating to dimensions and overall scores increase the reliability estimate within GBAs?

RQ3: Does the person main effect, dimensions or levels contribute the most to reliable sources of variance in GBAs?

RQ4: Does the situation contribute a considerable amount of variance to scores in GBAs?

Chapter 3 discusses the theoretical underpinnings of measurement, which is hotly contested in the literature. It outlines trait theory, situationism and the interactionist perspective, the evidence that supports each and how they compare to other theories. It explains that G theory can be used to provide evidence to support these theories and discusses how changing the theoretical perspective can affect reliability. This chapter outlines the different theoretical backgrounds associated with measurement, and sets the scene to pose my final RQ:

RQ5: Do the findings from the G study provide evidence to support trait theory, the situationist perspective or the interactionist perspective of personality

Chapter 4 summarises and concludes the findings and states the research questions addressed in this study.

Chapter 5 outlines the practical part of the study, its measures and analysis and the samples used. **Chapter 6** outlines the results from initial correlational analyses of the data and reports the results. **Chapter 7** summarises the findings and how they relate to our current understanding of behavioural theory. It discusses how the findings add to our current understanding of the psychometric properties of GBA and the reliability of GBAs. **Chapter 8** reviews the main limitations of this study and future research directions.

Chapter 2: The Reliability of Measurement

In the previous chapter, I introduced the concept of game-based assessment (GBA) and how these can be applied to selection contexts. There is currently a severe lack of research in the area of GBA, and some of the fundamental concerns of any form of measurement relates to reliability and the measurement structure of the assessment. Until both the reliability and measurement structure of a GBAs has been established, the interpretation of observations from this method of assessment are called into question. In this chapter, I will describe the history of reliability and how it is viewed as a foundational aspect of measurement. I will discuss some of the forms of reliability that can be estimated, such as internal consistency and test-retest (TRT) reliability, and how these have been applied to GBAs. Lastly, I will discuss generalisability theory (G theory) and how this approach to reliability is underused, and how G theory can be used to evaluate GBAs and their multifaceted measurement structure.

Reliability can be broadly defined as 'quantifying the consistencies and inconsistencies in observed scores' (Brennan, 2010, p.2.). It can be viewed as a necessary part of validity, in that one cannot have validity without reliability. Therefore, reliability affects our ability to interpret. scores in an assessment and is a fundamental aspect of measurement (Ritter, 2010). However, more recently, researchers have been debating this, noting that higher reliability can actually decrease validity, due to homogeneity expected between variables, and note that reliability and validity can be looked at as two separate aspects of an assessment (Strauss & Smith, 2009). Therefore, it is possible to judge both reliability and validity separately, instead of discounting the validity of an assessment due to the reliability, meaning that validity can be present, even with a lower reliability estimate.

This has been argued to help improve predictive validity because it is then possible to align variables within the assessment with the breadth of the multidimensional criterion construct, without limiting the assessment to items that are interrelated and reliable (Li, 2003). However, reliability is still a fundamental aspect of assessment. Although there has been evidence to suggest that validity is present within GBAs, it is also important to understand how reliable these measures are, and to ensure reliability is estimated robustly based on the multifaceted nature of the assessment method.

Furthermore, when estimating reliability through G theory, this also blurs the distinction between reliability and validity (Brennan, 2001), as when different facets are introduced, the findings from a G study can be used to provide an analogue to validity coefficients (Woher et al, 2012). Therefore, the findings from this thesis will help further our understanding of the validity of GBAs.

In occupational psychology (OP), gamification is being applied to the selection process to measure specific traits and abilities that can be used to predict job-relevant behaviours (Georgiou et al., 2019). Without understanding the extent to which these types of assessment provide evidence of reliability, one cannot determine how accurate and interpretable the results are. There have been a number of researchers that have called for more research into the reliability of gamified assessments (Armstrong et al., 2016; Chamorro-Premuzic et al., 2016). This study will add to the limited literature by investigating the extent to which GBAs are reliable and to further develop the literature by understanding the internal structure of GBAs from an academic perspective with the intention of using this knowledge to understand how and if GBAs can be used to guide high-stakes selection decisions. Different approaches exist relating to a consideration of reliability and measurement structure. Depending on the type of assessment and measurement design, some of these approaches are more appropriate than others.

2.1 The Operationalisation of Reliability

Some of the fundamental concepts in reliability theory as they apply to psychology are rooted in classical test theory (CTT). These ideas were pioneered by Charles Spearman in 1904 (Alagumalai et al., 2005). A key premise of CTT is that an observed score is made up of two components: true score variance and error variance (Brennan, 2010; Cronbach & Shavelson, 2004). True score variance is the hypothetical expected score for a trait that a person would achieve if they were tested an infinite number of times (Vispoel et al., 2017). A person's true score is an approximation of what a person would consistently report without their score being confounded by error. This differs from the observed score, which is assumed to be confounded by residual error. An estimate of reliability can be calculated by the proportion of true score variance and to total observed variance (Streiner, 2003).

In CTT, there is one single error variance term that is affected by random error. There are several ways to estimate reliability and, depending on these estimations, error variance will relate to different elements of the measurement condition (Schmidt et al., 2003). For example, internal consistency (most commonly estimated by Cronbach's alpha) is an estimate of reliability that assesses the consistency of responses in a single measurement occasion (Gnambs, 2015). This approach is the most popular method for estimating reliability (Brennan, 2010; Ziegler et al., 2014). The approach was popularised by Cronbach in 1951 and was an extension of the Kuder-Richardson 20 estimate of reliability that could be used on scaled items as opposed to just binary data (Kuder & Richardson, 1937). Cronbach's alpha can be interpreted as the average correlation of all possible split-half reliabilities of a scale and is estimated based on the covariance of items within the scale, and therefore, only needs a single administration (Webb et al., 2006a). Some authors argue that internal

consistency is a useful way to estimate reliability as it allows researchers to see if the content being assessed is consistent, which is one of the main sources of measurement error in assessment and should be applied to all measurement models (Nunnally & Bernstein, 1994). Any residual variance among the items contributes to error and all error variance is averaged among the items in the assessment. This is one of the many drawbacks of Cronbach's alpha, in that longer assessments with more items tend to have a higher internal consistency (Panayides & Karwowski, 2013; Taber, 2018). Therefore, if a test has lower internal consistency, the alpha value can be artificially inflated by increasing the number of items. An example of this can be found in Mun, Mun and Kim (2015)where sub-factor alpha scores of independent scales reach range from 0.25 and 0.6 for constructs with few items (2-4 items per scale), but when combined into sub-factor scores made up of more items, the range increases (0.55 to 0.67) and when all 20 items are combined, the alpha value increases to 0.79 even though the constructs are deemed to be independent.

Other authors have further criticised internal consistency as an estimate of reliability. One of the reasons is the recommended level of alpha needed for it to be deemed reliable. As previously discussed, the length of the assessment increases the reliability of the scale, this may encourage test developers to increase the number of items in an assessment to artificially inflate their alpha estimation. Although the value of 0.70 is generally used as the minimum for reliability, the often misquoted Nunnally and Bernstein paper suggests 0.50 to 0.60 for early research tools and 0.8 for basic research tools (see: Nunnally & Bernstein, 1994). However, others have suggested that there is no optimal level of alpha and that measures with lower alpha levels can still be useful (N. Schmitt, 1996).

In contrast, a high alpha value (>0.90) may also be problematic as it may indicate redundancy in the items and narrowness of the trait being measured, meaning that although the scale may be reliable, the coverage of its content may be less than optimal (Loevinger,

1954). For example, if a scale measures the trait 'conscientiousness', if all the items are highly correlated and the alpha estimate is above 0.90, one could assume that the coverage of these items may not relate to all elements of the broad construct conscientiousness as defined by other measures of the same trait with a lower alpha estimate.

2.2 Test-retest Reliability (TRT)

TRT is another form of reliability estimation that is frequently used in the measurement literature. This form of reliability requires at least two measures of the same assessment across two occasions. To estimate the reliability of the assessment, correlations between results are computed, producing a coefficient relating to temporal stability (Webb et al., 2006). Any residual variance unaccounted for by each scale is associated with transient error, which is defined as variations in participants responding across conditions, which is not related to the constructs being measured (Schmidt, Le & Ilies, 2003). This form of reliability coefficient requires two administrations of the assessment and is not as highly referenced in the literature as internal consistency, with researchers frequently concluding adequate reliability regardless of the size of the reliability coefficient (Watson, 2004). There is further discussion of TRT reliability in measurement theory in the following chapter.

When estimating the reliability of an assessment, error variance is operationalised slightly differently in more traditional approaches such as TRT and internal consistency. Because there is only a single error term in the model according to CTT, TRT and internal consistency operationalise error in different ways. A measure can thus have poor internal consistency but strong TRT reliability as error in internal consistency is related to the covariance between items, whereas TRT is related to differences in responding across conditions and does not take into consideration error associated with the items being

measured (Arterberry et al., 2014; Feldt & Brennan, 1989). Chmielewski and Watson (2009) found very little relationship between measures of internal consistency and TRT reliability (mean r=0.25) indicating only a small amount of shared variance. Therefore, when estimating reliability, it is necessary to take into consideration the different sources of error that can impact the reliability of a measure whereas, using a single measure like internal consistency or TRT in specific measurement designs could lead to the overestimation of reliability (Becker, 2000; Schmidt, Le & Ilies, 2003).

McCrae and colleagues (McCrae et al., 2011) proposed in their paper that there are many different aspects of measurement that can impact reliability. They note that different types of error relate to different estimates of reliability. For example, item irrelevance, which can be defined as items unrelated to the construct being measured, would affect internal consistency but not TRT as long as responses to the items remain consistent across occasions. Item heterogeneity also affects internal consistency, but not TRT reliability for the same reasons, as heterogeneity will have little impact on participant responses across occasions, whereas it could impact the covariance between items, and therefore reduce the internal consistency of a construct. Respondent error is another source of error in which participants carelessly respond to items. This would have an overall impact on TRT reliability but may not have a large effect on internal consistency. Accordingly, if a researcher focuses on one form of reliability over another, they may not be acknowledging the impact different sources of error have on the reliability of the assessment.

Internal consistency and TRT reliability estimates are not equivalent perspectives because they estimate reliability based on different sources of variance, and therefore do not consider the effect of all sources of variance relevant to measurement. This is a downside to traditional reliability estimation, particularly concerning multifaceted measures: there is only a single error term and a single term for true score variance in traditional reliability estimates

and not considering the different sources of error when estimating reliability introduces confounding (Suen & Lei, 2007). Using a more comprehensive reliability estimate would allow the researcher to decompose these multiple sources of variance and identify the extent to which they contribute to true score and error variance.

GBAs, by their nature, are multifaceted forms of measurement, meaning that multiple sources of variance can affect the reliability of the assessment. For example, GBAs can contain items that measure specific dimensions across many levels and these are sometimes measured across different occasions. As internal consistency and test-retest reliability only consider specific sources of variance in their reliability estimation, using these approaches with GBAs would mean that true score and error estimates are confounded with other sources of variance which may distort reliability estimates. Cronbach stated that reliability estimates should be matched to that of the measurement design and that the limiting design of alpha was not appropriate as an estimate of reliability for multifaceted measurement designs. Other methods such as G theory should be used to model additional sources of variance related to measurement (Vispoel et al., 2017).

2.3 Estimating Reliability Using G Theory

Cronbach and Shavelson (2004) suggest that G theory is more appropriate than Cronbach's alpha as an estimate for reliability for multifaceted measures because it allows a two-way interaction (person × item) to be acknowledged in a similar way to alpha, but also permits the introduction of additional components of variance when estimating reliability in assessments with multiple sources of variance such as GBAs. It is therefore is a more appropriate estimate of reliability (Shavelson et al., 1989).

G theory has been described as an expansion of CTT, in that it allows researchers to separate different sources of systematic variability and random error (Arterberry et al., 2014). Unlike internal consistency or TRT, G theory can model different sources of variability and the interactions between them at the same time and give a deeper insight into what variance components contribute to both true score and error variance (Putka & Sackett, 2010). The estimation of variance components in a G study is a critical aspect of understanding the quality of a measurement tool and how best to use and improve it (Le et al., 2009). In G theory studies, sources of variability are referred to as *facets* and can include such measurement conditions as persons, items, raters, occasions of assessment, and tasks (Shavelson et al., 1989). It allows researchers to quantify multiple sources of variance which gives a more detailed indication of reliability as it shows how these different facets contribute to overall variance.

G theory has not yet been applied to GBAs even though GBAs by their very nature are multifaceted. For example, it is possible to measure: persons, items, dimensions, levels (or mini-games) and, if the GBA is completed more than once, occasions. All these facets could confound the reliability estimation if they are not taken into consideration in the reliability estimate, resulting in inaccurate estimates of reliability. Internal consistency does not consider transient error when estimating reliability, nor does TRT reliability consider the effect of the items on constructs when used to estimate the reliability coefficient. However, with G theory, all facets of measurement can be modelled to identify how much the item and dimension-related variance and occasion contribute to overall variance and how these facets interact with the person and each other to account for variance. Therefore, using G theory makes it possible to gain a better and more robust understanding of the reliability of the measure.

When designing a G study, several parameters need to be addressed. Firstly, which facets of measurement can be modelled. Previous studies have attempted to include all facets of measurement, but others have noted that certain facets have been disregarded in the modelling of a G study, resulting in confounded effects (see Jackson et al., 2016). Therefore, careful consideration of the facets of measurements and interactions between these facets is needed before conducting a G study. It is also possible for the researcher to specify which effects specified in the model contribute to reliability, unreliability and irrelevant variance. This can be dependent on the object of measurement – in this case, the person, as they are not a facet of measurement (Fan & Sun, 2014) – and the effects their overall score in the assessment; what contributes noise to the results of their score on a dimension; and what makes no difference in their overall score on a particular construct. Therefore, it is the responsibility of the researcher to isolate which effects are related to reliability and which variance components are related to measurement error (Webb et al., 2006a). This can be based on the theoretical standpoint of the researcher, which is discussed in the following chapter.

Shavelson et al. (1989) noted that when assessing candidates across conditions, the researchers can do one of two things. Firstly, they can estimate both internal consistency and TRT reliability separately, and then they would have to decide which estimate best represents the reliability of their data, however, both reliability estimates would be highly confounded. Secondly, the researcher could conduct a G study that will allow them to model multiple different facets, including items and occasions and estimate the reliability and psychometric structure of the measure while unconfounding the different sources of variance.

In a person × item × occasion G study design, the person, item and occasion effects determine how much variance is due to these different facets without considering any interaction between them. The person × occasion effect relates to how much variance is

accounted for by the person depending on the occasion and relates to how differently a person responds across conditions. The person × item effect relates to how much variance is associated with the person depending on the items. The occasion × item effect is unrelated to the person and accounts for variance that differs across occasions depending on the items and the residual term is all the additional variance unaccounted for in the model. To estimate the reliability of this study, the researcher looks at the ratio between true score and error, or reliable and unreliable and reliability unrelated sources of variance.

Reliable variance can be defined as variance that is related to the object of measurement, in this case the person, so any variance associated with the person (apart from the error term) can be seen as sources of reliable variance. This would include the person main effect, person × occasion effect and the person × item effect. The occasion and item main effects and the occasion × item effect would be classed as irrelevant variance, as they do not affect the between-person comparisons of scores. The residual term would be the proportion of variance related to error. The G coefficient can be calculated as:

Estimated G = reliable variance subtotal/(reliable variance subtotal + unreliable variance subtotal).

2.4 Reliability in GBAs

Literature on reliability in GBAs is sparse, and the studies discussed here come from an array of different applications, which include those relating to learning and teaching, market research and exercise. Because reliability is fundamental to any form of measurement, it is still necessary to estimate reliability in a manner that is appropriate to the conditions of measurement faced by the researcher, regardless of the context in which that measurement takes place.

DiCerbo (2014) conducted a study in which participants completed a GBA that measured persistence. The measure was made up of two variables (quests completed and time spent in the level) across three different game levels. In assessing the internal consistency of the measure of persistence, the author did not consider how the different levels may have introduced an error that was not accounted for in their reliability estimation. Instead of assessing the internal consistency of all six items (person × items), the author might have considered the other sources of variance that contribute to true score and random error. This would involve treating the different levels as a facet, which would allow the author to identify how much variance is associated with this facet. A G study would have allowed the author to identify how much variance depends on the person, items and levels and the interactions between these facets. It would have been possible to identify the extent to which the person effect, regardless of the items or level, contributed to variance in the assessment. Interaction effects would have allowed the researcher to identify how much variance was associated with the levels in the assessment and further our understanding of how stable the scores are across different situations or levels. Although the alphas reported in this study are acceptable (.83 – .85), these results are confounded, and previous research suggests that reliability estimates can be overestimated by 12%– 33% (Vispoel et al., 2018) This research design and approach to estimating reliability has also been observed in other studies, which means that these estimates of reliability are likely to be confounded and overestimated (Quellmalz et al., 2010; Seufert et al., 2016).

Kim and Shute (2015) developed a physics GBA for children that assessed their physics knowledge. They developed two versions of the GBA, one with a linear game-like feature in which levels have to be completed sequentially and one in which participants were allowed to select the order in which they completed the levels. The authors used eight in-game metrics to calculate the internal consistency of the measure, even though they

suggested that this would be inappropriate as they expected the measure was not unidimensional. However, the internal consistency estimates ranged between .50 and .63, which is consistent with more multidimensional constructs (N. Schmitt, 1996). The authors could have estimated the reliability of the assessment using G theory and nested the participants in the conditions, allowing for additional effects, including the participant main effect, to be calculated which would have identified if some participants scored better than others regardless of the items or assessment format, or the effect of nesting the participants in samples. As this effect would have been nested, it would not have been possible to separate this effect from other effects. In a similar study using a similar physics test, Venture and Shute (2013) used a GBA to measure persistence but as the items were not crossed across participants, the authors were unable to compute the internal consistency of the scale. In a different study, Venture et al. (Ventura, Shute, Wright, et al., 2013) also tested a virtual spatial ability assessment in which they assessed the validity of the tool but did not estimate the reliability of the scale. G theory could have been used to estimate the reliability of the scale with items nested in people.

Dubbelt et al. (2014) developed a measure of Machiavellianism that assessed four subfactors in an SJT-like simulation. They investigated the factor structure of the tool and the validity of the assessment but did not estimate the reliability as the tool measured multiple constructs and they believed that due to the multiple constructs being measured it would be difficult to estimate the internal consistency of the measure which in any case was not relevant to their study. Although this is technically correct, the paper cited to support this assertion (McDaniel et al., 2007) does mention that SJTs are multifaceted and not unidimensional so internal consistency is not an appropriate estimation of reliability for an SJT, but it is possible to use G theory to estimate the reliability of this multifaceted measure (Jackson et al., 2017)

Additionally, another study estimated internal consistency differently when using a GBA. In a study in which a GBA assessment was developed to measure participants' pain levels (Stinson et al., 2015), daily scores were collected on 20 items measuring three distinct constructs (pain intensity, unpleasantness and interference). Scores on each scale were averaged across two weeks to formulate a weekly average score. To estimate the reliability of the measure, the authors correlated the average two scores of each of the constructs. They noted that this was a measure of internal consistency. Although internal consistency measures were found to be above .9, this form of reliability estimate is inappropriate for this method of assessment as the situation is likely to be confounding any internal consistency estimate. Furthermore, there is no evidence to suggest the items are internally consistent in regards to the distinct traits the items are measuring. Using G theory would allow the researcher to estimate the person × item interaction (nested in dimension), and identify the extent to which each occasion on which the participant completed the assessment was influenced by transient error and affected overall variance, and how this affects the reliability of the assessment.

In one study the authors developed a GBA to measure the trait personality trait honesty, they measured this trait using three different methods within the GBA, which included a cognitive task, SJT, and a decision making task (Barends et al., 2021). Internal consistency was used to estimate the reliability of the individual levels, as well as the overall honesty score (.45 – .78). This estimate of reliability is also likely to be confounded. It is not possible to isolate the impact of the level on the reliability of the assessment, but furthermore, SJTs alone are multifaceted measures that often measure dimensions across multiple situations (Jackson et al., 2017). Evidence that will be presented in more detail later show through the use of G theory that very little variance in SJTs are related to dimensions being measured (Jackson et al., 2017). Therefore, although there would be added complexity in

estimating the reliability using G theory for this GBA, the authors would likely have a better understanding of not only the reliability of the GBA, but a clearer understanding of what particular facets of measurement are contributing to the scores within this GBA.

Other studies have bypassed internal consistency and only focused on estimating TRT reliability. For example, Mavletova (2015) conducted a study in which participants were required to complete one of three assessments: (a) a text-based assessment, (b) a gamified assessment, and (c) an image-based assessment. The measures in each condition were the same, with the addition of images or gamified elements in the latter versions of the assessments. As this study only estimated the test-retest reliability, it is confounded by other facets that are not accounted for, such as dimension-related effects, or methods-related differences in scores, so the likelihood is that this estimate of reliability is confounded. For example, as previously mentioned when estimating the TRT reliability of a measure, the reliability estimate does not take into consideration the internal consistency of the construct being measured, so any variance accounted for by the items or dimensions within this study are confounded within the TRT reliability estimate, and therefore, we are unaware of how much variance in the overall score in this assessment is related to the construct of interest. Furthermore, the reliability of the assessment ranged between .23 to .85 with lower coefficients in the gamified version compared to the visual and traditional versions of the survey across all dimensions, indicating method-related differences across occasions may be impacting the dimension-related scores. However, this would need to be confirmed through the use of G theory to fully understand the variance associated with the multiple facets within this study design. This study design was also replicated in a study by Tong and colleagues (2016) when estimating the reliability for a cognitive screening tool. Therefore, because the TRT reliability estimates does not take into consideration the internal consistency of the dimensions being assessed, nor the impact of the person × dimesion × occasion interaction,

the reliability of these measure is likely to confounded and less accurate (Suen & Lei, 2007; Vispoel et al., 2017)

There is also evidence that reliability has not been estimated at all in a number of different studies using GBAs (Georgiou et al., 2019; Landers et al., 2017; Lopez & Tucker, 2017; Ninaus et al., 2017). Reliability is a fundamental concern of any type of measurement tool and should be estimated in any study that uses a form of measurement, especially under-researched assessments like GBA where there is a lack of evidence to suggest that behaviour can be measured consistently, and more evidence is required to understand the psychometric properties.

Lopez and Tucker (2017) sought to identify which gamified elements would motivate users to take part in physical activity. Participants were required to complete a number of tasks that involved moving around obstacles in a simulated environment where physical tracking was monitored using Microsoft Kinect. Performance data was captured and analysed, comparing two different gamified conditions. The authors suggest that their results would allow others to understand the effects of game elements on participant performance, but because no estimation of reliability was reported, there are several sources of variance that could have affected the reliability of the assessment and how to interpret the results. Therefore, researchers should be cautious when interpreting these findings. For example, the person effect could explain a large proportion of variance, which may indicate that the condition and tasks contribute less to reliable variance than anticipated. Due to the design of this study, it is also difficult to conclude that one condition is better than the other because participants are nested in conditions, and it is not possible to deconstruct the variance between participants and conditions. It is also important to determine if the task effect contributes more variance than anticipated indicating that, regardless of the person or condition, a large portion of variance could be explained by the task. This would mean that

the task itself may explain more variance rather than, as the authors believe, the condition. Using a more comprehensive approach to estimating reliability will be more beneficial because it will help researchers understand how the different facets of measurement contribute to reliable and unreliable variance.

GBA providers frequently report internal consistency coefficients ranging from .51 to .96 (Williams, 2019). Evidence of the reliability of GBAs has been presented in non-peer reviewed technical manuals for a number of GBA developers. For example, the test developer Arctic Shores (2019) report an average internal consistency of .72 for the personality traits measured in their GBA. The GBA developer Pymetrics (2015) report the internal consistency and TRT of the original non-gamified version of the cognitive tasks used within their GBAs, but do not report any reliability estimate for the scores calculated based on their own assessment, even though they purport to measure dimensions that are measured across multiple tasks/levels . These estimates of reliability are likely to be confounded and G theory is required to deconstruct these sources of variance to estimate a more robust reliability coefficient. In these GBAs, a number of different effects can be modelled using G theory and these are highlighted in Table 2.1. The interpretation of these effects are explained and, depending on your theoretical standpoint, the researcher must determine which effects contribute to reliable, unreliable and reliability unrelated variance. This will be discussed more in the next chapter.

Table 2.1 Decomposition of variance in a GBA

Source of variance	Interpretation
p	Some people will score higher than others regardless of the items, dime levels)
pl	People's scores are dependent on the level.
pd	People's scores are dependent on dimensions.
pld	People's score on a level is dependent on the dimension.
pi:d	People's score is dependent on the items within a particular dimension.
l	Scores are dependent on levels.
d	Scores are dependent on the item within dimensions.
i:d	Scores are dependent on items within dimensions.
dl	Scores on particular dimensions are dependent on levels.
li:d	Scores on items within dimensions are dependent upon levels.
pi:dl + residual	Error term based on four-way interaction between persons, items withi dimensions, levels +residual term.

Note. p = person, l = level, d = dimension, i = item, G = Generalizability Coefficient

2.5 The Application of G Theory to Other Multifaceted Measures

GBAs are a multifaceted measure and traditional approaches to estimating the reliability of such a measure are inappropriate. G theory is a more robust estimate of reliability for assessments that are multifaceted such as GBAs (Clarke-Midura & Dede, 2010). As G theory has yet to be applied to GBAs, this section will outline how it has been applied to other multifaceted measures, such as ACs, SJTs and simulations, and how the findings from this literature may relate to the reliability and psychometric structure of GBAs.

ACs are similar to GBAs in that researchers generally measure different dimensions based on items across different tasks or exercises based on ratings by trained assessors (International Task Force on Assessment Center Guidelines, 2000) . Through the use of GBAs, researchers measure dimensions based on items across a number of different levels or mini-games, differing only in that GBAs are automatically scored and do not require assessor

input. Due to the lack of assessors in a GBA, the effect of assessor-related variance contributing to overall variance is eliminated, meaning that this facet of measurement will not affect the reliability of the assessment, whereas in ACs it is a facet of measurement that is modelled in a G study.

In an AC, there are a number of different facets and interactions that can be modelled, far more than outlined in Table 2.1 for a GBA. Putka and Hoffman (2013) used four variance components as sources of variance: the candidate effect; candidate × dimension interaction; candidate × exercise interaction; and candidate × dimension × exercise interaction. These effects were deemed reliable as they all related to the object of measurement, the person. Each interaction considers the effect of the person on the different facets of measurement related to reliable variance, that being the dimension and exercise and the interaction between the person, and both these facets of measurement. Unreliable variance relates to sources of variance that are unrelated to the person or any assessor-related variance. This includes: candidate × assessor; candidate × dimension × assessor; candidate × exercise × assessor; assessor main effect; assessor × dimension; assessor × exercise; assessor × dimension × exercise; and the residual term. Other components that are unrelated to reliability include: dimension main effect; exercise main effect; and dimension × exercise effect. These effects do not impact a candidate's between-participant percent of variance and are deemed irrelevant to reliability for relative scores (Shavelson et al., 1989).

When investigating the reliability of ACs Jackson et al. (Jackson et al., 2016) noted 14 additional AC-related effects when they nested items in dimensions and candidates in samples that were not considered in the study by Putka and Hoffman (2013). Although these differences resulted in similar G coefficients (.77 and .80), the authors concluded very different interpretations of the findings. While variance accounted for by dimensions in the AC was relatively low across studies (0.54% – 1.10%), Putka and Hoffman (2013) concluded

that dimensions are still a fundamental component of an AC due to the three-way interaction between the candidate × dimension × exercise effect, which accounted for between 12.75% and 23.40% of variance across both studies. Jackson et al. (2016) attributed this to the 'exercise effect' frequently found in the AC literature, in which correlations tend to be higher between different dimensions scored in the same exercise in comparison to correlations between the same traits across different dimensions (Bowler & Woehr, 2006; Lance et al., 2004; Lievens, 1999; Lievens & Conway, 2001; Sackett & Dreher, 1982; Woehr, 2003). This means that dimension-related behaviour is less consistent across exercises, and that performance in one exercise is likely to be consistent across all dimensions measured within the same exercise, providing evidence of a situational effect (Jackson et al., 2016). This is supported by the candidate × exercise interaction in both papers accounting for the majority of reliable variance (35.20% – 36.76%). The exercise effect has been consistently replicated in the AC literature (Lance, 2008; Lievens & Christiansen, 2010) and affects the ability to meaningfully interpret scores in an AC (Bowler & Woehr, 2006; Kuncel & Sackett, 2014) assessor-related effects contributed little to overall variance in both studies, suggesting that, although this is not a source of error in GBAs, the effect it has on the reliability of ACs is small.

Within a study of psychometric structure of *multiple mini-interviews*, participants were asked to complete a range of tasks including interviews and role-plays across three different situations (Breil et al., 2020). The interview process resembled a type of AC in that multiple dimensions were measured over a number of different tasks and rated by assessors (Povah & Povah, 2012). Similar to the findings above, little overall variance was associated with dimensions (1.87%), while the person × situation (26.37%) and person × dimension × situation (9.47%) accounting for the majority of reliable variance. This provides further evidence of the exercise effect.

Some researchers do not perceive lack of consistency as a problem in ACs and, in some cases, believe it should be expected. They assert that across-exercise dimensional consistency is not the point of an AC and that multiple exercises are used to broaden the scope of the construct and make it more aligned with the domain of interest across a range of contexts that are job-relevant (Neidig & Neidig, 1984). The authors further argue that dimension related stability should not be expected, and is not associated with measurement error because parallel exercises should have stronger correlations, but with different exercises, variability in responding should be expected. Therefore, if ACs were made up of parallel exercises, they would be redundant and not adding additional variance to the criterion. A range of exercises may broaden the construct of interest to more fully measure the criterion construct, and should not be assumed to be measurement error, but true score variance (Howard, 1997; Lance et al., 2000).

Additional multifaceted measures such as SJTs have used G theory to estimate the reliability of this type of measure. Jackson et al. (2017) explain that in a standard SJT, there are seven effects, four of which are related to score: the candidate main effect; candidate × dimension effect; candidate × situation (nested in dimension) interaction; and the residual term. As the situations are nested in dimensions, it is not possible to separate the effects, but it may be possible in other SJTs that follow a different design.

Jackson et al. (2017) found that when considering the dimension-based approach to SJTs, the majority of variance was related to the residual term (50.68% – 71.82%) and that dimensions accounted for very little reliable variance (0.58% – 11.44%). Very little variance was related to the candidate × situation (nested in dimension) interaction (3.85% – 5.18%) suggesting that, unlike ACs, little variance was related to the situation. This has been noted in other papers investigating the impact of situations in SJTs (Krumm et al., 2014). However, within SJT participants respond based on how they would react to different situations,

whereas ACs put people in different situations to complete a task and this may account for the differences in observed variance. The main contribution to reliable variance was the candidate main effect (18.55% – 34.02%). This can be interpreted as a general factor in which that some participants score higher than others regardless of the item, dimension or situations associated with the assessment.

Simulations are similar to SJTs or ACs in which participants are scored on dimensions across different tasks are also multifaceted measures and therefore estimating the reliability of this type of assessment necessitates the use of G theory. Few studies have investigated the reliability or psychometric structure of this method of assessment using G theory and tend to rely on traditional measures, like internal consistency (Catano et al., 2012; Vispoel et al., 2017). However, Quellmalz et al. (2013) used G theory to decompose multiple sources of variance in a science simulation assessment. Three types of assessment were used, with varying elements of interactivity between the participant and simulation (no interactivity, some interactivity and strong interactivity). In each, three constructs were measured. Although all participants completed all three forms of the simulation, the authors estimated the G-coefficient for each assessment type and each assessment construct separately to ascertain a reliability estimate for each construct and assessment type (.53 to .79). The main source of variance associated with reliable variance was the person × item effect (63.09% – 78.00%). However, in doing so, the results were confounded. For example, as the scores on each construct were separated by method of assessment, it was not possible to determine the extent to which the method, or any interactions with the items or constructs, contributed to variance and compare this to any dimension-related effects. In conducting the study in this way, the authors were estimating a proxy of internal consistency for each dimension in each assessment format by conducting a p × i design G study. Although they

intended to highlight the differences across the methods of assessment for each construct, the design and estimates of reliability are confounded in this study.

Although not directly related to reliability, a study was conducted using a commercial videogame to investigate the effect of situational factors on performance (Jackson et al., 2016). Participants were separated into two conditions: those who completed the same videogame situation three times (condition 1) and those who completed three different game situations (condition 2). The authors found that situational-related variance (p × s) explained more variance in condition 1 (47.5% vs 31.03%) meaning that situational effects accounted for more variance when the situation was repeated, in comparison to when participants were subjected to different situations. This has implications for GBAs as participants may be subjected to repetitious situations in a level and scores may vary as a result of intraindividual behaviour, as opposed to consistently measuring performance.

The participant main effect was much lower in condition 1 (4.21% vs 24.78%) indicating that different situations result in an increased participant main effect, which relates to some people scoring higher regardless of dimensions and situations. This may affect GBAs as candidates complete several tasks and if the participant main effect contributes significantly to reliable variance, the effect of the dimensions may be greatly reduced in comparison which may hinder our ability to quantify what is being measured in a GBA. Dimensions related variance (p × d) also contributed to variance, albeit to a lesser extent in both samples (12.48% vs 17.53%). According to the AC and SJT literature, dimensions play a relatively small role in performance in a videogame further highlighting the need for more research into GBAs to identify if this method of assessment promotes the measurement of dimensions in a meaningful way.

Other multifaceted measurement designs have found a similar effect in which method effects (akin to situations) accounting for more variance than dimensions. For example, when comparing personality traits based on peer ratings (self, peer and parent) small correlations were found between the ratings of the same traits between groups, but higher correlations were found between different trait ratings within the same group indicating method-related variance (Biesanz & West, 2004). This has previously been investigated with self-report personality data using multi-trait multi-method (MTMM) analysis where monotrait heteromethod (same-trait, different-method) relationships were found to be lower than heterotrait monomethod (different trait, same method; Eid et al., 2003). This method has also been used to investigate the impact of the validity of ACs (see Bowler & Woehr, 2006). Although these findings indicate that hetero-trait mono-method relationship were stronger than monotrait heteromethod relationships, MTMM approaches to measuring AC data has been found to be problematic and G theory is a more preferred approach (Woehr et al., 2012).

The research from multifaceted measures such as SJTs and ACs indicate that the psychometric structure of these assessments are less aligned with dimensions, and relate to other facets of measurement, such as exercises or general factors. This has an impact on the interpretability of the scores, as if dimensions are not contributing variance to the scores, it does not seem appropriate to report scores, make decisions, or give feedback based on dimensions. The psychometric structure of GBA has yet to be investigated using G theory, and as GBAs can measure multiple dimensions across different levels in a similar way to ACs, then the impact of the dimensions on scores will have a direct impact on how these scores should be interpreted (Kuncel & Sackett, 2014). Further insight is required to identify the impact dimensions and situations have on scores within a GBA, and will be investigated directly within this thesis.

2.6 Aggregation of Scores

Results in psychometric assessments are rarely reported at the item level, and within multifaceted measures, are normally aggregated to different levels (see: Jackson et al., 2016, 2017; Putka & Hoffman, 2013). Furthermore, it is recommended to interpret the reliability of a scale based on the levels of aggregation that are of interest (Jackson et al., 2017; LoPilato et al., 2015; Vispoel et al., 2017). Aggregation is used in psychometric assessments to allow correlated variance to increase and uncorrelated variance to decrease (Kuncel & Sackett, 2014). Thus, as we calculate scale or dimension-based scores, construct-relevant covariance of individual items should increase, and construct-irrelevant variance should decrease. In SJTs, it is possible to aggregate scores based on items, dimensions, situations and overall scores. Jackson et al. (2017) found that G coefficients were low when aggregating across items, dimensions and situations (.02 to .49) whereas when aggregated to the overall level, the G coefficient increased to near-acceptable levels (.51 to .75). Although in some samples this may be seen as acceptable in terms of reliability, the ability to distinguish candidates between scores on situations or dimensions, as they contribute very little to overall variance, is not recommended due to the lack of contribution of these facets to overall scores. Aggregation has also been used in the AC literature and found to improve the reliability estimates, depending on the level of aggregation, but to a lesser extent than found in SJTs due to the larger G coefficients observed before aggregating (Jackson et al., 2016; Putka & Hoffman, 2013).

