Behaviour-Based Assessment: a proposed solution against response distortion in workplace assessment.

Lara C L Montefiori

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Faculty of Brain Sciences

University College London

February 16, 2025

I, Lara C L Montefiori, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Assessment faking, a volitional behaviour in which candidate engage to portray themselves more desirably to prospective employers, is a serious challenge that companies face in the context of candidate selection but which has also far-reaching consequences related to job performance and deviance.

Several theoretical models have been proposed to explain how and why faking emerges, and they have broadly identified three factors responsible for this behaviour: intention/motivation, ability, and opportunity to fake. While motivation should not be expected to radically change across assessment formats, it should be possible to target ability and opportunity to fake through assessment development strategies aimed at impeding faking by design.

This thesis explores how a relatively new assessment method, designed to elicit and observe behavioural differences stemming from the neurobiological mechanisms underlying personality traits, may offer a solution to this issue.

Five studies, featuring a mix of research designs and samples, provide strong evidence that it is not possible to fake Behaviour-Based Assessment. This thesis reports that intentions to fake BBAs are affected by low perceptions of control over these assessments and that people lack the ability to articulate strategies to fake them; it also provides evidence of failed attempts to achieve desired scores through faking in experimental setting. However, the most critical contributions of this thesis were to explain why the response options in BBAs do not lend themselves to faking, and to reveal that, contrary to what happens with other test formats, the data of real-life candidates show markers of optimal but not inflated performance.

Implication for practice in selection and future directions are discussed.

Acknowledgements

Thanks to all the people involved in this journey. To those who paid for it and to those who gained from it. To those who supported me and to those who exhausted me. To those who made me stronger with their love and to those who made me stronger with their other feelings. To those who believed in me and to those who had doubts. To those who cared and to those who didn't. To those who will read this thesis and to those who will never do that. To all the amazing people in my life, each and every one of them. And finally to me. The one who absolutely did not need to do this. The one who did not had the time to do it. Yet the one who ultimately did it.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Faking is a volitional behaviour [whereby] individuals knowingly choose to answer [...] questions in a manner that provides an inaccurate characterisation [of themselves]"

(Ellingson and McFarland 2011; p.323)

*****

Understanding and reducing faking is important for organisation using assessments for candidate selection, which are estimated to be 70% globally (Church & Rotolo, 2013), because not only the resulting assessment scores are biased and misrepresent the individual's real psychological profile (Griffith, Malm, English, Yoshita, & Gujar, 2006; Ziegler & Buehner,2009), they also interfere with hiring decisions by overly inflating the suitability of test-fakers at the detriment of better suited honest candidates (Berry & Sackett, 2009; Griffith, Chmielowski, & Yoshita, 2007; Peterson, Griffith, & Converse, 2009). This means that when faking is not policed, the validity of selection measures could be heavily reduced (Gilmore, Stevens, Harrell-Cook, & Ferris, 1999; Marcus, 2006; Rosse, Stecher, Miller, & Levin, 1998), and there is also a high risk of hiring people more likely to engage in counteractive workplace behaviours as assessments are not able to detect undesirable traits properly, but also because fakers tend to possess specific combinations of traits linked to deviance (i.e.: low morality and high risk propensity) that are not commonly assessed in candidates (O'Neil, Lee, Radan, Law, Lewis, and Carswell, 2013).

Some types of assessment are more susceptible than others to faking. Self-reported personality measures are type of psychological assessment typically consists in short series of statements such as "Worries get in the way of my success" that are rated for how well they apply to the respondent. Veracity of the ratings is usually unverifiable, and nevertheless contingent to respondents' willingness and/or ability to provide accurate responses – "neither of which can be assumed when there is a desirable outcome at stake" (Morgeson, Campion, Dipboye, Hellenback, Murphy & Schmitt, 2007, p. 685), as in the case of candidate assessment.

In fact, it is feared that results from those tests may carry less information about the constructs under scrutiny, and more about the psychological, sociological, linguistic, experiential and contextual factors surrounding the assessment event (e.g. Harrison, McLaughlin & Coalter, 1996; Lanyon & Goodstein, 1997; Morgeson et al., 2007). Despite this shortcoming, self-reported measures are widely employed in Talent Acquisition practices because their ease of administration generally offers a reasonable alternative to more reliable yet costly personality assessment methods (Chamorro-Premuzic, 2018).

The extent to which personality questionnaires may be susceptible to response distortion, and especially the extent to which this may affect their validity, has been the focus of a contentious debate ever since self-reported inventories were introduced in personality assessment (e.g.: Kelly, Miles & Terman, 1936; Murphy & Dzieweczynski, 2005; cf. Hogan, Barrett & Hogan, 2007). Whilst complete agreement is yet to be achieved, most evidence suggests that social desirability response bias (i.e.: positive responding that does not reflect true disposition; Edwards, 1953) is very common in self-reported inventories (e.g.: Morgeson et al., 2007). Recent evidence shows that the prevalence of faking should be a major concern as studies have found that up to 49% of applicants cheat on personality tests in real-life selection processes, and up to 81% admit to lying on job applications (Griffith & Converse, 2012; Weiss & Feldman, 2006). Furthermore, it has been demonstrated that even a small amount of faking (as little as 2% of respondents) can significantly impact hiring decisions (Holden, 2006). Consequently, organisations need

to develop hiring assessments that are resistant to applicant faking, as the cost of ineffective hiring decisions can be as high as 1.4 times an employee's annual salary (Fell & König, 2016).

Two types of bias have regularly been observed: Impressions Management, that is the voluntary, purposeful, and deceitful response distortion in which respondents engage when attempting to enhance themselves; and Self-Deception, that is the involuntary, accidental, and ingenuous counterpart caused by a lack of self-insight and the tendency to endorse descriptions of an idealised self (Paulhus, 1884; 1988; 1989; Paulhus & Reid, 1991; Ones, Viswesvaran & Reiss, 1996)). This thesis focuses on the former. Impression management has been detected to a higher extent in candidates than in employees (Barrick and Mount, 1996), suggesting that benefits conditional to test performance increase self-enhancement motivation. In fact, research has demonstrated that high stake situations, especially when coupled with certain personality characteristics, predictably results in social desirability response bias (e.g.: Karren & Zacharias, 2007; Morgeson et al 2007), and that response patterns systematically change to match the personality profiles required by the context in which the assessment is taking place (e.g.: Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Mueller-Hanson, Heggestad & Thornton, 2006).

Response distortion is condoned by some researchers who argue that test faking is per se an important individual difference, and suggest that impression management entails endorsing positive communal values and, for this, signals desirable psychological factors such as Agreeableness and Social Adjustment (Hogan, 2005; Hogan & Chamorro-Premuzic, 2012) or Emotional Intelligence (Pelt, van der Linden, & Born, 2018). Hogan and colleagues (e.g., Johnson & Hogan, 2006) applied socio-analytic theory to self-presentation, which offered a new perspective on the relationship between personality and faking. According to socio-analytic theory, personality and self-presentation are almost synonymous, as individuals convey a certain image of themselves in order to achieve getting ahead and getting along with others (Bakan, 1966). While this image may vary across situations, individuals have the tendency to strive for consistency, and their behaviour across situations

is influenced by their relatively stable self-image. Assessments can be conceptualised as one of many occasions in which individuals can showcase their desired self-image to others. Also, ability to fake could lead to better job performance since an individual who can fake being competent on an assessment might also be able to fake being competent in real-life situations (Chamorro-Premuzic & Furnham, 2010), and this is particularly relevant for jobs involving a high levels of social interaction such as customer service (Morgeson et al., 2007). Under this premises, the validity of personality tests in selection settings is maintained to the extent to which "the factors that facilitate skilful self-presentation in everyday life might also be applied to competent self-presentation on personality tests" (Johnson & Hogan, 2006; p. 217).

Those not against faking also believe that the main goal of personality assessment in the workplace is to predict organisational outcomes (e.g.: Schmidt & Hunter, 1998), and that this has long been achieved despite social desirability response bias (Hogan & Chamorro-Premuzic, 2012). Furthermore, an interesting study conceptualised the benefits of faking into a new construct and demonstrated how this predicted job performance above other measures (Marcus, Goldenberg, Fine, Hummert, & Traum, 2019). The authors argue that not only faking doesn't impact the diagnostics value of selection measures, but it can also offer important insights on an individual potential to successfully adapt their behaviour to situations. In their study, Marcus et al describe the development and validation of the Ideal Employee Coefficient (IEC) whereby, for any scale, it is possible to compute a IEC score through an omnibus index of similarity between expert ratings of the ideal responses of an assessment, and the candidate's responses of the same items. The rationale behind this index is that if a candidate can simulate the same profile that an expert deems ideal, this means that they know what behaviours are expected of them, and they can use this ability in the workplace.

There is a fundamental assumptions on which this types of rhetoric stands on, which make the arguments shaky. The ability to identify the right responses in an assessment do not automatically warrant the ability or predict the likelihood of

consistently exhibiting the same behaviours at work. Much of the variance in job performance explained by faking might just be accounted by general mental ability, which can also be expected to predict the ability to connect behaviour, traits, and situations (e.g.: Christiansen, Wolcott-Burnall, Janovics, Burns, and Quirk, 2005). Furthermore, most of the evidence suggests that faking leads to worse hiring decisions (e.g.: Donovan & Dwight 2011), which is not surprising because if candidates need to lie to appear more suitable for a role, and lying equates to suitability, then the only logical deduction is that they are not suitable for the role and they should not be hired.

Nevertheless, this thesis is based on the assumption that faking should be avoided, and that an assessment format able to prevent it from occurring should be a welcome addition to the field of Psychometrics. The following chapters will provide an overview of the current models of faking, and describe five experimental studies mapped on those models that investigate whether a specific type of psychometric assessment is able to prevent faking by impeding its emergence from several individual and situational angles.

Chapter 2

Literature Review

Faking in assessment is a widespread phenomenon which posits a real challenge for talent acquisition. Many researchers have attempted to understand and explain it, and many other have proposed solutions aimed at either preventing or identifying the occurrence of this behaviour, especially in the context of candidate assessment.

In order to fully understand how to prevent faking, it is necessary to understand how faking emerges and develops. A clear overview of the mechanisms underlying the faking process and of the interfaces at which situation-level differences may have an impact on faking would enable preventative interventions to be targeted to the right places. Furthermore, understanding how people fake would support the development of methods to identify faking.

2.1 Theoretical Models of Assessment Faking

Faking is not a new concept and it dates back to Aristotle's times, when the Philosopher pioneered the methodological study of *akrasia* - lack of self control - in the Seventh book of Nicomachean Ethics, attempting to provide a plausible rationale to explain why people act against what they know is best.

Several models and theories have being used to describe faking and explain its underlying processes; earliest attempts focused almost entirely on the individual, whereas more recent frameworks have added a multitude of situational factors and described more complex dynamic interactions between the individual, the situation, and the assessment itself.

In its most simplistic form, faking behaviour can be explained in terms of Rational Choice Theory (RCT), according to which people calculate the probabilistic overview of the costs, risks, and benefits associated with an action to evaluate whether to initiate it or not. The rationale underlying the RCT originated in an essay written by the neoclassic philosopher and economist Adam Smith's , *"An Inquiry into the Nature and Causes of the Wealth of Nations"* (Smith, 1776), but the theory was not adopted until the 50s and most importantly, not without controversies (Boudon, 2007). Nevertheless, the RCT would suggest that individuals fake on workplace assessment purely because they perceive it to be a rational decision that has the potential to increase their chances of being selected for a job, and the value of this potential outweighs the cost of producing the faking behaviour and the risks associated with it. Explaining faking in terms of the RCT is reductionist at best. First, it fails to take into account the fact that faking behaviour can also have negative consequences for the individual and the organisation, such as reduced job satisfaction, turnover, and decreased productivity (reference?); and second, because its laser focus on the individual's choice grossly underestimates the influence of factors external to the individual beyond the scope of the individual's interest and perception.

The self-presentation theory (Goffman, 1959) and Social desirability bias model (Edwards, 1957) emerged around the same time, building on an already established perspective of the person as a "performer" (e.g.: Burke, 1969), but complementing it with a layer of goal-directness whereby individuals are able to carefully and strategically curate the perceptions others develop of them by manipulating their behaviour to impersonate a contextually desirable profile. The importance of these early theoretical contributions is that they are heavily saturated with the social interactionist underpinnings of behavioural manipulation, yet, whilst describing the occurrence and intended outcomes of faking, they are severely limited in their ability to *explain* faking with sufficient preventative utility in that they don't take into account the cognitive and psychological processes that underlie the initiation and execution of faking behaviours, nor they incorporate the situational factors

that moderate them.

Subsequent theoretical models attempted to explain faking through an interaction-based narrative by incorporating individual and situational variables and their interplay as antecedents, mediators, and moderators of faking behaviours in their frameworks.

Snell, Sydell, and Lueke's (1999) Iteractional Model of Faking identifies two direct correlates of faking behaviour - motivation/intentions, and ability - each with several underlying antecedents. It postulates that dispositional factors such as general mental ability and emotional intelligence, experiential factors such as familiarity with specific job characteristics and related constructs, and test characteristics such as item type, response options, and scoring strategies dictate the level of ability to fake assessments. The model also proposed that motivation to fake is dependent on three main categories of individual differences: demographics variables such as age and gender, individual traits related to impression management such as integrity and machiavellianism, manipulativeness, organisational delinquency, locus of control and stage of moral development, and perceptual factors such as others' behaviour and attitudes, fairness, expectation of success, and importance of outcome. While Snell et al. (1999) provides a much more comprehensive theoretical explanation of why faking behaviour emerges than its predecessors, it falls short of actionable insights due to a heavy reliance on dispositional factors, and a questionable overemphasising of immutable demographic variables such as age and gender. Furthermore, situational factors are only modestly featured in the model, in fact only some test characteristics are included, and very little is speculated about the interaction of all the correlates of faking behaviour, making the model considerably static. These limitations highlight the need for a more comprehensive and truly integrative model of faking behaviour, which takes into account both individual and contextual factors, as well as their potential interactions.

McFarland and Ryan (2000) developed a model of applicant faking behaviour according to which individuals' beliefs towards faking underlie intentions to fake, but it is the ability to fake an assessment that dictates whether intentions to fake

successfully translate into faking behaviour. The initial model was offered as a heuristic to explain discrepancies in the faking literature whereby evidence around assessment faking was conflicting and discordant, and it offers a dynamic overview of how a range of individual differences underlie a person's beliefs towards faking, and how this is interacts with situational influences, such as the desire for the job and the likelihood of being caught faking, to predict intentions to fake. At this point, the model introduces two moderators: the first is the ability to fake, which includes the individual ability to successfully and purposefully manipulate their behaviour (i.e.: self-monitoring), the knowledge of the construct being measured, and the degree of item transparency (i.e.: the degree to which is possible to infer what an item is measuring); whereas the second moderator is the opportunity to fake, that is, the scope that the discrepancy between a person true score and target score offers for faking.

In 2006, the authors incorporated the Theory of Planned Behaviour (TPB; Ajzen, 1991) into the model, expanding the range of factors influencing faking behaviour. In the Integrated Model of Faking (McFarland & Ryan, 2006) the TPB components - attitudes, subjective norms, and the Perceived Behavioural Control (PBC) towards a behaviour - are adapted to the context of faking, substituting the earlier model's antecedents of intentions.

The most significant innovation of this model was the inclusion of PBC, that is, the extent to which an individual perceives that they have control over performing a particular behaviour. In the context of faking, PBC is the individual's perception of their ability to present themselves in a more favourable light on the basis of how easy it is to fake the the assessment. The reason why this constitutes a radical shift from earlier models is that the construct of PBC itself encompasses a dynamic interplay between the individual and the assessment; that is, PBC is the perceptual byproduct of the interface between the individual, the assessment, and the situation, and this adds a new dynamic twist to the model.

The Integrated Model of Faking outshines earlier models with a marked increase of utility; by leveraging the great empirical support of the TPB (Ajzen, 1991;

Ajzen & Madden, 1986; Reinecke, Schmidt,& Ajzen, 1996) it grounds itself as a stable framework upon which the influence of each of the multiple factors driving faking behaviour can be isolated from the others. And this allows targeted preventative action to be focused on the intersection at which the progressive flow of interactions leading to faking behaviour can be stopped.

More recently, Ellingson and McFarland's (2011) proposed a more parsimonious, yet inclusive, model of faking behaviour based on the Valence-Instrumentality-Expectancy (VIE) theory of behaviour motivation (Vroom, 1964) which narrows down the determinants of behaviour to three factors. The VIE theory postulates that individuals make behavioural choices on the basis of the valence, instrumentality and expectancy that are associated with the behaviour and the resulting outcomes. Valence refers to the perceived importance of the outcome, whereas instrumentality refer to the degree of causal relatedness that the behaviour has with the outcome, and expectancy refers to the individual's perceived ability to perform the behaviour in the way in which is associated with the intended outcome.

The VIE theory can easily and intuitively be adapted to the context of assessment faking, whereby an individual's valence for presenting themselves in a particular way may be influenced by the potential benefits of being hired, while the instrumentality of faking may be influenced by their perception of the importance of specific personal characteristics for the job, and the individual's expectancy that they can successfully present themselves in the desired way may be influenced by their belief in their ability to convincingly simulate the desired personal characteristics through the assessment.

Despite a debate on the multiplicative vs additive nature of the model in the original theory, in the context of assessment faking, the authors treat the model as multiplicative - that is, if any of the three factors is too close to zero, then the behaviour is unlikely to occur (see Valois et al, 2016 for evidence on the psychometric properties of PBC multiplicative scale). This is an important factor to consider in the context of preventing faking behaviours, as it implies that it is possible to avert the risk of faking by addressing even just one of the three determinants of motiva-

tion.