The findings on the reliability of multifaceted measures are mixed. The AC literature has consistently found that exercises account for the majority of variance, with person × exercise and person × exercise × dimension-related effects accounting for the most variance.

GBAs are similar in design to ACs and measure dimensions across different tasks (or levels) in the GBA. It will be a novel approach to investigate the extent to which the exercise effect is reflected in a GBA. To the best of my knowledge, no study has yet used G theory to deconstruct the multiple sources of variance related to GBAs and this is the first to examine if this effect is also observed in a GBA, thereby furthering the academic literature in the understanding of multifaceted measures.

Aggregating to dimensions, exercises and overall scores in ACs resulted in minor changes to G coefficients (.89 to .97) and (.74 to .90) in Jackson et al.'s (2016) and Putka and Hoffman's (2013) studies, respectively, indicating an increase in reliability depending on the level of aggregation. However, in the SJT findings, aggregating to situations, items and dimensions still resulted in sub-par reliability estimates (.02 to .49; Jackson et al., 2017) until aggregated to the overall level (.51 to .75) indicating a significant improvement in some samples. However, little is known about the effect of aggregating across different dimensions, levels or overall score in a GBA. No study has applied aggregation to GBAs, so this thesis will add to the literature by determining to what extent aggregation at the dimension, level and overall level will affect the reliability of a GBA.

Simulations and ACs have been found to predict job performance (Moser et al., 1999; Rene et al., 2011) . These tools are designed to reflect job-relevant tasks and it is likely that even though we may not know what is contributing to performance in these assessments, they are highly job-related. GBAs can be lower in fidelity and can sometimes show little overlap with job-related behaviour, meaning that if level-related variance explains the most variance in a similar way to ACs, it might show less job relevance and therefore be less predictive than ACs. If the tasks are unrelated to the role, unlike SJTs and simulations, aggregating to an overall level may increase reliability but show little job relatedness or predictive validity. Therefore, it is important to approach this gap in research and identify if GBAs are reliable

and to what extent different facets of the GBA, be they dimensions, or error, contribute to overall variance as this will allow researchers to gain a better understanding of the internal structure of a GBA. Further research should build on these findings to understand what is being measured and how this relates to criterion-related behaviour. With the use of G theory, it will be possible to advance our knowledge of the theoretical underpinnings of measurement by taking differing approaches to segment sources of variance in the G study. This will be discussed in more detail in the following chapter.

Chapter 3: The Theoretical Perspective of Trait Measurement

In the previous chapter I reviewed the importance of reliability, and the most common types of reliability estimates that are reported, and how it has been applied to GBA within the literature. I noted that there are large gaps within the literature in reference to how reliable GBAs are and specified that the few studies that do estimate the reliability of GBAs do not take into consideration their multifaceted nature, which introduces confounding into the reliability estimate.

In this chapter, I outline the theoretical underpinnings of behaviour and how this relates to measuring personality traits. I will then outline how alternative measures, such as ACs, have been used to further our theoretical understanding of behaviour. Furthermore, as I will be using generalizability theory to deconstruct the different sources of variance related to measuring behaviour within a GBA within this thesis, I will discuss how depending on the theoretical perspective you adopt, differences can be observed in their interpretation of what constitutes as a reliable source of variance. Additionally, I will highlight how the findings from this thesis could help further our understanding of theory related to trait measurement.

3.1 Trait Theory

The stability of personality is one of the fundamental concepts that underlie trait theory. In trait theory, it is assumed 'that people have stable and identifiable response dispositions that can be mapped by test scores' (Hogan et al., 1977; p.297). Authors have argued that personality traits are real attributes of individuals that can be measured (Funder, 1995), while others view traits as a description of behaviour that are used to predict future behaviour and not inherent dispositions (Hogan & Foster, 2016). Regardless of these views,

trait theorists contend that traits are stable and can be used to predict behaviour across different situations. The theory generally follows two assumptions: (a) that there will be differences between how people respond to the same situation; and (b) that people will show some level of consistency in their responses across situations (Ross & Nisbett, 1991).

In selection, according to trait theory it is possible to measure latent traits which can be used to predict how a participant will behave. Traits are a set of natural predispositions that are inferred through patterns of behaviour that are stable across time and between different observers (Costa & McCrae, 1991). For example, variability in extraversion can be used to discriminate between candidates who need to interact with others as part of their job. According to trait theory, those who score low in extraversion are likely to be more shy and timid regardless of the situation, while those who are more extroverted are more likely to display pro-social behaviours across situations. Evidence of this has been found in studies that involve interpersonal elements as part of a person's role. For example, extroversion was found to predict performance in police officers (ρ=0.20; Salgado, 1997) and people in sales (ρ = 0.15), customer service (ρ = 0.11) and managerial roles (ρ = 0.12; Hurtz & Donovan, 2000).

Reliability is one of the fundamental principles of measurement, and one way to assess the reliability of a measure is through temporal stability, in which an assessment is taken across two or more occasions, and the scores from each occasion are correlated. If, according to trait theory, traits are stable and consistent measures, temporal stability estimates should be high, then and any residual variance would be associated with error in measurement (Meade, 2004). Many studies have investigated the temporal stability of personality assessments and have found evidence of reliability. Gnambs (2014) analysed 84 studies that investigated test-retest reliability of personality traits scores and found evidence

of stability across occasions in the study (ρ=0.82). This suggests that personality traits are relatively stable.

3.2 Reliability of the FFM

Internal consistency estimate's reliability based on the correlations between the items in the scale and is the most popular estimate of reliability (Streiner, 2003). If we imagine that items in the scale are indicators of how a person will behave, in that someone high in a particular trait will score high in all the individual items and someone lower on the trait will score lower on these items, this is evidence that behaviour is consistent across all the items in the scale. For example, in the IPIP FFM measure (Johnson, 2014), two items; 'I love large parties' and 'I try to lead others' relate to extraversion. If the majority of respondents rates themselves highly in both questions, then this provides additional evidence of the stability of traits, and will also result in a higher internal consistency coefficient if the sample answers consistently across similar items. However, lower internal consistency would indicate that people respond less consistently to items intended to measure the same construct. As discussed earlier in this section and in the previous chapter, reliability of the FFM has been investigated in great detail within the literature and evidence of the temporal stability and internal consistency of the FMM has been found to be adequate or better. This provides further evidence that traits are stable across conditions and aligns with the theoretical underpinnings of trait theory.

3.2 Cross-situational Consistency of Traits

To understand the stability of traits across different types of situations, researchers have investigated the cross-situational consistency of measurement using multifaceted measures containing multiple sources of variance. In separating these sources, it is possible to identify if traits are stable across situations, and how much variance is associated with the situation. For example, in ACs, dimensions are akin to traits (e.g. leadership skills, communication skills, problem-solving) and are measured multiple times across different exercises. Trait theory assumes consistency of behaviour across situations; therefore, trait theorists posit that people who score high on a particular dimension will score high in the same construct across different exercises. For example, if a participant scores high in the dimension communication skills in a presentation exercise, they would be expected to score highly in the same construct in a different exercise such as a group discussion or a roleplay task. This type of method of assessment is similar to GBA, in that participants are asked to complete several tasks or 'levels' in which a number of dimensions are assessed on each level. Therefore, participants who score high on a construct in one level, according to trait theory, should also score highly on the same construct in a different task.

Previous research suggests that cross-situational consistency is not observed in ACs. Sackett and Dreher (Sackett & Dreher, 1982) used factor analysis to analyse the data from three different AC samples. In all three, the factors did not represent the dimensions being measured but represented the exercises in the AC. This was a result of high in-exercise ratings on different constructs and low correlations between the same dimensions across different exercises. For example, participants who scored high in 'communication' in the role-play task were more likely to score highly in the 'organisational skills' and 'leadership skills' in the same exercise, but these scores were less likely to be consistent in the same dimensions across different exercises. This is what is known as the 'exercise effect' and has been replicated consistently in the AC literature (Dilchert & Ones, 2009; Haland &

Christiansen, 2002; Lievens et al., 2006; Sackett & Dreher, 1982). This poses an issue as we are unsure of what is being measured in an AC. However, the criterion-related validity of ACs is promising with overall assessment results correlating with job-performance ratings (*r*= .25 – .39; Arthur et al., 2006).

In reference to GBAs, there has been no evidence to-date that has investigated if dimension scores are consistent across levels, or if the exercise effect manifests in a similar way across levels. Therefore, the findings of this study will contribute to our understanding of the consistency of behaviour. I will investigate the relationships observed between dimension scores across different levels within the GBA to see if dimension-related behaviour is consistent. Furthermore, through investigating the psychometric structure of the GBA, I will be able to ascertain the extent to which dimensions contribute to scores within a GBA.

Evidence of the exercise effect contradicts trait theory as dimension- related behaviour has been found to be inconsistent across situations. Further theories of personality, including situationism and the interactionist perspective add additional explanation to what has been observed within the literature in regard to the stability of assessment. In the following section I will discuss how situationism differs from trait theory and how this impacts our understanding of measurement.

3.3 The Situationist Perspective

Within a GBA, it is possible to measure dimensions across different levels or tasks, and these levels are akin to different situations. According to the trait perspective, traits should be consistent across situations, however, as will be discussed below, this is not always the case.

The classical theory of behaviourism suggests that individual behaviour is based on each person's unique learning history (Funder, 2006). Behaviour is reinforced through stimulus-response style learning, in that once a behaviour has been reinforced, when presented with similar stimuli the person is likely to behave in the same way. For example, when a person has worked under a manager that reinforces the need to be less dominant and more subservient, it is likely they will maintain this behavioural response when in this situation. The same person may behave more dominantly and take a leadership role when this behaviour is reinforced in a different situation, like when playing competitive team sports. In contrast to trait theory, behaviourism dictates that behaviour is related to antecedent cues based on the environment rather than stable and inherent latent traits that impact behaviour (Skinner, 1974). Although classical theories like behaviourism can explain why behaviour may vary depending on the situation and how people have been individually reinforced to behave in a specific way, it does not explain what personality is, how it is measured, nor what affects it (Staats, 1996). One of the biggest criticisms of behaviourism is its lack of theoretical underpinnings (Delprato & Midgley, 1992). Advocates of situationism such as Walter Mischel (1968) have argued that individual differences account for some of the variance in behaviour, but once this is taken into account, the amount of variance attributed to traits becomes trivial. It was Mischel's controversial stance on trait theory that brought attention to alternative perspectives to trait measurement.

Trait theory was one of the most widely accepted measurement theories until 1968 when Mischel became an advocate for situationism. Commentators defined his book as a 'full-blown frontal attack' on personality psychology (Van Heck et al., 1994). Mischel suggests that, although the premise of trait theory was logical and plausible, the empirical evidence to quantify the theory was lacking. He noted that the cross-situational consistency of trait measurement had a predictive ceiling effect of .3 when correlated with other

behavioural measures. Thus, although there is evidence that individual differences show some level of consistency across conditions, a large portion of the variance is explained by other factors. Previous studies have found Mischel's prediction of cross-situational consistency effect size to be accurate ranging between .19 and .20 (Richard et al., 2003).

Evidence of the effect of the situation has been found in many studies. For example, in the classic experiments by Stanley Milgram (1965), the relationship between obedience and inflicting pain on others was related to the level of victim isolation and proximity of commanding authority, rather than any internal trait related to moral standing and personal values. This study was rather controversial, and the findings reiterate the notion that situational characteristics play a more prominent role in observed behaviour as opposed to inherent traits. Similar results were also found in the Stanford prison experiment in which participants' behaviour was dependent on the role they were assigned to, be that prisoner or guard (Zimbardo, 2004). Festinger & Carlsmith (1959) also found when asking participants to lie, there was a relationship between lying behaviour and the amount of monetary incentive offered to the participant. These studies all subjected participants to different situations and the findings suggest that the situation affects individuals' behaviours and that traits are not as stable across situations as anticipated.

Other authors have investigated the effect of the situation on behavioural responses without relying on experimental manipulation. Funder and Ozer (1983) reviewed several different studies in which situational variables were correlated with behavioural responses. The average correlation between these variables across studies was .38, which suggests that a substantial proportion of variance can be explained by the situation. Researchers have also investigated the effect of the situation through multi-faceted measures, such as ACs, which measure behaviour across different situations. By separating each source of variance

associated with multifaceted measures, it is possible to investigate how much variance in measurement is associated with the situation.

Situationism in multifaceted measuresPutka and Hoffman (2013) used G theory to decompose the multiple sources of variance present in ACs, identifying additional sources of variance not accounted for in previous studies (Arthur et al., 2000; Bowler & Woehr, 2009). The candidate × dimension variance component is the amount of variance related to dimensions being dependent on the candidates, meaning that some candidates are likely to score higher on certain dimensions irrespective of other factors. If dimensions are perceived as stable, this would account for a large amount of variance. Previous authors have estimated this interaction to account for 18-27% of the variance (Arthur et al., 2000; Bowler & Woehr, 2009). However, these results are confounded due to a misspecified model that does not consider all possible sources of variance. In Putka and Hoffman's (2013) study, this estimate was much lower and accounted for 0.4% – 1.4% of the variance, indicating that a small amount of variance is associated with dimension scores.

The candidate × exercise interaction explains how much variance in the assessment is dependent on how participants respond in the exercise regardless of the dimensions being measured. This accounted for a small amount of variance in Arthur et al.'s (2000) study (6.8%) and between 32% and 40% across the additional samples (Bowler & Woehr, 2009; Putka & Hoffman, 2013). This suggests that a large amount of variance is dependent on how participants respond to the different exercises, rather than how they respond to specific dimensions. This provides further evidence of the effect of the situation, as different exercises represent different situations. If behaviours were consistent across exercises, only a small amount of variance would be explained by the candidate × exercise effect.

In a more recent study, Jackson et al. (2016) expanded on the 14 effects modelled in Putka and Hoffman's (2013) study and took into consideration 29 effects within their study, and used a more complex form of analysis that allows for a more accurate decomposition of variance. Some of the findings in both studies were very similar in that a large portion of variance was associated with the candidate × exercise interaction (35.76%), and little variance was accounted for by candidate × dimension interaction (0.54%; Jackson et al., 2016). In contrast, comparing the variance associated with the candidate × dimension × exercise interaction, Putka and Hoffman found with their confounded estimates around 23% of variance was associated with this interaction, whereas Jackson and colleagues found 12.75% of variance was associated with this interaction. This effect can be defined as a participant's behaviour on a specific dimension being dependent upon the exercise. Jackson et al. (2016) found a smaller amount of variance associated with this effect and concluded that any dimension related effects are highly dependent on the situation, which provides additional evidence that behaviour is situationally specific, but also raises the issue of interpreting dimensions as meaningful measures of behaviour if they account for very little overall variance. In contrast, Putka and Hoffman's (2013) attributed this to another theory of measurement, the interactionist perspective, and that even though dimension-based variance was low, that the importance of dimensions should not be undervalued.

As mentioned previously, dimensions within ACs can be expected to be stable behavioural categories that are distinct within exercises, while remaining consistent across exercises (Lance, 2008; Sackett & Dreher, 1982). However, the majority of variance has been found to be related to the exercises with very little related to dimensions. Some authors contend that variation in trait scores across exercises is to be expected and that participants are naturally going to differ in how they perform on different exercises (Howard, 1997; Neidig & Neidig, 1994). This could explain why we observe variance attributable to the

candidate × exercise, and candidate × exercise × dimension interactions, but it still doesn't solve the issue related to giving feedback or making high-stakes selection decisions on dimension-related performance when the situation has a larger impact on scores. This has led to proponents of mixed-model ACs that align more with interactionist theories of personality as opposed to staunch trait or situationist perspectives (Melchers et al., 2012).

3.4 The Interactionist Perspective

Although the evidence presented from the AC literature supports the situationist perspective being responsible for the exercise effect, no study has yet to model the facets of measurement associated to understand the relative contributions of dimension-related and situation-related variance. While Jackson et al. (2016) disagree with Putka and Hoffman's (2013) claim that their findings align with the interactionist perspective, GBAs may align more with this perspective. As part of this thesis, I will be exploring the interactionist perspective alongside the trait and situationist perspectives to understand how this impacts the reliability of GBAs.

Although trait theory is one of the more popular theoretical perspectives related to behaviour, key contributors of trait theory such as Alport and Cattell emphasised the importance of the interaction between the traits and situation when measuring behaviour (Griffo & Randall Colvin, 2009). However, the interactionist perspective can be traced back as far as 1926 when Kantor stated 'no biological fact may be considered as anything but the mutual interaction of the organism and the environment' (1924, p.369). The interactionist perspective does not favour the person or situation in isolation but assumes that the main source of behavioural variance is related to an interaction between them (Endler, 1975).

One of the early criticisms of the interactionist perspective was that using traditional analysis such as correlations and factor analysis does not allow the researcher to isolate situational characteristics that could contradict the results, therefore making the interactionist hypothesis harder to test (Ekehammar, 1974). Evidence of this can be found in early research that noted that correlation coefficients decreased as situations changed but never reached zero, indicating some underlying variable was still accountable for behaviour, but that this could not confirm the effect of the trait and situation interaction (May et al., 1928). However, more recent studies have focused on the interaction between the trait and situation. Researchers have investigated the extent to which situational characteristics can improve the predictive validity characteristics of personality assessments. Judge and Zapata (2015) coded previous studies based on the related job situational elements and found that this accounted for incremental validity in predicting job performance. These findings have also been observed in studies that have investigated contextual rather than overall performance (Hurtz & Donovan, 2000), indicating that the interaction between the person × traits × situation could be responsible for a large portion of the variance in behaviour.

Using variance components analysis allows researchers to deconstruct the sources of variance associated with a form of measurement. This can be used to estimate the reliability of a measure, but also to gain insight the psychometric structure and contributions of the different facets of measurement therefore allowing the researcher to test the contribution of trait, situation, and interactionist perspectives (Ekehammar, 1974). Early studies using this approach found similar findings to those in the AC literature, in that trait-related variance accounts for a small portion of overall variance, while the biggest contributor to overall variance was an interaction between the person × situation (Bishop & Witt, 1970; Endler & Hunt, 1968). With so little research surrounding the area of GBAs and little understanding of GBAs psychometric properties, using variance components analysis to investigate the

reliability of GBA will allow me to further the understanding of the consistency of behaviour across situations and identify the extent to which situations affect performance. As trait theory, situationism and interactionism are three distinct theories that allow for different operationalisations of what effects are deemed reliable or unreliable or even irrelevant to reliability, variance components analysis will also allow me to section off the variance associated with the different perspectives and estimate the reliability of GBA according to each perspective to see how different theoretical perspective can impact reliability.

Trait-activation theory is an interactionist perspective that outlines behavioural responses based on trait-relevant cues found in situations that are therefore dependent on the interaction between the trait and situation (Tett & Guterman, 2000). These trait-relevant cues can differ in strength where stronger situations tend to elicit specific responses, while lower strength situations are more ambiguous (Lievens et al., 2006). When applied to ACs, exercises are no longer viewed as equivalent situations to elicit consistent trait behaviour, but a set of different trait situational cues where different types of behaviour can be elicited based on situational strength (Lievens & Christiansen, 2010). For example, group exercises are more likely to elicit communication or teamwork-related behaviour than written exercises. Previous research has investigated this theory in relation to ACs by comparing convergence between high and low trait-activation exercises, finding higher convergence between the former and providing evidence to support the idea that exercises with similar levels of trait-activation will show higher consistency in performance (Haaland & Christiansen, 2002).

In Putka and Hoffman's study (2013), they treated candidate, candidate × dimension, candidate × exercise, and candidate × dimension × exercise effects as reliable sources of variance. This would suggest that they are taking an interactionist perspective to measurement, as any dimension related variance and exercise related variance associated with the candidate is treated as reliable. Across all three samples used within the study the G

score ranged between .76 – .77. However, when treating exercise related effects as irrelevant variance and adopting a purely trait perspective, the G score drops to .40 – .45. This has a severe impact on the reliability of the measure.

In this thesis, to ascertain the extent to which source of variance contributes the most to reliability in a GBA, it will be necessary to use variance components analysis to look at the data in a similar way to Putka and Hoffman's (2013) study, but to build on our understanding of measurement, I will also be estimating and comparing the reliability coefficients of the measure across theoretical perspectives. This will allow me to see how reliability is impacted depending on the perspective taken, and identify how this can further our understanding of what aspects of measurement contribute the most to reliability, be that the trait, the situation, or an interaction between both.

As noted previously, Jackson et al. (2016) conducted a similar study to Putka and Hoffman (2013), but considered sample and item-related effects. This allowed for a less confounded estimate of the variance components related to ACs. Also, due to the way the authors modelled the effects in both studies, one could argue that they also took an interactionist perspective when deciding what was deemed to be reliable, unreliable, and irrelevant variance. This meant that any exercise or situation effect related to the person was also treated as reliable variance. Their original reliability estimate was .8 when taking the interactionist perspective when partitioning the variance, however, in a similar way to Putka and Hoffman's results, this drops to .41 when treating exercise related variance as irrelevant variance, again showing the impact of theoretical justification of variance components.

3.5 Theoretical Implications of Aggregation

In both of the aforementioned AC papers by Putka and Hoffman (2013) and Jackson et al. (2016), the authors also aggregated their scores to different levels. This can be defined as summing dimension scores across exercises to arrive at an overall dimension score where correlated variance accrues while error is reduced (Kuncel & Sackett, 2014). This is common within the AC literature, in that scores can be aggregated into an overall dimension scores (Putka & Hoffman, 2013; Jackson et al., 2016), or can be aggregated across exercises, where an overall exercise score is calculated based on scores from within each exercise (Jackson & Englert, 2011). Scores can even be aggregated to arrive at an overall AC score, in which scores are aggregated across exercises and dimensions. This has also been applied to other multifaceted measures, such as SJTs (Jackson et al., 2017). However, the level of aggregation has an impact on how we can interpret the data and the psychometric properties of the measurement.

In GBAs, there are several levels of aggregation that can also be considered. Traditionally, as a proposed measure of personality dimensions, aggregating to a dimension-based score would be logical. The levels are independent tasks that are designed to measure aspects of specific dimensions and are therefore aggregated into overall dimension scores. Aggregating to an overall score in the GBA could be beneficial as it would be possible to ranked participants based on overall performance. When considering the different levels of aggregation, it is important to understand how this affects our ability to understand what is being measured. When aggregating to level-based, exercise-based or overall scores, one cannot rely on the dimension as the focal point of measurement and giving candidates feedback based on their dimension-based scores becomes irrelevant. In these instances, feedback should be given at the exercise or overall level. In ACs, this seems more logical as the tasks tend to be high fidelity or job-relevant, whereas, in GBA, tasks are more abstract

and gamified, meaning task-based or overall score related feedback may not be job-relevant for candidates.

Jackson et al. (20170) found that when aggregating an SJT to items, situations or dimensions, the G coefficient did not surpass .49 across three samples, but when aggregating to the overall score level, it ranged between .51 and .75. When aggregating scores, error is reduced and reliable variance percentage increases, resulting in higher G coefficients. However, any feedback on the scores based on the items, situations or dimensions is not appropriate as the structure of the data does not support these factors contributing to reliable variance. Both Putka and Hoffman (2013) and Jackson et al. (2016) found G scores were higher when aggregating to exercises and an overall score, indicating that aggregation level can affect the reliability of a measure.

The impact of aggregating GBA scores to different levels will be explored as part of this thesis. I will present the findings of the reliability of the aggregated scores according to the interactionist, situationist, and interactionist perspectives to add to our current understanding of how aggregation impacts reliability for the different theoretical perspectives.

In this chapter, I have discussed the different theoretical approaches related to measurement and how this can affect our understanding of what we deem reliable and unreliable sources of variance. Currently, there is a debate surrounding which contributes more to behaviour: the trait or the situation. There is evidence to suggest that both contribute to behaviour, but the effect of that contribution seems to be minimal for traits and the situation seems to be much more important (Jackson et al., 2016). However, the interactionist perspective notes that neither is more significant and that the interaction between both that is important to how someone will behave (Griffo & Colvin, 2008). To

date, there is little understanding of the measurement structure of a GBA and therefore I have discussed the findings related to other multifaceted measures and how this may relate to GBAs.

In this thesis, I will look at the contribution of the dimensions measured in the GBA (akin to traits), the contribution of the levels (akin to situations) and the interactions between these elements to understand the psychometric properties of the GBA to see which theory is most fitting to the data. I will also partial the reliable and unreliable variance according to the different perspectives to see how reliability coefficients differ depending on your theoretical perspective. I will also present the G coefficients associated with the different levels of aggregation possible to understand how aggregation impacts reliability, and how this information can be used to optimize reliability and the use of GBA for selection decisions. Due to the lack of literature in this area, this thesis will help further our understanding of measurement properties of a GBA, how reliable it is and how best to operationalise what is being measured in a GBA.

Chapter 4: Research Aims

In the previous chapters, I outlined what GBAs are, and how using GBAs as a psychometric tool requires that fundamental psychometric properties such as reliability need to be understood for this measurement so results can be meaningfully interpreted. Even with the lack of research into GBAs, I highlighted how reliability has been applied to GBAs, and how due to their multifaceted nature, traditional forms of reliability are not appropriate for GBAs and that other, more nuanced measures of reliability, such as G theory may be more appropriate to use to estimate their reliability. I also explained different theories of behaviour, and how they relate to measurement, and highlighted how these can affect reliability.

In the following chapter, I will summarise the research presented in the past three chapters and highlight the main research questions I will address in my thesis. I will finish the chapter by outlining how the findings in this thesis will add to the literature and increase the knowledge base surrounding GBAs.

4.1 The Reliability of a GBA

As outlined in the previous chapters, reliability is a fundamental aspect of measurement and without reliability, one cannot meaningfully interpret scores on an assessment (Ritter, 2010). This has both academic and practical implications. From an academic perspective, without reliability, it is not possible to interpret the results or findings from research based on unreliable tools. Any findings garnered from these tools are likely to be highly confounded. Therefore, without understanding the reliability of GBAs, it is not possible to understand how they may be used as a meaningful selection tool in high-stakes

recruitment. A number of authors have researched the validity of such selection tools without taking into consideration the reliability of these assessments (Georgiou et al., 2019; Landers et al., 2017; Lopez & Tucker, 2017; Ninaus et al., 2017). As commonly quoted within the psychological literature, one cannot have validity without reliability, the lack of reporting of the reliability of GBAs is a massive gap in the literature, considering the little research that is available in this area (Armstrong et al., 2016; Chamorro-Premuzic et al., 2016).

From a practitioner perspective, using an assessment without evidence of its reliability is not best practice (ITC, 2001) as it can result in good participants being sifted out based on data that is not interpretable or is inaccurate. This has issues both from a legal and ethical standpoint. Without reliability, one cannot be sure that a person is being sifted based on job-relevant criteria or just random chance. Giving feedback to a candidate based on unreliable scores can leave the candidate feeling negative about the process which can harm how candidates view the organisation (Hausknecht et al., 2004). Furthermore, hiring candidates without the relevant dispositions could mean that organisations spend more time and money training candidates, have to hire frequently based on termination or retention issues, or could result in a reduction in efficiency and organisational performance (Crook et al., 2011; Gatewood et al., 2008; Guest, 1997).

Reliability is a fundamental concern for any form of measurement but has wide-reaching implications when little is known about the reliability of a selection tool. Due to the relative novelty of GBAs in comparison to traditional forms of assessment, there is little research in this area. As part of this thesis, I aim to investigate the reliability of a GBA to understand the extent to which scores derived from this method of assessment can be meaningfully interpreted and used to sift candidates in a high-stakes recruitment process.

Several authors have investigated the reliability of GBA, albeit not those used for selection; a review of the literature is presented in a previous chapter. Most have not considered the multifaceted nature of this method of assessment and therefore the results are flawed. For example, some only estimate the internal consistency of a measure and not how other facets of measurement may affect their result (DiCerbo, 2014; Kim & Shute, 2015; Venture & Shute, 2013; Stinson et al., 2015).

Internal consistency also has limitations. From a classical test theory perspective, measurement consists of a ratio between true score and error, (Brennan, 2010; Cronbach & Shavelson, 2004). However, with only a single true score and error terms available in the model, it is not possible to untangle the different sources of variance that can contribute to them. This becomes more of a problem with multifaceted measures such as GBAs, where facets relating to measurement can be classified either as a true score or error (Suen & Lei, 2014). Further issues related to internal consistency include misquoted and arbitrary cut-off values (Nunnally and Bernstein, 1994); the effect of multidimensionality (Schmidt, 1996); increased length of test inflating reliability coefficient (Panayides, 2013); and item redundancy of alpha value being too high (Loevinger, 1954).

Other authors estimate the TRT reliability, but this does not consider any effect related to the items or dimensions, but only the effects of transient error related to variance associated between scoring across occasions. It is thus likely to be confounded. McCrae et al. (2011) noted that, when considering only internal consistency or TRT reliability, different sources of error can affect the results in different ways depending on what is classified as error variance. For example, when considering internal consistency, item irrelevance can have a significant effect on reliability coefficients due to the lack of shared variance between the item and the other items in the measure. In contrast, when only considering TRT reliability, as long as the item is scored consistently across occasions, the reliability

coefficient will not be reduced. The relationship between TRT and internal consistency is small at best (*r*=.25), which further suggests that they are predominantly measuring two different things (Chmielewski & Watson, 2009). If a single scale can have multiple estimates of reliability, it is difficult for the researcher to know which form of reliability to use when, for example, correcting for things like attenuation or range restriction, and can result in an overcorrection (Vispoel et al., 2017) The researcher will need to pick which form of reliability to use and when they can have such different results, this can have real implications for their use (Shavelson et al., 1989).

GBAs are multifaceted measures that consist of items, dimensions, levels/mini-games and it is necessary to consider how these different sources of measurement affect reliability. GBAs are similar in design to ACs and when estimating the reliability of ACs, previous authors have used G-theory to deconstruct the different sources of variance related to this measure (Jackson et al., 2016; Putka & Hoffman, 2013). However, the sparsity of research in the area of GBAs has resulted in more of a focus on construct validity. Although the findings are promising, without a reliable measure, results cannot be meaningfully interpreted and (Ritter, 2010), therefore additional insight in needed to understand the reliability of GBA.

Some authors have estimated the reliability of GBAs using traditional approaches such as TRT or internal consistency. As noted in the previous section, GBAs are multifaceted measures, with different components of the assessment contributing to both reliable, reliability unrelated, and irrelevant variance. Using traditional estimates of reliability result in confounded estimates that do not take into consideration the different sources of error that can impact reliability. For example, DiCerbo (2014) conducted a study investigating persistence in a GBA. Participants completed a game comprised of 3 levels in which variables were captured across each. The authors estimated reliability using Cronbach's alpha, without considering the effect that the level would have on the reliability

of the measure. If G theory was used to estimate the reliability of the measure, the researchers would have been able to deconstruct the sources of variance related to each level, item and interaction. Without taking this approach, the reliability estimation is confounded by level-related variance. Other studies have estimated reliability in a GBA using internal consistency and are therefore likely to have confounded reliability estimates (Quellmalz et al., 2010; Seufert et al., 2016; Kim & Schute, 2015). This issue is also present in several studies that estimate the reliability of GBA using TRT reliability (Mavletova, 2015; Tong et al., 2016) which means that reliability estimates are likely to be confounded as the potential effects related to the measure have not been considered.

Test developers have also shared the average internal consistency of their GBA constructs (.72; Arctic Shores, 2019). However, these estimates of reliability are also confounded as internal consistency does not consider all the facets of measurement associated with a GBA. Another test developer reports the internal consistency and TRT reliability of the original non-gamified version of the tasks used within their assessment (Pymetrics, 2015). However, with the addition of game-elements and changes to how the assessment is scored, these estimates of reliability are likely to be inaccurate.

G theory builds on CTT and considers how different facets of measurement affect the reliability of a multifaceted measure and model these sources (Arterberry et al., 2014; Vispoel et al., 2017). This form of reliability estimate is an improvement over the traditional estimate of reliability as one can model additional sources of variance not considered in traditional approaches (Cronbach & Shavelson, 2004). Reliability is a fundamental concern of any type of measurement and should be estimated in any study that uses a form of measurement, especially under-researched assessments like GBA where more evidence is required to understand the psychometric properties. To my knowledge, no previous study has applied this type of analysis to GBAs, and this is one of the first to estimate the reliability of

a GBA used for selection, in a way that incorporates the sources of variance that could affect GBAs.

RQ 1: Are GBAs a reliable measures of personality dimensions?

4.2 The impact of Aggregation in a GBA

It is possible to aggregate scores to different levels. Aggregation is used to sum scores across different facets of measurement, be they dimensions, exercises, levels or situations, to arrive at an overall score. When doing this, correlated variance increases and error variance decreases (Kuncel & Sackett, 2014). For example, in ACs, it is possible to aggregate scores to dimensions, exercises or overall scores (Jackson et al., 2016; Putka & Hoffman, 2013). In SJTs, it is possible to aggregate scores to dimensions, exercises and overall scores (Jackson et al., 2017). However, aggregating scores to different levels can affect the interpretability of scores, in that, if scores are aggregated to dimensions, results on dimensions can be interpreted but not task-based or exercise-based scores. The reverse is true if scores are aggregated to exercises, then exercise-related scores are interpreted, but judgements about dimensions become harder to interpret. If overall scores are interpreted, then the effect of the exercise and dimensions become meaningless as the scores are determined based only on overall performance.

In the AC literature, aggregation has little effect on the reliability coefficient as that this has already been found to be quite high (Jackson et al., 2016; Putka & Hoffman, 2013). However, in the SJT literature, aggregating to dimensions results in a low reliability coefficient (Jackson et al., 2017) which improves slightly when aggregated to situations, but approaches acceptable reliability levels when aggregated to the overall level (.51 to .75). In GBAs, it is possible to aggregate scores to dimensions, levels and an overall score. There are a number of different cognitive tasks in GBAs that have been developed separately to

identify individual differences. Combining scores from different tasks to form dimension scores in this GBA is a novel concept and no other study has investigated how they can be combined to form dimension-based scores. This study will be the first of its kind and will identify the effect on aggregation has on the reliability of the GBA.

RQ 2: Does aggregating to dimensions and overall scores increase the reliability estimate within GBAs?

4.3 The Psychometric Structure of a GBA

When decomposing the multiple sources of variance related to a multifaceted measure, the researcher becomes able to identify what facets of measurement contribute to reliable variance. In the case of GBAs, I will be able to identify the extent to which the person-main effect, dimension, level, and interactions between these facets. With this information, it is possible to see what is contributing to reliable variance, and to understand the internal structure of a measure. No study has yet to apply G theory to further understand the internal structure of a GBA, this study will be the first of its kind.

Previous studies which investigated the internal structure of multifaceted measures, such as ACs and SJTs. Although both types of assessment tend to focus on measuring specific dimensions, when taking into consideration how much variance contributes to reliable variance, this tends to only be a small proportion. This means that other facets of measurement contribute to what is being measured in the tool. In ACs, this tends to be an interaction between the person × dimension × exercise interaction. While overall, the person × dimension effect accounts for a small amount of variance (Jackson et al., 2016; Putka & Hoffman, 2013). The variance with the biggest contribution to reliable variance in SJTs is

the person main effect, with dimensions contributing little to overall reliable variance (Jackson et al., 2017). The lack of dimension related variance in both measures indicate that other effects are related to how people respond to these measures. As noted, no other study has yet to optimise G theory to further understand the internal structure of GBAs, and therefore, there is currently no understanding of what contributes to reliability in GBAs. As mentioned, little is known about the internal structure of GBAs, so more research is needed to fully understand if dimensions can be meaningfully interpreted, or if, like other multifaceted measures, do other facets of measurement contribute more variance to reliability, and what does that mean when interpreting GBA scores.

RQ 3: Does the person main effect, dimensions or levels contribute the most to reliable sources of variance in GBAs?