Similarly to the Integrated Model of Faking, the model stresses that motivation to fake must interact with the *objective* ability to produce the required faking behaviour. Vroom (1964) also points that, when considering antecedent of behaviour, it is sometimes difficult to disentangle ability and motivation. For example, knowing what assessment scores match the selection criteria for a job can facilitate behaviour by increasing expectancy, and with that motivation, but also by increasing the objective ability to produce the correct faking behaviour. Furthermore, in the model, ability is divided between two distinct sources: the individual and the situation.

Namely, expectancy occurs both at the individual and situational level. Candidates must believe in their ability to fake but also encounter a situation that would facilitate the intended faking behaviour in order to develop high expectancy beliefs. From an individual perspective, knowing how to fake, that is, to have a clear strategy pertaining a successful faking behaviour is an key determinant of expectancy judgements. Similarly, the objective ability to fake also resides within the individual and the situation in parallel. At the individual level, self-monitoring and intelligence have been consistently demonstrated to facilitate faking behaviour (Tett, Freund, Christiansen, Fox, and Coaster, 2012) for equipping the individual with the skills necessary to fake assessments; whereas at the situational level, the objective ability to fake is determined by the actual scope that an assessment offers for faking. That is, to warrant objective ability to fake, the format of the assessment should not prevent faking by design, and it shouldn't have in-built mechanisms to detect faking which would undermine the effectiveness of faking behaviour.

Furthermore, expectancy perceptions are determined by situational factors of which individuals are aware, whereas objective ability is determined by situational factors of which they are not aware. For example, assessment features and formats that leave little or no room for faking weaken the perception of expectancy, and so are faking warnings. Both reduce expectancy by hampering the perceived likelihood of successful faking, and this results in lower motivation to fake. On the other hand, situational factors affecting the objective ability to fake must exist beyond

the awareness of the candidate or they would have to be considered determinant of expectancy, and they act by invalidating the faking attempt by either not allowing faking behaviour to occur or by detecting it.

This aspect is what differentiates the VIE-based model of faking and the Integrated Model of Faking the most. Whilst the two models fit in with each other without major contradictions, the concept of PBC is split into individual and situational factors. This provides a better framework from which to isolate the contribution of assessment format from all the factors surrounding faking behaviours, and, in addition to this, the assessment format itself is able to target both expectancy and objective ability, yet through different features, meaning that different characteristics of the assessment can be targeted to prevent faking individually and independently from the others.

2.1.1 Ability to Fake

As explained above, in the last 25 years, several models of applicant faking have been proposed, each of which emphasising slightly different factors influencing faking behaviour. One factor that has incrementally emerged as important in these models is the individual's ability to fake, both in terms of their perceived ability and their actual, objective ability. Interventions aimed at reducing faking behaviour need to target both by removing the ostensible and actual scope for manipulation from the format through purposeful assessment design. Highly complex, knowledge-free, ability-based assessments should successfully signal less susceptibility to distortion and also lend themselves a lot less to manipulation.

As the individual and situation-level motivational factors surrounding the emergence of faking behaviours are unpredictable in their occurrence and idiosyncratic interplay, and because motivation can legitimately expected to be consistently high in situations like job applications, the choice of assessment format is the only real option through which employers can exert enough control to successfully thwart the influences that drive faking.

In addition to the situational factors featured in the models of faking, the recent emergence of highly developed and user-friendly open source AI software based on

natural language processing (NLP) is opening a new avenue for assessment faking, granting assessment formats with the least reliance on language an even more striking advantage over language-based assessments. Assessments such as self-reported questionnaire, situational judgement tests, and similar types, are much more at risk of being perceived as easier to fake and are also considerably more at risk of becoming *objectively* easier to fake, meaning that both driving and enabling forces are at play to facilitate faking like never before.

Whether AI-based models are all that skilful at simulating the language-based response patterns of successful candidate, and whether candidates are actually able to take advantage of their full potential is a very open question. Yet, this is the kind of open questions likely become obsolete rather swiftly, and the advent of accessible NLP-based AIs might be already seriously impacting the validity and utility of language-based assessments, that is, most assessment on the market.

Assessments not based on language, such as game-based assessment, or behaviour based assessment are by their very nature well equipped to withstand this challenge. This is because their underlying scientific rationale, format, interface, and response options do not rely on language by design and this makes them immune to NLP-based AIs as a consequence. As per above, the below may soon become an obsolete point as other types of AI are bound to emerge which might change this completely. Yet, for now, we are still rather comfortably far from mass-marketed user-optimised AI software able to:

1. identify complex cross-task variable-level behavioural correlates of individual differences at the trait level,

2. decipher their relationships and interactions within multi-layered machine learning scoring models,

3. account for the overarching adaptive nature of behavioural assessment implying a very large number of combinations of behaviours, reaction times, response outcomes, and task variations,

4. define the optimal 30/40 minutes of time-based behavioural performance to

generate all the possible combinations of trait scores the assessment supports,

5. map those scores on job descriptions containing unpredictable and inconsistent quantity and quality of language-based information with no discernible coherence across them with regards to the semantic association between descriptive words and the intended behavioural and/or dispositional outcomes they aim to represent, completely lacking stable sector-specific contextualisation and realistic insights of the cultural landscape in which the role is set,

6. define the exact 30/40 minutes behavioural sequence to produce the required set of scores according to those job descriptions,

7. enable the user to faithfully and painstakingly reproduce the same behavioural pattern, in real time from their own devices,

8. and do so without leaving the same exact blueprint for each user yet guaranteeing a competitive advantage to any given user

Nevertheless, despite and regardless the ongoing threat under which most of the talent acquisition industry is currently heading, this thesis was designed to evidence the recommendation of GBAs (or Behaviour-Based Assessment, BBA, as will become clearer later) as a solution against response distortion in workplace assessment. This is on the basis that this type of assessment format represents a huge obstacle to candidates' objective ability to fake, and is expected to prevent motivation, no matter how strong, from acting as catalyst to faking behaviour.

2.2 Game-Based Assessment

A game-based assessment (GBA) is a relatively new type of assessment in which "gameplay" data is used to compute psychometric scores. GBAs are the byproduct of a much bigger trend advocating for the use of concepts traditionally associated with gaming outside the immediate context of gaming - that is, "gamification". Gamification is an umbrella term that is agnostic to the field to which is applied; in assessment context, gamification broadly refers to the use of game elements within the selection process and/or the assessments themselves.

2.2.1 Gamification

The application of Gamification to selection assessment fits under the broader concept of Organisational Gamification, which seeks to reinvent job tasks and company processes throughout the workplace, blurring the boundaries between work and play (Chou, 2015). However, the push to gamify a variety of tools and programs using game design principles has not always produced successful interventions, and has oftentimes left empirical research playing catch-up to determine the proper guidelines for Gamification (Landers, 2018). This issue is compounded by a lack of consensus surrounding the definition of Gamification. Deterding, Sicart, Nacke, O'Hara, & Dixon provide perhaps the most widely cited definition, stating that Gamification is "the use of game design elements in nongame contexts" (2011, p. 1). The authors specified that it was kept vague intentionally, to avoid limiting "gamification to specific usage contexts, purposes, or scenarios" (Deterding, Khaled, et al., 2011, p. 3). Thus, while Deterding, Sicart et al.'s (2011) definition of gamification is appropriate as an umbrella term for the field, due to its ambiguity, this thesis will adopt a more specific definition of Gamification in selection and assessment contexts. This will help provide a better understanding of the core mechanisms of Gamification for selection.

Gamification in selection involves adding game elements (e.g. badges, points or storylines) throughout the testing process in order to increase its attractiveness and ease of use, thereby increasing the engagement and motivation of the individuals completing the assessment. The implementation of game elements looks to harness the sense of 'fun' present during gameplay (Malone, 1980) and the sense of accomplishment and autonomy that games can provide (Ryan et al., 2006), in order to place applicants into a similar state of mind to the one they have when playing actual games. Gamification of selection processes can be versatile. It does not necessarily seek to gamify the psychological assessment itself, and can range from enhancing the appeal of recruitment campaigns to making onboarding activities more engaging. Furthermore, although Gamification can involve a large variety of different game elements, a common selling point has been that one can apply as

many or as few game elements as one wants, which can facilitate its implementation.

The Gamification literature now contains a range of taxonomies for game elements, categorising them based on a number of different criteria (Bedwell et al., 2012; Deterding, Dixon, et al., 2011; Fetzer, McNamara, et al., 2017; Jia et al., 2016; Robinson and Bellotti, 2013). While a full classification of game elements is beyond the scope of this thesis, several key features are worth noting to contextualise current studies of Gamification. The three most common game elements examined in the Gamification literature are points, badges and leaderboards, which is primarily due to their ease of use (Hamari et al., 2014). However, games can be deconstructed into a wide range of other elements, which vary significantly in both their complexity and the potential outcomes that they can generate. They can be specific features designed to induce a sense of competition or convey a sense of achievement (Hamari et al., 2014), but they can also refer to more intangible aspects of games – such as the level of uncertainty present in the game's responses to players' actions (Fetzer, Mcnamara, et al., 2017).