4.4 The impact of the situation in a GBA

The theoretical understanding of behavioural measurement has been hotly debated within the literature (see Epstein & O'Brien, 1985). Trait theorists ascertain that traits are stable, and that measuring them across situations will result in some form of stability (Hogan, DeSoto, & Solano, 1977). Based on this reasoning it is therefore possible that measuring a person's inherent dispositions in an assessment, will allow hiring managers to understand how they will behave on the job. For example, those scoring higher in a measure of extraversion are likely to be outgoing and social within the role. It is also assumed within trait theory, that people will differ in how they respond to a situation (Ross & Nisbett, 1991). For example, when put into an environment that involves high stress situations, those lower in neuroticism will be able to function efficiently under pressure, while those higher in

neuroticism will struggle, and this can be used to infer how a person will respond in situations relevant to the workplace.

Consistency of behaviour has been found within the literature in regard to TRT reliability, and through longitudinal studies across the lifespan (Costa & McCrae, 1994; Roberts and DelVecchio, 2000). This means that people tend to respond to personality assessments consistently, without much deviation in inherent dispositions, although personality has been found to fluctuate until later adulthood. However, when measuring the consistency of personality across different behaviours situations, like within ACs, the traits have been found to be less consistent. In alignment with trait theory, one would assume that dimension related behaviour in one task would be consistent with dimension related behaviour in another task. However, this is rarely observed within the AC literature, with correlations between different dimensions within the same task resulting in stronger correlations than the same dimensions across different tasks (Sackett and Dreher, 1982). This is known as the exercise effect and provides evidence to suggest that behaviour may not be stable, but differ across situations (Dilchert & Ones, 2009; Haland & Christiansen, 2002; Lievens et al., 2006; Sackett & Dreher, 1982).

Concerning other forms of multifaceted assessment, this effect has not been replicated as consistently. For example, when looking at SJTs, the impact of the situations does contribute to the variance within an assessment, but nowhere near as much as observed within ACs with the majority of reliable variance being associated with the person-main effect (Jackson et al., 2017). This is also observed within a study by Jackson and colleagues (2016), in which participants completed a first-person shooter style videogame. Participants were separated into two different conditions, in the first condition, participants were exposed to the same situation three times, while the second condition were exposed to three different levels. The authors found that the largest source of variance in both conditions was the

person × situation interaction, which was significantly higher for the condition in which participants completed the same level, and less variance was accounted for by this effect in condition two. This suggests participants respond more consistently when presented with new environments, but less consistently when presented with the same situation multiple times. As the results of the impact of the situation is inconsistent across different methods of assessment, this study will add to the literature by presenting the impact the situation has on variance within a GBA.

RQ 4: Does the situation contribute a considerable amount of variance to scores in GBAs?

4.5 The Theoretical Understanding of Behaviour Within a GBA

Considering the effect that the exercise effect has on our ability to interpret scores in an AC, little is known how this effect may be replicated in GBAs or if this method of assessment reduces the likelihood of observing this effect. As GBAs measure dimensions across levels, it is possible to identify to what extent the situation affects the psychometric structure and reliability of GBA. From a situationist theoretical perspective, behaviour does not take place in a situational vacuum and the situation can affect how people respond (Deinzer et al., 1995). This theory is aligned with early behavioural theories that assume that behaviour is related to antecedent causes from the environment (Skinner, 1974). For example, regardless of a person's personality, most people will behave in a similar fashion when attending a funeral. This perspective became popularised thanks to the work of Walter Mischel in the 1960s. Mischel (1968) reviewed the existing literature and found a predictive

ceiling effect of personality on behaviour of .3 and concluded that the majority of variance in behaviour relates to things other than personality.

Important contributions to the effect of the situation on behaviours include the Stamford prison experiment (Zimbardo, 2004) and Milgram's obedience studies (1965) which both found that situational effects accounted for more variance in behaviour than inherent dispositions. These studies were highly controversial and resulted in participants carrying out heinous acts that they would not normally do, providing evidence that situations can affect a person's behaviour and that traits are less stable across situations. Evidence of the effect of the situation on behaviour was observed in a literature review by Funder and Ozer (1983). They concluded that situational characteristics explained a large portion of the variance in behaviour (r=.38).

In the AC literature, evidence of the situation explaining the variance in behaviour is abundant. For example, Jackson et al. (2016) found that dimensions alone accounted for a small amount of variance in behaviour, and the exercises, which can be interpreted as different situations, explain more variance. Furthermore, the authors attribute the person × dimension × exercise effect, in that participants' behaviour is dependent upon the interaction between the exercise and dimension, is also attributable to the exercise effect. This has implications for our understanding of behaviour, as we try and measure behaviour in one instance in an attempt to predict how someone will act in another situation. However, even though ACs can be highly job-relevant and more closely aligned to the role in which they are being used to assess suitability (if designed properly), situational variance is likely to relate to the job. However, lower fidelity measures like GBA, if used to indicate a candidate's suitability for the role, may have fewer situational elements related to the role and therefore would not predict how a participant would behave in that particular situation. Evidence of this has been found in research that shows high-fidelity simulations are much better at

predicting future job performance (Rene et al., 2011). With GBAs, the situational variance explained by the assessment could be a lot less job-relevant and therefore be an issue when trying to predict job performance. Therefore, it is important to understand the extent to which taking a situationist perspective when classifying facets of variance would affect the reliability of the assessment.

Although Jackson et al. (2016) stated that the person × dimension × exercise effect was related to the exercise effect, Putka and Hoffman (2013) attributed it to the interactionist effect. The interactionist perspective to measurement contends that behaviour is made up of an interaction between the person and the situation and does not put more emphasis on one over the other (Endeler, 1975). Further research into the interactionist perspective has found evidence that small amounts of variance are attributable to the person and that the majority of variance is related to an interaction between the person and the situation (Bishop & Witt, 1970; Endler & Hunt, 1968). Situational elements have also been found to add incremental validity over and above that of personality in predicting job performance (Judge & Zapta, 2005) indicating that, both the innate disposition of the individual and the situation they are in, effectively contribute to behaviour. Although Walter Mischel was one of the first proponents of the situationist perspective, he became an advocate for the interactionist perspective stating 'the question of whether individual differences or situations are more important is an empty one that has no general answer' (Mischel, 1977, p. 340).

GBA offers a unique perspective in which candidates can be presented with information across a number of different situations that can be standardised across all candidates. The data gathered throughout a GBA can be used to identify how candidates react to different situations. It can offer researchers a unique and relatively unexplored medium in which to investigate behaviour. Furthermore, through using G theory to identify the psychometric structure of GBA, it is possible to align the effects with different theoretical

perspectives. This will allow me to further expand our current understanding of what extent do different theoretical perspectives impact the reliability of a GBA.

RQ 5: Do the findings from the G study provide evidence to support trait theory, the situationist perspective or the interactionist perspective of personality.

Chapter 5: Methodology

This chapter outlines the methodology of the study. It begins with an overview of the data and the implications associated with using secondary data, including ethical considerations. The participants in each of the three samples are described and the justification for the sample size examined.

The materials used in the study refer to the GBA and the specific levels that were retained for use in this study are described in the context of the original task they were based on, and how these tasks were used to measure individual differences. The constructs used in this study and how they relate to job-relevant behaviour are also explained. Finally, I outline the data analysis procedure, including how the variables were selected, how the data were cleaned and how the statistical analyses were run. It finishes with an overview of the G-study models used in this study and how different levels of aggregation and different perspectives were calculated.

5.1 Secondary Data Analysis and Ethical Considerations

In this study, secondary data from three separate samples were analysed. Secondary data is data that is used to answer a research question that was not the focus of the original data collection (Heaton, 2008; Koziol & Arthur, 2011). The anonymised datasets used were provided by the GBA test publisher and consisted of GBA data collected as part of three separate high-stakes recruitment processes. The use of secondary data is common in research, especially in an applied area such as organisational psychology, but it comes with some important considerations. In the following section, I will outline the secondary data

used and the positive aspects of using this type of data in applied research, while explaining some of the issues associated with using secondary data and how I overcome them.

There are many reasons why researchers use secondary datasets. Samples needed for primary data collection can be difficult for researchers to access and using secondary data can grant a researcher access to data that may not have otherwise been available (Cowton, 1998). Using secondary data can also be more efficient, less labour intensive and less costly than collecting primary data. It can give researchers access to a larger breadth of data, resulting in more variables in the analysis and a larger number of participants (Koziol & Athur, 2011). This can help researchers increase the power of their analysis and replicate their data on multiple samples. It can also afford the opportunity to answer a specific research question that has not yet been addressed using the same data and is thus more likely to reduce the chances of the researcher replicating previous findings (Tripathy, 2013).

In this thesis, secondary data was used to gain access to context-specific data-sets that would have been difficult to collect using primary research, due to the sample sizes, breath of data, and time in which the data were collected. Previous research has also found that participant motivation in low-stakes testing situations can affect how a participant responds to an assessment and can potentially reduce the validity of the findings (Cole et al., 2008; Wise & DeMars, 2005). Using data collected in a high-stakes recruitment process from candidates applying for a real job allows the findings to reflect real situations in which the GBA is used. Access to multiple samples also offers insight into the replicability of the findings across samples and allows for the further investigation of differences that may be related to a specific sample and to understand how generalisable the findings are. Accessing primary data of this detail with the required sample size to run this level of analysis would have been very difficult and would have jeopardised the quality of the results.

There are many implications associated with using secondary data. For example, the data were originally gathered for a specific reason (in this instance, as part of a recruitment process), but in this study are being used to estimate the reliability of the measure, which was outside of the intended purpose of the original data collection. The data were not originally collected to answer additional research questions beyond that of the original purpose – to make informed selection decisions based on the results from the GBA. This can cause problems when attempting to analyse secondary data, especially as the researcher has no control over how the data were collected, which may affect the quality of the data and any further analysis (Thornhill, Saunders & Lewis, 2009).

Regarding the purpose of data collection, the aims of this thesis are relevant regardless of the context. Reliability is a fundamental aspect of measurement. This means that even though the original intention of the data collection was not to estimate how reliable the measure is, the data can still be analysed to explore this question. The context in which the data were collected is a perfect example of how it will be used in an applied context and is therefore not only relevant, but the findings from this thesis are more generalizable to the GBA's intended purpose.

Concerning the ethical considerations associated with secondary data, all participants who completed the GBA agreed to a set of terms and conditions before completing the assessment and were aware of the original purpose for which they were completing the assessment. Although there has been debate in the literature about whether participant consent can be fully informed if the participants are unaware of what they are consenting to in the future (Morrow et al., 2014), this is a common practice of gaining informed consent which still complies with ethical standards (van den Eynden et al., 2009). Any data shared by the test developer complies with EU GDPR in that no personal or identifiable information is shared with third parties (Tikkinen-Piri et al., 2018). All data were anonymised before being

shared by the test publisher ensuring that individuals could not be identified. Thus, it was not possible to follow up with participants by questionnaire as there was no way of tracking them. Participants were able to remove their data from the test developer's database at any time. If they had requested this before the date on which the developers shared their data, they were not included in this study. If this was requested after this data had been received, due to the anonymisation of the data it would not have been possible to identify and remove their data from the dataset. A nondisclosure and partnership agreement were signed by the researcher to ensure compliance with the test developer's data sharing policy (see Appendix VI). The findings of this analysis are based on archival datasets and any decision made to select an employee had already been made. Therefore, the outcomes of this study will not affect the decisions for which the data were originally collected. Furthermore, this study was granted ethical approval by the Department of Organizational Psychology Ethics Chair.

5.2 Power and Sample Size

In the literature, there are differing views on adequate sample sizes in a reliability study. Inadequate sample sizes can result in unstable G-coefficients (Shavelson et al., 1989). Kline (Kline, 1986) suggests that adequate sampling must be over 200 participants, whereas Nunnally and Bernstein (1994) suggest a minimum of 300 to minimise any error. Other authors have gone on to suggest sample sizes over 400 are needed to adequately estimate the population's reliability (Segall, 1994).

For this G study, a sample of 50 – 300 participants would allow for a robust estimate of reliability (Atilgan, 2013). Therefore, as the samples used in this study ranged from 398 to 702, they could provide a robust and unbiased reliability estimate.

5.3 Participants

The participants were taken from three independent samples who completed the GBA as part of the application process for three unrelated roles. The original purpose of the data collection was to identify how well each candidate scored on job-relevant dimensions measured within the GBA. The data were originally used to sift out participants who were less suited for the role and progress higher-scoring participants through to the next stage of the recruitment process. The roles and job-relevant traits used to sift the candidates were different for each role. Furthermore, the samples may have completed the assessment after initial sifting by the employer, meaning that there may be more of a range restriction in some of the samples. Therefore, each sample was treated independently, and results compared across samples to remove bias from sample-specific effects. The demographic variables from each sample can be found in Table 5.1

In Sample 1, the majority of the participants were White (34.9%) or Asian (26.9%), and the majority of the sample were male (81.9%). The role was heavily science-dependent and in the area of construction, both of which are areas dominated by men (Wang, Eccles & Kenny, 2013; Dainty et al., 2004). Therefore, although this sample is heavily skewed towards men, it may be representative of the population from which it is sampled (Glass & Minnotte, 2010). In Samples 2 and 3, the majority of participants were female (65.4% and 64.3%, respectively), and white (57.9%) in Sample 2 and Asian in Sample 3 (41%). Concerning age, the mean age in Samples 1 and Sample 2 was within that which would be anticipated for graduate roles, and no data were available in Sample 3. However, as the role also targeted the graduate population, it is likely to have been similar to Samples 1 and 2.

Table 5.1 Demographic characteristics of each sample

Demographic		Sample		
		1	2	3
Sex				
	Male	326	243	-
	Female	70	459	-
	Other	1	-	-
	Unknown	1	-	-
	Total	398	702	429
Age				
	Mean	24. 59	21. 23	-
	SD	4. 359	2. 87	-
Ethnicity				
	Asian	107	152	-
	Black/African/Caribbean	57	49	-
	Other	37	51	-
	Unknown	58	47	-
	White	139	411	-
	Total	398	710	429

Sample 1 consisted of applicants applying for a graduate engineering role in a British construction company. Participants were required to register their interest online for the position, and complete an application form. Then, based on meeting the relevant criteria for the role, they were sent an email providing details of the assessment, how long it would take to complete and instructions for downloading and completing it. Participants were required to complete eight levels of the GBA and two additional assessments of numerical and spatial reasoning. They were ranked based on a combination of job-relevant traits from the GBA, deemed to be job-relevant through in-depth job analysis. Once they had completed the GBA, they were asked to complete a short self-report questionnaire in which they were asked to rate their experience of completing the GBA.

Sample 2 was of applicants applying for a UK graduate recruitment scheme to work with a large European energy provider for an engineering position. The programme was

wide-reaching and was used to give graduates experiences in different placements in the organisation across Europe. Participants registered online and if they met the minimum criteria, were sent an invitation email explaining the assessment and what was expected of them. They were given instructions on how to download and complete the assessment by email. The same post-assessment feedback form was given to participants online once they completed the GBA. A feedback report was then sent to the candidate outlining their results in each of the dimensions measured in the GBA.

Sample 3 consisted of applicants applying for an early-career or graduate role in a New Zealand engineering company. Participants registered their interest on the company's online application form and were instructed to download and complete the GBA by email. Similar to Sample 1, participants completed eight levels of the GBA and two aptitude assessments of numerical and spatial reasoning. Following completion of the GBA, participants were directed to the same post-assessment survey as Samples 1 and 2.

Although each sample completed the same GBA, there are some differences between them. Samples 1 and 2 were UK graduates, but the roles and responsibilities for each target job differed as both the industries, and type of engineering roles were distinct. Sample 3 was focussed on a non-UK population and the industry differs from that in Samples 1 and 2. Furthermore, demographic information is not available for Sample 3, so it is not possible to identify how much Sample 3 diverges from the other two samples. Differences may be observed due to the specific selection process implemented for each role. Given the lack of control over the collection of secondary data, instead of amalgamating the samples into a single sample, I opted to treat each sample independently, to allow for comparisons between them. Each sample was thus analysed independently, and the results compared across the three. This approach has been used to compare samples in previous studies using a similar methodology (Jackson et al., 2017; Putka & Hoffman, 2013).

5.4 Materials

The measure used in this study was a GBA called Skyrise City. This GBA is made up of nine levels containing cognitive tasks that have been widely researched in the cognitive neuroscience literature and been found to relate to individual differences (Arctic Shores, 2019). It is set in a futuristic office space and the storyline places the candidate as an employee. As part of their role, they are required to complete some work-related tasks. The GBA assesses 28 cognitive and personality traits across all nine levels and takes 35 to 40 minutes to complete.

The GBA development process is slightly more complex than traditional assessments due to the amount of data collected within the assessment. According to the assessment developers (Arctic Shores, 2018), cognitive tasks were gamified that had shown previous evidence in identifying individual differences. For example, one of the tasks that were gamified for the GBA and reported on within this study (Iowa Gambling Task) had been found to correlate with self-report measures of the trait Sensitivity to Punishment (*r*=.31;Franken et al., 2008). This task and measure will be discussed in more detail later.

The next step involves data collection of both the GBA and traditional self-report metrics, and test the hypothesis to see if the relationship between variables is consistent with the literature. Furthermore, as it is possible to collect more data than traditional tasks, data were then subjected to principal components analysis, and other, similar procedures to understand the dimensionality of the data, and to take a data driven approach to include other variables related to the constructs of interest in the scoring of the dimension. From then, the developers assess the relationship between the self-reported traits and GBA dimensions to gauge the extent to which they show evidence of validity.

To date, there has been no peer-reviewed research of this particular assessment tool, and little understanding on the most appropriate way to design and validate GBAs. However, the test publishers provide information on the validity of the GBA and report average correlations of .33 to .51 with self-reported measures of the same or similar constructs (Arctic Shores, 2018). A recent analysis of an updated scoring method developed for the GBA has shown evidence of internal consistency using stratified alpha (Osburn, 2000) with an average coefficient of 0.72 (Arctic Shores, 2019).

Three traits were analysed across three different levels. Log data was captured when completing the GBA and when the maximum number of levels was administered to a participant, up to 5,000 data points could be captured in a single session (Arctic Shores, 2018). Not all of these data were relevant to the participant interacting with the GBA. The final selection criteria were determined by variables contingent on a response. and were therefore, related to performance (Jackson et al., 2016). These variables are defined as behavioural biomarkers and relate to specific variables that are generated throughout the assessment through interactions within the game (Mandryk & Birk, 2019). Furthermore, some constructs within the GBA were only measured via one level and therefore could not be used to investigate the situational variance associated with measurement, which is one of the key research aims of this study. Therefore, these dimensions were omitted from this study.

Each task in the GBA was labelled as a different level. These were: Balloon Burst (BB), Focus Group (FG) and Power Source (PS). The balloon analogue risk task (BART) was taken from the cognitive psychology literature and used as the foundation for the BB level (Arctic Shores, 2018). The original BART shown in Figure 5.1 is very similar to the BB level in Figure 5.2. The original task required participants to inflate a balloon. They were rewarded with a small cash prize for each successful inflation, but if they inflated the balloon too much it popped, resulting in the participant losing their money. They had the

opportunity to ‘bank’ their money at any point. There were three different types of balloons, and each would burst after a certain number of inflations. The object was to earn as much money as possible.

In the GBA, participants were informed that they were required to inflate the balloons. As the balloons were branded, and the larger the logo, the more money participants earned. The GBA also has animated features, background music and a loud pop echo when the balloon bursts. Although the logic was the same as that of the original BART, the situation had a more gamified appearance. In both, participants could select the inflate button to blow up the balloon or the bank button to save their money and move onto the next balloon.

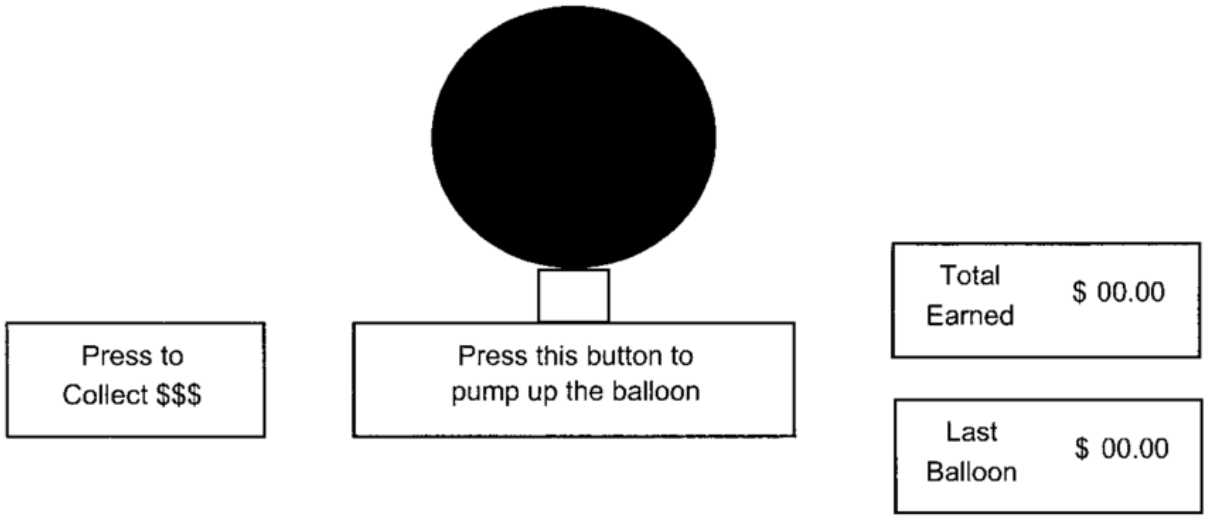


Figure 0.1: Original image of the Balloon Analogue Risk Task (Lejuez et al., 2002).

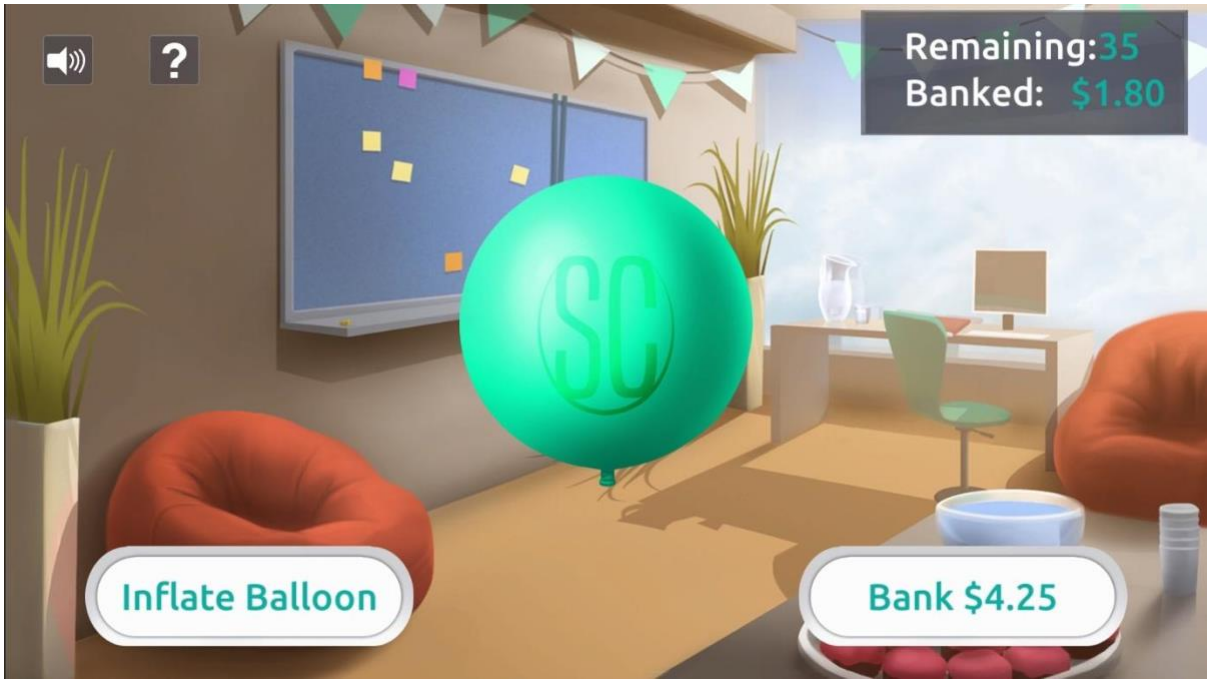


Figure 0.2: Image of the Balloon Burst Level from the Skyrise City GBA (Arctic Shores, 2018). Reprinted with permission.

The BART presents participants with a situation in which they must balance their behaviour based on the rewards incentivised by the cash prize in the game and the threat of loss from bursting their balloon and losing money. It has been found to predict individual differences in sensation seeking, risk-taking and impulsiveness and real-life occurrences of risk-taking behaviour (Lejuez et al., 2002; Vigil-Colet, 2007). Variables collected within this level include money banked within the level or balloons popped.

The FG level is based on a classic emotion recognition task (Montagne et al., 2007) in which participants are required to identify a subject's emotions in video clips that range in intensity across six emotions (Figure 5.3). The GBA differs slightly from the original. Participants are asked to rate the emotions of participants taking part in a marketing focus group, so their facial responses can be categorised accordingly. They are presented with images of subjects that have been computer rendered and the emotions vary in intensity

across six emotions, with an additional option for neutral expressions. Scores on this task have been found to relate to measures of personality, including extraversion and neuroticism (Canli et al., 2004; Yoon & Zinbarg, 2008). Variables collected within this level relate to accuracy in predicting facial emotions, and accuracy in predicting emotion intensity.

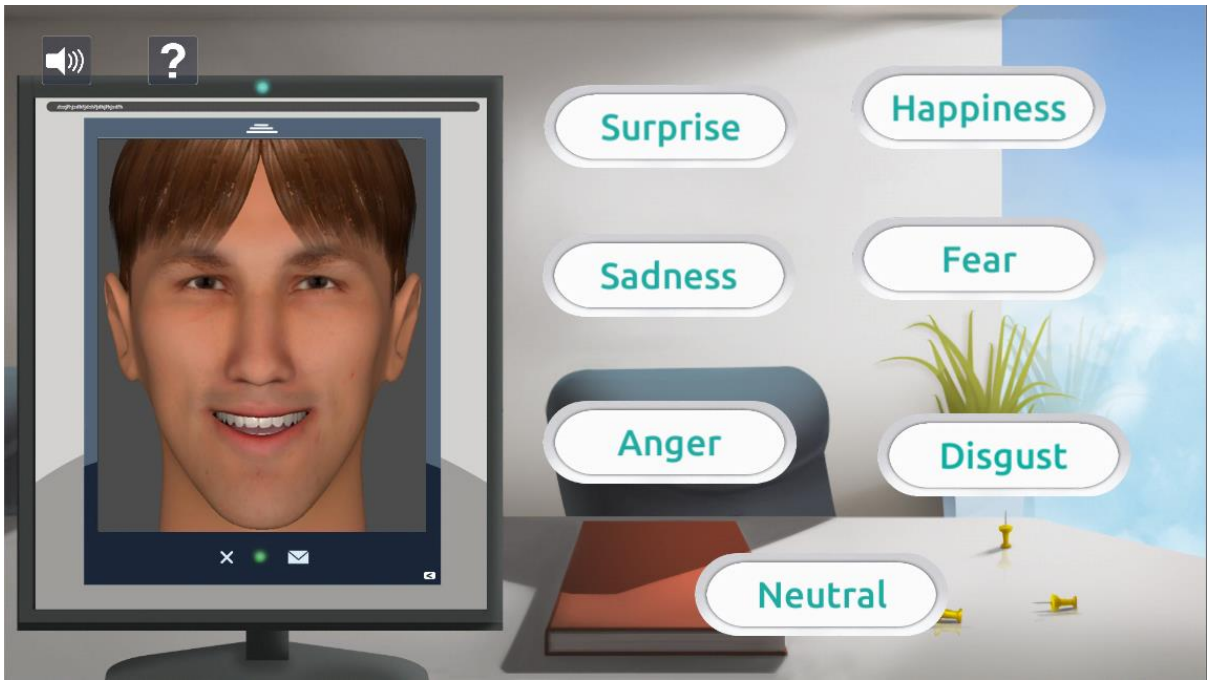


Figure 0.3: Image of the Focus Group level from the Skyrise City GBA (Arctic Shores, 2018). Reprinted with permission.

The last level used in the GBA, PS, is based on the Iowa Gambling Task (IGT). The original task was designed to measure real-life decision-making in a laboratory but was criticised for its lack of reliability and the effect of personality and mood on performance (Buelow & Suhr, 2009). In the original task, participants were given a fake sum of $200 and a choice between four decks of cards. The participant was required to pick a card from one of the four and each card gained or lost them money. The aim was to increase the amount of money as much as possible. The first two decks yielded a larger average profit ($100) after each selection but incurred a steeper penalty for selecting from these decks consistently

throughout the task. The other two had smaller average cash rewards ($50) but lower losses overall. The original task was carried out using real decks of cards but was subsequently computerised (see Figure 5.5). Variables collected within this level include deck chosen, and penalties gained.

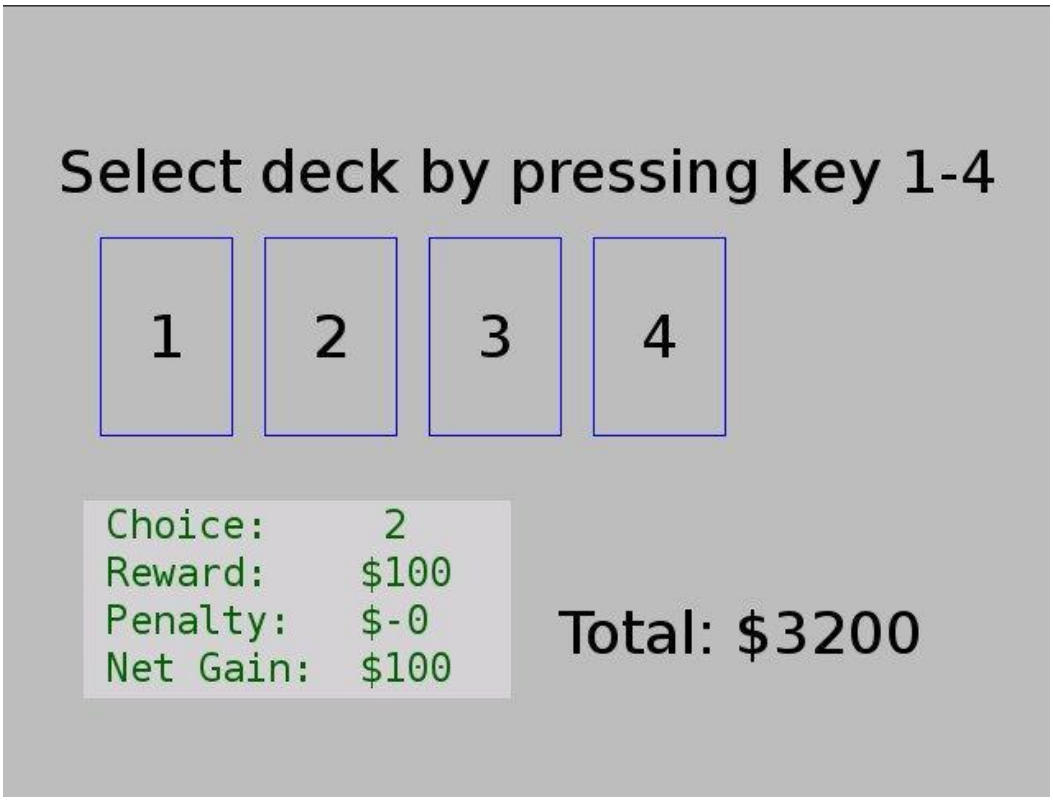


Figure 5.4 Image of Iowa Gambling Task (PEBL, 2018).

In the GBA, the task is similar, except participants are informed that the building in which they work has experienced a power cut. They are told that they must restore power to the building by pulling the levers on four generators (see Figure 5.4). Each generator produces a different amount of energy, as expressed in the power bar along the top of the screen. As they pull a lever, the power bar will either increase or decrease just as the amount of money increased or decreased in the original task. The aim is to restore as much power as possible.

Figure 0.4: Image of the Power Source level from the Skyrise City GBA (Arctic Shores, 2018). Reprinted with permission.

Different behavioural reactions in this cognitive paradigm have been found to correlate to individual differences. For example, making riskier decisions by selecting the higher reward has been linked to traits such as sensitivity to loss (Franken & Muris, 2005). Correlations have also been found between specific variables measured in this task and measures of drive, impulsiveness, sensations seeking and affective state (Buelow & Suhr, 2013; Franken & Muris, 2005).

The cognitive tasks in the GBA were designed to replicate the original task in a more game-like format. However, there is little evidence to suggest how these tasks relate to work-relevant behaviour. Although the studies referenced above assert that behaviour in these tasks relates to variance in individual differences, there is little known about how that manifests to behaviour in the workplace, especially since these tasks are generally

administered in a controlled environment. Although there is some evidence to suggest that this GBA can predict performance in the workplace, it is unclear how these behaviours can be used to predict behaviour reliably and consistently. The original tasks were never treated as specific trait measures, relying more on test-retest reliability when the tasks were reviewed in isolation to gauge consistency in behaviour (Buelow & Barnhart, 2018; Matsumoto et al., 2000; White et al., 2008)). Additionally, with the advances in the way data are collected in the GBA in comparison to the original cognitive tasks, it is now possible to capture a lot more data than was previously possible, which means that there are variables used in the GBA that have yet to be researched that may affect the scoring of the constructs in the GBA. With this form of dimension-based scoring, the test developers used a combination of variables from across the levels to arrive at a dimension-based score. With GBA, linear games that follow a standardised format are optimized to use more traditional estimates of reliability (Barends et al., 2021). However, due to the multifaceted nature of the assessment, G theory is more appropriate for this assessment. In this instance, variables were then treated like items in a measure to arrive at a composite score. It is thus important to understand the reliability of the instrument and its psychometric properties.

5.4.1 Constructs measured in the GBA

Three constructs were retained for analysis in this study: *novelty-seeking*, *creativity* and *sensitivity to punishment*. This section outlines these constructs, how they relate to behaviour and how they may be assessed in the GBA.

As mentioned in the first chapter, there is much controversy surrounding the definition and measurement of creativity. However, to put in simple terms (although the term is thought to be more complex and multidimensional) it is linked to the generation of original

and novel ideas (Runco & Jaeger, 2012). Those who score higher in this trait are likely to generate more ideas and think more conceptually, whereas someone low in this dimension is more likely to think in logical and definitive terms. In terms of its relationship to personality, creativity has been found to strongly correlate with traits such as extraversion (Furnham et al., 2008; Kiran Singh & Kaushik, 2015) or extraversion and openness to experience (Sung & Choi, 2009). Many roles rely on creative individuals to generate novel ideas, come up with creative ways to solve problems or innovate and disrupt specific industries. Accurately measuring creativity could help identify applicants who can use their creative abilities in a role and those who may struggle. However, measuring creativity is difficult as it is a broad and multidimensional construct that can be situationally specific and dependent on the knowledge and experience of the individual (Cropley et al., 2013). In the workplace, creativity has been found to correlate strongly with organisational citizenship behaviours, task performance and negatively with counterproductive work behaviours (Harari et al., 2016).

Creativity as measured within the GBA has shown strong convergent validity with self-reported measures of creativity (r=.50) indicating a large portion of variance is shared between the traditional self-report measure and the GBA (Arctic Shores, 2018). Previous evidence has found a link between emotional ambivalence (the inability to detect emotion, which can be linked to the FG task) and creative thinking (Fong, 2006). Risk-taking has also been linked to creativity as measured in the IGT and BART (Agarwal & Kumari, 1982) showing further evidence of the relationship between creativity in the tasks used within the GBA.

Novelty-seeking can be defined as the tendency to wish to explore new situations and experiences (Gocłowska et al., 2018). Those higher in novelty-seeking are likely to desire new and exciting situations, compared to those low in this trait who tend to prefer regularity

in situations, consistency and few surprises (Arctic Shores, 2018). Novelty-seeking is similar to the openness to experience a facet of the 'big 5', which has been found to rarely show strong evidence of criterion-related validity (Barrick & Mount, 1991; Griffin & Hesketh, 2004; Salgado, 1997). Evidence of a link has been found between novelty-seeking and both extraversion and openness to experience and those who are more likely to show creative thinking (Gocłowska et al., 2018), however, novelty-seeking was also found to hinder job performance in certain roles (Reio & Sanders-Reio, 2006).

Novelty-seeking can be relevant to specific roles. For example, roles that are subject to frequent change or require adapting to new tasks or environments may be better suited to those high in novelty-seeking, while those that require more repetition and regularity may not. Novelty-seeking and impulsivity have also been linked to tasks measured in the GBA, specifically the BART and IGT (Buelow & Suhr, 2009; Lauriola et al., 2013; Suhr & Tsanadis, 2007). This indicates that variance in the tasks used in the GBA is shared with traditional measures of novelty-seeking. The GBA measure of novelty-seeking has also been found to have moderate construct validity when compared to self-reported measures of the same scale (r=.49; Arctic Shores, 2018).

Sensitivity to reward can be defined as an overall sensitivity, or negative reaction to adverse situations or stimuli (Sava & Sperneac, 2006). Those high in sensitivity to punishment are more likely to experience nervousness and trepidation in both the anticipation and presence of negative feedback or punishment, whereas someone low in this trait is less likely to let fear of negative repercussions affect how they approach a task or make decisions. (Arctic Shores, 2018). The trait can be linked to neuroticism and has been found to negatively relate to extraversion (Torrubia et al., 2001). In the workplace, sensitivity to punishment can affect performance in, for example, roles where criticism and feedback are high. Those higher in this trait may struggle with the regular feedback received but may

succeed more in a role that requires the avoidance of risky actions such as surgeons or stockbrokers.

Sensitivity to punishment is linked to neuroticism with highly neurotic individuals displaying more frustration in tasks in which reinforcement levels change abruptly (Ball & Zuckerman, 1990), which is similar to the reward contingencies experienced in the IGT in which punishment levels are higher in riskier decks, and the BART in which participants lose their reward for bursting the balloon. Construct validity of the GBA measure of sensitivity to punishment with a self-reported measure of a similar construct shows moderate levels of construct validity (r=.46).

5.5 Procedure

Once each participant registered their intertest for a role online, they were sent a personalised email which differed slightly in terms of the organisation's branding and the information provided. An example of an invitation email can be found in Appendix I. All emails contained a link to download the GBA app and participants were required to enter a username and password given to them in the email. They were then taken through a list of terms and conditions which included information on what their data would be used for. After they had agreed, they were then asked to complete the GBA. This information can be found in Appendix III. Participants could not proceed to the GBA without first accepting the terms and conditions.

At the beginning of each level, participants were given instructions on how to complete the task. They were informed of the objective task and instructed on how to respond. For example, in the BB level (see Figure 5.6), participants were instructed to inflate the balloons by pressing the inflate button and that each inflation would earn them a

monetary reward. After reading the instructions, participants were asked if they wanted to reread the instructions or continue. On some levels, participants also had the opportunity to practice before moving on to the real test. Once the level had been completed, the candidate was taken to a different screen in which they received abstract points for completing the level. The points were essentially meaningless and did not affect any dimension score.

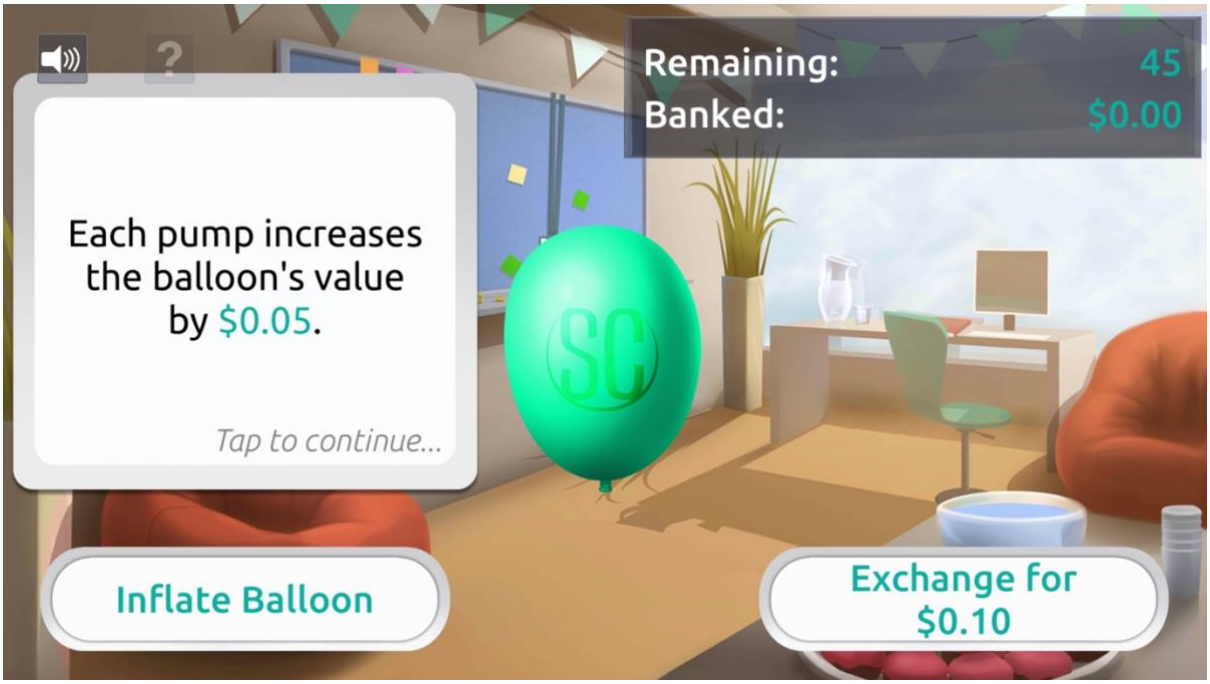


Figure 0.5: Instruction screen image taken from BB level in Skyrise City. Reprinted with permission.

This format was repeated for each of the 8 or 9 GBA levels. After all levels were completed, the candidate was given an overall score based on points they earned from each level and thanked for taking part. They were then transferred to an external survey outside the GBA app hosted on SurveyGizmo. They were asked to fill in a short survey of their experience and to rate items on a Likert scale or asked for a qualitative response via an open

text box. The survey also asked participants for basic demographics such as age, gender, experience of playing video games and educational background. However, this information was not collected for all participants as the questionnaire was optional. Some organisations asked the test developer not to collect demographic data in the GBA or post-assessment survey, hence the absence of age data in Sample 3.

5.6 Data Analysis

IBM SPSS Version 23 was used to clean and run descriptive statistics and initial correlations on the data from each sample. The R package lme4 (Bates et al., 2015)was used to run the variance components analysis using restricted maximum likelihood (REML) procedures. These are favoured over the traditional ANOVA methods, as they have all the added benefits associated with maximum likelihood estimates such as unbiased and small standard errors (Bollen, 1989) and are designed to handle missing data or unbalanced designs, which makes it an ideal for G studies (Jackson et al., 2017). Before running any analysis, the data was cleaned.

5.6.1 Data cleaning and transformation.

As GBA data are more complex than traditional self-reported data due to the many variables, the data needs to be transformed. The raw scores were reversed as required to ensure that the variables were measuring the construct in the intended direction as outlined by the test developer and then winsorized. Winsorizing is the process of converting data that is extreme (either high or low) to a value that is the highest or lowest data point not to be considered extreme. It preserves the data without hindering the analysis by removing the

data case (Reifman & Keyton, 2010). There can be a large variation in scores when using GBA data, especially at the extremes of the range. However, as the data was extracted from live recruitment data, it would be better not to remove outliers it is likely that extreme scores will be seen in real-life situations in which this assessment is used.

When looking at behavioural data such as reaction time and latencies, winsorizing has been found to improve both reliability and validity as a form of transformation, while removing data negatively affecting reliability and validity (Richetin et al., 2015). Studies using similar tasks in the cognitive science literature frequently uses winsorizing as a transformation tool to deal with a small number of outliers (Devlin et al., 2015; Hew et al., 2016). It has been used to reduce the effect of extreme responses, by truncating scores that fell below the 10^{th} or above the 90^{th} percentiles (Pek et al., 2017). This is a procedure that can help reduce standard error and the likelihood of a type I error (Liao et al., 2016).

Additional transformation procedures were used to improve the normality of the data. As the GBA uses items measured on different scales and tends to be varied in comparison to the Likert-scale data often used in organisational psychology, the data was found to still be non-normally distributed. The next step was to apply log, inverse, Box-Cox and power transformations. These did not result in normalised data, but square root transformation resulted in data that was closer to normality. This replaces the original variable with a square root of the score and is frequently used with data that are non-Gaussian (Osborne, 1964).

Multivariate normality is an assumption REML, real data are rarely meet the requirements of multivariate normality (Byrne, 2013). Even though it violates variance components assumptions, REML is robust against this particular violation (Banks et al., 1985; Lumley et al., 2002; Ma & Mazumdar, 2011). A simulation study by Burch (2011)

found that, with larger sample sizes, the effect of non-normality is reduced when using REML, with larger sample sizes resulting in smaller and more accurate confidence intervals.

Although winsorizing was used to help normalise data and eradicate extreme responses, there is still a chance of outliers. This may be a result of participants not paying attention in the assessment or clicking randomly to progress through the assessment, or of not understanding the instructions. These participants are likely to bias the results. Therefore, multivariate outliers were identified and removed from the data as they diverged from the natural way the data was formed (Filzmoser, 2004). Manhobis distance (MD) is a common method used to identify outliers. Observations are compared to the rest of the data distribution and the more the observation deviates from the centre, the larger the MD value. Those with a larger MD value are outliers and removed from the sample (Ben-Gal, 2006). In this study 26, 35 and 20 participants met the threshold for MD and were removed from Samples 1, 2 and 3 respectively.

5.6.2 Descriptive and correlational analysis

To gain an initial indication of the extent to which the same dimensions are related across the different levels and the extent to which different dimensions in the same level show convergence or divergence, correlational analysis was run on each sample. To do this, composite scores were calculated for each dimension and each level and the overall score and overall dimension-based scores in each level. However, the results of this analysis will only offer an initial insight into the data and the results may still be affected either by dimension or level-related variance. Therefore, additional analysis was required.

5.6.3 Psychometric structure and reliability

Participants, dimensions and levels were specified as crossed random facets in this G-study and the items nested in dimensions. This level of nesting is frequently observed in the G-study literature as items are unlikely to be fully crossed with any dimension or exercise because if items are measuring something unique, they are unlikely to overlap across different levels or exercises and any overlap could mean that items measure the same construct and may therefore be redundant. There were 11 effects in total modelled in this design which are outlined in Table 5.2. A G coefficient of >.80 is generally regarded as acceptable (Jackson et al., 2017; Jean et al., 1976; Mushquash & O'Connor, 2006)

5.6.4 Comparison of different theoretical perspectives

This study treats dimensions as analogues of traits. In trait theory, any effect unrelated to the person main effect or interaction between person × dimension is not a reliable source of variance. In contrast, any effect related to the level is perceived as reliability-unrelated (person × level, person × level × dimension). The situationist perspective treats these effects as reliable and any dimension related variance as unreliable (person × dimension, person × item: dimension). The interactionist perspective treats both dimensions related variance and level related variance as reliable.

Depending on the theoretical perspective taken, the remaining effects were categorised. For example, in the trait perspective, situational related variance is characterised as unrelated to reliability, therefore the person × level interaction was treated as reliability-unrelated. Situational and interactionist perspectives treat it as related to reliability, as

behaviour is seen to be dependent on the situation. Therefore, this variance component will factor into the reliability score when the G-coefficient is calculated for these different theoretical perspectives, but will not be included in the G-coefficient for the trait-based perspective

Table 5.2 The classification of effects across each perspective

	Perspective		
Effect	Trait	Interactionist	Situational
p	R	R	R
pl	X	R	R
pd	R	R	U
pld	X	R	R
pi:d	X	R	U
pi:dl + residual	U	U	U
l	X	X	X
d	X	X	X
i:d	X	X	X
dl	X	X	X
li:d	X	X	X

Note. p = person, l = level, d = dimension, i = item, R =reliable, U=unreliable, X = reliability-unrelated

5.6.5 Levels of aggregation

Further analysis was conducted to investigate the effect aggregation has on reliability coefficients. Aggregating the sources of variance can result in dramatic changes in the reliability coefficient in a multifaceted measure, but can also affect how we interpret what is being measured, especially when aggregating to an overall level (Kuncel & Sackett, 2014). For example, SJTs have been found to have poor reliability when aggregated across dimensions or levels, but when aggregated to an overall level, their reliability coefficient ranges between .51 and .75, which is much higher than when aggregating across dimensions or levels (Jackson et al., 2016). Therefore, to investigate the effect of aggregation on

reliability coefficients in this study, the results are expressed as between-subject comparisons and aggregation to dimensions, levels and an overall score.

Between-subject sources of variance are reported in this study as they directly relate to variance associated with differences between participants (Putka & Hoffman, 2013). Aggregation to dimensions, levels and the overall level were calculated using formulae reported in the literature (Putka & Hoffman, 2013, Jackson et al., 2016). In estimating the proportion of variance related to reliable and unreliable variance associated with each level of aggregation, it is possible to calculate G-coefficients for each level of aggregation and compare them across samples and perspectives. As there has been little research in the area of GBA, a comparison between different approaches will help further understanding in the area and identify which level and perspective result in the most reliable outcomes for GBAs.

Chapter 6: Results

In the previous chapter I outlined the methodology of this study, and outlined the analysis techniques I planned to use. In this chapter I will present the results of the generalizability study. The aim of this chapter is to present the findings from three separate samples to identify the extent to which different sources of variance contribute to overall reliability in a GBA, and to analyse the results in comparison to different theoretical perspectives.

Firstly, I will present the descriptive statistics from the study along with correlational analyses, which can give an initial indicator of the relationship between the dimensions and levels measured within the GBA. Next, I will present the results of the generalizability study based on different levels of aggregation to the dimensions, levels and overall scores. Finally, I will present the findings based on the dimension-based, situation-based perspectives and compare these to the interactionist perspective to further identify how different levels of aggregation relate to different theoretical perspectives, and how classifying the different sources of variance as reliable and unreliable variance impacts the reliability of the GBA.

6.1 Descriptive Statistics

The descriptive statistics for all three samples are presented in Table 6.1. The mean scores and standard deviations (SD) for each item are broadly consistent across all three samples, but there is evidence of some differences across samples for specific items. For example, Sample 2 mean scores are higher for Lev1 Dim1 Itm1. All additional items are similar to Sample 1. Sample 3 has a larger proportion of mean scores that are higher when

comparing to Samples 1 and 2. For example, the mean score on Lev3 Itm3 Dim3 is more than double the mean score in the same item in Samples 1 and 2. Additional analysis is required to identify to what extent these differences affect the overall reliability of the assessment across the samples.

6.2 Correlational Analysis

To identify the extent to which the constructs measured in the GBA share variance, a Pearson's correlation analysis was performed. Although the correlational analyses are going to be explained in detail, it is important to note that these results are purely descriptive due to the highly confounded scores discussed. Strong positive correlations between the dimensions would suggest that the dimensions being assessed were measuring a similar construct. Due to the nature of GBA, items measured within the GBA tend to be on variable scales, therefore, before conducting the correlational analysis, items within the level and dimension-based scores were standardised before being combined to form a composite score. Table 6.1 presents the descriptive statistics of the items prior to standardisation, whereas the following tables present the means and SDs of the composite scores.

No significant correlations were identified between the dimensions being measured within the GBA, indicating that whatever is being measured within the three dimensions does not overlap and provides evidence of a lack of redundancy. A lack of significant correlations between the overall dimensions scores was also found in the third sample, which replicates the findings from the previous sample. However, as presented in Table 6.1, a small positive correlation was found between Dimensions 3 and 2 (r=.26) in the second sample, albeit the magnitude of the correlation does not suggest evidence of multicollinearity.

Table 6.1 Descriptive Statistics of Item Scores Across Each Sample

	Sample 1		Sample 2		Sample 3	
Item	Mean	SD	Mean	SD	Mean	SD
Lev1 Dim1 Itm1	1.718	0.317	1.777	0.293	7.702	0.245
Lev1 Dim1 Itm2	5.614	0.200	5.359	0.204	6.085	0.172
Lev1 Dim1 Itm3	5.121	0.169	4.788	0.180	5.606	0.156
Lev1 Dim2 Itm1	7.931	2.074	7.597	1.951	8.284	1.797
Lev1 Dim2 Itm2	24.207	8.280	24.050	8.145	24.807	7.923
Lev1 Dim2 Itm3	9.704	2.058	9.349	1.983	9.997	1.798
Lev1 Dim3 Itm1	5.675	1.014	5.884	0.991	6.797	1.060
Lev1 Dim3 Itm2	2.707	0.794	2.719	0.829	5.075	0.618
Lev1 Dim3 Itm3	5.693	1.130	5.783	1.183	6.744	1.164
Lev2 Dim1 Itm1	1.475	0.179	5.614	0.177	1.407	0.138
Lev2 Dim1 Itm2	4.816	0.050	4.486	0.041	5.306	0.035
Lev2 Dim1 Itm3	1.210	0.109	1.252	0.111	5.726	0.090
Lev2 Dim2 Itm1	2.222	0.406	2.049	0.399	6.020	0.384
Lev2 Dim2 Itm2	5.440	0.372	5.120	0.385	5.863	0.364
Lev2 Dim2 Itm3	5.168	0.221	4.861	0.221	5.626	0.191
Lev2 Dim3 Itm1	6.405	0.826	6.218	0.888	6.733	0.767
Lev2 Dim3 Itm2	6.312	0.952	6.043	1.044	6.607	0.868
Lev2 Dim3 Itm3	3.183	0.519	2.869	0.509	6.791	0.468
Lev3 Dim1 Itm1	5.242	0.160	4.993	0.167	5.742	0.129
Lev3 Dim1 Itm2	5.179	0.135	4.874	0.146	5.661	0.115
Lev3 Dim1 Itm3	5.013	0.138	4.709	0.150	5.497	0.134
Lev3 Dim2 Itm1	11.253	2.078	11.670	2.033	8.490	1.957
Lev3 Dim2 Itm2	11.592	1.817	11.395	1.866	11.841	1.915
Lev3 Dim2 Itm3	2.067	0.380	2.252	0.399	6.012	0.349
Lev3 Dim3 Itm1	6.415	1.014	6.071	0.991	6.797	1.060
Lev3 Dim3 Itm2	33.406	12.088	32.708	10.807	31.135	8.784
Lev3 Dim3 Itm3	3.183	0.519	2.869	0.509	6.791	0.468

Note. Lev = Level, Dim=Dimension, Itm=Item.

Table 6.2 Correlations Between the Composite Dimension Scores Measured in Sample 2

Level	Mean	SD	1	2	3
1	.00	3.80	-		
2	.00	2.37	.82	-	
3	.00	3.46	-.11	.26**	-

**. Correlation is significant at the 0.01 level (2-tailed).

To further identify the extent to which variance is shared across levels, a composite level-based score was also calculated. Correlations between these composite scores are explained below, however, it is imperative to note that these correlations are confounded by every other effect not modelled in the analysis, and are only presented to give an initial indication to the extent to which the dimensions and levels used within this study share variance. In Sample 1, small negative and positive correlations were observed between the levels as presented in Table 6.3. This composite score is confounded by dimension-related variance that may be apparent across all levels, therefore an overlap in variance is unsurprising. Furthermore, small correlations indicate that a large portion of the variance in each level is still unaccounted for, indicating that the levels used within the GBA individually contribute unique variance unaccounted for in the other levels, and that the different levels used within the GBA are not redundant in that they all measure the same behaviour equally across each level.

The magnitude of the correlations in Sample 2 and Sample 3 are similar to those of Sample 1, the relationships between the levels and the direction of the relationships differ. For example, in Sample 1, a small but negative correlation was observed between Dimension 2 and Dimension 3 (r=-.22), whereas in Sample 2, the relationship between the same two dimensions is positive (r=.30) which was also replicated in Sample 3 (r=.34). This suggests that the relationships between levels differ across samples.

Gender differences were investigated using independent t-tests for Sample 1 and Sample 2, as the data were not available for Sample 3. For Sample 2, no gender differences were found for all level-based or dimension-scores. Within Sample 1, there were gender differences observed for Level 3 in the GBA (t(356)=-2.24, p=0.26) with females (M=0.89, SD=.39) scoring higher than males (M=-0.03, SD=.38). However, this effect was small (d=.30;Cohen, 1988).

Finally, the relationship between age and performance on level-based and dimension-based scores were also investigated for Sample 1 and Sample 2 using correlational analysis. Small positive relationships were found between age and Dimension 2 (r=.21, $p<0.05$) and age and Level 2 (r=.19, $p<0.05$). However, within Sample 2, these relationships were not significant.

Small negative relationships were also found in Sample 1 with age and Dimension 1 (r=-.16, $p<0.05$), and age and Level 3 (r=-.14, $p<0.05$). Similar relationships were found between he same variables in Sample 2 (r=-.11 - .12). These findings indicate that there could be age-related effects impacting scores within the GBA, however, it is important to note that these findings are an initial indication of an age-related effect, and are not conclusive as scores are confounded. Furthermore, the age range is very restricted with less than 8% of participants aged over 30 in Sample 1, and less than 2% aged over 30 in Sample 2. Therefore, these findings should be interpreted with caution.

Table 0.3 Correlations Between the Composite Level Scores

Level	Mean	SD	1	2	3
Sample 1 (N= 398)					
1	.00	3.81	-		
2	.00	2.92	.15**	-	
3	.00	2.60	-.20**	-.22**	-
Sample 2 (N=702)					
1	.00	3.80	-		
2	.00	2.37	.082	-	
3	.00	3.46	-.11*	.27**	-
Sample 3 (N=429)					
1	.00	3.80	-		
2	.00	2.37	.082	-	
3	.00	3.46	-.11*	.27**	-

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

In the literature review, I discussed the exercise effect in relation to ACs. Previous studies have replicated this effect consistently, in that the same dimensions measured in different AC exercises frequently correlate less than those between different dimensions measured in the same task (Bowler & Woehr, 2006; Lance et al., 2004; Lievens, 1999; Lievens & Conway, 2001; Sackett & Dreher, 1982; D. Woehr, 2003). As GBA measures dimensions across multiple tasks (in this case, levels), correlational analysis was conducted between the three composite dimension scores in each level to identify the extent to which this phenomenon was observed in the GBA. Table 6.6 presents these correlations for Sample 1.

6.2.1 Monotrait heteromethod correlations

The correlation between the 3 levels used to measure Dimension 1 shows low-to-strong correlations. This suggests that they might be tapping into the same underlying construct. However, the direction of the relationship between these measures of the same dimension suggests that the relationship between the variables may be in the opposite direction. This suggests that certain dimension scores correlate in the opposite direction to that intended. Although these findings would be a sign of the need to reverse score items, to ensure that the relationships between the dimensions were positive .For example, there is a moderate correlation between the first dimension scores between Levels 1 and Level 3 (r=.35), but there is a moderate negative relationship between the scores on the first dimension between Levels 2 and Level 3 (r=-.52), but a small positive relationship between Levels 1 and Level 2 for the same dimension (r=.12). Measuring the internal consistency of the items could help identify if the items are more indicative of a unidimensional construct when reverse scored. However, due to the confounding associated with the different facets of

measurement in this study, this would be inappropriate. It is also difficult to conclude any findings based on these analyses alone as the results are heavily confounded.

For the second dimension, the findings were similar, in that a positive relationship was found between the dimension score for Level 1 and Level 2 (r=.31), but a negative relationship was found between Levels 1 and Level 3 (r=-.20). A non- significant relationship was found between Level 2 and Level 3 measuring the same second dimension This suggests that these dimension scores are unrelated and are unlikely to be measuring the same construct.

For the third dimension, negative and non-significant relationships were observed across levels, indicating that this particular construct may not be measuring a single unidimensional construct. However, it is not possible to draw reliable conclusions from the correlational analyses due to the variation in the scores across different samples. Further analysis of the relationship between measured within the same level (heterotrait monomethod) will show the extent to which the exercise effect is observed in GBAs and give further insight into the extent to which the variation affects candidates' overall scores on the dimensions being measured.

6.2.2 Heterotrait monomethod correlations

For Level 1 and Level 2, there were no significant correlations between the dimensions measured within the same level and thus no correlation. However, for the third level, a small correlation was found between Dimensions 1 and 3 (r=.16) and a strongly negative correlation between Dimensions 2 and 3 (r=-0.51). This suggests that although these dimensions are unrelated, there is an overlap in scores of different traits measured at the same level. However, the strength of these correlations appears to be lower than the majority

of within-dimension correlations, offering an initial suggestion that any observation of the exercise effect might be less substantial than that observed in the AC literature.

6.2.3 Replicability of findings across samples

Results of the correlational analysis between the three composite dimension scores in each level are presented in Table 6.7 and Table 6.8. The previous correlational analysis highlighted a general inconsistency in the relationships across the samples. There are several reasons why this might be, including differences in how participants responded in the different samples, but these results are highly confounded as all the sources related to the measure have not been taken into consideration. This inconsistency is apparent when comparing the monotrait heteromethod correlations in each sample. The heterotrait heteromethod correlations seem to differ in both the strength of relationship and their direction. For example, there was a moderate negative relationship in Dimension 1 measured in Level 2 and Level 3 in the first sample (r=-.52), it was a moderate positive relationship in Sample 2 (r=.31) and a small negative relationship in Sample 3 (r=-.24). The inconsistency between in-dimension correlations is prevalent across all other dimensions in all samples, suggesting that there are sample-related differences in how participants respond to the assessment, resulting in differences in the strength and direction of correlations.

Inconsistencies were also observed between heterotrait monomethod correlations across all samples. For example, in Sample 1, there were no significant correlations found between different dimensions in the same level for traits Level 1 and Level 2, however, in Sample 3, a small negative correlation was observed between Dimensions 1 and Dimension 2 measured across the Level 2 (r=-.08).

Table 6.4 Correlations of Composite Dimension Scores Measured Across All Levels Within the GBA.

Construct	Mean	SD	1	2	3	4	5	6	7	8	9
Sample 1 (N=398)											
1. Dimension 1 Level 1	.00	.95	-								
2. Dimension 1 Level 2	.00	2.00	.12*	-							
3. Dimension 1 Level 3	.00	2.07	.35**	-.52**	-						
4. Dimension 2 Level 1	.00	2.90	.02	-.03	.02	-					
5. Dimension 2 Level 2	.00	1.74	<-.01	.02	-.06	.31**	-				
6. Dimension 2 Level 3	.00	1.28	.21**	.51**	-.51**	-.20**	.01	-			
7. Dimension 3 Level 1	.00	2.14	.03	.04	-.05	.03	-.04	.09	-		
8. Dimension 3 Level 2	.00	.97	.06	.05	-.09	.03	.07	.20**	-.07	-	
9. Dimension 3 Level 3	.00	1.74	.03	-.11*	.16**	-.07	-.01	-.14**	-.56**	-.18**	-
Sample 2 (N=702)											
1. Dimeson 1 Level 1	.00	0.992	-								
2. Dimeson 1 Level 2	.00	1.562	-.186**	-							
3. Dimeson 1 Level 3	.00	2.032	.311**	.333**	-						
4. Dimeson 2 Level 1	.00	2.896	-.025	.022	.085	-					
5. Dimeson 2 Level 2	.00	1.602	.028	-.017	.043	.240**	-				
6. Dimeson 2 Level 3	.00	1.314	-.295**	.296**	.460**	-.022	-.034	-			
7. Dimeson 3 Level 1	.00	2.176	.069	-.015	-.019	.012	.029	-.056	-		
8. Dimeson 3 Level 2	.00	0.889	.048	.067	.105*	-.058	-.094	.112*	-.092	-	
9. Dimeson 3 Level 3	.00	1.689	-.010	.012	.094	.064	.042	.043	-.592**	-.033	-
Sample 3 (N=429)											
1. Dimeson 1 Level 1	.00	2.472	-								
2. Dimeson 1 Level 2	.00	1.545	.041	-							
3. Dimeson 1 Level 3	.00	1.940	.800**	-.241**	-						
4. Dimeson 2 Level 1	.00	2.886	.065	-.004	<0.01	-					
5. Dimeson 2 Level 2	.00	1.559	.044	-.082*	.061	.126**	-				
6. Dimeson 2 Level 3	.00	2.008	-.135**	.173**	-.209**	-.206**	.226**	-			
7. Dimeson 3 Level 1	.00	2.172	-.043	<0.01	-.032	-.026	-.001	.030	-		
8. Dimeson 3 Level 2	.00	2.460	.109**	.024	.057	-.084*	.034	<0.01	-.108**	-	
9. Dimeson 3 Level 3	.00	1.855	.054	-.043	.037	-.062	.012	-.009	.334**	.583**	-

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

6.3 Generalizability Study

Table 6.9 shows all 11 effects modelled in this study for the first sample. Variance estimates are shown as percentages of variance in relation to the total variance explained in the model. The variance estimates reported in the subsequent columns are further categorised as between-participant variance and are relevant to reliability and also presented concerning aggregation related to dimensions and overall scores. Less than 3% of the overall variance was associated with reliable variance and almost 13% was associated with unreliable variance. This suggests that the majority of the total variance from the GBA is unrelated to the measurements' reliability. When only considering the variance related to the reliability of the GBA, this has substantial implications for the assessments' reliability. Almost 85% of between-participant percent of variance is associated with unreliable related effects.

The results from the first sample show that the majority of variance in the GBA tends to be associated with effects that are irrelevant to reliability. For example, Table 6.9 shows that 73.45% of variance in the assessment is associated with the level × item (nested in dimensions) interaction. This effect is based on the interaction between items and levels, which are nested in levels across dimensions and has no effect on the participants' performance in the assessment and thus cannot be associated with either reliable or unreliable variance.

In relation to Research Question 3, which questions the extent to which the person main effect, dimension, and level contributes to scores, from Table 6.9 it appears that although the GBA assesses dimensions, <0.01% of the related variance is accounted for by the person × dimension interaction. Similarly, the person × level interaction also accounted for a small proportion of variance (0.1%). A larger proportion of variance is associated with a three-way interaction between person × level × dimension. This effect includes variance

that is related to how a person scores on a particular dimension being dependent on what level they are completing. Depending on the level of aggregation being considered, this effect also contributes significantly to reliability-related variance (14-56%).

In relation to Research Question 1 and 2 which queries the reliability of the GBA and how aggregation impacts reliability, when aggregating across dimensions, there is always a small effect associated with the person × dimension interaction which results in a low G-coefficient (.16). In contrast, aggregating to an overall GBA score greatly improves the G-coefficient (.63). However, aggregating to an overall score has implications for interpretation of the results from the GBA that need to be considered depending on how they are used. Even with the dramatic improvement of the G-coefficient when aggregating to an overall score, the result is still low.

6.3.1 Replicability of G-Study findings across samples

As in Sample 1, the majority of variance in Sample 2 can be attributed to effects unrelated to reliability. The results for Sample 2 are presented in Table 6.10. The percentage of variance associated with reliability-unrelated effects does not differ significantly from Sample 1. For example, in Sample 1, the level × item (nested in dimensions) interaction accounted for 73.45% of overall variance and in Sample 2, it accounted for 74.72%. A small increase is observed in the variance associated with reliability-related variance. For example, in Sample 1, less than 0.01% of overall variance was associated with the person × dimension interaction. In Sample 2, this rises to 0.56%. Although this may not seem like a large contribution to overall variance when only considering the amount of reliability-related variance, this increases to almost 4.1% and when comparing between-participant variance, the increases explain almost 9% of variance when aggregating across dimensions and almost

34% when aggregating to an overall score. In Sample 1, due to the small figure associated with the person × dimension interaction, even when considering variance, this interaction accounted for little in each level of aggregation (<0.01%).

Although the person × dimension interaction in Sample 2 does account for more variance than that reported in Sample 1, the percentage of variance is still low when considering between-participant percent of variance and when aggregating to dimensions. This effect, in theory, should account for the majority of variance due to the nature of the assessment, in that the assessment is designed to measure dimensions. The person × level interaction (0.1%) also accounts for a small amount of total between-participant variance like the effect found in Sample 1. The largest source of variance related to reliability is the person × level × dimension interaction, which accounts for almost 15% of reliable variance. This percentage is similar to that observed in Sample 1. These findings are consistent and provide evidence that although a large amount of variance is not dependent on the level alone, but is related to the dimension-based scores being dependent on the person and which level they are playing

The small increase in reliability-related variance has resulted in an increase in the overall G-coefficient when aggregating scores across dimensions (.42) and the overall level (.75). However, although the G-coefficients are quite low, when aggregating to an overall score, this begins to approach an acceptable level.

Table 6.5 *Variance Decomposition of Game-based Assessment Scores*

	Sample 1					Sample 2					Sample 3				
Source of variance	VE	Total (%)	BW (%)	D-Score (%)	O-Score (%)	VE	Total (%)	BW (%)	D-Score (%)	O-Score (%)	VE	Total (%)	BW (%)	D-Score (%)	O-Score (%)
p	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	-	-	-	-	-
pl	.059	.101	.670	4.501	7.796	<0.01	.027	.124	.802	1.019	.874	1.641	11.04	45.326	58.009
pd	<0.01	<0.01	<0.01	<0.01	<0.01	.243	.562	4.094	8.86	33.76	.014	.027	.181	.247	.949
pld	1.263	2.168	14.346	32.104	55.604	.874	2.278	14.719	31.855	40.461	1.208	2.269	15.264	20.89	26.735
pi:d	<0.01	<0.01	<0.01	<0.01	<0.01	-	-	-	-	-	<0.01	<0.01	<0.01	<0.01	<0.01
pi:dl + residual	7.482	12.840	84.984	63.395	36.600	4.812	11.125	81.064	58.482	24.760	5.818	10.927	73.515	33.537	14.307
l	1.644	2.821	-	-	-	.218	.505		-	-	.609	1.142	-	-	-
d	.876	1.504	-	-	-	.491	1.136	-	-	-	.112	.2109	-	-	-
i:d	1.887	3.238	-	-	-	.620	1.434	-	-	-	.956	1.795	-	-	-
dl	2.258	3.875	-	-	-	3.668	8.48	-	-	-	.956	1.795	-	-	-
li:d	42.800	73.452	-	-	-	32.320	74.722	-	-	-	42.7	80.194	-	-	-
Estimated G			.145	.361	.634			.189	.415	.752			.189	.664	.857

Note. p = person, l = level, d = dimension, i = item, G = Generalizability Coefficient

In Sample 1, the person × item (nested in dimensions) effect accounted for little overall variance, regardless of the level of aggregation (<0.01). However, in Sample 2, this resulted in a fenced estimate. Fenced estimates occur when the model estimates the variance related to an effect to be of a negative value. An effect cannot be responsible for a negative amount of variance, therefore, the estimator sets this figure to zero (Brennan, 2003). Small effects can return negative variance due to the small amount of variance they are responsible for in more complex models. When comparing the variance to Sample 1 where this effect did not return a fenced estimate, the percentage of variance associated with this effect is almost negligible (<0.01). One of the concerns associated with fenced estimates is that it may bias the results of the model (Jackson et al., 2016) however when compared to the results from Sample 1 and Sample 3, the results are fairly similar.

Table 6.11 displays the G-study for Sample 3. The results are consistent with Sample 1 and 2, however, there are some differences. For example, in Sample 2 there was a fenced estimate for the person × item (nested in dimensions) effect. However, in Sample 3 this effect results in a non-fenced estimate of variance attributable to this effect (<0.01%). This finding is consistent with Sample 1, providing further evidence that this fenced estimate may be a result of a small amount of variance. By contrast, in Sample 3 a fenced estimate was returned for the participant main effect. This is unlike any other sample in this study. This could also be the result of a small amount of variance being associated with this effect. When looking at how much between-participant variance is explained by this effect in Samples 1 and 2, we see that this proportion of variance is almost negligible (<0.01%), indicating that this may also be the cause of the estimate in Sample 3.

Other similarities arise between Sample 3 and the other two samples in the overall variance being related to reliability-unrelated effects, such as the level × item (nested in dimensions) interaction which accounted for around 80% of variance. The specific GBA used in this study measures dimensions, but the amount of overall variance associated with person × dimension-related effects was low (0.03%), indicating that dimensions had very little effect on how participants responded in the GBA for this sample. The person × item (nested in dimensions) interaction also accounts for an insignificant amount of overall variance (<0.01%) which suggests that item-based variance is not significantly responsible for overall variance.