2.2.2 Theoretical Models Relevant to Gamification

Three of the main theoretical frameworks which have been proposed to try and explain the psychological underpinnings of Gamification: Csikszentmihalyi's theory of 'Flow' (Csikszentmihalyi, 1990); Need satisfaction theories, of which the most notable is Self-determination theory (Deci & Ryan, 2000; Ferrell, Carpenter, Vaughn, Dudley, & Goodman, 2016); and thirdly, Goal-setting theory (Locke, 1968).

Flow theory hypothesises that an individual enters a condition called 'flow' when they work on a task that has a level of complexity and difficulty which is appropriate for their skillset (Csikszentmihalyi, 1990). When this flow state is achieved, the task elicits a state of interest, concentration, and enjoyment for the individual (Csikszentmihalyi, 1990). Csikszentmihalyi (1990) outlines certain elements which are deemed crucial for achieving and maintaining a flow state, including clear goals, feedback and the possibility for control. These general 'flow ele-

ments' have been integrated into a model of flow adapted specifically for "computer-mediated environments" (Kiili, 2005), which has demonstrated promising correlations with the experience of flow states (Kiili, 2006). The fact that these 'flow elements' have overlapped significantly with the game elements identified by Hamari et al. (2014) in their literature review of Gamification also suggests that there is a direct alignment between Gamification and the enhancement of flow. One can therefore hypothesize that Gamification can facilitate key features of flow states, which in turn promotes concentration and enjoyment.

On the other hand, Need satisfaction theories draw upon Maslow's Hierarchy of Needs (Ferrell et al. 2016), and make the assumption that activities which satisfy fundamental psychological needs are intrinsically motivating. Self-determination theory outlines three of these fundamental needs: the need for competence, the need for autonomy, and the need for relatedness (Deci and Ryan, 2000). They have been linked to game elements in several ways. For example, elements such as points or badges are hypothesized to reinforce our sense of competence by providing feedback on the success of in-game actions and tracking user progress (McDaniel and Fanfarelli, 2016; Sailer et al., 2013, 2017). The need for autonomy is thought to be addressed through game elements that provide people with choices and flexibility – such as branching gameplay options or customisable avatars (Ferrell et al., 2016; Sailer et al., 2017). Gamification is also thought to strengthen our sense of relatedness, with the social aspect of games such as badges, and message-boards helping to foster a sense of social connectedness (Andrade, Mizoguchi, & Isotani, 2016; Ferrell et al., 2016).

However, the experimental evidence for this theory is mixed. Mekler, Brühlmann, Opwis and Tuch (2013) found that the individual application of either points, badges or leaderboards in an image annotation task led to a significant increase in task performance compared to a control group, but did not lead to a significant difference in intrinsic motivation or task enjoyment between groups. In this study game elements appeared to act solely as extrinsic motivators, providing feedback on participant performance without affecting intrinsic motivation. The au-

thors replicated these results in a follow up study, finding that while points, badges and leaderboards significantly increased users' task related output, there was no significant difference between the gamified and control condition on measures of intrinsic motivation and competence satisfaction (Mekler, Brühlmann, Tuch, and Opwis, 2015).

Mekler et al.'s (2015) findings are contradicted by the work of Sailer and colleagues (2017), who investigated the effects of Gamification on an order-picking task in a virtual environment. The task was gamified using two different sets of game elements, with need satisfaction levels evaluated for participants in each condition. The first gamified condition implemented badges, leaderboards and performance graphs (which evaluate a player's performance over time), and was found to significantly improve ratings of competence need satisfaction compared to a control condition. The second gamified condition, containing avatars, meaningful stories and teammates, significantly improved social relatedness need satisfaction compared to a control condition. However, including several game elements in each condition limits the study in its ability to isolate the effect of each individual game element. This is also a potential reason for Sailer et al. (2017) observing a significant change in competence need satisfaction, while Mekler et al. (2015) did not, where Sailer employed multiple Gamification elements targeting users need for competence as opposed to testing each variable individually. Employing multiple elements simultaneously could have had the benefit of creating a more engaging Gamification experience. If this does play a crucial role, it would suggest that a more immersive Gamification environment is needed to replicate the effect that games have on intrinsic motivation and wellbeing.

Finally, Goal setting theory outlines the relationship between a person's goals and their actions. It theorises that goals direct a person's effort towards goal-related activities and away from goal-unrelated activities (Locke et al., 1981). Furthermore, it hypothesises that setting goals increases persistence when performing goal related actions and can lead to the implementation of goal-related knowledge and task-related strategies (Locke & Latham, 2002). There are several key factors me-

diating the relationship between goal setting and an individual's actions, including goal difficulty and a person's commitment to the goal (Landers, Bauer, & Callan, 2017; Locke & Latham, 1990). Goal difficulty has been shown to be linearly proportional to goal-related effort (Locke & Latham, 1990; Locke & Latham, 2006), and goal commitment becomes increasingly influential for success as goal difficulty increases (Klein et al., 1999; Latham and Locke, 1991). The implementation of points, badges, leaderboards and other Gamification benchmarks can integrate specific goals into workplace activities. Gamification can also help by providing goal-related feedback, and by maintaining attention and interest they can enhance goal commitment, improving goal related efficacy (Ferrell et al. 2016; Landers, Bauer, Callan, & Armstrong, 2014).

Landers, Bauer and Callan (2017) provided evidence for the applicability of goal setting theory to Gamification by gamifying a task using points and leaderboards. They found that the participants in a points-based leaderboard condition performed significantly better on an idea generation task than both an easy goal condition and a 'do-your-best' goal condition. Performance in the gamified condition was on par with groups who were told to achieve difficult or impossible goals. Interestingly, points and leaderboards did not actually provide specific goals, instead they seem to give users a scale upon which to set their own goals (i.e. an amount of points that they want to obtain, or a position they want to reach on a leaderboard). Goals are thus set with a certain level of autonomy – albeit with influence from the user's environment. This could be an explanation for why Sailer et al. (2017) also found that a gamified condition with badges, leaderboards and performance graphs significantly improved task meaningfulness. Being able to set one's own goals in the competitive context provided by a leaderboard could motivate people to perform well without diminishing their sense of autonomy and the contextualisation of their scores could increase their sense of relatedness. However, although the results from Landers et al. (2017) suggest that people naturally set goals at the top of the leaderboard, they did see that participants' goal commitment strongly mediated performance in the leaderboard condition. Goal commitment has been shown to vary

from person to person, with some people being satisfied with gaining a place on the leaderboard, while others aim for the top (Hamari et al., 2014). As such, it seems plausible that the effect of game elements can vary depending on the user and context. If some participants find a feature more motivating than others, then it adds bias to the assessment and lowers the validity of the measure, which is especially problematic in high stakes assessment contexts. As will be seen in sections 3.1 and 7, one way this variability might occur is through personality differences. This variability in individuals' responses to game elements is a useful tool for measuring personality constructs, a factor often measured in selection (Zibarras and Woods, 2010).

2.2.3 Gamification and Psychometrics

Within the broader context of Gamification within assessment and selection, there are some important definitions to be made. Landers and Sanchez (2022) provide a clear description of three key concepts in this space:

- *Gameful design*, which is an assessment design strategy that combines psychometrics and game mechanics and concepts to develop novel assessments

- *Assessment gamification*, which refers to the practice of redesigning existing assessments by adding game elements without changing much about the assessment itself. This has also been called *framification* (Collmus & Landers, 2019) or *storyfication* (Landers & Collmus, 2021) to convey the idea that *gamified assessment*s are still the same assessment they were before the gamification, and they are just "framed" as a game, and sometimes this is achieved by means of embedding them into a storyline.

- *Game-Based Assessment*, which is an assessment methods in which candidates actually have to "play a game" and their psychometric profile is computed from data generated through the core *gameplay loop*. This is very unique in that the psychometric assessment occurs by means of tasks not traditionally used for this purpose, and the development of GBA is not different

> from the development of any other entertainment game - with the exception of being designed with the sole purpose of measuring individual differences.

According to Landers and Sanchez (2022), and equally to the evidence available in this field and the existing candidate selection assessments commercially available, there are two types of psychometric assessments that can be nested under gamification to date. Those are Gamified Assessments, and Game-Based Assessment.

2.2.3.1 Gamified Assessment

The term Gamified Assessment refers to a traditional measure that has been altered to appear more like a game. Those solutions maintain the psychometric properties of the existing assessment but also add game elements around it. The simplest way in which the contextual effect of Gamification has been studied is through 'shallow gamification', in which a task is framed as a game without actually adding any game elements to it (Lieberoth, 2015). This has been shown to increase participant ratings of interest and enjoyment when completing a tedious task (Lieberoth, 2015). This framing effect has also been seen to reduce candidates' perception of test length when applied to a cognitive assessment (Collmus & Landers, 2017). Thus, gamifying assessments could allow them to be made longer and more detailed without causing negative applicant reactions or increasing dropout rates. 'Shallow gamification' has limitations and it cannot offset the negative impact of failing a task on motivation and enjoyment (Brühlmann, 2016).