Sample 3's main source of reliability-related overall variance is associated with the person × level × dimension interaction (2.27%). This variance specifies that any dimension-related score is dependent on the level they are completing. When considering the amount of variance related to between-participant percent of variance, this figure increases to account for 15.26% of variance. In Sample 3, there is also an increase in the amount of variance associated with the person × level interaction. Although this interaction only accounts for a small amount of overall variance (1.64%), this figure rises exponentially when considering any effect related to between-participant comparisons (11.04%), dimension-based aggregation (45.33%) and overall scores (58.01%). Due to the relatively small amount of overall variance associated with this effect in Sample 1 (0.1%) and Sample 2 (0.02%), the effect of aggregation did not result in large increases in the variance associated with this effect. This indicates that participants' behaviour is dependent on the level they are completing within this particular sample.

When considering the effects only related to between-participant percent of variance, person × level interaction (11.04%) accounts for less variance than the person × level × dimension interaction (15.26%). However, when aggregating scores, there is a shift in the amount of variance associated with both these effects. When aggregating across dimensions, the person × level interaction raises to account for 45.33% of variance, while the person × level × dimension interaction raises to account for 20.9%, which indicates that when aggregating across dimensions, the person × level interaction accounts for more variance. This then increased when aggregating to an overall score as the person × level × dimension interaction accounts for 26.74% of variance, whereas the person × level interaction accounts for 58.01%. The effect of aggregation in Samples 1 and 2 has increased the amount of variance associated with reliability-related variance increasing the G-coefficient dependent on the level of aggregation. However, even though aggregation raises the G-coefficient, the aggregated results are still low. This is also reflected in the findings of Sample 3 (.66) although it is higher than reported in Samples 1 and 2 (0.36-0.42). When aggregating to an overall score, increases in G-coefficients are observed for Sample 1 (.63) and Sample 2 (.75); however, in Sample 3, the G-coefficient increases to .86 indicating that, at the overall level, for sample 3 the GBA can be interpreted as reliable.

6.4 Alternative Theoretical Perspectives Results

In the previous section, I presented the results of the G-study in which the basis of what was deemed to be reliable, unreliable and reliability-unrelated variance was based on the interactionist perspective. This perspective is based on the theoretical understanding that behaviour is a product of an interaction of trait and situation, and that both trait (in this instance, dimensions) and situation (levels) contribute to reliability-related variance. In this

section, to address Research Question 5 which relates to what impact that aligning to different theoretical frameworks has on the reliability of the GBA, I will present the findings from the G-studies based on trait and situationist theories. This will help identify the extent to which different theoretical perspectives can change what we class as reliable and unreliable effects and what effect this has on reliability estimates. The results for this part of the analysis are presented in Table 6.12.

The majority of variance is not related to dimensions and when removing any variance associated with the level and classing this as irrelevant, this reduces all G-coefficients across dimensions to less than .05 across all three samples in comparison to the G-coefficient observed for the interactionist perspective analysis. This is because the largest contributing factor to reliability-related variance was level-based and as this has been removed from the model, there is little variance left in the model, with the largest portion of variance still associated with the residual term. This is also observed when aggregating across levels and, to some extent, when aggregating to an overall score. However, in Sample 2, when aggregating to an overall score, the G-coefficient is significantly higher (.58), but all G-coefficients are still low, which indicates that when taking a strongminded dimension-based perspective, reliability estimates decrease in the GBA.

This theoretical perspective of the situationist perspective implies that situation-related effects such as the level will be responsible for the largest portion of variance in the GBA. As noted in the previous section, the majority of variance was associated with level-related effects which can be defined as different situations. When considering the situationist perspective, any dimension-related effect is treated as error, apart from the person × level × dimension interaction which can be observed as a level-related effect, as any dimension-based score is dependent on the level.

In comparison to the dimension-based perspective, the G-coefficients are higher when taking a staunch situationist perspective regardless of the level of aggregation. This is a result of the minimal contribution dimension-based scores have at explaining overall variance. However, the majority of G-coefficients are still not high enough to be deemed reliable. All G-coefficients are smaller than that observed in the interactionist model although, for Sample 3, when aggregating to an overall score the G-coefficient is large enough to be classed as reliable (.85). These findings suggest that the interactionist perspective provides stronger evidence of reliability and that behaviour in the GBA can be explained by both the interaction between the dimension and situation. The impact of the results discussed in this section will be explored in the next chapter.

Table 6.6 Composition of Variance and Generalization for Dimension and Situationist based perspectives

		Variance Composition		Eρ2			Interpretation
Perspective/Level		Reliable	Unreliable	Sample 1	Sample2	Sample 3	
Dimension							
	Dimension	p, pd/nd	pi:dl + residual/nd	<0.01	.048	.002	Expected coefficient when dimensions are treated as reliable variance, and aggregated across dimensions.
	Level	p,pd	pi:dl + residual/nl	<0.01	.131	.007	Expected coefficient when dimensions are treated as reliable variance, and aggregated across different levels.
	Overall	p, pd/nd	pi:dl + residual/ni:dl	<0.01	.577	.062	Expected coefficient when dimensions are treated as reliable variance, and aggregated across items, levels and dimensions.
Situation							
	Dimension	p, pl,pld	pi:dl + residual/nd, pd/nd, pi:d/ni:d	.059	.157	.311	Expected coefficient when levels are treated as reliable variance, and aggregated across dimensions.
	Level	p, pl/nl,pld/nl	pi:dl + residual/nl, pd, pi:d	.441	.355	.263	Expected coefficient when levels are treated as reliable variance, and aggregated across different levels.
	Overall	p, pl/nl,pld/nld	pi:dl + residual/ni:d, pd/nd, pi:d/ni:dl	.16	.626	.847	Expected coefficient when levels are treated as reliable variance, and aggregated across items, levels and dimensions.

Note. Perspective/Level = aggregation to dimension, level or overall score for both the dimension and level based perspectives, p = person, l = level, d = dimension, i = item, Eρ2 = expected reliability, estimated as the proportion of reliable between-participant variance, nd = number of dimensions in this study (in this study = 3), nl = number of levels (in this study = 3), i:dn = number of items nested in dimensions (in this study = 9).

Chapter 7: Discussion

As noted within the literature review chapters, GBAs have been rising in popularity as an alternative method to measuring individual differences, however, the lack of academic and robust research within this area has resulted in a call for research to explore the psychometric properties into this method of assessment (Armstrong et al., 2016; Chamorro-Premuzic et al., 2016; Church & Silzer, 2016; Horn et al., 2016; Lievens & Van Iddekinge, 2016). Early research into these tools have focused on validity over and above that of reliability and have found evidence of a relationship between constructs measured in a GBA and that of constructs measured through traditional methods (Barends et al., 2021; Buford & O'Leary, 2015; DiCerbo, 2014; Georgiou et al., 2019; Landers et al., 2017; Lopez & Tucker, 2017; Ninaus et al., 2017; Peters et al., 2021; Shute et al., 2016). These findings show initial insight into how GBAs can be used as an alternative form of assessment than their traditional counterparts, however, research into the application of GBA to selection contexts has yet to be investigated, and without evidence of predictive validity, it is unclear how well this method of assessment relates to job-performance, which is one of the most important criteria for an assessment (Schmidt & Hunter, 2004). Furthermore, as discussed within previous chapters, evidence of the reliability of GBAs are sparse, and where reported, does not take into consideration the multiple sources of variance associated with assessment, and therefore, the reliability estimates are likely to be confounded. Without a robust understanding of the reliability of these methods of assessment, validity evidence should be interpreted with extreme caution.

In the previous chapter, I outlined the results from this thesis in regard to the psychometric structure and reliability of a psychometric GBA. The study reported in this thesis

is the first of its kind, in that G theory has been applied to a GBA to deconfound the different sources of variance associated with GBA, and is therefore, to my knowledge, the first study that has been reported to show accurate and unconfounded estimates of reliability for a GBA. I compare three different theoretical viewpoints to measurement, and contrasted these approaches based on the psychometric structure and theoretical interpretation of what sources of variance can be deemed reliable. I also presented the findings associated with the level of aggregation and impact this has on the reliability coefficient.

The results from all three samples share a consistent theme. When considering between-participant percent of variance, the residual term accounts for most of the variance. In answer to Research Question 1, aggregation across dimensions increases the G-coefficient across all samples (.36 – .66), but not to the extent that the assessment can be deemed a reliable measure of dimensions. When aggregating to an overall performance score the G-coefficient approach the boundary to be deemed reliable for Samples 1 and 2 (.63 – .75), however, only when aggregating to an overall score, Sample 3 shows evidence of reliability with a G-coefficient of .86. There were also fluctuations in scores across samples in the variance accounted for by each effect, but also in G-coefficients. In response to Research Question 2, these findings support the notion that reliability of a GBA is dependent on the level of aggregation, with aggregation to an overall score being the most reliable method. However, only one sample approached adequate levels of reliability when aggregated to an overall score.

The primary purpose of the GBA is to measure dimensions. A large amount of variance associated with the person × dimension interaction would indicate that the dimensions contributed to how people responded in the assessment. Consistently across samples, this effect accounted for an insignificant amount of variance in the GBA, regardless of level of aggregation

considered, indicating that dimensions do not contribute to how people perform in the assessment. In answer to Research Question 3 and 4, the largest amount of overall reliability-related variance was found to be related to the person × level × dimension interaction across all samples regardless of the level of aggregation. In Sample 3, this effect also accounted for a large amount of reliability-related variance, but the person × level effect accounted for more variance than the person × level × dimension interaction when aggregating across dimensions or to an overall score. In Samples 1 and 2, this effect explained less variance across all levels of aggregation.

When comparing these findings across different theoretical perspectives, the results provide evidence that neither the dimension-based perspective nor the situationist perspective improve reliability. To address Research Question 5, the results support the interactionist theoretical standpoint in that behaviour is dependent on an interaction between both dimension and situations, and resulted in higher reliability estimates in comparison to the dimension-based or situationist perspective. Coupling these findings with the evidence that the majority of reliability-related overall variance was explained by the interaction between person × level × dimension provides further evidence of the interactionist perspective, as any score on dimensions is dependent on the level being completed, indicating behaviour may be a product of both the level and dimension. The person × dimension effect explained little overall variance in all samples regardless of level of aggregation, while the person × level effect explained a small amount of overall variance in all samples, but explained a substantial amount of variance when aggregating Sample 3 to the dimension and the overall performance level. However, in Samples 1 and 2, this amount of variance was much small. This is further evidence that, when taken singularly, any contribution to overall variance by the level and dimension is minimal and it is

only when both are considered in the same interaction does it explain a meaningful amount of overall variance.

Overall, the findings suggest that when comparing participants, the GBA used in this assessment is not reliable across all samples regardless of perspective. For Samples 1 and 2, even when aggregating across dimensions and overall scores, the G-coefficient were still too low to be considered reliable. In Sample 3, dimension-based aggregation was higher than the other two samples but still resulted in a G-coefficient below 0.8 across all perspectives. However, when aggregating to an overall score for both the interactionist and the situationist perspective, the G-coefficient was above 0.8 and could be deemed reliable. The implications for these findings are discussed within the following sections in more detail.

7.1 Dimensions vs Levels

To investigate the relationship between dimension-based scores and level-based scores correlational analysis was used. Previous studies have used correlation-based analysis to investigate the relationship between multifaceted measures (Chan, 1996; Fleenor, 1996; Jansen & Jongh, 1997; Schneider & Schmitt, 1992) however, it is important to note that these results will be confounded by the additional facets of measurement and offer only an initial insight into the psychometric properties of the assessment . If we are to assume traits are stable, as highlighted within the trait-theory literature (Ross & Nisbett, 1991), one would assume that there would be a strong correlation between the same traits being measured across the different dimensions. However, when looking at the AC literature, the results suggest something different in that (a) dimensions tend to show lower correlations across different exercises, and (b) dimensions within the same level show larger correlations (see Sackett & Dreher, 1982). This has been characterised as the exercise effect (Dilchert & Ones, 2009; Haland & Christiansen,

2002; Lievens et al., 2006; Sackett & Dreher, 1982), and provides evidence that dimensions are less stable across exercises, and variance is more aligned with similar scores across different dimensions within an exercise.

The scoring procedures used in most ACs are dimension orientated (Lance, 2008), yet we know that the psychometric structure of ACs are reflected better in shared variance within exercises, rather than across dimensions. This presents a problem due to the fact that specific targeted dimensions are measured, and if dimension-related variance accounts for little variance in overall scores, it is inappropriate to base decisions or feedback on these scores (Bowler & Woehr, 2006). Therefore, it is not possible to conclude that meaningful individual differences in personality dimensions can be garnered from the scores within ACs.

Within GBAs, dimension scores are also measured across levels. Scores are aggregate to formulate an overall dimension score. However, no research has yet fully accounted for the multifaceted measurement design of GBAs and how this impacts the psychometric structure of the assessment. However, within GBAs, like ACs, one would expect that cross-level correlations of the same traits (monotrait heteromethod) would be higher than within-level correlations of different traits (heterotrait monomethod). However, if there are stronger correlations between heterotrait monomethod relationships, this would provide evidence of an analogue of the exercise effect in ACs, and would have implications for the interpretation of GBA dimension scores.

To compare the relationship between the dimension-based scores and level-based scores, firstly, scores on each dimension were formed into a composite score for each level. This resulted in nine different composite values. The results from Sample 1 illustrate a small to moderate correlation (r=-.52 – .35) for all but two composite scores.

What is important to note is that the relationships between the variables are not all positive, meaning that some items have a negative correlation. This is known to have an impact on reliability (Hughes, 2009). This indicates that there could be some inconsistencies in the relationships between the same dimensions.

Correlations were also found to be lower for heterotrait monomethod composite scores. For Level 1 and Level 2, there were no significant correlations between the dimensions measured. However, for Level 3, there were moderate correlations found between the dimension scores (r=-.52 – .31). Furthermore, these correlations being of a similar value, if not higher than those reported within the between monotrait heteromethod composite scores. There was also a moderate correlation between *Dimension 2 Level 3* and *Dimension 1 Level 2* indicating that there was evidence of a large overlap in variance between different traits measured across different levels (r=.51). These findings are mixed in terms of interpretation, on the one hand, for *Dimensions 1* and *Dimension 2*, scores seem to cluster a lot more organically around dimension-based scores, in comparison to the level. This suggests that the relationship between the dimensions are more consistent, and less impacted by the level, which is the opposite to what is observed in ACs in regard to the exercise effect (Sacket & Dreher, 1982). It is important to note that these are just initial suggestions, and further evidence will be presented in the G study results. However, the initial findings do not suggest that a similar phenomenon to the exercise effect is replicating within GBAs. These findings provide insight into the Research Question 4, in which the impact of the situation (represented in this study by the level) has less of an impact on scores than dimensions, as stronger relationships between heterotrait monomethod variables would suggest this. However, these findings are just an initial insight, therefore, further confirmation is required from the G-study to confirm these findings.

7.1.1 Cross-Sample Consistency

The replication of these findings across the different samples used in this study returned mixed results. There seems to be some level of consistency with the findings. For example, the relationships between the dimension scores are similar within Sample 2 and 3. For example, all composite scores for *Dimension* 1 show small correlations (*r*=-.186 – .33). However, there were some differences noted across samples, for example, within *Dimension 2* only one small correlation was found between all three composite scores (*r*=.24), while the others were found to be nonsignificant. This was also found in *Dimension 3*, with a moderate correlation being found between only two composite scores (*r*=-.59). This is different to the Sample 1 results reported in the previous section. This cross-sample variability shows initial evidence that there may be differences in how participants respond to the assessment across samples, meaning further research into cross-sample stability of GBAs may be required. Furthermore, the issues with the directionality of the relationships between monotrait heteromethod dimensions is consistent across samples.

However, for *Level 1* and *Level 2*, there were no significant correlations between any of the composite scores. For *Level 3* a moderate correlation found between two composite scores (*r*=.46) that was also found in Sample 1, however, within Sample 1, the relationship was reversed. These findings provide further evidence of potential sample-related differences in responding within the GBA.

In Sample 3, the results are more aligned with Sample 1. The majority of correlations were found between the same dimension composite scores across the different levels with only slight differences noted in the strength and directionality of the relationships (r=-.25 – .08).

The results across all three samples suggest that the exercise effect may not be as prominent within GBAs as the correlations between the dimension-based composites tend to be higher. Generally, the correlations tend to be much lower for different dimensions within the same level, and different dimensions across different levels. However, it is not completely cut and dry, as specifically within the first sample, there was evidence of some strong correlations between dimensions measured within the same level. This indicates that some levels may rely less on measuring individual differences but may be measuring a more general level factor. This means that the exercise effect, although less prominent than that observed within the AC literature, could still be an issue within GBAs depending on the type of level. However, results from the G study will confirm this.

A reason for this initial finding may lie in an interactionist theory called trait-activation theory (TAT; Tett & Guterman, 2000). According to this theory, the strength of the situation can determine how participants respond. For example, if a group of participants are put into a high strength situation it is likely they will respond in a similar way (Judge & Zapata, 2015). If GBA levels are extreme in terms of situational strength, this may result in participants responding similarly within a level, while more variance in individual differences appear in levels with lower situational strength. This would suggest that heterotrait monomethod correlations may appear higher for specific levels, while monotrait heteromethod strong correlations could also appear when comparing different levels. These findings have been replicated within the AC literature in regard to specific personality traits (Lievens, et al., 2006) and may explain these

initial findings. However, it is important to note that these findings are discussed to give an initial insight into the GBAs psychometric structure, but the results are highly confounded, and this needs to be considered when interpreting the results. Further analysis presented later will discuss the exercise effect in relation to the findings from the G study analysis.

Another key finding from this analysis relates to the correlations between composite scores seem to differ in strength and direction between samples. As already noted, three samples were analysed individually for this thesis, and upon inspection, it appears that GBA behaviour is not consistent across samples. This may be a result of sample-related differences (different occupations, educational backgrounds and country of origin). Further studies should investigate the stability of the relationships observed within the GBA across similar samples, as well as the differences observed across different samples. These findings will have practical implications for the suitability of the assessment within specific populations.

The correlational analysis reported within this study gives an initial indication of the psychometric properties of the GBA, however a different form of analysis is required to build on these findings and present unconfounded results of the psychometric structure of this GBA. This will allow for a deeper and more robust insight into the reliability of the GBA and to understand what it contributing to reliable and unreliable variance. These analyses reported so far are only an indication of the psychometric properties associated with GBA, whereas the following section will outline how G theory was used to further understand the psychometric properties and reliability of the GBA.

7.2 Variance Components Analysis of the GBA

Variance components analysis was used to identify the psychometric properties of the GBA, and to estimate the reliability of this assessment. When conducting a G study, it is at the researcher's discretion on how to partial the different effects modelled in the study into reliable and unreliable variance. In this study, 11 sources of variance were modelled based on the facets of measurement associated with the GBA. The results from this G study are presented from the interactionist perspective, taking into consideration the effects that have been modelled previously as reliable and unreliable variance in other studies investigating the psychometric properties of multifaceted measures (Jackson et al., 2017; Jackson et al., 2016, Putka & Hoffman, 2013). Any variance component related to the person (apart from the residual term) was counted as reliable variance. One of the reliable sources of variance was related to the participant main effect, also interpreted as a general performance effect (Jackson et al. 2016). There was also one source of variance associated with reliable dimension-related variance (pd) and one source of variance associated with reliable level-related variance (pl), with the final reliable source of variance associated with an interaction effect between the person, level, and dimension (pld). Any source of variance component unrelated to the participant (e.g. level main effects, dimension main effects) were classified as unrelated to reliability.

The purpose of the GBA used within this study, along with many different types of assessment, is to measure job-relevant traits or dimensions in an attempt to predict job performance (Hogan & Foster, 2016; Schmidt & Hunter, 2004). Within self-report personality inventories and dimension-based assessment centres, scores tend to be aggregated to specific traits or dimensions. As this GBA measures specific dimensions (novelty-seeking, creativity, sensitivity to punishment) one would expect that the largest portion of variance in the G study

results to be related to the pd interaction, and less variance accounted for by any level related effects.

7.2.1 Person × Dimension Effect

Research Question 3 focused on which effect out of the participants main effect, level effects, and dimension effects contributed most to reliable variance. To identify the impact of the dimension on overall scores, it is necessary to first look at the pd interaction. From Sample 1, the pd interaction accounts for <0.01% of variance when considering between-participant variance. This means that within this GBA, little variance from a participant is dependent on the dimensions being measured. This finding is consistent with research investigating the psychometric properties of other multifaceted measures like ACs and SJTs. For example, within the AC literature, dimensions were found to account for very little variance (.5% – 1.87%; Jackson et al., 2016; Breil et al., 2020). In SJTs, this figure was found to be similar (.4% – 1.1%; Jackson et al., 2017). Therefore, within GBAs, the fact that dimensions contribute very little to overall scores is consistent with the findings within the literature. These findings are also consistent in the other samples explored within this study where the pd interaction accounts for between .18% – 4.1% of variance.

These findings have practical implications for the use of dimensions within GBA. If hiring decisions are being made, and feedback are being given to candidates based on dimensions, then it is necessary that dimensions are contributing to the scores. However, within the GBA, and with other multifaceted measures such as SJTs and ACs, this is not the case.

Therefore, other approaches need to be considered to scoring that do not rely on dimensions only. These will be explored in further sections.

7.2.2 Person × Level Effect

Within the GBA, participants are exposed to a number of different situations in the form of levels. If level-related variance contributes the most to scores, an approach to scoring and interpreting the results from a GBA would need to shift from a dimension-based perspective to a level-based perspective. This has been suggested within the literature for ACs and SJT (Jackson et al., 2017; Jackson, et al., 2016). However, with GBAs, using situational-based scores be more problematic due to the low fidelity nature the assessments can take. For example, within AC exercises, SJTs and simulations, these should be designed to be more job-related, therefore meaning that performance on these assessments is more likely to be aligned with the job being applied for (International Taskforce on Assessment Center Guidelines, 2015). In contrast, GBAs can be quite abstract with little overlap with job-relevant behaviour. For example, the tasks within this study relate to earning money to blow-up balloons on a computer screen, or restoring power to a building. These tasks would have a limited application to the workplace. This means that if level-related variance is a significant contributor to scores, then this may raise questions about what the levels are measuring if not dimension-related behaviour, and how this relates to job-performance.

Within Sample 1 the variance associated with the pl effect accounts for .7% of between-participant related variance, indicating that although the situation does in fact account for more variance in comparison to dimensions, overall, the variance associated with this effect is

marginal. In Sample 3, there is a higher percentage of variance being accounted for by the pl interaction (11%). These findings are consistent with the AC literature, in that behaviour within an AC is accounted for more by the person × exercise interaction, in comparison to the person × dimension interaction, and forms the foundations of the exercise effect (Bowler & Woehr, 2006; Lance et al., 2004; Lievens, 1999; Lievens & Conway, 2001; Sackett & Dreher, 1982; Woehr, 2003). However, it must be noted that the percentage of variance accounted for by both effects in ACs is much higher on average than found within this study (Jackson et al., 2016; Putka & Hoffman, 2013), and contributes considerably more to reliable variance within ACs. Therefore, based on the findings from this study, a level-based approach would not be appropriate to GBAs as the level contributes little variance to the GBA score.

7.2.3 Participant Main Effect

The participant main effect in this GBA contributed a very small amount of variance in two of three samples (<0.01%) and resulted in a fenced estimate for the third sample. Fenced estimates occur when a model estimates the variance associated with a particular effect as negative, and therefore sets this value to zero (Brennan, 2003). This is a known limitation of REML but can occur when estimates are extremely small, as noted within the other two samples within this study in the other two samples (<0.01%). The scores are set to zero and treated in the same way as fenced estimates in Putka and Hoffman's (2013) study.

The lack of variance associated with the main effect indicates that participants score on GBA is less likely to be related to a general performance factor. In comparison to other multifaceted measures, the participant main effect was found to explain a lot more reliable

variance (Breil et al., 2020; Jackson et al., 2017; Jackson, et al., 2016; Putka & Hoffman, 2013; Viswesvaran et al., 2005). This can be viewed as a positive, as it means that participants are less likely to score higher than others due to a general GBA effect.

7.2.4 Person × Dimension × Level Effect

As noted so far, both levels and dimensions are responsible for a small amount of variance within GBA. The lack of dimension-related variance aligns with the literature on multifaceted measures, but the lack of level-related variance is unique and answers *Research Question 4*, highlighting that the situation does not contribute much variance within GBAs. However, in comparison, the pdl interaction effect accounted for a lot more variance in all three samples for between-participant percent of variance (14.7% – 15.26%). This effect has accounted for a large portion of variance within the AC literature; however, this effect was found to be smaller than the person × exercise effect (Jackson et al., 2016). Comparatively, within GBAs, this would suggest that participant responding on dimensions is dependent on the level to a larger extent than any interaction between the pd interaction, and the pl interaction. Putka and Hoffman (2013) explained the variance associated with the person × dimension × exercise interaction in ACs reinforced the importance of dimensions. However, within their study, the variance associated with the person × dimension interaction was minimal. For this reason, Jackson et al. (2016) attributed the variance attributed to the person × dimension × level effect to further evidence of the exercise effect, alluding to the fact that situational characteristics were more salient in impacting behaviour than dimensions. Putka and Hoffman (2013) argued that even though only a small amount of variance was related to the person × dimension effect,

evidence of the interaction between person × dimension × exercise provided evidence of the interactionist perspective (which will be discussed in more detail later) and therefore, the importance of the dimension could not be underestimated.

The findings in the SJT literature are different to what has been observed within the AC literature, mainly because within the study by Jackson and colleagues (2017), situations were nested in dimensions, so it was not possible to look at the participant × situation effect separately. However, the amount of variance associated with the participant × situation-nested dimensions was small (1.3% – 1.8%) and the majority of variance being unrelated to the dimension or situation.

In regard to this study, although the pl effect did account for more variance than dimensions, the overall impact of both pl and pd effects did not contribute much variance to the overall score, whereas the majority of variance associated with scores relates to the pdl interaction. Therefore, the results seem more aligned with the interactionist perspective as described by Putka and Hoffman (2013). Furthermore, there seems to be less evidence of a GBA equivalent to the exercise effect observed in ACs. This means that even when removing the impact of the dimensions, levels still accounted for relatively little variance. The levels are akin to situations in this study, and in answer to *Research Question 4*, situations do not account for much variance in GBAs, especially in comparison to ACs (Jackson et al., 2016), however levels do contribute variance as part of the interaction effect, meaning that there is situation-specific dimension-related variance captured within the assessment, but this cannot be disentangled from dimensions. This also aligns with the interactionist theoretical standpoint, which does not favour the trait or the situation above the other, but believes that the majority of variance is due to an interaction between both the trait and the situation (Endler, 1975).

The findings reported furthers our understanding of psychometric structure of GBA, and outlines the contribution of the person main effect, dimensions, and levels contribute to reliable variance, which address *Research Question 3*. Furthermore, the results outline how behaviour in GBA can be mainly attributed to an interaction between the person × dimension × level. With such little variance being associated with the dimensions alone, this raises some important consideration regarding the interpretability of the scores generated from the GBA, and their use within selection. With dimension-related behaviour being dependent on the level, it becomes inaccurate to report on dimension scores alone. It would be more appropriate to base decisions or give specific feedback based on within-level dimension-specific performance. This has been researched previously within ACs (Jackson & Englert, 2011). However, the usefulness of this feedback would need to be investigated as mentioned earlier, level-based insights may be less useful for GBA due to the low fidelity and lack of job-related characteristics.

7.3 GBA Reliability

The findings discussed above give an overview of the psychometric structure of the GBA in terms of measurement facets related to reliable variance. When taking into consideration all the sources of variance associated with reliability, the pd and pl interactions were overshadowed by the portion of variance associated with the pld effect, however, the largest percentage of variance associated with the between-participant variance was accounted for by the residual term in all studies (73.5% – 85%).

In comparison to GBAs, the error associated with between-participant percentage of variance in ACs ranged between 20.29% – 44.8% (Jackson et al., 2017; Putka & Hoffman,

2013; Bowler & Woehr, 2009; Arthur et al., 2000). However, it is important to note, only Jackson et al. (2016) correctly modelled all the potential effects associated with ACs, while the other authors estimates were still cofounded. Jackson et al.'s (2016) study had the lowest percentage of variance associated with unreliable variance in comparison to the other studies. However, within the SJT literature, the majority of variance was associated with the error term (54% – 97%; Jackson et al., 2017). This aligns more with the findings from the GBA analysed within this study.

7.3.1 Aggregation of scores within a GBA

GBAs are multifaceted measures, and therefore, it is possible to aggregate scores to different levels like other multifaceted measures (Jackson et al., 2017; Jackson, et al., 2016; Putka & Hoffman, 2013). Aggregation is used within psychometric assessments to create a composite score that allows correlated variance to increase, and uncorrelated variance to decrease (Kuncel & Sackett, 2013). In regard to GBAs, it is possible to aggregate to a dimension-based score. This requires dividing dimension-based variance by the number of dimensions measured within this study (3). However, as the pd interaction accounts for relatively little variance (<0.01%) this has little impact on reliable variance. On the other hand, the residual term accounts for the most variance, and is made up of dimension related variance. As a result, unreliable related variance decreases, which in turn increases the amount of variance associated with the pl (4.5%) and pdl (32.1%) and results in a G coefficient for Sample 1 of .36. This is far below the appropriate value to deem the measure reliable (Arterberry et al., 2014), and as dimension-based scores are the primary indication of a person's inherent traits and suitability

for a role, it is important to note that the majority of variance is still associated with the residual term (63.4%).

In Sample 2, the findings are similar with an increase in pld related variance (31.85%) and pl related variance (8.87%) and a nominal increase in pd related variance (0.8%). This increase results in a higher G coefficient of .42 for this sample, but it is still low in terms of acceptability of reliability.

The results from Sample 3 show a larger increase in dimension-based score within the GBA when aggregating to dimensions. When taking into consideration the impact of dividing the amount of variance across the number of dimensions, the percentage of variance associated with the pl effect (45.33%) increased enough to account for the majority of reliable variance, in comparison to the pld effect (20.9%). This increase in reliability-related variance, coupled with the decrease in variance associated with the residual-term (33.54%) resulted in a G coefficient of .66. This figure is approaching acceptable standards associated with reliability, but it is still low and furthermore, the strength of the coefficient seems to be inconsistent across the different samples.

Aggregation, regardless to which level, results in higher G coefficients reported across all samples, and this has been observed in other studies that have used similar methods to investigate other multifaceted measures. For example, within the AC literature aggregating scores to dimensions resulted in minor improvements to G coefficients, which were already high to begin with (Putka & Hoffman; Jackson et al., 2016). In comparison, when aggregating to dimensions in SJTs, the G coefficient increased from an average of 0.04 when aggregating to the item-level, to an average of .36 (Jackson et al., 2017). However, aggregating scores to

dimensions within GBAs result in G coefficient falling below the acceptable level to deem this assessment a reliable measure of dimensions (Arterberry et al., 2014).

Within SJTs, it is also possible to aggregate to the situation, and for ACs, it is possible to aggregate scores to exercises. This improved the G coefficient slightly over and above that of the item-level aggregation within SJTs, but this still resulted in a low G coefficient (.12 – .20). Aggregating to exercises was also found to increase the G coefficient within ACs in comparison to dimension-related aggregation (.97; Jackson et al., 2016). Even though GBAs measure dimensions across different levels, level-based scores are not used to infer individual differences or suitability for a job within this assessment. As these scores aren't used for selection purposes, it was beyond the scope of this thesis to calculate level-based aggregated score. However, this is something that can be explored further in different types of GBAs that are designed to measure level-based scores. An example of this can be found in Landers et al.'s (2017) study. The authors used a GBA designed to measure GMA in which levels were designed to measure specific facets related to cognitive ability. Scores from each level were aggregated to form a level-based score. Therefore, this level of aggregation, and the impact this has on the reliability of the measure may be more applicable for different types of GBAs.

Finally, GBAs scores can be aggregated to the overall level, in which participants' scores across all the levels and dimensions are aggregated to form an overall score. In this instance, the amount of variance associated with each relevant effect is divided by; the number of levels (3), the number of items associated with the dimensions (3), and the number of dimensions measured within the GBA (3). This results in changes to both the amount of variance associated with each effect, and the overall G coefficient. Within the GBA, this resulted in G coefficients ranging between .63 to .86. This figure approaches the minimum figure for acceptable reliability and

exceeds it in one of the samples. Whereas previously, aggregating scores to dimensions had resulted in sub-par reliability coefficients, the results from this study suggest that reliability can be optimized for GBAs by reporting results at an overall level. It is important to note that the G coefficient differs between samples, being acceptable in one, borderline in another, and exceeding coefficient guidelines in the third sample, therefore this emphasises that reliability is not a property of an assessment, but rather is sample specific (Taber, 2018). Therefore, it is necessary, to estimate the reliability of GBA for particular samples rather than relying on reliability generalizations.

The findings from this thesis align with what has been found within the SJT literature, in which aggregating results to overall scores was also found to significantly improve G coefficients across all samples (.54 to .75; Jackson et al., 2017). Each sample within the SJT completed different types of SJTs, and therefore differences observed between samples could be a result of the different SJTs used within the study. In comparison, within this study, all participants received the same GBA, indicating sample-related differences in how participants respond to the same assessment. However, when aggregating to an overall score, the largest proportion of variance accounted for in SJTs was associated with the participant main effect. However, within this study, the participant main effect contributes only a trivial amount of variance in all three samples (<0.01%).

In terms of the GBA used within this study, the outcome is similar to that of SJTs, interpreting dimension specific, or even situationally specific differences in behaviour are not recommended as they are not psychometrically warranted (Jackson et al., 2017). In response to Research Question 2 aggregation level does impact the reliability of the GBA. From a practical perspective this means that it is possible and recommended to aggregate to an overall level and

isolate these reliable effects, because aggregation to dimensions or levels does not result in a adequate reliability estimate. However, what this score represents within a GBA is unknown and will require further investigation. However, due to the low fidelity of the GBA assessment it is unknown how this will relate to job performance. SJTs and ACs tend to be more job-relevant and show predictive validity with job performance (Lievens & Sackett, 2006; Christian et al.,2010; McDaniel et al., 2001) regardless of the fact that dimension-based scores in both methods of assessment contribute very little variance.

In regard to *Research Question 1* (to what extent can GBAs be deemed reliable measures of personality dimensions) the findings show that dimensions account for very little variance in scores across all samples within this study. Furthermore, in regard to dimension-based scores, reliability coefficients are low and cannot be deemed reliable. However, to also address the *Research Question 2* (to what extent does aggregating to dimensions and overall scores affect the reliability of GBAs) the results show that when aggregating to an overall score, reliability coefficients increase, and either approach or exceed standard reliability benchmarks. Therefore, scoring GBAs based on dimensions is not appropriate as dimensions contribute very little to overall variance. Furthermore, when aggregating to dimensions the G coefficients were found to be unreliable. Therefore, based on these findings, it is suggested that GBA scores should be aggregated to the overall level, as this was found to be more reliable. However, caution should be exercised when interpreting this score, as further evidence is required to understand the validity of the overall GBA performance score, and how this can be used to predict job-performance.

7.4 Theoretical Perspectives to Reliability

In the previous section, I investigated the impact of aggregating scores to different levels (dimension and overall scores) in order to optimize the reliability of this GBA. However, the variance components that were attributed to reliable variance, unreliable variance, and reliability unrelated variance were based on previous studies investigating multifaceted measures (Jackson et al., 2017; Jackson, et al., 2016; Putka & Hoffman, 2013) and reflects the interactionist perspective, as both dimension and level related variance is interpreted as being related to reliable variance. However, there are a number of different theories that relate to measurement, and due to the flexibility associated with G theory, it is possible to classify the effects modelled within the G study in different ways to reflect different theoretical perspectives. In the following section, I compared alternative theoretical approaches to measurement that have been hotly contested within the literature to explore the impact this has on the reliability of GBAs, and to ascertain how reliability estimations can be impacted depending on your theoretical persuasion.

7.4.1 The trait perspective

Theoretically speaking, if you abide by the trait-perspective, any variance accounted for by the situation (or level) should be treated as irrelevant variance. When using a GBA, dimension-based scores are measured to identify a participant's score on that particular dimension (Arctic Shores, 2018). The expectation from a measurement perspective would be that the dimensions should account for the majority of variance, with dimension-based variance (and the participant main effect) reported as the only sources of reliable variance. However, based on the previous findings reported in this thesis, we know that this is not the case, and that dimensions account for a trivial proportion of total variance. Also, taking into consideration that the pdl effect should be treated as irrelevant variance when considering the trait-theory

perspective, it is not surprising that the G coefficients for dimension-based aggregate scores across samples are so low when aggregated to dimensions (<.001 to .05). When comparing the results from a trait-theorist perspective the reliability is lower than that of the interactionist perspective reported earlier.

Some improvements to the G coefficient can be found in two of the three samples when aggregating scores to levels (<.01 – .13). However, this is still much lower than the expected level to conclude that the measure is reliable. However, when aggregating scores to the overall level, and taking a trait-theory perspective, the G coefficient increases dramatically for one of the samples (.58) but has little effect on the other two samples (<0.01 – .06). This is mainly due to the fact that dimensions account for such a small amount of variance prior to aggregation with the majority of variance still being associated with the residual term post-aggregation.

Previously it has been noted that trait-theory makes logical sense (Mischel, 1968), however, there is a lack of support for this theory, as outlined in research related to other multifaceted measures such as SJTs and ACs, and is further supported by the findings in this study. Dimensions do not account for a substantial amount of variance, and the variance associated with dimensions is not enough provide evidence that behaviour is impacted by traits. As noted by Putka and Hoffman (2013), the dimensions are important and their value should not be discounted, but as many other authors have noted, behaviour does not take place in a situational vacuum (Deinzer et al., 1995), and the impact of the situation cannot be underestimated. Therefore, taking dimensions into consideration without also considering the impact of the situation is an issue with this type of measurement, and even when aggregating to different levels (dimension, level, or overall), this still results in low reliability estimates.

GBAs have not been researched widely, but games provide an optimal environment to measure the consistency of traits across situations (Jackson et al., 2016). The results from the correlational analysis show some relationship between dimensions across certain levels, but this result is highly confounded by level-related variance. The G study extends on these findings by showing that the majority of reliable variance is associated with the pld interaction, and little variance is associated with pd interaction. Finally, the results of the trait-perspective G study also show that when taking a trait-theory perspective, reliability is negatively impacted, even when aggregating to an overall score. The evidence presented in this thesis replicates a key issue found within the literature regarding trait-theory, in that the consistency of behaviour across multifaceted measures is low. This aligns fully with the interactionist perspective, and from a practical perspective these findings highlight that when multifaceted assessments are designed to measure dimensions, the importance of the situation needs to be taken into consideration, especially within GBA, to ensure that the constructs that are being measured are reliable, and the scores generated from these assessments reflect the constructs of interest. Without this consideration, it is likely that your assessment will be measuring something different and may be less reliable and a less valid measure of a construct. Therefore, early consideration for the impact of the situation should help reduce this issue.

7.4.2 The situationist perspective

According to the situationist perspective, situations are the main contributor to behaviour and dimension-related variance accounts for little variance in behaviour (Ekehammar, 1974; Mischel, 1968). Therefore, dimension-related variance can be classified as unreliable variance when taking a situationist perspective to measurement. This would mean that any dimension

related effects are treated as unreliable and contributing to error within GBAs, alongside the residual term. However, in the design of this study, the pdl effect could be interpreted as a situational effect, as any effect of the dimension is related to level, as seen in Jackson et al. (2016), as well as an effect related to the interactionist perspective. Furthermore, situationist believe that dimension-related variance does account for some variance in behaviour, however, after the situation has been taken into consideration, the impact of dimensions becomes redundant (Mischel, 1968). Therefore, in the situational G study analysis, the pdl interaction term is treated as a reliable source of variance, and any dimension related effect not associated with the situation (or level) is treated as unreliable. In the previous analysis, the main contribution of reliable variance was found to be the pdl interaction, and as this is being treated as reliable variance from the situationist perspective, it is no surprise that at each level of aggregation there were consistent increases in the G coefficient. For aggregation to the dimension-based scores, considering the only form of dimension-related variance that is deemed reliable is the pdl effect, G coefficients range from .16 to .31. This value is still low across all samples, but is a marked improvement over the trait-based perspective.

In regard to the aggregating to level-based scores within the situationist perspective, we see a further improvement in the G coefficient in two samples (.36 – .44), however, within Sample 3, this is lower when aggregating to level-based scores as there was more variance in this sample that was associated with the pd interaction, which was classed as an unreliable source if variance within the situationist perspective. Therefore, this results in a decrease in the G coefficient from .31, when aggregated to dimensions, to .26 when aggregated to level-based scores. This effect is not as impactful within the other samples as the amount of variance associated with the pd effect is trivial in size. However, even though this may be the case within

one sample, both coefficients fall well below the desired level to be deemed reliable. This means that even when taking into consideration the situationist perspective, and aggregating to level-based scores, that the scores are still not reliable, and therefore cannot be meaningfully interpreted.

When aggregating to the overall scores the G coefficients increase substantially, with one particular sample exceeding that of adequate reliability (.55 – .85). These findings are aligned with previous research that have found that situations are more important than dimensions at explaining behaviour (Mischel, 1968; Van Heck et al., 1994). However, these findings do not exceed what was found when aligning the variance components with the interactionist perspective.

In terms of the theoretical contribution from the findings within this study, it is possible to see how reliability is impacted depending on which particular theoretical approach is taken. In response to *Research Question 5*, there is very little evidence to support trait-theory. Furthermore, when aggregating scores to dimensions, this results in low G coefficients regardless of the theoretical perspective taken, especially so in the trait-perspective analysis. This goes against our current understanding of trait-theory, in that behaviour should remain consistent across different situations, and that behaviour predicted in one situation (i.e. a personality test or GBA) should predict future behaviour (Ross & Nisbett, 1991). This is not a new finding, as this theory has widely been contested within the literature, however this was the first study to investigate the impact of traits on behaviour (or dimension) within a GBA. These novel findings add to our understanding of behavioural measurement, and further align with the notion that traits do not behave consistently. However, as noted by authors who have replicated these findings with different forms of measurement, low dimension-related variance poses problems of

interpretability (Jackson et al, 2016; Jackson et al, 2017). Because dimensions play such a small role in the overall variance measured in a GBA, it is not appropriate to use dimension scores to make high-stakes selection decisions or use these scores to give candidates feedback based on dimensions.

Within the AC literature, Putka and Hoffman (2013) argue that even though dimension-related variance explains such a small amount of variance, one cannot overlook the importance of dimensions, in that there is still a large amount of variance explained by the person × dimension × exercise interaction. However, within ACs, there has been a large amount of variance associated with the person × exercise effect, and as this is much higher than the person × dimension interaction, this serves to show that the situation is far more important to explain behaviour rather than the dimension (Jackson, 2016). This suggests that the person × dimension × exercise interaction is ACs could be more aligned to the situationist perspective, in that it is the situation that explains the most variance, and that dimension related variance is trivial.

Within this study, the pdl interaction effect accounted for the most reliable variance within the GBA, and due to the minor contribution of the pd effect, the interactionist perspective had the highest G coefficients across all samples when scores were aggregated to the dimension-based scores or to an overall score. As noted previously, the interactionist perspective takes into account the interaction between the situation and dimensions (Griffo & Colvin, 2008). Both are deemed to be more important than any individual contribution of the trait or situation. Previous research has noted the importance of both in explaining behaviour, and the results from this study show that behaviour is not contingent on only dimensions, or only situations, but how these two facets interact with each other. This study illuminates our understanding of the theoretical underpinnings of what is being measured within GBAs, and adds to the literature in

providing a comparison between three different theoretical standpoints, and acknowledging that the interactionist perspective to behaviour results in the most reliable and most interpretable understanding of behaviour within a GBA.

Chapter 8: Limitations and Future Research Directions

In this thesis I have contributed to the literature by outlining the reliability and psychometric structure of a GBA. I have explored how different levels of aggregation impact the reliability of the GBA, and I have investigated how reliability coefficients differ depending on alignment to the trait, situationist, and interactionist perspective of measurement. The implication of these findings will be discussed below from both a practical and theoretical perspective. I will finish the chapter by outlining the limitations to this study, future research directions, and key messages from the thesis.

8.1 Practical Implications

The research literature on GBAs is limited, and it is clear from the research that is available, that there has been a strong focus on validity of GBA with a number of authors finding moderate to strong relationships with traditional assessments methods (Barends et al., 2021; Buford & O'Leary, 2015; DiCerbo, 2014; Georgiou et al., 2019; Landers et al., 2017; Lopez & Tucker, 2017; Ninaus et al., 2017; Peters et al., 2021; Shute et al., 2016). However, less research has been conducted on the reliability of GBA, and where it has been investigated, it is likely to be confounded due to measurement facets not being controlled for (Peters et al., 2021; Quellmalz et al., 2010; Seufert et al., 2016). It is best practice, and the responsibility of the test provider to ensure assessments that are used are both valid and reliable (ITC, 2001). Furthermore without reliability, a test cannot be considered valid (Ritter, 2010). Therefore, without reliability, it is important to be cautious when interpreting the results. To my knowledge, this study is the first

of its kind, and presents an unconfounded estimate of reliability for a GBA, while also giving an insight into the psychometric structure of the assessment, and therefore has implications for how to use GBA within selection.

The results presented in this thesis raise concerns about the practical use of dimensions within GBA. To use a GBA to report on dimensions, give feedback to candidates, or make hiring decisions is not appropriate, as dimensions contribute very little variance to scores, and when aggregating to dimensions, G coefficients fall below the acceptable standard associated with a reliable measure (Vispoel et al., 2017). Therefore, making selection decision, or giving feedback based on dimensions is not recommended.

Aggregating GBA scores to an overall performance level was found to result in the highest reliability coefficient. Therefore, to optimize the reliability of the GBA, it is recommended that overall performance scores should be used instead of dimension-based scores. However, if this approach is taken, then further research is needed to identify the validity of this measure, and how this relates to job-performance.

Research within the AC literature found a similar issue in regard to low dimension-related variance with the majority of variance associated with the exercise effect. This has resulted in authors suggesting an alternative approach to dimension-based ACs. Melchers et al. (2012) discuss mixed-model ACs in which scores on both dimensions and exercises can be used to help make selection and developmental decisions. Jackson and Englert (2011) explored this method and found that this type of AC was more psychometrically sound when compared to similar research using dimension-based ACs.

Within this study, the pdl interaction was found to contribute the most to reliability related variance, meaning that participant responding on a dimension was dependent on the level. Therefore, a mixed-model approach to reporting the scores within a GBA could be explored. However, this approach may be less appropriate with some types of GBAs, specifically those using low-fidelity and non-job-related tasks (levels). This is because level-based feedback is less likely to relate to job-relevant situations. This may be more applicable for different types of gamified or GBA assessments noted within the literature, for example, Georgiou et al. (2019) developed a gamified SJT to measure dimensions. If the assessment measures responses to job-relevant situations, then a mixed approach to scoring and feedback based on situation-related and dimension-related scores could be a better approach than reporting on dimensions. Similarly, Barends et al. (2021) developed a GBA to measure the personality trait honesty. In their GBA, the personality trait honesty was measured over a number of different tasks. Therefore, results based on level-related and dimension-related scores could be presented to help make selection or development decisions. However, as neither study investigated the variance components of their assessment, it is unclear if this approach would be appropriate due to the psychometric structure of their specific assessments. Further research would need to be conducted to identify how dimensions and levels contribute to scores within these specific assessments.

Researchers should not rely on traditional estimates of reliability for multifaceted measures as it can result in confounded estimates. Using G theory to estimate the reliability can provide more accurate and more useful information to help in the development and scoring of a measure. Based on the results of this thesis, reporting scores on individual dimensions and giving feedback to participants based on their dimension scores is inappropriate, as dimensions contribute very little variance to dimension-based scores, and they are not reliable, even though

the test developers report adequate levels of stratified alpha for the dimensions measured within the GBA (Arctic Shores, 2018). Instead, it is recommended that assessment developers use G theory to understand what is being measured reliably (in this case, overall GBA aggregated scores), and then present the results in this way so selection decisions and feedback can be based on reliable scores, and feedback is relevant. However, this can only be done if a solid evidence base and understanding of what is being measured at this level of aggregation is available.

Furthermore, practitioners and users of GBAs should familiarise themselves with G theory, and hold test developers accountable for more robust estimates of reliability as a standard consideration before adopting GBAs within selection (and other multifaceted measures), or conduct G studies themselves to gain better understanding of the reliability of the assessment. A Cronbach's Alpha score of above .7 is not appropriate for all measures and may result in confounded estimates of reliability if used inappropriately.

Within this thesis, there were mixed results in terms of the consistency of the findings across assessments. In terms of the psychometric structure of the assessment, the largest contribution of variance to scores was for the pdl effect, and this was consistent across findings. However, within the correlational analysis, the relationship between the dimension and level composite scores was found to differ across the samples in both size and direction, even though all participants completed the same assessment. One reason for this may be due to variation in participant responding across samples. Therefore, when using a GBA, it is necessary to ensure that the assessment has evidence of measurement invariance with the population group you are investigating. This is important as cultural and language differences has been found to impact the measurement invariance in traditional assessments (Schmitt & Kuljanin, 2008). Until this has been established, it is unclear if scores within a GBA are likely to be consistent across

different samples. Therefore, more research is required to establish if a GBA is suitable for use within specific populations.

8.2 Theoretical Implications

The results from this study add additional insight into the trait vs situation debate that has been prolific within the personality literature (Epstein & O'Brien, 1985). According to the trait perspective, behaviour should remain consistent over situations (Hogan et al., 1977). The correlational analysis provides some evidence of a stability of dimension-based scores across different levels. However, these relationships are highly confounded by other sources of variance. Therefore, once the dimension-related variance was isolated within the variance components analysis, we see that dimensions account for very little variance in GBA scores. This contradicts the basic assumption of trait theory that specifies that behaviours are consistent across situations (Ross & Nisbett, 1991).

Further evidence opposing the trait perspective was evident in the reliability coefficients that were generated for the different theoretical perspectives. In this study, when aggregating to dimensions within the trait, situationist and interactionist perspective, G coefficients were consistently low and deemed unreliable.

The situationist perspective ascertains that traits contribute contribution to behaviour is redundant and that situations are responsible for how people behave (Mischel, 1968). This would suggest that the majority of variance should be associated with the pl effect within this thesis. Although more variance was associated with this effect in comparison the pd interaction, the overall contribution of the levels was minimal. When aggregating scores based on the

situationist perspective, improvements were observed in comparison to the trait perspective, however, when aggregated to levels, the G coefficients were still unreliable.

The results from this thesis provide evidence that the interactionist theory is best suited to explain behaviour within a GBA. The interactionist perspective posits that individually, the situation and trait have less impact on behaviour, but it is the interaction between the trait and the situation which accounts for behaviour (Griffo & Randall Colvin, 2009). This was observed within the variance components analysis results when the pd and pl effects accounted for small amounts of variance, and the majority of reliable variance was explained by the pdl interaction. Furthermore, when taking into consideration all the different levels of aggregation available, the results from the interactionist G study was found to have higher G coefficients compared to trait and situationist results. However, it is important to reiterate that the G coefficients were only reaching acceptable levels for the overall GBA performance score level of aggregation. Jackson et al. (2016) noted that games offer researchers a unique environment to measure participant performance across different situations. In measuring dimensions across different behaviours within a GBA, the findings within this thesis help extend our understanding of behaviour further, and develop our knowledge of what contributes to behaviour and measurement within GBAs – which is an interaction between dimensions and levels.

These findings can be used to help improve assessment design, especially for multi-faceted measures like GBAs. When situational-related variance contributes to variance in an assessment, then ideally, situations should be somewhat job-relevant to help predict performance. ACs tend to use job-relevant exercise (International Taskforce on Assessment Center Guidelines, 2015) , and exercises within ACs have been found to predict job performance (Hoffman & Lopilato, 2015). Without taking into consideration the situation, and only focusing

on dimensions within a measure, it risks confounding the scores with potentially irrelevant variance. This could have a negative impact on the validity of the measure. Therefore, based on the findings from this thesis, assessment developers and researchers need to consider the impact of the situation when designing the tasks within their GBA to ensure that the variance associated with scores is job relevant.

8.3 Limitations

One of the key limitations associated with this study was related to the dimensions selected to include in this analysis. Dimensions were selected based on two key criteria. Firstly, that dimensions were measured across multiple levels. Therefore, dimensions that did not meet this criterion were excluded from the study. This was necessary as the key aim of my thesis was to identify the psychometric structure of the GBA, and one of my research questions required the investigation of the variance associated with situational (level) effects. If levels were nested within dimensions, it would not be possible to deconstruct the variance associated with levels and identify how much level-related variance contributes to overall variance. Although the findings from this study suggest that level-related variance contributed little to the overall score, additional studies could build on the findings of this study to investigate the reliability of additional GBA dimensions not used within this study, to help develop the literature further.

Secondly, the final selection criteria for dimensions were determined by variables being contingent on a response, and were therefore, related to performance in the GBA (Jackson et al., 2016). The variables selected are what is known as behavioural biomarkers and relate to specific variables that are generated throughout the assessment through interactions within the game

(Mandryk & Birk, 2019). Within GBA and other technology enhanced assessments, there is an opportunity to collect a lot more data in comparison to traditional assessments (DiCerbo, 2014; Hao et al., 2015). For example, within one GBA generates up to 5,000 data points (Arctic Shores, 2018). This study is limited by the number and types of variables that were used within this study. Therefore, further research could build on this study to include more of the variables generated through GBA to identify how the use of additional variables could impact the reliability and psychometric properties of this GBA.

As noted in the previous section, there were mixed results in terms of the consistency of the findings across assessments. The relationship between the dimension and level composite scores was found to differ across the samples in both size and direction, even though all participants completed the same assessment, however, a limitation of this study is that the specific roles participants were applying for, and details of the selection process prior to the participants completing the assessment were not shared with the researcher, and therefore, this could contribute to the differences observed between samples. Therefore, this makes it difficult to generalise the findings to alternative samples. Further research into the stability of the psychometric structure across similar samples would help add to the literature to establish if the inconsistencies are accountable to sample-related differences.

Another limitation associated with this study are the samples used. Firstly, as the data used within this study was not primary data, specific limitations were not able to be controlled for. Firstly, limited information about the roles, and the exact process that candidates followed in the selection process prior to completing the GBA was shared, meaning that sample related differences could be related to the process that is out of the researcher's control. Future research using primary data where it would be possible to limit any sample related differences, or account

for these through the analysis. Secondly, demographic data for Sample 3 was not shared, meaning sample related differences observed could be related to demographic differences between the samples. Thirdly, there was a small correlation found between age and scores within the GBA, however, this was small, and the sample was heavily restricted in age. Future research on a less restricted sample would be recommended to investigate if this effect increases in magnitude when the age range is less restricted.

Furthermore, previous research into GBA shows the variability in the types of GBA that can be used for research. For example, some authors have used GBAs that have been specifically designed with certain elements to measure targeted constructs (Godwin-Jones, 2016; Keuning et al., 2019; Landers et al., 2017), whereas others have taken advantage of commercial videogames to create a GBA (Buford & O'Leary, 2015; DiCerbo, 2014; Peters et al., 2021). Therefore, there are multiple game-elements that can be added to gamify an assessment (Fetzer et al., 2017). This study represents one specific type of a GBA with specific game-elements. Therefore, a limitation to this study lies in the difficultly with generalizing these findings to other assessments. Further research is needed to help develop the GBA literature further to understand the psychometric structure of different types of GBA. Furthermore, with the ability to manipulate situational characteristics within a game, it would be interesting to investigate how the extent to which situational strength can be controlled for within a GBA, and what impact this has on situation-related variance in performance. This could help contribute to the literature regarding TAT.

Another limitation of this study was that the REML estimate resulted in two fenced estimates across all three samples. This is a common issue associated when using REML to conduct variance components analysis, in which variance components are set to zero when a

negative variance has been calculated (Brennan, 2001). Although the fenced estimates are likely to be a result of a trivial amount of variance related to these effects within this study, as witnessed in the effect size for the same components in the different samples, it is unknown what impact this could have on the other effects within the G study (Jackson, et al., 2016). However, this has not stopped other authors from using REML estimates in their G theory studies (Putka & Hoffman, 2013). Authors have used Bayesian estimates to overcome the issues associated with REML estimators (Jackson et al., 2020; Jackson, et al., 2016). Future research could use Bayesian G theory to explore GBA to see if this impacts the findings reported within this study.

As noted previously, GBAs generate more data than traditional assessments, and this data can be a lot more complex. The variables used within this study were found to have non-normal multivariate distributions. One of REMLs assumptions is that the data should show normal multivariate distributions. Although this is a limitation to this study, evidence suggest that REML estimators are robust against the violation of the normality assumption, especially in datasets with a larger sample size (Banks et al., 1985; Lumley et al., 2002; Ma & Mazumdar, 2011).

8.4 Future Research Directions

One of the main challenges of conducting this thesis was operating in a research area that has not been heavily developed. Psychometrically speaking, the research presented in this thesis has just scratched the surface of the psychometric properties associated with GBA. As a result, there is an open playing field in future research directions. Researchers who aim to further develop the GBA literature should take into consideration the findings from this thesis and use it

as a foundation to build on, as the findings from this thesis present additional questions that need to be explored to further our understanding of GBAs.

I have already presented a number of potential research areas that could be further explored based on the findings from my thesis in earlier sections of this chapter. However, some research directions are more important than others. I will present future potential research directions that are the biggest concern in regard to GBA and could help contribute to gaps within the knowledge base.

The use of assessment tools for selection purposes should adhere to best practice guidelines. The means that assessments should be valid, reliable, and fair (ITC, 2001; Evers et al., 2013). Currently, there is not enough evidence to suggest that GBA meets these guidelines, so future directions should aim to address these directly.

The research presented in this thesis provides evidence of the unconfounded reliability of a GBA, and evidence of the psychometric structure of assessment scores. Future research should build on these finding where possible. As discussed previously, there are many different formats a GBA can take and further evidence of the psychometric structure and reliability of different types of GBA would benefit the literature. Furthermore, the validity evidence of the GBA used within this assessment reflect that of dimension-based scores, in which correlations were found between GBA dimensions scores and similar constructs measured by a traditional self-report assessment (Arctic Shores, 2019). The findings from this thesis suggest that dimension-based scores are not reliable, and overall performance scores from the GBA are more reliable. Therefore, reporting dimension-based scores is not recommended, as dimensions were found to contribute little variance to overall scores. Validity refers to the extent to which an assessment measures what it intends to measure (Borsboom et al., 2004). However, as the focus of research

has been addressing the validity of dimension-based scores withing gamified assessments used for selection (Pymetrics, 2015; Arctic Shores, 2018), future research would benefit from investigating the validity of the overall performance scores generated scores from a GBA to further understand what is being measured.

Previous evidence has found that cognitive ability is related to performance within video-games (Baniqued et al., 2013; Quiroga et al., 2009; Quiroga et al., 2011). Therefore, overall GBA performance scores could overlap with GMA. It would be beneficial to understand if overall GBA performance scores relate to cognitive constructs, as this would has been found to strongly predict job performance (Tett et al., 1991). Furthermore, it would be pertinent to understand to what extent a GBA that was designed to measure personality dimensions can be used to measure cognitive ability, and how this compares to GBAs that were purposefully designed to specifically measure cognitive ability (see Landers et al., 2017; Peters et al., 2021). This research could help identify key features within a GBA that could be used to maximise the validity of future assessments.

One of the most important aspects of an assessments is to predict job performance (Schmidt & Hunter, 2004). There has also been little research into the predictive validity of the scores associated with GBAs. As outlined in the literature review, convergent validity has been found with GBA scores other traditional measures, although the reliability of the scores investigated in these studies is likely to be confounded, and reliability is a necessary percussor to validity (Ritter, 2010). Therefore, future research should focus on investigating reliable scores, in the case of this study, overall GBA performance, to provide evidence of what is being measured within this score. Landers et al., (2017) did investigate the relationship between overall performance in a cognitive ability GBA, and found that this was a stronger predictor of

GPA than traditional cognitive ability assessments. Even though GPA is often used as a proxy of job performance, further evidence of the predictive validity of GBAs is required.

Fairness is another requirement of assessments used for selection purposes. For an assessment to be considered fair, it must show evidence of validity, scores should relate to essential requirements for the role, and it must not show any adverse impact on subgroups (Higuera, 2001). In the section above, I discussed the need for further research investigating the validity of GBAs. In relation to fairness, there is a gap in the literature to address how reliable scores from a GBA differ across subgroups. This is especially important from a legal perspective for protected groups, which include those with disabilities, females, and ethnic minorities (Bartram, 2008). Gender differences have been found in regard to performance in gamified assessments used within the learning literature (Christy & Fox, 2014; Y. J. Kim & Shute, 2015). Furthermore, group differences have been observed in cognitive ability assessments in regards to ethnic minorities and those with disabilities such as dyslexia (see Beidas et al., 2013; Bobko et al., 2005). With a potential overlap with videogame performance and cognitive ability (Baniqued et al., 2013; Quiroga et al., 2009; Quiroga et al., 2011) , further research should prioritise investigating the subgroup differences for overall GBA scores to ensure that GBA are fair and do not cause adverse impact across groups.

The reliability coefficients presented within this thesis were lower than recommended for the majority of scores, regardless of aggregation, across all samples (>.80; Mushquash & O'Connor, 2006). The findings from this study give insight into how a GBA can be scored to optimize reliability, in that aggregating to an overall performance score is preferred. However, the results of this finding can be expanded upon to also include decision studies. Decision (D) studies can be used to give added insight into how to reduce the error associated with an

assessment (Webb et al., 2006b). A D study is comparable to the Spearman-Brown prediction formula that is suitable for multifaceted measures, such as GBA (Shavelson et al., 1989) . D studies allow the researcher to identify what the G coefficient would be if changes were made to the assessment (Mushquash & O'Connor, 2006). In relation to GBAs, this could involve the measure of additional levels or dimensions. Therefore, a D study could be used to identify how a GBA could be modified to improve the reliability. When applied to SJTs, this resulted in increased variance explained by the participant main effect when adding more dimensions into the assessment. However, the contribution of variance made by dimensions was still low (Jackson et al., 2017). In contrast, when additional dimensions were added using a D study, this resulted in minimal changes to the reliability coefficient (Jackson et al., 2016). It would be beneficial for researchers and assessment developers to explore this method further to help improve the reliability of GBAs, and to see what changes can be made, and how this can result in additional variance explained by dimension or level-related effects.

Finally, GBAs, and other technology enhanced assessments present a new frontier to Occupational Psychologists. The data used within this thesis was limited to the variable information shared by the developer. However, the data generated by GBA is much larger than traditional assessments. For example, up to 5,000 data points can be generated from a single assessment session (Arctic Shores, 2018). Therefore, alternative analysis techniques taken from the data science arena may be more suited to handle this type of data and offer additional insights into the dimensionality of the data collected within a GBA (Mandryk & Birk, 2019). An example of this includes the use of Bayesian network (Bayes nets) analysis that can be used to optimize the scoring within a GBA. Bayes nets allow for interactions between variables and use multi-source data to arrive at scores (Hamilton, 2012). This approach have already been

explored in relation to a learning GBA, with resulting GBA scores showing convergent validity with traditional assessment methods (Y. J. Kim et al., 2016) . These approaches are used less within OP, but they are becoming more popular due to the increase in the amount of data, and complexity of data that is being generated based on enhanced technological assessments.

Another example of further data science analysis methods include data mining, which is an umbrella term that describes different approaches that can be used to identify patterns of behaviour within large data sets, which can be used to predict behaviour (Hand, 2007). Peters et al. (2021) compared scores on a cognitive ability GBA based on the use of outcome data (action data that is contingent on a response) with log data scores generated through different data mining and other machine learning techniques. Log data includes a large number of variables and timestamps collected throughout the entire assessment. The author found that there was a strong correlation between the outcome data and log data scores (r=.67), and a small correlation with the log data scores and a traditional measure of cognitive ability (r=.21). Whereas scores between the outcome data scores and the traditional assessment were slightly higher (r=.39). However, the authors did not investigate if the scores from the log data accounted for additional variance in scores on the traditional assessment over and above that of the scores generated from the outcome data, so it cannot be concluded that this data explains additional variance in the criterion measurement, However, this methodology could be researched further to identify the extent to which log data elements can be used to improve GBA scoring.

However, little is known about how these approaches could impact the psychometric properties of an assessment. Therefore, future research should explore how these analysis methods can be used to improve our current understanding of behaviour and optimize our ability to predict behaviour within the workplace. Furthermore, it would be critical to understand how

these approaches compare to traditional methods used within OP to identify dimensionality of the data. Additionally, it would be pertinent for researchers to consider the psychometric properties associated with assessment when using these methods, such as reliability and validity, and to consider how different variance components could be impacting the scores.

8.5 Final Summary

The findings presented in this thesis contribute well needed insights into a scant literature surrounding GBA. I presented the first unconfounded reliability estimates associated with GBA, and present key insights into what contributes to behaviour within a GBA. A summary of the findings discussed in the thesis is presented below:

To answer *Research Question 1*, G theory was used to identify if GBAs provide reliable measures of dimensions. The findings suggest that this is not the case. Reliability coefficients were incredibly low for scores aggregated to dimensions. Furthermore, dimensions were found to contribute little variance to scores within a GBA. Therefore, dimensions cannot be reliably measured which means that dimension-based scores within a GBA cannot be meaningfully interpreted.

Alternative levels of aggregation were also explored within this study to identify what impact this has on the reliability of the GBA (*Research Question 2*). Aggregating scores to dimensions was found to result in poor reliability estimates (.36 – .66). In contrast, when aggregating to an overall performance score, the G coefficient approached and exceeded acceptable levels depending on the sample (.63 – .86). Aggregated overall performance scores were found to be a more reliable measure of behaviour, but the validity of this needs additional

investigation to help understand what GBA overall performance scores measure from a psychological perspective, and how this relates to job performance.

Research Question 3 and *4* were related to contribution of person, dimension, and levels to reliable variance within the GBA. The largest contribution to variance within the assessment was the Person × Level × Dimension effect (14.35% –15.26%), with little variance being accounted for by the Person × Level (0.67% – 11.04%) or Person × Dimension effect (<0.01% – 4.09%). These findings support the argument that dimensions have little impact within GBAs.

G theory allows researchers to choose which particular facets of measurement are deemed reliable, unreliable, or irrelevant to reliability. To address *Research Question 5*, the variance components of the G study were aligned with the trait-perspective, situationist-perspective, and interactionist-perspective to identify which theoretical perspective had the most reliable scores within the GBA. Evidence from this study support the interactionist perspective with the pdl interaction accounting for the majority of variance. This results in lower reliability estimates when selecting reliable and unreliable sources of variance based on the trait-theory perspective, which is further dependent on the level of aggregation (<0.01 – 0.58). Furthermore, with level-related variance also found to account for a small proportion of variance, resulted in lower reliability estimates until aggregating to an overall score (0.06 – 0.85). The pdl interaction supports the interactionist theoretical perspective, which posits that behaviour is determined by an interaction between the trait and the situation (Griffo & Randall Colvin, 2009), and that reliability was higher than the trait and situational perspectives when selecting reliable and unreliable sources of variance based on the interactionist perspective (.36 – .86).

The findings from this thesis call for further investigation into the psychometric structure of GBAs, with an emphasis on investigating overall performance scores generated from the

GBA, as these were found to be the most reliable. The research presented here also supports the use of G theory to estimate reliability when using multifaceted measures like GBA in comparison to traditional reliability estimates (internal consistency and TRT), which are more likely to result in confounded reliability estimates. Furthermore, the results from this thesis align with previous studies on multi-faceted measures that have found that dimensions contribute very little variance to scores. The biggest contribution to scores was found within the pdl interaction which further supports the interactionist perspective to behaviour. The findings suggest that dimension and levels contribute very little variance in GBA scores, which does not support the trait-perspective and situationist-perspective to behaviour. These findings should be considered when designing future GBAs and the scoring approaches used within this type of assessment.

References

Acquisti, A., Science, C. F.-M., & 2020, undefined. (2020). An experiment in hiring discrimination via online social networks. *Pubsonline.Informs.Org*, *66*(3), 1005–1507. https://doi.org/10.1287/mnsc.2018.3269

Agarwal, S., & Kumari, S. (1982). A correlational study of risk-taking and creativity with special reference to sex differences. *Indian Educational Review*, *17*(3), 104–110. https://psycnet.apa.org/record/1984-17787-001

Alagumalai, S., Curtis, D. D., & Hungi, N. (2005). *a Pplied Rasch Measurement : a Book of Exemplars Education in the Asia-Pacific Region : Issues , Concerns and Prospects*.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*(1), i–171. https://doi.org/10.1037/h0093360

Armstrong, M. B., Ferrell, J. Z., Collmus, A. B., & Landers, R. N. (2016). Correcting Misconceptions About Gamification of Assessment: More Than SJTs and Badges. *Industrial and Organizational Psychology*, *9*(03), 671–677. https://doi.org/10.1017/iop.2016.69

Arterberry, B. J., Martens, M. P., Cadigan, J. M., & Rohrer, D. (2014). Application of generalizability theory to the big five inventory. *Personality and Individual Differences*, *69*, 98–103. https://doi.org/10.1016/j.paid.2014.05.015

Arthur, W., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical re-examination of the assessment center construct-related validity paradox. *Journal of Management*, *26*(4), 813–835. https://doi.org/10.1016/S0149-2063(00)00057-X

Ashton, M. C., Perugini, M., De Vries, R. E., Boies, K., Lee, K., Szarota, P., Di Blas, L., & De Raad, B. (2004). A Six-Factor Structure of Personality-Descriptive Adjectives: Solutions from Psycholexical Studies in Seven Languages. *Journal of Personality and Social Psychology*, *86*(2), 356–366. https://doi.org/10.1037/0022-3514.86.2.356

Ashton, M., & Lee, K. (2005). A defence of the lexical approach to the study of personality structure. *European Journal of Personality*, *19*(April 2003), 5–24. https://doi.org/10.1002/per.541

Atilgan, H. (2013). Sample Size for the Estimation of G and Pho Coefficients in Generalizability Theory. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, *51*, 215–228. https://files.eric.ed.gov/fulltext/EJ1059904.pdf

Ball, S. A., & Zuckerman, M. (1990). Sensation seeking, Eysenck's personality dimensions and reinforcement sensitivity in concept formation. *Personality and Individual Differences*, *11*(4), 343–353. https://doi.org/10.1016/0191-8869(90)90216-E

Baniqued, P. L., Lee, H., Voss, M. W., Basak, C., Cosman, J. D., DeSouza, S., Severson, J., Salthouse, T. A., & Kramer, A. F. (2013). Selling points: What cognitive abilities are tapped by casual video games? *Acta Psychologica*, *142*(1), 74–86. https://doi.org/10.1016/j.actpsy.2012.11.009

Banks, B. D., Mao, I. L., & Walter, J. P. (1985). Robustness of the Restricted Maximum Likelihood Estimator Derived Under Normality as Applied to Data with Skewed Distributions. *Journal of Dairy Science*, *68*(7), 1785–1792. https://doi.org/10.3168/jds.S0022-0302(85)81028-6

Barends, A. J., de Vries, R. E., & van Vugt, M. (2021). Construct and Predictive Validity of an Assessment Game to Measure Honesty–Humility. *Assessment*. https://doi.org/10.1177/1073191120985612

Barrick, MURRAY R, & Mount, M. K. (1991). the Big Five Personality Dimensions and Job Performance: a Meta-Analysis. *Personnel Psychology*, *44*(1), 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Barrick, Murray R, Mount, M. K., & Judge, T. A. (2001). Personality and Performance at the Beginning of the New Millennium: What Do We Know and Where Do We Go Next? *International Journal of Selection and Assessment*, *9*(1&2), 9–30. https://doi.org/10.1111/1468-2389.00160

Bartram, D. (2001). The Development of International Guidelines on Test Use: The International Test Commission Project. *International Journal of Testing*. https://doi.org/10.1207/s15327574ijt0101_3

Bartram, D. (2008). Testing on the Internet: Issues, Challenges and Opportunities in the Field of Occupational Assessment. In *Computer-Based Testing and the Internet: Issues and Advances*. https://doi.org/10.1002/9780470712993.ch1

Bates D, Maechler M, Bolker B, & Walker S. (2015). Package"lme4". *Journal of Statistical Software*.