Greater benefits can be seen when actual game mechanics are implemented around assessments. For example, when applied to a mathematics assessment, points were found to increase the speed of responses as well as participants' enjoyment of the assessment (Attali and Arieli-Attali, 2015). This effect was significant for both adults and children, and gender was not found to have a significant impact. However, the implementation of game elements is not always so clean cut. Participant gender was found to mediate the effect that leaderboards had on assessment performance (Christy and Fox, 2014). Christy and Fox found that females performed worse when exposed to a female dominated leaderboard versus a

predominantly male leaderboard, which was thought to be because of the negative effects of comparisons with higher achieving females. Game elements can also have different motivational effects on candidates. Two studies have shown that more extraverted individuals were more likely to enjoy using points, levels and leaderboards (Jia et al., 2016; Nov and Arazy, 2013) who are more likely to have an increased sensitivity to reward (Smillie, 2013). Jia et al. (2016) also found that individuals who were rated higher on emotional stability and openness measures were less likely than others to find certain game elements motivating. Therefore, in the case of Gamified Assessments, particular care needs to be taken that the effects of game elements are constant for all individuals.

Furthermore, one cannot simply reduce the connection between the game elements and the assessment items, as it is also necessary to ensure that the effects of game elements are aligned with the aims of the assessment (Hughes & Lacy, 2016; Kim, 2015; Mislevy et al. 2012). If the actions rewarded by game elements do not correspond to actions that are beneficial for the assessment, then one may motivate actions which are unsuited for the purposes of the assessment. Kim and Shute (2015) exemplified this by taking a game that assessed physics understanding and creating two versions of it, which were identical except for the amount of autonomy participants had when choosing assessment levels. In one version, participants could complete the assessment levels in whichever order they liked, whereas in the other condition the order of the levels was predetermined. This was found to alter users' goals when completing the assessment, as the group playing the non-linear game concentrated on finding the best solution for each level and spent more time on each one, whereas the linear group concentrated on getting through the levels as fast as possible. If the way in which the user completed the game was not aligned with the way the assessment was scored, then even if their latent physics understanding was similar in both cases, they would obtain different results.

However, when constructed correctly, Gamified Assessments can provide several benefits over more traditional assessments used within selection, such as questionnaires and situational judgement tests. In particular, there is some initial evi-

dence to show that Gamified Assessments are able to mitigate cheating behaviour and test anxiety. When items from a personality test assessing Extraversion and conscientiousness were administered periodically throughout a platform style video game, participants were less able to fake their scores than when participants simply completed it as an online survey (Ramsay, 2017). The impact of Gamified Assessments on test anxiety was investigated by Mavridis and Tsiatsos (2016). They converted a multiple-choice assessment into a 'treasure hunt' game by implementing avatars, the ability to navigate through a virtual world, and aspects of a multiplayer experience, and subsequently observed a significant reduction in test anxiety. Since anxiety has been found to be negatively correlated with assessment performance (Cassady & Johnson, 2002; Mavridis & Tsiatsos, 2016) and evidence has shown that minority groups are more likely to experience test anxiety (Putwain, 2007), the ability for game mechanics to decrease test anxiety may also be beneficial in lowering the potential for adverse impact.

2.2.3.2 Game-Based Assessments

Game-Based Assessments embed the Gamification within the core of their assessment model, harnessing the full scope of game-thinking, not just for better applicant reactions, but also to capitalise on the inherent psychometric properties of games. Games are well suited for assessment purposes as they naturally present players with a stream of choices during gameplay. By having users interact with the GBA, they are forced to demonstrate their knowledge and ability directly, as opposed to more traditional assessments (e.g. self-report measures) where they are responding based on self-reflection and meta-cognition. Moreover, recording both player choices and the game's paradata (i.e. data about how the player arrived at their choice; Stieger & Reips, 2010) allows GBAs to analyse information concerning the users decision-making process, which is difficult to measure through traditional psychological assessments (Landers, 2015; Shute and Ventura, 2013).

In the past decade, several studies have been conducted using GBAs which measure a variety of cognitive traits and spatial reasoning measures(Ángeles Quiroga et al., 2015; Buford and O'Leary, 2015; Foroughi et al., 2016; Lim and

Furnham, 2018). These studies present some initial evidence of construct validity for the GBAs, which although limited in scope, looks to be promising. Several research projects have created GBAs using the commercially available Portal series as a measure of fluid intelligence, finding significant correlations with Ravens Progressive Matricies ranging from r = 0.4 to r = 0.65 (Buford and O'Leary, 2015; Foroughi et al., 2016; Lim and Furnham, 2018). Foroughi et al. (Foroughi et al., 2016) also demonstrated that participants' Portal 2 performance was significantly correlated (r = 0.78) with a latent variable of fluid intelligence extracted from two previously validated intelligence measures. There have also been several successful attempts to use other GBAs to assess aspects of general cognitive ability. Ángeles Quiroga et al. (2015) found that a performance measure created from a Wii game called Big Brain Academy was highly correlated (r = 0.93) with a general intelligence measure (g) extracted from a battery of conventional tests. A GBA assessing cognitive ability had greater incremental validity than a measure of Spearman's g (derived from a battery of cognitive ability tests) when both were used to predict GPA (Landers, Armstrong, et al., 2017).

Studies examining the construct validity of GBAs measuring other performance constructs have also been promising, although as with the rest of the Gamification literature, this evidence is still sparse. Shute, Wang, Greiff, Zhao, & Moore (2016) found that their problem solving GBA had a correlation of approximately r= 0.4 with both RPM and another problem-solving measure called MicroDYN, while Godwin, Lomas, Koedinger, & Fisher (2015) found that their GBA measures for sustained attention and memory were significantly correlated with existing measures of these constructs (r = 0.62 and r = 0.52, respectively) .

Most of these studies surpass the acceptable threshold for correlations in construct validity studies, which is set at 0.45 for the British Psychological Society and 0.55 for the European Federation of Psychologists' Associations (Kurz, 2016). Note, while it seems logical to strive for very high convergent validity, correlations between GBAs and traditional measures tend to be weaker than convergence between measures of similar formats. This is because the latter are inflated by

common method variance, which is the variance attributable to the similarities in the measurement method of two assessments, rather than the constructs of interest (Doty & Glick, 1998). The concurrent validity evidence for GBAs can also be negatively affected because traditional assessments often lack the ability to measure behaviours which are observable in GBAs, such as candidates' decision-making processes. Additionally, GBAs often measure more variables than their traditional counterparts to increase content validity. As a result, comparisons between different assessment formats can be troubled by alignment issues (Rupp et al., 2010). The qualifier is of course that any new facets added into an assessment to increase construct representation are well linked with existing research in order to avoid simply adding construct-irrelevant variance.

To determine whether the variables measured by a GBA increase construct representation rather than construct-irrelevant variance, it is critical to evaluate multiple sources of validity evidence of the assessment. Many of the studies outlined above lack a sufficiently broad examination of validity, which is a problem affecting much of the Gamification literature (Landers, 2015). However, criterion validity and evidence of GBA predictive validity is one commonly omitted source of evidence that is especially important to include in assessment contexts. Initial predictive validity evidence has been gathered for GBAs in educational contexts, where a GBA measuring learning behaviours was found to correlate significantly with the students' depth of understanding both directly after the assessment ($r = 0.51$), and one week later ($r = 0.31$; Snow, Likens, Allen, & McNamara, 2016). However longitudinal studies also need to be conducted in workplace contexts. The inclusion of criterion validity evidence would also be a good way to overcome the limitations mentioned previously with concurrent validity studies. In particular, comparing GBAs against relevant criterion measures would lessen the current reliance on concurrent self-report measures when assessing construct representation.

The application of game elements to assessments has been facilitated by the prevalence of technologically rich environments that are able to support new assessment formats. The ubiquity of smartphones and the widespread acceptance of

mobile gaming has provided GBAs with an essential conduit to get into the hands of a large number of users (Lawrence & Kinney, 2017). Furthermore, current technological capabilities enable user actions to be assessed at multiple levels of granularity, providing a large number of variables to measure traits such as processing speed and learning agility, as well as develop an in-depth personality profile of users (Fu et al., 2014). This is critical because GBAs measure large amounts of data, from a candidate's responses to visual stimuli through to how they make complex decisions. This data is stored in a 'log-file', which is configured to seamlessly record all the data that is generated as the player uses the game (Hao et al., 2016). This data can include information about button presses, how long they take for each action and the consistency of these factors over time (Hao et al., 2016). The ability to extract this data and connect it to the traits and constructs being assessed forms the crux of the 'stealth assessment' model underlying the majority of GBAs (Shute, 2011). Stealth assessment allows for a direct measurement of user behaviour, as opposed to the indirect methods underlying self-report. This provides an uninterrupted experience that enhances immersion and increases the player's sense of flow and engagement (Csikszentmihalyi 1990; Shute, Ventura, Bauer and Zapata-Rivera, 2009). However, depending on how the game is configured, it is not always obvious what information should be extracted from the large amount data logged by the application (DiCerbo, 2014; Hao et al., 2016).