Batey, M., & Furnham, A. (2006). Creativity, intelligence, and personality: A critical review of the scattered literature. *Genetic, Social, and General Psychology Monographs*, *132*(4), 355–429. https://doi.org/10.3200/MONO.132.4.355-430

Bedwell, W. L., Pavlas, D., Heyne, K., Lazzara, E. H., & Salas, E. (2012). Toward a Taxonomy Linking Game Attributes to Learning: An Empirical Study. *Simulation {&} Gaming*, *43*(6), 729–760. https://doi.org/10.1177/1046878112439444

Beidas, H., Khateb, A., & Breznitz, Z. (2013). The cognitive profile of adult dyslexics and its relation to their reading abilities. *Reading and Writing*, *26*(9), 1487–1515. https://doi.org/10.1007/s11145-013-9428-5

Ben-Gal, I. (2006). Outlier Detection. In *Data Mining and Knowledge Discovery Handbook* (pp. 131–146). Springer-Verlag. https://doi.org/10.1007/0-387-25465-x_7

Biesanz, J. C., & West, S. G. (2004). Towards understanding assessments of the big five: Multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. In *Journal of Personality*. https://doi.org/10.1111/j.0022-3506.2004.00282.x

Bishop, D. W., & Witt, P. A. (1970). Sources of behavioral variance during leisure time. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/h0030067

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, *117*(2), 187–215. https://doi.org/10.1037/0033-2909.117.2.187

Bobko, P., Roth, P. L., & Buster, M. A. (2005). Work sample selection tests and expected reduction in adverse impact: A cautionary note. *International Journal of Selection and Assessment*, *13*(1), 1–10. https://doi.org/10.1111/j.0965-075X.2005.00295.x

Bollen, K. A. (1989). A New Incremental Fit Index for General Structural Equation Models. *Sociological Methods & Research*. https://doi.org/10.1177/0049124189017003004

Borgatta, E. F. (1964). The structure of personality characteristics. *Behavioral Science*, *9*(1), 8–17.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. In *Psychological Review* (Vol. 111, Issue 4, pp. 1061–1071). https://doi.org/10.1037/0033-295X.111.4.1061

Bowler, M. C., Bowler, J. L., & Phillips, B. C. (2009). The Big-5 ± 2? The impact of cognitive complexity on the factor structure of the five-factor model. *Personality and Individual Differences*, *47*(8), 979–984. https://doi.org/10.1016/J.PAID.2009.08.002

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *The Journal of Applied Psychology*, *91*(5), 1114–1124. https://doi.org/10.1037/0021-9010.91.5.1114

Bowler, M. C., & Woehr, D. J. (2009). Assessment center construct-related validity: Stepping beyond the MTMM matrix. *Journal of Vocational Behavior*, *75*(2), 173–182. https://doi.org/10.1016/J.JVB.2009.03.008

Breil, S. M., Forthmann, B., Hertel-Waszak, A., Ahrens, H., Brouwer, B., Schönefeld, E., Marschall, B., & Back, M. D. (2020). Construct validity of multiple mini interviews–Investigating the role of stations, skills, and raters using Bayesian G-theory. *Medical Teacher*, *42*(2), 164–171. https://doi.org/10.1080/0142159X.2019.1670337

Brennan, R. L. (2003). Generalizability Theory. *Journal of Educational Measurement*. https://doi.org/10.1111/j.1745-3984.2003.tb01098.x

Brennan, R. L. (2010). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Brown, J. N. (2011). *The complete guide to recruitment : a step-by-step approach to selecting, assessing and hiring the right people*. Kogan Page Publishers. https://books.google.co.uk/books?hl=en&lr=&id=eJ_llfzKACYC&oi=fnd&pg=PR5&dq=The+Complete+Guide+to+Recruitment:+A+Step-by-step+Approach+to+Selecting+...&ots=JerGXHl1Kr&sig=i6ILzj_jTr_kXCNo2GE8qxVzMhE

Buelow, M. T., & Barnhart, W. R. (2018). Test–Retest Reliability of Common Behavioral Decision Making Tasks. *Archives of Clinical Neuropsychology*, *33*(1), 125–129. https://doi.org/10.1093/arclin/acx038

Buelow, M. T., & Suhr, J. A. (2009). Construct validity of the Iowa gambling task.

Neuropsychology Review. https://doi.org/10.1007/s11065-009-9083-4

Buelow, M. T., & Suhr, J. A. (2013). Personality characteristics and state mood influence individual deck selections on the Iowa Gambling Task q. *Personality and Individual Differences*, *54*, 593–597. https://doi.org/10.1016/j.paid.2012.11.019

Buford, C. C., & O'Leary, B. J. (2015). Assessment of Fluid Intelligence utilizing a computer simulated game. *International Journal of Gaming and Computer-Mediated Simulations*. https://doi.org/10.4018/IJGCMS.2015100101

Burch, B. D. (2011). Confidence intervals for variance components in unbalanced one-way random effects model using non-normal distributions. *Journal of Statistical Planning and Inference*, *141*(12), 3793–3807. https://doi.org/10.1016/J.JSPI.2011.06.015

Byrne, B. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. https://books.google.co.uk/books?hl=en&lr=&id=8vHqQH5VxBIC&oi=fnd&pg=PR5&dq=Structural+equation+modeling+with+Amos:+Basic+concepts,+applications,+and+programming+(3rd+ed.&ots=yn8NKuuDPc&sig=fdEZqK4SLHvu_nGDmIGsTDbQmUg

Canli, T., Amin, Z., Haas, B., Omura, K., & Constable, R. T. (2004). A double dissociation between mood states and personality traits in the anterior cingulate. *Behavioral Neuroscience*. https://doi.org/10.1037/0735-7044.118.5.897

Carvalho, M. B., Bellotti, F., Berta, R., De Gloria, A., Sedano, C. I., Hauge, J. B., Hu, J., & Rauterberg, M. (2015). An activity theory-based model for serious games analysis and conceptual design. *Computers and Education*, *87*, 166–181. https://doi.org/10.1016/j.compedu.2015.03.023

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the Reliability of Situational Judgment Tests Used in High-Stakes Situations. *International Journal of Selection and Assessment*, *20*(3), 333–346. https://doi.org/10.1111/j.1468-2389.2012.00604.x

Cattell, R. B. (1947). Confirmation and clarification of primary personality factors. *Psychometrika*, *12*(3), 197–220. https://doi.org/10.1007/BF02289253

Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2016). New Talent Signals: Shiny New Objects or a Brave New World? *Industrial and Organizational Psychology*, *9*(03), 621–640. https://doi.org/10.1017/iop.2016.6

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology*, *69*(2), 167–181. https://doi.org/10.1111/j.2044-8325.1996.tb00608.x

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, *97*(1), 186–202. https://doi.org/10.1037/a0015618

Christy, K. R., & Fox, J. (2014). Leaderboards in a virtual classroom: A test of stereotype threat and social comparison explanations for women's math performance. *Computers and Education*, *78*, 66–77. https://doi.org/10.1016/j.compedu.2014.05.005

Church, A. H., & Silzer, R. (2016). Are We on the Same Wavelength ? Four Steps for Moving From Talent Signals to Valid Talent Management Applications. *Industrial and Organizational Psychology*, *9*, 645–654.

Clarke-Midura, J., & Dede, C. (2010). Assessment, Technology, and Change. In *Technology, and Change JRTE* | (Vol. 42, Issue 3). www.iste.org/jrte

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. In *Statistical Power Analysis for the Behavioral Sciences (*. Routledge. https://doi.org/10.4324/9780203771587

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*. https://doi.org/10.1016/j.cedpsych.2007.10.002

Commission, I. T. (2001). International Guidelines for Test Use. *International Journal of Testing*.

Costa, P. T., & Mccrae, R. R. (1991). Four ways five factors are basic. *Personality and Individual Differences*, *13*(6), 653–665.

Cowton, C. J. (1998). The Use of Secondary Data in Business Ethics Research. *Journal of Business Ethics*, *17*(4), 423–434. https://doi.org/10.1023/A:1005730825103

Cronbach, L. J., & Shavelson, R. J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, *64*(3), 391–418. https://doi.org/10.1177/0013164404266386

Crook, T. R., Todd, S., Combs, J. G., & Woehr, D. J. (2011). Does Human Capital Matter? A Meta-Analysis of the Relationship Between Human Capital and Firm Performance Organizational Routines in Franchising View project Optimizing Student Team Skill Development using Evidence-Based Strategies View project. *Article in Journal of Applied Psychology*. https://doi.org/10.1037/a0022147

Cropley, D. H., Cropley, A. J., Chiera, B. A., & Kaufman, J. C. (2013). Diagnosing Organizational Innovation: Measuring the Capacity for Innovation. *Creativity Research Journal*, *25*(4), 388–396. https://doi.org/10.1080/10400419.2013.843330

De Beuckelaer, A., & Lievens, F. (2009). Measurement Equivalence of Paper-and-Pencil and Internet Organisational Surveys: A Large Scale Examination in 16 Countries. *Applied Psychology*, *58*(2), 336–361. https://doi.org/10.1111/j.1464-0597.2008.00350.x

de Vries, R. E., de Vries, A., de Hoogh, A., & Feij, J. (2009). More than the Big Five: Egoism and the HEXACO model of personality. *European Journal of Personality*, *23*(8), 635–654. https://doi.org/10.1002/per.733

Deinzer, R., Steyer, R., Eid, M., Notz, P., Schwenkmezger, P., & Ostendorf, F. (1995). Situational effects in trait assessment: The FPI, NEOFFI, and EPI questionnaires. *European Journal of Personality*, *9*, 1–23.

Delprato, D. J., & Midgley, B. D. (1992). Some fundamentals of B. F. Skinner's behaviorism. *American Psychologist*, *47*(11), 1507–1520. https://doi.org/10.1037/0003-066X.47.11.1507

DeRosier, M. E., & Thomas, J. M. (2018). Establishing the criterion validity of Zoo U's game-based social emotional skills assessment for school-based outcomes. *Journal of Applied Developmental Psychology*. https://doi.org/10.1016/j.appdev.2017.03.001

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining gamification. *Proceedings of the 15th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '11*, 9–11. https://doi.org/10.1145/2181037.2181040

Devlin, H. C., Johnson, S. L., & Gruber, J. (2015). Feeling Good and Taking a Chance? Associations of Hypomania Risk with Cognitive and Behavioral Risk Taking. *Cognitive Therapy and Research*, *39*(4), 473–479. https://doi.org/10.1007/s10608-015-9679-3

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between Facets and Domains: 10 Aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5), 880–896. https://doi.org/10.1037/0022-3514.93.5.880

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Journal of Educational Technology {&} Society*, *17*(1), 17–28. http://ezp-prod1.hul.harvard.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true%7B&%7Ddb=psyh%7B&%7DAN=2014-06607-003%7B&%7Dsite=ehost-live%7B&%7Dscope=site%7B%25%7D0Ahttp://kristen.dicerbo@pearson.com

Digman, J. M., & Takemoto-Chock, N. K. (1981). Factors In The Natural Language Of Personality: Re-Analysis, Comparison, And Interpretation Of Six Major Studies. *Multivariate Behavioral Research*, *16*(2), 149–170. https://doi.org/10.1207/s15327906mbr1602_2

Dilchert, S., & Ones, D. S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytic ncremental validity. *International Journal of Selection and Assessment*, *17*(3), 254–270. https://doi.org/10.1111/j.1468-2389.2009.00468.x

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics*, *35*(10), 1297–1326. https://doi.org/10.1080/00140139208967394

Downes-Le Guin, T., Baker, R., Mechling, J., & Ruylea, E. (2012). Myths and realities of respondent engagement in online surveys. *International Journal of Market Research*, *54*(5), 1–21. https://doi.org/10.2501/IJMR-54-5-000-000

Dubbelt, L., Oostrom, J. K., Hiemstra, A. M. F., & Modderman, J. P. L. (2014). Validation of a digital work simulation to assess Machiavellianism and compliant behavior. *Journal of Business Ethics*, *130*(3), 619–637. https://doi.org/10.1007/s10551-014-2249-x

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating Trait Effects from Trait-Specific Method Effects in Multitrait-Multimethod Models: A Multiple-Indicator CT-C(M-1) Model. In *Psychological Methods*. https://doi.org/10.1037/1082-989X.8.1.38

Ekehammar, B. (1974). Interactionism in personality from a historical perspective. *Psychological Bulletin*, *81*(12), 1026–1048. https://doi.org/10.1037/h0037457

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. *Quality of Life Research*, *13*, 715–716.

Endler, N. S., & Hunt, J. M. (1968). S-R Inventories of Hostility and Comparisons of the Proportions of Variance From Persons, Responses, and Situations for Hostility and Anxiousness. *Journal of Personality and Social Psychology*, *9*(4), 309–315. https://doi.org/10.1037/h0026100

Endler, Norman S. (1975). The case for person-situation interactions. *Canadian Psychological Review Psychologie Canadienne*, *16*(1), 12–21.

Epstein, S., & O'Brien, E. J. (1985). The person–situation debate in historical and current perspective. *Psychological Bulletin*, *98*(3), 513–537. https://doi.org/10.1037/0033-2909.98.3.513

Evers, A., Muñiz, J., Hagemeister, C., Hstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*. https://doi.org/10.7334/psicothema2013.97

Eysenck, H. J. (1991). Dimensions of personality: 16, 5 or 3?—Criteria for a taxonomic paradigm. *Personality and Individual Differences*, *12*(8), 773–790. https://doi.org/10.1016/0191-8869(91)90144-Z

Eysenck, H. J. (1992). Four ways five factors are not basic. *Personality and Individual Differences*, *13*(6), 667–673. https://doi.org/10.1143/JJAP.17.1679

Fan, X., & Sun, S. (2014). Generalizability Theory as a Unifying Framework of Measurement Reliability in Adolescent Research. *The Journal of Early Adolescence*, *34*(1), 38–65. https://doi.org/10.1177/0272431613482044

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)* (pp. 105– 146). Macmillan.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58*(2), 203–210. https://doi.org/10.1037/h0041593

Fetzer, M., McNamara, J., & Geimer, J. (2017). Gamification, serious games, and personnel selection: Challanges and potential benefits. In *The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection, and Retention*.

Filzmoser, P. (2004). A multivariate outlier detection method. *Seventh International Conference on Computer Data Analysis and Modeling*, *1*(1989), 18–22. https://pdfs.semanticscholar.org/7e9e/ef87c456b643e2ab6a5856c1bb1c9a01c3d2.pdf

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *The Journal of Abnormal and Social Psychology*, *44*(3), 329–344. https://doi.org/10.1037/h0057198

Fleenor, J. W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. *Journal of Business and Psychology*, *10*(3), 319–335. https://doi.org/10.1007/bf02249606

Fong, C. T. (2006). The Effects of Emotional Ambivalence on Creativity. *Academy of Management Journal*, *49*(5), 1016–1030. https://doi.org/10.5465/amj.2006.22798182

Franken, I. H. A., & Muris, P. (2005). Individual differences in decision-making. *Personality and Individual Differences*, *39*(5), 991–998. https://doi.org/10.1016/j.paid.2005.04.004

Franken, I. H. A., van Strien, J. W., Nijs, I., & Muris, P. (2008). Impulsivity is associated with behavioral decision-making deficits. *Psychiatry Research*. https://doi.org/10.1016/j.psychres.2007.06.002

Fu, J., Zapata, D., & Mavronikolas, E. (2014). Statistical Methods for Assessments in Simulations and Serious Games. In *ETS Research Report Series* (Issue December). https://doi.org/10.1002/ets2.12011

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

Funder, D. C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality*, *40*(1), 21–34. https://doi.org/10.1016/j.jrp.2005.08.003

Funder, D. C., & Ozer, D. J. (1983). Behavior as a function of the situation. *Journal of Personality and Social Psychology*, *44*(1), 107–112. https://doi.org/10.1037/0022-3514.44.1.107

Furnham, A., Batey, M., Anand, K., & Manfield, J. (2008). Personality, hypomania, intelligence and creativity. *Personality and Individual Differences*. https://doi.org/10.1016/j.paid.2007.10.035

Furnham, A., & Fudge, C. (2008). The Five Factor Model of Personality and Sales Performance. *Journal of Individual Differences*, *29*(1), 11–16. https://doi.org/10.1027/1614-0001.29.1.11

Galton, F. (1884). Measurement of character. *Fortnightly Review*, *36*, 179–185.

Gatewood, R. D., Feild, H. S., & Barrick, M. R. (2008). *Human resource selection*. Thomson/South-Western.

Geimer, J. L., Sanderson, K., & Popp, E. (2015). Effects of Gamification on Test Performance and Test Taker Reactions. *Symposium Presentation at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, April*.

George, J. M., & Zhou, J. (2001). When openness to experience and conscientiousness are related to creative behavior: An interactional approach. *Journal of Applied Psychology, 86*(3), 513–524. https://doi.org/10.1037/0021-9010.86.3.513

Georgiou, K., Gouras, A., & Nikolaou, I. (2019). Gamification in employee selection: The development of a gamified assessment. *International Journal of Selection and Assessment*, *27*(2), 91–103. https://doi.org/10.1111/ijsa.12240

Gnambs, T. (2014). *A meta-analysis of dependability coefficients (test-retest reliabilities) for measures of the Big Five*. https://doi.org/10.1016/j.jrp.2014.06.003

Gnambs, T. (2015). Facets of measurement error for scores of the Big Five: Three reliability generalizations. In *Personality and Individual Differences* (Vol. 84, pp. 84–89). https://doi.org/10.1016/j.paid.2014.08.019

Gocłowska, M. A., Ritter, S. M., Elliot, A. J., & Baas, M. (2018). Novelty seeking is linked to openness and extraversion, and can lead to greater creative performance. *Journal of Personality*. https://doi.org/10.1111/jopy.12387

Godwin-Jones, R. (2016). Augmented reality and language learning: From annotated vocabulary to place-based mobile games. *Language Learning and Technology*.

Goldberg, L. R. (1990). An Alternative "Description of Personality": The Big-Five Factor Structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Gopinath Bharathi, A. K. B., Singh, A., Tucker, C. S., & Nembhard, H. B. (2016). Knowledge discovery of game design features by mining user-generated feedback. *Computers in Human Behavior*, *60*(July), 361–371. https://doi.org/10.1016/j.chb.2016.02.076

Gosling, S. D., Rentfrow, P. J., & Jr, W. S. (2003). A Very Brief Measure of the Big Five Personality Domains. *Journal of Research in …*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Griffin, B., & Hesketh, B. (2004). Why Openness to Experience is not a Good Predictor of Job Performance. *2International Journal of Selection and Assessment*, *12*(3), 243–251. https://doi.org/10.1111/j.0965-075X.2004.278

Griffo, R., & Randall Colvin, C. (2009). A brief look at interactionism: Past and present. *Journal of Research in Personality*, *43*(2), 243–244. https://doi.org/10.1016/j.jrp.2008.12.038

Guest, D. E. (1997). Human resource management and performance: a review and research agenda. *The International Journal of Human Resource Management*, *8*(3). https://s3.amazonaws.com/academia.edu.documents/40990409/095851997341630.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1529327503&Signature=k0bRx97cVpqMHeKq8shcaiJd0h4%3D&response-content-disposition=inline%3Bfilename%3DHuman_resource_management_and_pe

Guidelines and ethical considerations for assessment center operations: International task force on assessment center guidelines. (2000). In *Public Personnel Management*. https://doi.org/10.1177/009102600002900302

Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology*. https://doi.org/10.1111/j.1744-6570.2002.tb00106.x

Haland, S., & Christiansen, N. D. (2002). Implications of Trait-Activation Theory for Evaluating the Construct Validity of Assessment Center Ratings. *Personnel Psychology*, *55*(1), 137–163. https://doi.org/10.1111/j.1744-6570.2002.tb00106.x

Hamilton, S. H. (2012). *Article in Environmental Modelling and Software*. https://doi.org/10.1016/j.envsoft.2012.03.012

Hand, D. J. (2007). Principles of data mining. *Drug Safety*. https://doi.org/10.2165/00002018-200730070-00010

Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing Process Data from Game/Scenario-Based

Tasks: An Edit Distance Approach. *JEDM - Journal of Educational Data Mining*, *7*(1), 33–50. http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/JEDM072

Harari, M. B., Reaves, A. C., & Viswesvaran, C. (2016). Creative and innovative performance: a meta-analysis of relationships with task, citizenship, and counterproductive job performance dimensions. *European Journal of Work and Organizational Psychology*, *25*(4), 495–511. https://doi.org/10.1080/1359432X.2015.1134491

Harwell, D. (2019). *Rights group files federal complaint against AI-hiring firm HireVue, citing 'unfair and deceptive' practices*. The Washington Post. https://www.washingtonpost.com/technology/2019/11/06/prominent-rights-group-files-federal-complaint-against-ai-hiring-firm-hirevue-citing-unfair-deceptive-practices/

Hausknecht, J., Day, D. V, & Thomas, S. C. (2004). *Applicant Reactions to Selection Procedures: An Updated Model and Meta-Analysis*. http://digitalcommons.ilr.cornell.edu/articles

Heaton, J. (2008). Secondary analysis of qualitative data: An overview. *Historical Social Research*. https://doi.org/10.12759/hsr.33.2008.3.33-45

Hew, K. F., Huang, B., Chu, K. W. S., & Chiu, D. K. W. (2016). Engaging Asian students through game mechanics: Findings from two experiment studies. *Computers and Education*, *92–93*, 221–236. https://doi.org/10.1016/j.compedu.2015.10.010

Higuera, L. A.-Z. (2001). Adverse Impact in Personnel Selection: The Legal Framework and Test Bias. *European Psychologist*, *6*(2), 103–111. https://doi.org/10.1027//1016-9040.6.2.103

Hoffman, B. J., & Lopilato, A. C. (2015). A Review of the Content, Criterion-Related, and Construct-Related Validity of Assessment Center Exercises. *Article in Journal of Applied Psychology*. https://doi.org/10.1037/a0038707

Hogan, R., DeSoto, C. B., & Solano, C. (1977). Traits, Tests, and Personality Research. *American Psychologist*, *32*(4), 255–264. https://doi.org/10.1037//0003-066X.32.4.255

Hogan, R., & Foster, J. (2016). Rethinking personality. *International Journal of Personality Psychology*, *2*(1), 37–43. http://ijpp.rug.nl/article/viewFile/25245/22691

Horn, R. G., Kaminsky, S. E., & Behrend, T. S. (2016). Don't forget to properly use your signal: Driving down new roads to selection decisions. *Industrial and Organizational Psychology*, *9*(03), 666–671. https://doi.org/10.1017/iop.2016.68

Houser, K., & Voss, W. (2018). GDPR: The end of Google and Facebook or a new paradigm in data privacy. *Rich. JL & Tech*, *25*(1). https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/jolt25&section=5

Howard, A. (1997). Assessment centers: Research and applica-tions. In *Journal of Social Behavior and Personality* (Vol. 12, Issue 5).

Hughes, G. D. (2009). The Impact of Incorrect Responses to Reverse-Coded Survey Items. In *RESEARCH IN THE SCHOOLS Mid-South Educational Research Association* (Vol. 16, Issue 2).

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology*, *85*(6), 869–879. https://doi.org/10.1037/0021-9010.85.6.869

Ihsan, Z., & Furnham, A. (2018). The new technologies in personality assessment: A review. *Consulting Psychology Journal: Practice and Research*, *70*(2), 147–166. https://doi.org/10.1037/cpb0000106

International Taskforce on Assessment Center Guidelines. (2015). Guidelines and Ethical Considerations for Assessment Center Operations . *Journal of Management*, *41*(4), 1244–1273. https://doi.org/10.1177/0149206314567780

Jackson, D. J. R. R., LoPilato, A. C., Hughes, D., Guenole, N., & Shalfrooshan, A. (2017). The Internal Structure of Situational Judgment Tests Reflects Candidate Main Effects: Not Dimensions or Situations. *Journal of Occupational and Organizational Psychology*, *90*(1), 1–27. https://doi.org/10.1111/joop.12151

Jackson, D. J. R. R., Michaelides, G., Dewberry, C., & Kim, Y.-J. J. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology*, *101*(7), 976–994. https://doi.org/10.1037/apl0000102

Jackson, Duncan J.R., Michaelides, G., Dewberry, C., Schwencke, B., & Toms, S. (2020). The implications of unconfounding multisource performance ratings. *Journal of Applied Psychology*, *105*(3), 312–329. https://doi.org/10.1037/apl0000434

Jackson, Duncan J R, & Englert, P. (2011). Task-based assessment centre scores and their relationships with work outcomes. *New Zealand Journal of Psychology*, *40*(2), 37–46. http://www.psychology.org.nz/wp-content/uploads/NZJP-Jackson.pdf

Jackson, Duncan J R, Kim, S., Lee, C., Choi, Y., & Song, J. (2016). Simulating Déjà Vu: What happens to game performance when controlling for situational features? *Computers in Human Behavior*, *55*, 796–803. https://doi.org/10.1016/j.chb.2015.10.031

Jansen, P. G. W., & Jongh, F. de. (1997). *Assessment centres : a practical handbook*. John Wiley & Sons. https://books.google.co.uk/books/about/Assessment_Centres.html?id=_lRyQgAACAAJ&redir_esc=y

Jean, C., Tourneur, W., & Allal, L. (1976). THE SYMMETRY OF GENERALIZABILITY THEORY: APPLICATIONS TO EDUCATIONAL MEASUREMENT. *Journal of Educational Measurement*, *13*(2), 119–135. https://doi.org/10.1111/j.1745-3984.1976.tb00003.x

Jia, Y., Xu, B., Karanam, Y., & Voida, S. (2016). Personality-targeted Gamification: A Survey Study on Personality Traits and Motivational Affordances. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2001–2013.

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*. https://doi.org/10.1016/j.jrp.2014.05.003

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. M. R. M. R. (1999). The big five

personality traits, general mental ability, and career success a. *Personnel Psychology; Autumn*, *52*(3), 621–652. https://doi.org/10.1111/j.1744-6570.1999.tb00174.x

Judge, T. A., & Zapata, C. P. (2015). The person–situation debate revisited: Effect of situation strength and trait activation on the validity of the Big Five personality traits in predicting job performance. *Academy of Management Journal*, *58*(4), 1149–1179.

Kantor, J. R. (1924). *Principles of psychology.* Alfred A. Knopf. https://doi.org/10.1037/10752-000

Keuning, J., Schouwstra, S., Scheltinga, F., & van der Lubbe, M. (2019). Game-Based Spoken Interaction Assessment in Special Need Children. In *Nation and Snowling* (pp. 361–379). Dougherty. https://doi.org/10.1007/978-3-030-18480-3_19

Kim, J. Y., & Heo, W. G. (2021). Artificial intelligence video interviewing for employment: perspectives from applicants, companies, developer and academicians. *Information Technology and People*. https://doi.org/10.1108/ITP-04-2019-0173

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying Evidence-Centered Design for the Development of Game-Based Assessments in Physics Playground. *International Journal of Testing*. https://doi.org/10.1080/15305058.2015.1108322

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers and Education*, *87*(October), 340–356. https://doi.org/10.1016/j.compedu.2015.07.009

Kiran Singh, T., & Kaushik, S. (2015). A Study of Creativity In Relation To Big 5 Personality Traits. *The International Journal of Indian Psychology ISSN*, *3*(19), 2348–5396. http://www.ijip.in

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. Methune & Company.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805. https://doi.org/10.1073/PNAS.1218772110/SUPPL_FILE/ST01.PDF

Koziol, N., & Arthur, A. (2011). An Introduction to Secondary Data Analysis MA CYFS Statistics and Measurement Consultant. *CYFS Reserch Methodolgy Series*. http://r2ed.unl.edu/presentations/2011/RMS/120911_Koziol/120911_Koziol.pdf

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2014). *How "Situational" Is Judgment in Situational Judgment Tests?* https://doi.org/10.1037/a0037674

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. https://doi.org/10.1007/BF02288391

Kuncel, N. R., & Sackett, P. R. (2014). Resolving the assessment center construct validity problem (as we know it). *Journal of Applied Psychology*, *99*(1), 38–47. https://doi.org/10.1037/a0034147

Lance, C. E. (2008). Why Assessment Centers Do Not Work the Way They Are Supposed To. *Industrial and Organizational Psychology*. https://doi.org/10.1111/j.1754-9434.2007.00017.x

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., Conway, J. M., Lance, C. E. ;, Lambert, T. A. ;, Gewin, A. G. ;, & Lievens, F. ; (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology*, *89*(2), 377–385. https://doi.org/10.1037/0021-9010.89.2.377

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment Center Exercise Factors Represent Cross-Situational Specificity, Not Method Bias. *Human Performance*. https://doi.org/10.1207/S15327043HUP1304_1

Landers, R., Armstrong, M., Collmus, A., Mujcic, S., & Blaik, J. (2021). Theory-Driven Game-Based Assessment of General Cognitive Ability: Design Theory, Measurement, Prediction of Performance, and Test Fairness. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000954

Landers, R N, & Landers, A. K. (2014). An Empirical Test of the Theory of Gamified Learning: The Effect of Leaderboards on Time-on-Task and Academic Performance. *Simulation {&} Gaming*, *45*(6), 769–785. https://doi.org/10.1177/1046878114563662

Landers, Richard N. (2015). An introduction to game-based assessment: Frameworks for the measurement of knowledge, skills, abilities and other human characteristics using behaviors observed within videogames. *International Journal of Gaming and Computer-Mediation Simulations*, *7*(4), iv–viii.

Landers, Richard N., Armstrong, M. B., & Collmus, A. B. (2017, April). Empirical validation of a general cognitive ability assessment game. *SIOP Conference*.

Landers, Richard N, & Armstrong, M. B. (2017). Enhancing instructional outcomes with gamification: An empirical test of the Technology-Enhanced Training Effectiveness Model. *Computers in Human Behavior*, *71*, 499–507. https://doi.org/10.1016/j.chb.2015.07.031

Lauriola, M., Panno, A., Levin, I. P., & Lejuez, C. W. (2013). Individual Differences in Risky Decision Making: A Meta-analysis of Sensation Seeking and Impulsivity with the Balloon Analogue Risk Task. *Journal of Behavioral Decision Making*, *27*(1), 20–36. https://doi.org/10.1002/bdm.1784

Le, H., Schmidt, F. L., & Putka, D. J. (2009). *Estimating Construct-Level Relationships The Multifaceted Nature of Measurement Artifacts and Its Implications for On behalf of: The Research Methods Division of The Academy of Management can be found at: Organizational Research Methods Additional services and information for*. *12*. https://doi.org/10.1177/1094428107302900

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). *Evaluation of a Behavioral Measure of Risk Taking: The Balloon Analogue Risk Task (BART)*. https://doi.org/10.1037/1076-898X.8.2.75

Lemons, G. (2011). Diverse perspectives of creativity testing: Controversial issues when used for inclusion into gifted programs. *Journal for the Education of the Gifted*, *34*(5), 742–772.

https://doi.org/10.1177/0162353211417221

Liao, H., Li, Y., & Brooks, G. (2016). Outlier Impact and Accommodation Methods: Multiple Comparisons of Type I Error Rates. *Journal of Modern Applied Statistical Methods*, *15*(1), 23. https://doi.org/10.22237/jmasm/1462076520

Lievens, F. (1999). Development of a Simulated Assessment Center. *European Journal of Psychological Assessment*. https://doi.org/10.1027//1015-5759.15.2.117

Lievens, F., & Christiansen, N. (2010). Core debates in assessment center research: Dimensions 'versus' exercises. *The Psychology of Assessment Centers*, 1–36. http://ebookcentral.proquest.com/lib/bbk/detail.action?docID=4277914.

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology*. https://doi.org/10.1037//0021-9010.86.6.1202

Lievens, F., Dilchert, S., & Ones, D. S. (2006). The Importance of Exercise and Dimension Factors in Assessment Centers: Simultaneous Examinations of Construct-Related and Criterion-Related Validity. *Lievens & Human Performance*, *22*, 375–390. https://doi.org/10.1080/08959280903248310

Lievens, F., & Van Iddekinge, C. H. (2016). Reducing the Noise From Scraping Social Media Content: Some Evidence-Based Recommendations. *Industrial and Organizational Psychology*, *9*(03), 660–666. https://doi.org/10.1017/iop.2016.67

Loevinger, J. (1954). Effect of distortions of measurment on item selection. *Educational and Psychological Measurement*. https://doi.org/10.1177/001316445401400301

Lopez, C. E., & Tucker, C. S. (2017). A quantitative method for evaluating the complexity of implementing and performing game features in physically-interactive gamified applications. *Computers in Human Behavior*, *71*, 42–58. https://doi.org/10.1016/j.chb.2017.01.036

LoPilato, A. C., Carter, N. T., & Wang, M. (2015). Updating Generalizability Theory in Management Research. *Journal of Management*, *41*(2), 692–717. https://doi.org/10.1177/0149206314554215

Lowman, G. H. (2016). Moving beyond identification: Using gamification to attract and retain talent. *Industrial and Organizational Psychology*. https://doi.org/10.1017/iop.2016.70

Lukacik, E. R., Bourdage, J. S., & Roulin, N. (2022). Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews. *Human Resource Management Review*, *32*(1). https://doi.org/10.1016/j.hrmr.2020.100789

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. In *Annual Review of Public Health* (Vol. 23, pp. 151–169). Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA . https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Ma, Y., & Mazumdar, M. (2011). Multivariate meta-analysis: A robust approach based on the theory of U-statistic. *Statistics in Medicine*, *30*(24), 2911–2929. https://doi.org/10.1002/sim.4327

Mandryk, R. L., & Birk, M. V. (2019). The Potential of Game-Based Digital Biomarkers for Modeling Mental Health. *JMIR Mental Health*, *6*(4), e13485. https://doi.org/10.2196/13485

Matsumoto, D., LeRoux, J., Wilson-Cohn, C., Raroque, J., Kooken, K., Ekman, P., Yrizarry, N., Loewinger, S., Uchida, H., Yee, A., Amo, L., & Goh, A. (2000). A new test to measure emotion recognition ability: Matsumoto and Ekman's Japanese and Caucasian brief affect recognition test (JACBART). *Journal of Nonverbal Behavior*, *24*(3), 179–209. https://doi.org/10.1023/A:1006668120583

Maurer, R. (2021). HireVue Discontinues Facial Analysis Screening. *SHRM Review*. https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/hirevue-discontinues-facial-analysis-screening.aspx

Mavletova, A. (2015). Web Surveys Among Children and Adolescents: Is There a Gamification Effect? *Social Science Computer Review*, *33*(3), 372–398. https://doi.org/10.1177/0894439314545316

May, M. A., Hartshorne, H., & Welty, R. E. (1928). Personality and character tests. *Psychological Bulletin*. https://doi.org/10.1037/h0070653

Mccrae, R. R., & John, O. P. (1992). An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28–50. https://doi.org/10.1177/1088868310366253

McDaniel, M., Hartman, N., & Whetzel, D. (2007). Situational judgment tests, response instructions, and validity: a meta-analysis. *Personnel*. http://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.2007.00065.x/full

McPherson, J, & Burns, N. (2007). Gs Invaders: Assessing a computer game-like test of processing speed. *Behavior Research Methods*. http://link.springer.com/article/10.3758/BF03192982

McPherson, Jason, & Burns, N. N. R. (2008). Assessing the validity of computer-game-like tests of processing speed and working memory. *Behavior Research Methods*, *40*(4), 969–981. https://doi.org/10.3758/BRM.40.4.969

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative me ... *Journal of Occupational and Organizational Psychology*, *77*(4), 531–552.