Whilst the applications and benefits behind using GBAs are relatively tangible, it is perhaps less apparent what really underlies their measurement of individual differences, particularly of personality. In assessment and selection, GBAs typically use touch responses in tasks designed to measure a range of traits. The overarching concept of button taps providing insight into an individual's personality is unconventional and may appear lacking in face validity *prima facie*. In addition to this, GBA is largely informed by findings in the field of cognitive neuroscience, which seldom ventures into occupational sectors, further adding to the challenge. Seminal authors of biological personality theories have sometimes focused on lab-based research or educational and clinical applications, meaning occupational uses have

often been neglected (Furnham, 2016). Practitioners and occupational researchers unfamiliar to concepts of cognitive neuroscience may therefore fail to immediately appreciate the logic behind cognition-personality interactions. This obscurity is compounded by the added granularity and intrinsic nature of GBAs, which differs to preceding assessment methods. As such, practitioners and occupational researchers may be unaware of the context, theory and mechanisms that drive GBAs used throughout the workplace. This is a barrier for those seeking to research the benefits and limitations of this assessment method or to indeed develop their own GBAs.

Much can be understood from personality through an individual's response to basic stimuli. Neuropsychological models of personality generally define four systems, termed the behavioural inhibition system (BIS), behavioural activation system (BAS), the fight-flight-freeze system (FFFS) and the Constraint System (Corr et al., 2013; Gray, 1981; Kennis et al., 2013). The BAS is responsive to positive, rewarding stimuli and is thought to underlie exploration, reward-seeking, risk-taking and other similar behaviour. The BAS is suggested to overlap with two traits under the Five Factor Model taxonomy of personality (Goldberg, 1990; McCrae and Costa, 1987), Extraversion and openness to experience. The BIS is sensitive to uncertainty, negative feedback, threat and omission of an expected reward. The BIS regulates the BAS and FFFS and is proposed to trigger more frequent anxiety, shyness, worry and guilt if overly sensitive. The FFFS responds to negative, punishing stimuli and is involved in avoidance behaviour. Both the FFFS and the BIS correspond with neuroticism because of their relevance to harm avoidance and withdrawal respectively. Finally, the Constraint System responds to working memory tasks and relates to individual persistence, reliability and impulse inhibition. This final system is related to conscientiousness and agreeableness in the Five Factor Model. Crucially, for the context of GBAs and the purposes of this paper, individual differences are found in how sensitive individuals are within each system, therefore impacting how they respond to and are affected by external stimuli. Advances in smartphone technology have provided a suitable platform for presenting various stimuli in GBA

format that are sensitive to an individual's stimulus response tendencies.

Within the systems outlined above there are several 'levels of processes', from low-level perceptual and attentional processing through to higher-level, goal-directed decision making that are modulated by individual differences. Starting with lower level processing, greater attentional control has been associated with attenuation of workplace stressors that lead to job tension and depressed mood (Parker et al., 2013). Through measuring behaviour in a GBA, it is possible to reveal how elevated neuroticism is related to differences in mechanisms facilitating visual attention. For example, it has been consistently found that more neurotic individuals have a reduced ability to disengage from stimuli. Bredemeier et al. (2011) used a rapid serial visual presentation (RSVP) task to understand how an effect called the attentional blink correlated with neuroticism. The attentional blink occurs when two targets are rapidly presented in close succession, causing accuracy rates to be lower for the second target. Bredemeier et al. (2011) used a single letter RSVP task to show that neurotic individuals detected the first target but were less accurate for successive targets within the attentional blink window (approximately 250-500 ms after the first target). More neurotic individuals were therefore less able to disengage from the first stimulus in order to detect the second target letter. A later study by Dhinakaran et al. (2014) further confirms this disengagement effect, showing that more neurotic groups are less able to disengage from a stream of more salient stimuli (larger word font) to a separate stream of less-salient (smaller word font), despite both categories being equally task relevant.

Visual search tasks are another type of cognitive task that require attentional shifts to identify a target amongst distractors and have proven sensitive to individual differences. Performance on the visual search task has been related to employment performance, including in aviation security and radiography (Hope, 2012; Maeda et al., 2013; Mitroff et al., 2018). Conscientiousness has been found to relate to visual search accuracy (Biggs et al., 2017), where the target letter "T" was hidden amongst pseudo-"L" distractors. Given the visual nature of touch devices and the body of evidence reporting personality interactions with attention processes, visual

perceptual and attention-based paradigms are well placed in measuring personality domains in selection environments. Moving on from lower-level processing, higher level decision making, particularly in reward and punishment contexts is a core component of the neuropsychological systems outlined above (Corr et al., 2013; DeYoung, 2013; DeYoung et al., 2010; Gray, 1981).

A task commonly used for measuring risk taking and reward seeking is the balloon analogue risk task (BART) (Lejuez et al., 2002). The BART requires individuals to collect points or monetary reward based on the amount of times they pump balloons. Pumping a balloon past its limit will result in the balloon popping and the reward being lost. Individuals are able to bank when they are comfortable a balloon has been pumped enough before it explodes. Importantly, the explosion thresholds of the balloons differ, meaning that a risk-reward trade-off exists for each trial. The BART has been replicated amongst several commercial selection assessment providers (see Arctic Shores, 2018) because of its particular sensitivity to neuroticism and Extraversion, which are predictive of workplace performance across multiple disciplines (Rothmann and Coetzer, 2003; Salgado, 1997). Based on neuropsychological personality theories (DeYoung, 2013; DeYoung and Gray, 2009), one might expect that individual differences in responses to rewarding and punishing stimuli (i.e., a balloon bursting from over-pumping) could be related to neuroticism. The insula is a brain region that has been associated with both behavioral inhibition and activation systems, depending on stimulus type (Kennis et al., 2013). Research has shown the insula plays a role in BART performance, where functional magnetic resonance imaging (fMRI; a technique that measures brain activity, most commonly using blood oxygenation level) signal is able to predict a safe or risky move with up to 70.4% accuracy from right anterior insula and right lateral orbitofrontal cortex activation alone (Helfinstein et al., 2014). Studies of the BART and similar tasks have observed fMRI signal responses in the right insula following punishment (points deduction) to risk taking, and these responses show correlations with harm avoidance, anxiety symptoms and neuroticism (Hoffmann et al., 2018; Paulus et al., 2003).

A challenge to GBA is that research in cognitive neuroscience has traditionally favoured highly replicable mechanisms that minimise any individual differences (Hedge et al., 2018). Any remaining between-participant variance is often quashed through sample-wide averaging, meaning that effects revealed by individual differences are overlooked (Kanai and Rees, 2011). Conversely, for the study of individual differences, between-subject variance is crucial to finding meaningful effects and is certainly not considered as noise. The transition of tasks used in cognitive labs to GBA format in selection and other occupational settings are therefore inherently affected by this approach. This means that many classical paradigms have low between-participant variability, which is required to understand how different individuals interact with GBAs. As such, classical tasks may require adjustments to be used within GBAs to better reveal personality differences. Hedge et al. (2018) present between and within (repeated measures) participant variance for seven cognitive tasks that measure attention, information processing and cognitive control. Their findings indicate a problem for GBAs based upon classical cognitive tasks for measuring individual differences. Methods for enhancing the sensitivity of paradigms to forms of individual differences are beyond the intentions of this thesis; however, Hedge et al. (2018) proposes composite response time & accuracy scores and alternative statistical approaches as potential avenues for improvement. Collecting a broader range of variables using GBAs can facilitate such an approach. Other authors have adjusted their experimental design in order to maximise individual differences. Whilst investigating attention processing and trait openness to experience, Wilson et al. (2016) avoided using highly discrete spatial target areas used in previous research in order to maximise individual differences. Continued research is required to realise the full potential for cognitive paradigms, and in turn GBAs that are based upon such paradigms to infer individual differences.

Nevertheless, a combination of brain imaging findings, task performance and neuropsychological models of personality paint a picture of how GBAs index personality traits. Whilst this is by no means a comprehensive review of personality interactions with each relevant brain system and cognitive task, it does offer an in-

troduction to the underlying logic of GBAs. The ultimate goal of this literature review is to provide enough insight to the reader to appreciate why this thesis has been designed to test the hypothesis that GBA are resistant to faking.

2.3 Game-Based Assessment and Faking

So why should we expect GBAs to be resistant to faking?

Landers and Sanchez (2022) offer a theoretical explanation of why and how GBAs may reduce faking based on the established models of faking references in the previous section. As described, faking is broadly believed to depend on three major antecedents to faking behaviour: opportunity, ability, and motivation. The authors suggest how gameful design and gamification can target them to inhibit faking.

The first antecedent is motivation to fake, which, according to the authors, can be influenced by game mechanics that increase immersion with the goal of triggering a flow state. As described in the previous section, flow is a positive mental state characterised by feelings of engrossment in an experience, effortless continuation of behaviour, and a perception of intrinsic reward from the behaviour. Landers and Sanchez (2022) suggest harnessing this psychological mechanism to shift the state of immersion to the game elements and away from faking, thereby replacing the individuals' motivation to fake with motivation to engage with the game more deeply.

The second antecedent is ability to fake, which refers to knowledge, skills, and abilities that enable faking. In terms of knowledge, this can include knowledge specific to the organisation, the job, or the assessment that can be used to align assessment responses to job requirements and expectations. Whilst the first two are not relevant to this point, knowledge about the assessment is key. As Sanchez and Langer (2020) point out, game elements in an assessment might be too novel for candidates to have enough knowledge of what faking behaviour would be appropriate to use. Even beyond the game elements, as described in the previous section, GBAs' psychometric scores are generated through very complex theory-

and data-driven algorithms based on biological models of individual differences which reduce candidate knowledge of potential faking behaviours even further. For eample, Landers, Auer, Mersy, Marin and Blaik (2022) described how they used machine learning to model trace data from GBA to compute psychometric scores, and this shows that the level of complexity pertaining the human-machine interactive behaviours linked to psychometric scores is too high for candidates to be able to purposely reproduce to mimic a specific profile. In terms of skill and ability, the authors mention that faking is a cognitively demanding task that requires complex mental processing, and the authors suggest leveraging cognitive load to reduce the cognitive bandwidth left available for faking. However, while a assessment gameful design and gamification targeting cognitive load might be able to inhibit the ability to fake, it might also result in unnecessary adverse impact. Furthermore, the author include pre-existing game experience, such as knowledge of game mechanics, skill with game mechanics, and attitudes to games as antecedents of the ability to fake. Whilst this would be relevant in a context of a commercial game, even the most complex GBAs on the market do not require any gaming experience whatsoever and there are no differences in performance between gamers and non gamers, and as a function of the type of games people play recreationally (Arctic Shores, 2016).

The third antecedent is opportunity to fake, which can be influenced by the design of the assessment itself. Gameful design and gamification can be used to create assessments that are more difficult to fake by leveraging the effect of low transparency. The authors explain that the format, presentation, and response modes in the assessment can be purposely designed avoiding clear desirable response patterns. Furthermore, they highlight the highly digitalised nature of gamefuly design, gamified, and game-based assessment which should lends itself to house sophisticated behavioural analytics able to detect faking as it happens. Considering the evidence summarised in the previous section with regards to the nature of the tasks typically used in GBAs and the complexity of their scoring models, it can legitimately be expected that opportunity to fake would have the most significant impact on reducing faking. The type of tasks used in GBAs are mostly measures of maxi-

mal performance, which by definition don't lend themselves to deliberate improvement, and this is possibly the biggest hurdle to opportunity to fake. A very recent paper has reported a study that is relevant to this point, albeit not related to faking directly. Simons, Wohlgenannt, Weinmann, Chneider, and vom Brockle (2023) correlated the performance on a commercial VR game called Job Simulator with a measure of intelligence, and found that they could predict intelligence and processing capacity through the speed at which the participants solved the VAR game challenges.

Landers and Sanchez (2022) provide an excellent theoretical rationale to explain why we should expect GBAs to offer a good degree of resistance to faking, albeit possibly overly emphasising the role of game elements and game mechanics in faking considering the extent to which they are actually used in commercial GBAs used for candidate selection. As it will become clear as this document unfolds, this thesis is also mapped on theoretical models of faking, however, it is much more heavily rooted in ability and opportunity to fake than it is in intentions and motivation.

In addition to the promising theoretical outlook just described, the little experimental evidence published to date on faking in gamified assessment and GBA is also auspicious.

Harman (2022) compared assessment performance between the IPIP-50 and a gamified measure assessing the same constructs, which they defined as a text-based game-like personality measure (GPM; Harman & Purl, 2022)) and consisted in embedding the question within a story. They found that the GPM, even if not radically different from the self-reported measure, recorded fewer instances of faking and careless responding. This suggest that in addition to preventing faking, gamification can help maintaining a better level of test performance.

The GPM described and used by Harman (2022; and Harman & Purl, 2022) falls into the *storification* category defined by Landers and Collmus (2021) in their seminal paper which also reports an experiment in which participants were asked to fake their conscientiousness and openness to experience scores in a self-reported

assessment and a storified version of it. In this case, the results showed a significant effect of reducing faking through storification for conscientiousness but not for openness to experience. Interestingly, faking instructions did not change the conscientiousness scores at all in the storified version whilst the manipulation had a very large effect in the non storified measure. These results are particularly remarkable because Conscientiousness is the most successfully faked construct in candidate assessment (Viswesvaran and Ones, 1999) and the only one consistently faked across job categories and sectors (see: Birkeland et al, 2006).

Barend and DeVries (2022) reported a study in which what the call a personality assessment game, which is a combination of GBA and gamified assessment, was used in a simulated selection situation to establish its resistance to faking. The assessment measures the Honesty-Humility HEXACO factor, which is similarly as high in social desirability as conscientiousness, and hence more likely to be a faking target. They found that, when participants were instructed to approach the assessment as if during the application for a job in which honesty and humility was overtly mentioned as a requirement, their scores were not different from the control condition. Furthermore, they found that the experimental faking condition did not alter the convergent validity of the tool. The assessment that they used in the study, Building Docks, was developed by the same research group (Barends, de Vries, and van Vugt, 2022) and it was previously demonstrated to be a valid measure of honestly and humility but also to possess incremental validity over self-reported scores of honesty humility in predicting important organisational outcomes such as cheating for financial gains.

The only evidence entirely based on GBA and faking is a paper from Waters, Sanchez, Garcia, and Rueda (2022) which describes an experimental study in which participants were labelled as *good* and *poor* fakers on the basis of their ability to fake a self-reported inventory to match a specific job profile. Good and poor fakers were asked to fake a GBA to achieve the same results, however, none of them were able to do so. This shows that even people who might normally be able to fake very successfully are not able to do the same if the assessment is a GBA.

Overall, the evidence consists in very few small scale experimental studies in which faking was instructionally manipulated. Much more research is needed that includes real candidates on one hand, but also that explores faking considerably more deeply than just observing its emergence. It is necessary to expand our investigation of the interface between GBA and faking by assessing the mechanisms through which this format impedes faking. The existing theoretical models of faking provide an excellent frame of reference to guide research question where the most critical shifts occur as a function of the idiosyncratic features of a new assessment format. It is important to build a holistic and comprehensive understanding of how those shifts happen and why in order to identify and leverage the most impactful mechanisms to prevent faking.

2.4 This Thesis

As far as both the theoretical and the experimental evidence reported above is very encouraging in suggesting that GBAs might not be as susceptible to faking as other methods, it is necessary to understand the mechanisms by which game-based assessment prevent faking by combining the two. This is for two reasons, first, it cannot be taken for granted that just because faking doesn't occur in low stakes and lab settings this would also be true in high stakes real life settings because the contribution of intentions and motivation to fake on faking behaviour are not quantitatively nor qualitatively comparable across contexts. It's therefore important to fully understand the *how* and the *why* in order to safely assume that the effect would replicate. The second reason is that by virtue of knowing the *how* and the *why*, it might be possible to democratise knowledge abut faking resistance by revealing the assessment features that are most successful at impeding faking.

This thesis was designed combine theoretical and experimental evidence on assessment faking to test the hypothesis that workplace assessments using experimental laboratory tasks to generate psychometric scores are not susceptible to faking, and this is because the format they use does not leave room for candidates to control the way in which they approach the assessment well enough to manipulate

the results they obtain.

To investigate this, exploring the type of behavioural differences elicited by GBAs and the behavioural constraints this format imposes to candidates is crucial, but none of the game elements mentioned in the gamification literature are. Whilst the gamification contextualisation of this type of assessment is key to fully appreciate its innovation and potential within the remits of talent acquisition, when it comes to faking and faking prevention, the gamification aspects are completely redundant to the discourse, whereas the behavioural nature of the assessment is where the most insightful discoveries can be made.

Whilst Landers and Sanchez (2022) attribute much of GBAs' resistance to faking to the effect of immersion, flow and enjoyment, which in theory would make a lot of sense, they overlook the fact that most proprietary GBAs used for candidate selection are nothing but glorified cognitive tasks, and it is unlikely that the theoretical concepts known to be triggered during gameplay would be...at play. So it is very likely that motivation might play a much less crucial role on faking compared to what they expect. This thesis does not intend to provide evidence against the role of gameplay-related states on reducing faking, also due to the fact that Landers and Sanchez (2022) was published months before this thesis was completed; instead it focuses on how BBAs might prevent faking purely from an assessment method perspective. That is, no matter the state and motivation, the format is not expect to "allow" faking.

For this reason, and also to reflect the fact that most psychometric providers of this type of assessment have increasingly reduced both their use of game elements in their assessment propositions and their references to games and gamification semantics in their brands, this thesis will use the term Behaviour-Based Assessment in lieu of Game-Based Assessment to describe assessments, whether gamified or not, which use data collected through replicating experimental laboratory tasks in digital form to build machine learning models of individual differences. This thesis consists of five studies, which are loosely mapped onto the Integrated Model of Faking (McFarland & Ryan 2006) and the VIE model of faking (Elling- son &

McFarland, 2011). Both models are broadly divided into three main blocks - intentions/motivation to fake, ability to fake, and faking behaviour - and so is this thesis.