Melchers, K., Wirz, A., & Kleinmann, M. (2012). *Dimensions AND exercises: Theoretical background of mixed-model assessment centers Self-presentation/faking of applicants View project Job insecurity View project*. https://www.researchgate.net/publication/273949972

Messick, S. (1984). The psychology of educational measurement. *ETS Research Report Series*, *1984*(1), i–55. https://doi.org/10.1002/j.2330-8516.1984.tb00046.x

Milgram, S. (1965). Some Conditions of Obedience and Disobedience to Authority. *Human Relations*, *18*(57), 57–76. https://doi.org/10.1177/001872676501800105

Mischel, W. (1968). Personality and assessment. In *International Encyclopedia of Social Behavioral Sciences*.

Montagne, B., Kessels, R. P. C., De Haan, E. H. F., & Perrett, D. I. (2007). The Emotion Recognition Task: A Paradigm to Measure the Perception of Facial Emotional Expressions at Different Intensities. *Perceptual and Motor Skills*, *104*(2), 589–598. https://doi.org/10.2466/pms.104.2.589-598

Morrow, V., Lives, Y., Boddy, J., & Lamb, R. (2014). *The ethics of secondary data analysis: Learning from the experience of sharing qualitative data from young people and their families in an international study of childhood poverty*. www.younglives.org.uk

Moser, K., Schuler, H., & Funke, U. (1999). The Moderating Effect of Raters' Opportunities to Observe Ratees' Job Performance on the Validity of an Assessment Centre. *International Journal of Selection and Assessment*, *7*(3), 133–141. https://doi.org/10.1111/1468-2389.00113

Mun, J., Mun, K., & Kim, S. W. (2015). Exploration of Korean Students' Scientific Imagination Using the Scientific Imagination Inventory. *International Journal of Science Education*, *37*(13), 2091–2112. https://doi.org/10.1080/09500693.2015.1067380

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behaviour Research Methods*, *38*(3), 542–547. www.education.uiowa.edu/casma/GenovaPrograms.htm.

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology*. https://doi.org/10.1037/0021-9010.69.1.182

Ninaus, M., Kiili, K., McMullen, J., & Moeller, K. (2017). Assessing fraction knowledge by a digital game. *Computers in Human Behavior*, *70*, 197–206. https://doi.org/10.1016/j.chb.2017.01.004

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, *66*(6), 574–583. https://doi.org/10.1037/h0040291

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: are they equivalent? *Ergonomics*, *51*(9), 1352–1375. https://doi.org/10.1080/00140130802170387

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.

Osborne, J. W. (1964). *Improving your data transformations: Applying the Box-Cox transformation*. *15*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.7417&rep=rep1&type=pdf

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*. https://doi.org/10.1037/1082-989X.5.3.343

Panayides, P., & Karwowski, M. (2013). Coefficient Alpha Interpret With Caution. *Europe's Journal of Psychology*, *9*(4), 687–696. https://doi.org/10.5964/ejop.v9i4.653

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism,

Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. https://doi.org/10.1016/S0092-6566(02)00505-6

Paunonen, S. V, & Jackson, D. N. (2000). What Is Beyond the Big Five? Plenty! *Journal of Personality*, *68*(5). https://pdfs.semanticscholar.org/f2c2/ffa097e1da83c3df6e0d6bd97ebeb9e0a741.pdf

Pek, J., Wong, O., & Wong, C. M. (2017). *Data Transformations for Inference with Linear Regression: Clarifications and Recommendations. Practical Assessment, Research & Evaluation,*. *22*. https://pareonline.net/getvn.asp?v=22&n=9

Pervin, L. A. (1994). A critical analysis of current trait theory. *Psychological Inquiry*, *5*(2), 103–113. https://doi.org/10.1207/s15327965pli0502_1

Peters, H., Kyngdon, A., & Stillwell, D. (2021). Construction and validation of a game-based intelligence assessment in minecraft. *Elsevier*. https://doi.org/10.1016/j.chb.2021.106701

Piffer, D. (2012). Can creativity be measured? An attempt to clarify the notion of creativity and general directions for future research. *Thinking Skills and Creativity*, *7*(3), 258–264. https://doi.org/10.1016/J.TSC.2012.04.009

Plucker, J. A., & Makel, M. C. (2012). Assessment of Creativity. In *The Cambridge Handbook of Creativity* (pp. 48–73). https://doi.org/10.1017/cbo9780511763205.005

Porter, D. B. (1995). Computer games: Paradigms of opportunity. *Behavior Research Methods, Instruments, & Computers*, *27*(2), 229–234. https://doi.org/10.3758/BF03204737

Povah, N., & Povah, L. (2012). What are assessment centers and how can they enhance organizations? In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 3–24). Routledge/Taylor & Francis Group.

Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology*, *98*(1), 114–133. https://doi.org/10.1037/a0030887

Putka, D. J., & Sackett, P. R. (2010). Reliability and validity. In *Handbook of Employee Selection*. https://doi.org/10.4324/9780203809808

Pymetrics. (2015). *Pymetrics Technical Manual*.

Quellmalz, E. S., Davenport, J. L., Timms, M. J., DeBoer, G. E., Jordan, K. A., Huang, C. W., & Buckley, B. C. (2013). Next-generation environments for assessing and promoting complex science learning. *Journal of Educational Psychology*, *105*(4), 1100–1114. https://doi.org/10.1037/a0032220

Quellmalz, E. S., Timms, M. J., & Buckley, B. (2010). Calipers project. In *Int. J. Learning Technology* (Vol. 5, Issue 3). http://simscientist.org/downloads/IJLT050302 QUELLMALZ.pdf

Quiroga, M. A., Herranz, M., Gómez-Abad, M., Kebir, M., Ruiz, J., & Colom, R. (2009). Video-games: Do they require general intelligence? *Computers and Education*, *53*(2), 414–418.

https://doi.org/10.1016/j.compedu.2009.02.017

Quiroga, M. Á., Román, F. J., Catalán, A., Rodríguez, H., Ruiz, J., Herranz, M., Gómez-Abad, M., & Colom, R. (2011). Videogame Performance (Not Always) Requires Intelligence. *International Journal of Online Pedagogy and Course Design*, *1*(3), 18–32. https://doi.org/10.4018/ijopcd.2011070102

Rdz-Navarro, K. (2019). Latent variables should remain as such: Evidence from a Monte Carlo study. *Https://Doi.Org/10.1080/00221309.2019.1596064*, *146*(4), 417–442. https://doi.org/10.1080/00221309.2019.1596064

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modelling for evaluating questionnaire item and scale properties. *Assessing Quality of Life in Clinical Trial: Methods and Practice*.

Reifman, A., & Keyton, K. (2010). Windsorize. In N. Salkind (Ed.), *Encyclopedia of Research Design* (pp. 1636–1638). SAGE Publications.

Reio, T. G., & Sanders-Reio, J. (2006). Sensation seeking as an inhibitor of job performance. *Personality and Individual Differences*, *40*(4), 631–642. https://linkinghub.elsevier.com/retrieve/pii/S0191886905002825

Rene, F., Lievens, O., & Patterson, F. (2011). *The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection*. https://doi.org/10.1037/a0023496

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*. https://doi.org/10.1037/1089-2680.7.4.331

Richetin, J., Costantini, G., Perugini, M., & Schönbrodt, F. (2015). Should We Stop Looking for a Better Scoring Algorithm for Handling Implicit Association Test Data? Test of the Role of Errors, Extreme Latencies Treatment, Scoring Formula, and Practice Trials on Reliability and Validity. *PLOS ONE*, *10*(6), e0129601. https://doi.org/10.1371/journal.pone.0129601

Richter, G., Raban, D. R., & Rafaeli, S. (2015). Studying gamification: The effect of rewards and incentives on motivation. In *Gamification in Education and Business* (pp. 21–46). https://doi.org/10.1007/978-3-319-10208-5_2

Richterich, A. (2018). How data-driven research fuelled the Cambridge Analytica controversy. *Partecipazione e Conflitto*, *11*(2), 528–543. https://doi.org/10.1285/i20356609v11i2p528

Ritter, N. L. (2010). Understanding a widely misunderstood statistic: Cronbach's alpha. *Understanding a Widely Misunderstood Statistic: Cronbach's α*, 1–17. https://files.eric.ed.gov/fulltext/ED526237.pdf

Ross, L., & Nisbett, R. E. (1991). *The Person and the Situation: Perspectives of Social Psychology.* McGraw-Hill. https://doi.org/10.2307/2075489

Rothmann, S., & Coetzer, E. (2003). The Big Five Personality Dimensions and Job Performance. *SA Journal of Industrial Psychology*, *29*(68–74). https://www.ianrothmann.com/pub/psyc_v29_n1_a9.pdf

Runco, M. A., & Jaeger, G. J. (2012). The Standard Definition of Creativity. *Creativity Research Journal*, *24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion*, *30*(4), 344–360. https://doi.org/10.1007/s11031-006-9051-8

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology*, *67*(4), 401–410. https://doi.org/10.1037/0021-9010.67.4.401

Said-Metwaly, S., Noortgate, W. Van den, & Kyndt, E. (2017). Methodological Issues in Measuring Creativity: A Systematic Literature Review. *Creativity. Theories – Research - Applications*, *4*(2), 276–301. https://doi.org/10.1515/CTRA-2017-0014

Salgado, Jesus F. (1998). Big Five Personality Dimensions and Job Performance in Army and Civil Occupations: A European Perspective. *Human Performance*, *11*(2–3), 271–288. https://doi.org/10.1080/08959285.1998.9668034

Salgado, Jesús F. (1997). The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology*, *82*(1), 30–43. https://doi.org/10.1037/0021-9010.82.1.30

Sava, F., & Sperneac, A. (2006). Sensitivity to reward and sensitivity to punishment rating scales: A validation study on the Romanian population. *Personality and Individual Differences*, *41*, 1445–1456. https://doi.org/10.1016/j.paid.2006.04.024

Schmidt, F. L., & Hunter, J. (2004). General Mental Ability in the World of Work: Occupational Attainment and Job Performance. *Journal of Personality and Social Psychology*, *86*(1), 162–173. https://doi.org/10.1037/0022-3514.86.1.162

Schmidt, F. L., & Hunter, J. E. J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond Alpha: An Empirical Examination of the Effects of Different Sources of Measurement Error on Reliability Estimates for Measures of Individual Differences Constructs. *Psychological Methods*, *8*(2), 206–224. https://doi.org/10.1037/1082-989X.8.2.206

Schmidt, F. L., Oh, I.-S., & Shaffer, J. A. (2016). The validity and utility of selection methods in personnel psychology: Practical and theoretical Implications of 100 Years. *Working Paper*.

Schmitt, N. (1996). *Psychological Assessment Uses and Abuses of Coefficient Alpha* (Vol. 8, Issue 4). Psychological Association, Inc. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.615.4053&rep=rep1&type=pdf

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222. https://doi.org/10.1016/j.hrmr.2008.03.003

Schneble, C. O., Elger, B. S., & Shaw, D. (2018). The Cambridge Analytica affair and Internet-mediated research. *EMBO Reports*, *19*(8), e46579. https://doi.org/10.15252/EMBR.201846579

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology*, *77*, 32–41. https://doi.org/10.1037//0021-9010.77.1.32

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, *8*(9). https://doi.org/10.1371/JOURNAL.PONE.0073791

Seddigh, A., Berntson, E., Platts, L. G., & Westerlund, H. (2016). Does personality have a different impact on self-rated distraction, job satisfaction, and job performance in different office types? *PLoS ONE*, *11*(5), 1–15. https://doi.org/10.1371/journal.pone.0155295

Segalin, C., Celli, F., Polonio, L., Kosinski, M., Stillwell, D., Sebe, N., Cristani, M., & Lepri, B. (2017). What your facebook profile picture reveals about your personality. *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, 460–468. https://doi.org/10.1145/3123266.3123331

Segall, D. O. (1994). The reliability of linearly equated tests. *Psychometrika*. https://doi.org/10.1007/BF02296129

Seufert, M., Burger, V., Lorey, K., Seith, A., Loh, F., & Tran-Gia, P. (2016). Assessment of subjective influence and trust with an online social network game. *Computers in Human Behavior*, *64*, 233–246. https://doi.org/10.1016/j.chb.2016.06.056

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922–932. https://doi.org/10.1037/0003-066X.44.6.922

Shores, A. (2018). *Cosmic Cadet Technical Manual*.

Shores, A. (2019). *Skyrise City Technical Manual*.

Shute, V. J. (2011). Stealth Assessment in Computer-Based Games To Support Learning. *Computer Games and Instruction*, 503–524. www.ncbi.nlm.nih.gov

Shute, V. J., & Ventura, M. (2013). Stealth Assessment: Measuring and Supporting Learning in Video Games. In *The John D. and Catherine T. MacArthur Foundation Reports on Digital Media and Learning*. MIT Press. https://doi.org/http://dx.doi.org/10.4135/9781483346397.n278

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

Simonton, D. K. (2003). Expertise, competence, and creative ability: The perplexing complexities. In *The Psychology of Abilities, Competencies, and Expertise*. https://doi.org/10.1017/CBO9780511615801.009

Simpson, D. (2005). Phrenology and the neurosciences: Contributions of F. J. Gall and J. G. Spurzheim. *ANZ Journal of Surgery*, *75*(6), 475–482. https://doi.org/10.1111/j.1445-2197.2005.03426.x

Skinner, B. F. (1974). About behaviourism. In *Analysis* (Vol. 18).

Staats, A. W. (1996). *Behavior and personality : psychological behaviorism*. Springer. https://books.google.co.uk/books?hl=en&lr=&id=TIyd4NBHGxAC&oi=fnd&pg=PR7&dq=behaviorism+and+personality&ots=uewluw8Y-I&sig=INi9Y_DhZaNZqj7mwfVfP5fv0s0#v=onepage&q=behaviorism and personality&f=false

Stieger, S., & Reips, U. (2010). What are participants doing while filling in an online questionnaire : A paradata collection tool and an empirical study. *Computers in Human Behavior*, *26*, 1488–1495. https://doi.org/10.1016/j.chb.2010.05.013

Stinson, J. N., Jibb, L. A., Nguyen, C., Nathan, P. C., Maloney, A. M., Dupuis, L. L., Gerstle, J. T., Hopyan, S., Alman, B. A., Strahlendorf, C., Portwine, C., & Johnston, D. L. (2015). Construct validity and reliability of a real-time multidimensional smartphone app to assess pain in children and adolescents with cancer. *PAIN*, *156*(12), 2607–2615. https://doi.org/10.1097/j.pain.0000000000000385

Streiner, D. L. (2003). Construct Validity of the Relationship Profile Test : A Self-Report Measure of Dependency-Detachment Construct Validity of the Relationship Profile Test : A Self-Report Measure of Dependency – Detachment. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001

Suen, H. K., & Lei, P. (2007). Classical versus Generalizability theory of measurement Classical and Generalizability Theories. *Educational Measurement*, *4*, 1–13.

Suen, H. K., & Lei, P. (2014). *Classical versus Generalizability theory of measurement Classical versus Generalizability theory of measurement Classical and Generalizability Theories*. *June*.

Suhr, J. A., & Tsanadis, J. (2007). Affect and personality correlates of the Iowa Gambling Task. *Personality and Individual Differences*, *43*(1), 27–36. https://doi.org/10.1016/J.PAID.2006.11.004

Sung, S. Y., & Choi, J. N. (2009). Do Big Five Personality Factors Affect Individual Creativity? the Moderating Role of Extrinsic Motivation. *Social Behavior and Personality: An International Journal*, *37*(7), 941–956. https://doi.org/10.2224/sbp.2009.37.7.941

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Taris, T., Alie, N., Loxt, J., Clark, D., & Clark, D. (2007). Reinforcement sensitivity theory at work: punishment sensitivity as a dispositional source of job-related stress. *Journals.Sagepub.Com*, *21*(7), 889–909. https://doi.org/10.1002/per.660

Tenorio Delgado, M., Arango Uribe, P., Aparicio Alonso, A., & Rosas Díaz, R. (2014). TENI: A comprehensive battery for cognitive assessment based on games and technology. *Child*

Neuropsychology, *22*(3), 276–291. https://doi.org/10.1080/09297049.2014.977241

Tett, R. P., & Guterman, H. A. (2000). Situation Trait Relevance, Trait Expression, and Cross-Situational Consistency: Testing a Principle of Trait Activation. *Journal of Research in Personality*, *34*(4), 397–423. https://doi.org/10.1006/jrpe.2000.2292

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, *44*, 703–742. https://doi.org/10.1111/j.1744-6570.1991.tb00696.x

Thornhill, A., Saunders, M., & Lewis, P. (2009). Research methods for business students. In *Essex: Pearson Education Ltd.*

Tikkinen-Piri, C., Rohunen, A., & Markkula, J. (2018). EU General Data Protection Regulation: Changes and implications for personal data collecting companies. *Computer Law and Security Review*. https://doi.org/10.1016/j.clsr.2017.05.015

Tong, T., DeGuzman, C., Chignell, M., Tierney, M. C., & Lee, J. (2016). Feasibility of Using a Game-Based Cognitive Assessment for Older Adults in Emergency Care. *Iproceedings*, *2*(1), e35. https://doi.org/10.2196/iproc.6236

Torrubia, R., Ávila, C., Moltó, J., & Caseras, X. (2001). The Sensitivity to Punishment and Sensitivity to Reward Questionnaire (SPSRQ) as a measure of Gray's anxiety and impulsivity dimensions. *Personality and Individual Differences*, *31*(6), 837–862. www.elsevier.com/locate/paid

Tripathy, J. P. (2013). Secondary Data Analysis: Ethical Issues and Challenges. *Iranian Journal of Public Health*, *42*(12), 1478–1479. http://www.ncbi.nlm.nih.gov/pubmed/26060652

Tupes, E. C., & Christal, R. C. (1958). *Stability of Personality Trait Rating Factors Obtained Under Diverse Conditions (No. WADC-TN-58-61)* (Issue May 1958). https://doi.org/10.1002/nav.3800080206

van den Eynden, V., Corti, L., Wollard, M., & Bishop, L. (2009). *Managing and sharing data; a best practice guide for researchers*. UK Data Archieve. https://doi.org/10.1016/0370-2693(94)91481-8

Van der Merwe, R. P. (2002). Psychometric testing and Human Resource Management. *SA Journal of Industrial Psychology*, *28*(2), 77–86. https://doi.org/10.4102/sajip.v28i2.52

Van Heck, G. L., Perugini, M., Caprara, G. V., & Fröger, J. (1994). The big five as tendencies in situations. *Personality and Individual Differences*, *16*(5), 715–731. https://doi.org/10.1016/0191-8869(94)90213-5

Ventura, M., Shute, V., Wright, T., & Zhao, W. (2013). An investigation of the validity of the virtual spatial navigation assessment. *Frontiers in Psychology*, *4*(DEC), 1–7. https://doi.org/10.3389/fpsyg.2013.00852

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

Vernon, P. E. (1933). THE RORSCHACH INK-BLOT TEST1. I. *British Journal of Medical*

Psychology, *13*(2), 90–118. https://doi.org/10.1111/j.2044-8341.1933.tb01094.x

Vigil-Colet, A. (2007). Impulsivity and decision making in the balloon analogue risk-taking task. *Personality and Individual Differences*, *43*(1), 37–45. https://doi.org/10.1016/j.paid.2006.11.005

Vispoel, W. P., Boo, J., & Bleiler, T. (2001). Computerized and paper-and-pencil versions of the Rosenberg self-esteem scale: A comparison of psychometric features and respondent preferences. *Educational and Psychological Measurement*, *61*(3), 461–474. https://doi.org/10.1177/00131640121971329

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2017). Applications of Generalizability Theory and Their Relations to Classical Test Theory and Structural Equation Modeling. *Psychological Methods*. https://doi.org/10.1037/met0000107

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Practical Applications of Generalizability Theory for Designing, Evaluating, and Improving Psychological Assessments. *Journal of Personality Assessment*, *100*(1), 53–67. https://doi.org/10.1080/00223891.2017.1296455

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? a meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*. https://doi.org/10.1037/0021-9010.90.1.108

Washburn, D. A. (2003). The games psychologists play (and the data they provide). *Behavior Research Methods, Instruments, and Computers*, *35*(2), 185–193. https://doi.org/10.3758/BF03202541

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*(4), 319–350. https://doi.org/10.1016/j.jrp.2004.03.001

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006a). 4 Reliability Coefficients and Generalizability Theory. *Handbook of Statistics*, *26*(July 2014), 81–124. https://doi.org/10.1016/S0169-7161(06)26004-8

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006b). Reliability Coefficients and Generalizability Theory. In *Handbook of Statistics* (Vol. 26, Issue 4, pp. 81–124). https://doi.org/10.1016/S0169-7161(06)26004-8

West, S. G., & Finch, J. F. (1996). Reliability and Validity Issues. *Handbook of Personality Psychology*, 143–164. https://doi.org/10.1016/B978-012134645-4/50007-X

Whissell, C. (2010). Emotion and the humors: scoring and classifying major characters from Shakespeare's comedies on the basis of their language. *Psychological Reports*, *106*(3), 813–831. https://doi.org/10.2466/pr0.106.3.813-831

White, T. L., Lejuez, C. W., & de Wit, H. (2008). Test-retest characteristics of the Balloon Analogue Risk Task (BART). *Experimental and Clinical Psychopharmacology*, *16*(6), 565–570. https://doi.org/10.1037/a0014083

Williams, B. (2019). The Evidence For Game-Based Assessments. *Association of Business Psychologists Annual Proceedings*.

Winsborough, D., & Chamorro-Premuzic, T. (2016). Talent Identification in the Digital World: New Talent Signals and the Future of HR Assessment. *People {&} Strategy*, *39*(2), 28–31. http://bit.ly/1lkv5gB.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*. https://doi.org/10.1207/s15326977ea1001_1

Woehr, D. (2003). The Construct-Related Validity of Assessment Center Ratings: A Review and Meta-Analysis of the Role of Methodological Factors. *Journal of Management*, *29*(2), 231–258. https://doi.org/10.1016/S0149-2063(02)00216-7

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An Examination of G-Theory Methods for Modeling Multitrait–Multimethod Data. *Organizational Research Methods*, *15*(1), 134–161. https://doi.org/10.1177/1094428111408616

Yoon, K. L., & Zinbarg, R. E. (2008). Interpreting Neutral Faces as Threatening Is a Default Mode for Socially Anxious Individuals. *Journal of Abnormal Psychology*, *117*(3), 680–685. https://doi.org/10.1037/0021-843X.117.3.680

Ziegler, M., Poropat, A., & Mell, J. (2014). Does the Length of a Questionnaire Matter? *Journal of Individual Differences*, *35*(4), 250–261. https://doi.org/10.1027/1614-0001/a000147

Zimbardo, P. (2004). A situationist perspective on the psychology of evil: Understanding how good people are transformed into perpetrators. In A. G. Miller (Ed.), *The social psychology of good and evil: Understanding our capacity for kindness and cruelty.* (pp. 21–50). The Guilford Press. https://doi.org/10.1097/01.nmd.0000207370.95835.47

Appendices

Appendix I: GBA Invitation Email

Hi [candidate name]

We're excited to invite you to take part in our brand new Game-Based Assessment on your mobile device. We are using this assessment because it provides a fair and objective way to help us understand your suitability for the role alongside other information.

Not a gamer? Don't worry, no prior gaming experience is required and it won't affect your results. This is still a fair and reliable psychometric assessment that is just presented in an easy-to-use, interactive format.

You have been asked to complete the following assessment:

Skyrise City

This is a personality assessment so there is no right or wrong way to approach it; simply follow the instructions and react to the scenarios naturally. On average, this assessment takes around 25-35 minutes to complete, but everyone is different and there is no time limit so please complete it at your own pace.

If you need us to consider any reasonable adjustments for learning difficulties, please make us aware before you start the assessment.

Here's how to play:

1. **Download**. Follow the links provided below to download the app on to an Android/iOS smartphone or tablet **manufactured within the last four years**. Detailed device requirements are listed on the relevant app stores. Please ask a friend or family member if you can borrow theirs if you do not have an appropriate device. Do not use emulators.
2. **Set-up**.
    - Find a quiet place free from distractions.
    - Make sure you have a good, stable WiFi or 4G/LTE connection and ample battery life.

    - Set your device to "do not disturb" and ensure distracting push notifications are disabled for social media and messaging apps.
3. **Log in.** On commencing the game, please enter your unique player key:

[login details]

Skyrise City: Android Download | iOS Download | FAQs

If you have any issues, please email support@arcticshores.com and include your player key, details about the make and model of your device, and the organisation you have applied to.

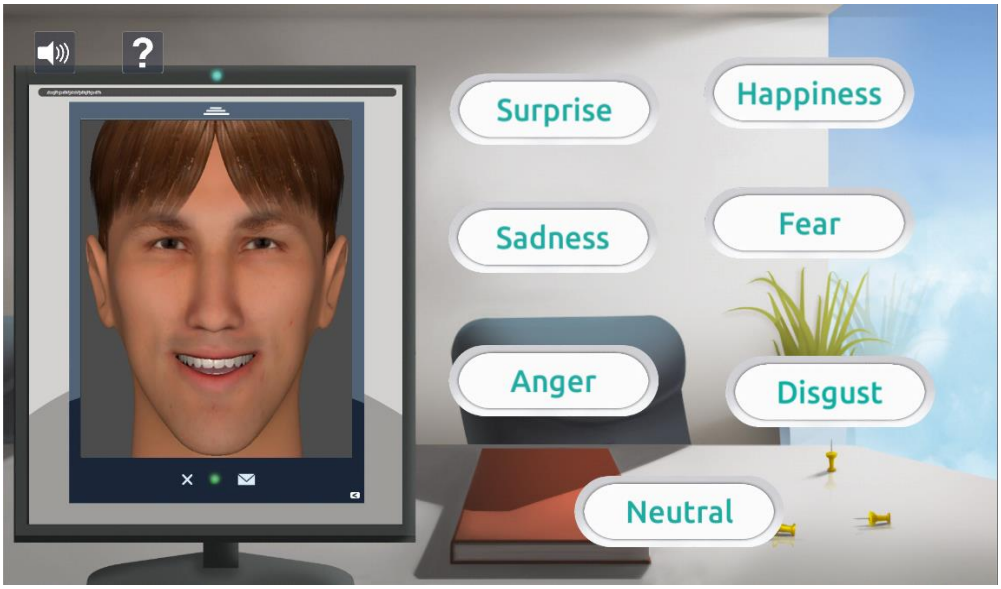
Kind regards

Arctic Shores Assessments

Appendix II: GBA Levels

Balloon Burst level

Focus Group



Power Source

Appendix III: Candidate Terms and Conditions

Terms and Conditions

Introduction

These are the terms of supply for services provided through this mobile application (Application). The Application is operated by or on behalf of Arctic Shores Limited (Arctic Shores, we, us and our). We are a limited company, registered in England. Our registered company number is 08589048, and our registered office is at Lowry House, 17 Marble Street, Manchester, M2 3AW. Our VAT number is 165182604.

Your use of any of the services offered by this Application (Services) is subject to these terms and by using any Service you agree to be bound by them. You should print a copy of these terms for future reference. We have been engaged by your employer, potential employer or recruitment company ("Requesting Business") to provide you with access to our Services. As part of this, both the Requesting Business and Arctic Shores may process your personal data. Please read our Privacy and Cookies Policy (https://www.arcticshores.com/privacy-policy) to learn how your personal data is used.

Arctic Shores provides psychometric Assessments (Assessment) to organisations who wish to recruit or develop existing or potential member of staff. We enable employers to differentiate their hiring or development process, raise brand awareness, engage prospective talent and make fair and objective people related decisions. The Assessment permits us to develop a set of personality traits of the user (you, your) through responding to psychometrically designed tasks made available through the Application.

These terms were last updated on 1st February 2019.

1. Access

Your Requesting Business may ask you to access or download the Assessment as part of selection or development process, and shall provide you with a username and password to enable you to do so.

Where you download the Assessment on to a device which does not belong to you, you must obtain their permission to download the Assessment prior to using the device.

Use of the application results in information and analytics which will be passed on to your requesting business.

2. Our Services

You acknowledge and agree that you are solely responsible for all use you make of any Service. You must not allow anyone else to complete the Assessment on your behalf. You must faithfully follow the Assessment instructions at all times and you acknowledge that any manipulation of your actions in the Assessment will render the results invalid. Further, you must not attempt to copy the Assessment, in whole or in part (including, for example, by taking screenshots of the Assessment), or attempt to distribute the Assessment, or any part of it, unless expressly allowed by these terms.

3. Changes to the Service and terms

As it is our policy continually to review and update our Service offerings, we reserve the right to make changes to any Service and/or to these terms from time to time.

4. Service suspension and termination

We may, from time to time, with or without prior notice, temporarily suspend the operation of any Service and/or the Application (in whole or in part) for repair or maintenance work or in order to update or upgrade any contents, features or functionality.

We may suspend and/or terminate any Service and/or your use of your account in the event that you have breached any of these terms.

If you have breached these terms, we may take such action as we deem appropriate. Such a breach by you may result in our taking, with or without notice, all or any of the following actions:

issue of a warning to you;

immediate, temporary or permanent withdrawal of your right to use any Service;

legal proceedings against you for reimbursement of all recoverable loss and damage resulting from the breach; and/or

disclosure of all relevant information to law enforcement authorities as we reasonably feel is necessary.

The responses described above are not limited, and we may take any other action we deem appropriate.

Upon termination of the Service or your account, for any reason:

all rights granted to you under these terms will immediately cease; and

you must promptly discontinue all use of the relevant Service.

5. Access to the Application

It is your responsibility to ensure your equipment (computer, laptop, netbook, tablet or other mobile device) meets all the necessary technical specifications to enable you to access and use the Application and is compatible with the Application.

We cannot guarantee the continuous, uninterrupted or error-free operability of the Application. There may be times when certain features, parts or content of the Application, or the entire Application, become unavailable (whether on a scheduled or unscheduled basis) or are modified, suspended or withdrawn by us, in our sole discretion, without notice to you. You agree that we will not be liable to you or to any third party for any unavailability, modification, suspension or withdrawal of the Application, or any features, parts or content of the Application.

6. What you are allowed to do

You may only use the Application for non-commercial use and only in accordance with these terms. Additional terms may also apply to certain features, parts or content of the Application and, where they apply, will be displayed on-screen or accessible via a link.

7. What you are not allowed to do

Except to the extent expressly set out in these terms, you are not allowed to:

'scrape' content or store content of the Application on a server or other storage device

connected to a network or create an electronic database by systematically downloading and storing all of the content of the Application;

remove or change any content of the Application or attempt to circumvent security or interfere with the proper working of the Application;

copy any part of the Application by way of “reverse engineering”, “disassembling” etc.

You must only use the Application and anything available from the Application for lawful purposes (complying with all applicable laws and regulations), in a responsible manner, and not in a way that might damage our name or reputation or that of any of our affiliates.

All rights granted to you under these terms will terminate immediately in the event that you are in breach of any of them.

Misuse of the application or breach of these terms may result in us notifying the Requesting Business who may take action against you with regards to your hiring or employment by them.

8. Intellectual property rights

All intellectual property rights in any content of the Application (including text, graphics, software, photographs and other images, videos, sound, trademarks and logos) are owned by us or our licensors. Except as expressly set out here, nothing in these terms gives you any rights in respect of any intellectual property owned by us or our licensors and you acknowledge that you do not acquire any ownership rights by downloading the Application. In the event you print off, copy or store information from the Application (only as permitted by these terms), you must ensure that any copyright, trade mark or other intellectual property right notices contained in the original content are reproduced.

9. Application features and content

We may change the format, features and content of the Application from time to time. You agree that your use of the Application is on an 'as is' and 'as available' basis and at your sole risk.

Whilst we try to make sure that content on the Application consisting of information of which we are the source is correct, you acknowledge that the Application may make content available which is derived from a number of sources, for which we are not responsible. In all cases, information on the Application is not intended to amount to authority or advice on which reliance should be placed. You should check with us or the relevant information source before acting on any such information.

We make or give no representation or warranty as to the accuracy, completeness, currency, correctness, reliability, integrity, quality, fitness for purpose or originality of any content of the Application and, to the fullest extent permitted by law, all implied warranties, conditions or other terms of any kind are hereby excluded and we accept no liability for any loss or damage of any kind incurred as a result of you or anyone else using the Application or relying on any of its content.

We cannot and do not guarantee that any content of the Application will be free from viruses and/or other code that may have contaminating or destructive elements. It is your responsibility to implement appropriate IT security safeguards (including anti-virus and other security checks) to satisfy your particular requirements as to the safety and reliability of content.

10. External links

The Site may, from time to time, include links to external sites. We include these to provide you with access to information that you may find useful or interesting. We are not responsible for the content of these sites or for anything provided by them and do not guarantee that they will be continuously available. The fact

that we include links to such external sites does not imply any endorsement of or association with their operators.

11. Our liability

Nothing in these terms shall limit or exclude our liability to you:

for death or personal injury caused by our negligence;

for fraudulent misrepresentation;

for breach of any term implied by the Consumer Rights Act 2015 and which, by law, may not be limited or excluded; or

for any other liability that, by law, may not be limited or excluded.

We will not be liable or responsible for any failure to perform, or delay in performance of, any of the Services that is caused by events outside our reasonable control.

12. General

You may not transfer or assign any or all of your rights or obligations.

All notices given by you to us must be given in writing to the address set out at the end of these terms. We may give notice to you by email.

If we fail to enforce any of our rights, that does not result in a waiver of that right.

If any provision of these terms is found to be unenforceable, all other provisions shall remain unaffected.

These terms may not be varied except with our express written consent.

These terms and any document expressly referred to in them represent the entire agreement between you and us. We are required by law to advise you that this contract may be concluded in the English language only and that no public filing requirements apply.

These terms shall be governed by English law, except that if you live in Scotland or Northern Ireland there may be certain mandatory applicable laws of your country which apply for your benefit and protection in addition to or instead of certain provisions of English law.

You agree that any dispute between you and us regarding these terms will only be dealt with by the English courts, except if you live in Scotland or Northern Ireland, you can choose to bring legal proceedings either in your country or in England, but if we bring legal proceedings, we may only do so in your country.

The European Online Dispute Resolution platform http://ec.europa.eu/consumers/odr/ provides information about alternative dispute resolution which may be of interest.

Contacting us

Please submit any questions you have about these terms by email to support@arcticshores.com.

Appendix VI: Collaboration Agreement

Background

At Arctic Shores, we are committed to support students wishing to use our tools and technological facilities as part of their research; since our inception in 2014, we have collaborated with several universities around the world, and we are proudly witnessing a developing body of evidence taking shape. We are interested in undertaking projects adding to the theoretical understanding of game-based assessment (GBA) as a psychometric approach, and those providing evidence of its applied use in the workplace and beyond.

There are several factors for any student to meet before we will commit to a supervision and/or collaboration:

1. Students must have a clear interest in the area and show a good degree of motivation about the project they are planning to commence.
2. A detailed research proposal must be agreed between the student, the supervisor, and Arctic Shores before any data can be collected.
3. Students will be assigned an Arctic Shores member of staff for support, depending on the nature of the project.

If the above criteria are met and approved in writing by Arctic Shores, the following conditions will apply and be accepted in writing below:

1. Surveys and online tests will be managed by Arctic Shores and automatically linked to a game based assessment (GBA).
2. All data from the GBA will be managed by Arctic Shores and shared with the students weekly or as otherwise defined in writing but no more frequently than on a weekly basis.
3. Participant support requests concerning technical issues with, and/or feedback about the GBA, app, or feedback report will be handled by Arctic Shores.
4. The student must contact Arctic Shores immediately in the event of any change to the purpose of the study.
5. The student will forward all relevant participant report requests to Arctic Shores support.
6. Support: We guarantee a response within three days. Unless otherwise agreed, communication must occur via email.
7. Other project requests: Responses to general requests will be made within a reasonable time and Arctic Shores project deadlines will be prioritised over all else.
8. The student is acting independently and in no way represents Arctic Shores or should imply any representation of Arctic Shores.
9. At the end of the study, the student must share a full report of their findings with Arctic Shores.
10. Arctic Shores requires that all students sign this agreement and the enclosed Non-Disclosure Agreement.
11. Failure to adhere to this agreement's terms and conditions and/or the NDA's terms and conditions will result in the GBA data and access to Arctic Shores information and support be withdrawn immediately.

By signing below, the students and/or supervisor agree to all the terms specified above.

Date 25/01/2021