The ultimate goal of this thesis is to demonstrate that the objective ability to fake is so heavily obstructed by the format in which BBAs are designed to prevent behavioural expression of faking no matter how high the intentions or motivation to fake might be.

The five studies are summarised below:

- **Intentions** This study is inspired on the Integrated Model of Faking, and it explores the individual-level factors affecting intentions to fake, investigating whether the same factors are correlated with each other in the same way for all assessment formats. It then goes to investigate whether the perceived behavioural control people have (or think they have) on a specific assessment format impacts their intentions to fake it, and whether the presence of a selection of faking warnings affects that perception and with it the intentions to fake. The aim of this study is to demonstrate that people have less strong intentions to fake BBAs compared to other types of assessment because they perceive less behavioural control over them, and this is true even when faking deterrents are factored in.

- **Ability** Two studies were designed address two different aspects of ability. This chapter untangles the dissociation between expectancy and objective ability to fake proposed in the VIE model of faking, but it also echoes Aizen's (2002) paper on the conceptual ambiguity of perceived behavioural control, deconstructing the concept of perceived behavioural control examined in the previous chapter:

    1. *Perceived Ability* - In this qualitative study participant were asked one single question for each of four different assessment formats: "how would you cheat on this assessment?" The faking strategies they described across formats were categorised to understand whether peo-

ple were able to suggest viable cheating strategies, and whether they adapted their strategies to the assessment format. The aim of this study is to demonstrate that people don't know how to fake BBAs, that is, that they don't perceive that they have the ability to fake BBAs.

2. *Objective Ability* - In this study a BBA model of Extraversion was unpacked, and its variables divided between measures of maximal performance and measure of typical performance. The rationale behind this study is that measures of maximal performance by definition don't lend themselves to faking, and their presence in a model of personality would substantially reduce the objective ability to fake. The aim of this study is to demonstrate that there is a lot of variance explained by maximal performance in BBAs, and for this they don't offer people the objective ability to be faked.

- **Behaviour** This chapter also includes two studies, both focusing on the emergence faking behaviour in situation where high intentions/motivation to fake can expected to be high:

  1. *Candidates vs Employees* - It is quite a well-known fact that candidates tend to score higher on assessments than employees, and this has traditionally been explained in terms of faking. This study investigates score differences between candidates and employees to establish whether there are differences in the data patterns between the two condition groups that might suggest faking. The aim of the study is to demonstrate that while there might be differences in scores between candidates and employees, this is not due to faking but rather to the effect of low motivation (in employees) on cognitive tasks which equates to lower scores.

  2. *Job Application Simulation* - The last study is framed as a simulated job application in which people pretend to apply to one of five different jobs for which they have been told the selection criteria. The selection

criteria are used to increase perceived ability, or expectancy, as much as possible so to isolate the role of (lack of) objective ability in preventing faking behaviour.

Chapter 3

INTENTIONS

The first block of the Integrated Model of Faking (McFarland & Ryan, 2006) is *intentions*, that is, the model implies that in order to engage in faking behaviour, the individual must have the intentions to do so. The model includes several individual-level antecedents of intentions to fake and several situation-level moderators of the relationship between the antecedents and intentions to fake, such as the valence towards doing well in the assessment and the presence of faking detection strategies.

The following study (Study 1) explores the predictive validity of the intentions antecedents featured in the model across different assessment formats, with the expectation to demonstrate that BBAs are less likely to be the target of intentions to fake.

3.1 Study 1 - Faking Intentions

Across all the models that have attempted to describe, explain, and understand faking, a universal *conditio si ne qua non* for faking is intentions. Whether as an assumption or as an overt factor within a framework, intentions to fake is at the core of any theory of faking. This is because faking is conceptualised as a ”volitional behaviour” (Ellingson & McFarland, 2011, p.323) in which individuals deliberately provide misleading information with a specific goal in mind, and intentions are the both the genesis and the directional force of goal-directed action. This is, of course, not just limited to faking and it applies to any deliberate - or planned - behaviour.

The Theory of Planned Behaviour (TPB; Ajzen, 1985) builds on the Theory of

Reasoned Action (TRA; Fishbein & Ajzen, 1975) which, resting on the assumption that intention is the strongest determinant of behaviour, illustrates how attitudes and subjective norms are at the basis of intentions and therefore the earliest antecedent of behaviour. Attitudes broadly refer to a person's perception of the desirability of a behaviour, whereas subjective norms refers to a person's perception of the attitudes of trusted individuals on the same behaviour. In other words, intentions toward a given behaviour are enabled by a combination of how we and other people we respect and trust feel toward such behaviour. Meta-analytic evidence provide strong support for the TRA showing that attitudes and subjective norms explain between 33% and 50% of variance in intentions which in turn explains between 19% and 38% of variance in behaviour (Armitage & Conner, 2001).

This early theory, however, focused entirely on the individual, and did not take into account one important factor which was subsequently added in the TPB: the degree to which the behaviour might be within the actor's control, and most importantly the person's perception of that level of control. Perceived Behavioural Control (PBC; Ajzen & Madden, 1986) is the subjective perception of estimated likelihood of succeeding in a behaviour, and it's a pivotal addition to the TRA because it anchors intentions to the perceived feasibility of a target behaviour, adding a more realistic dynamism to how intentions are formed. Godin and Kok (1996) identified 76 studies investigating the relationship between PBC and intentions, and out of the 65 of them in which PBC predicted intentions, the average correlation was r=.46, and PBC explained 13% additional variance over the TRA.

However, not all volitional behaviours are necessarily dependent on volitional control, and the influence of PBC on intentions is linked to the degree to which they are. Madden, Ellen, and Ajzen (1992) compared the additional variance of PBC over TRA between several behaviours differing on the basis of how much control the actor had over them, and they found that the magnitude of additional variance explained by PBC was linked to the degree of controllability of that behaviour. In their study, TPB did not add any value over TRA in predicting intentions for highly controllable behaviours such as listening to music, whereas it showed substantial

incremental validity in the context of intentions to engage in less controllable behaviours such as getting a good night sleep. This evidence suggests that the magnitude of incremental variance in intentions explained by TPB over TRA for a given behaviour might be a marker of the degree of controllability of that behaviour, above and beyond PBC itself.

In the context of faking assessment, the TPB has been used to complement existing models of faking, which focused almost entirely on the individual, with a new emphasis on the assessment itself, adding an important layer of interaction between the person and the target of the behaviour which was previously underestimated. Cronan, Mullins, & Douglas (2018) found that a model based on TPB explained 35-50% of the variance in intentions to commit academic integrity violation, that is, to plagiarise or share homework with other students. They found evidence that attitudes, subjective norms, PBC, but also moral obligations (i.e.: sense of guilt towards cheating) and past cheating behaviour were at the basis of declared intentions to engage in violation. Anther study, focusing on interviews, tested whether personality and TPB interacted in predicting faking; they found that TPB predicted faking and, in addition to this, attitudes and subjective norms mediated the relationship between Honesty-Humility and Conscientiousness and interview faking (Bourdage, Schmidt, Wiltshire, Nguyen, Lee, 2019).

With the addition of the TBP, the Integrated Model of Faking (McFarland & Ryan, 2006) shifted the conceptualisation of faking from a purely volitional behaviour to a complex behaviour encompassing volition, perception, and situational factors in which the assessment ceases to be a passive recipient of faking behaviour and acquires the role of an active and interactive influence able to moderate the intentions and successful execution of that behaviour with its identity. This new role that the Integrated Model of Faking granted to assessments is pivotal in enabling comparisons between assessment types. Previous models of faking were very limited in this scope as the exclusion of the assessment's identity as a source of variable level of susceptibility to faking did not incorporate enough of the idiosyncratic interplay between the assessment and the person that is needed to compare assessments

on the basis of how well and easily people think they can be faked.

In this study, the assessment's identity gains the centre of the stage, as the emphasis is on comparing how much participants intend to fake different types of assessment depending on the level of PBC linked to them specifically.

Whilst some assessment format lend themselves more than others to faking (e.g.: Marcus, 2009; McFarland & Ryan, 2000, 2006), it's worth noting that PBC refers to a person's perception of the level of control they have over a specific assessment, which may or may not reflect their effective faking potential as this is constrained by the scope for faking an assessment actually affords. This means that interventions aimed at manipulating perceptions of control might be able to impact PBC, and with that faking intentions, regardless of the control a person really has on faking that assessment.

A popular situational moderator of PBC tapping on this interface is the use of warnings to deter faking (McFarland & Ryan, 2006). In the Integrated Model of Faking, warnings are featured as situational influences directly impacting PBC and indirectly moderating the influence of TRA on intentions. Roulin et al (2016) also proposed a Dynamic Model of Applicant Faking which leans on their earlier work on signalling (Bangerter, Roulin, and König, 2012), and incorporates an element of comparison of opportunities and risks to fake within their model which is contingent upon the potential consequences of faking specific to a given assessment situation. Their model is sensitive to an organisation's investment in making faking hard or costly, and it describes two avenues to reduce faking: the first, is to use assessment which are in nature harder to fake, while the second pertains to the use of faking deterrents such as warnings and/or leveraging reputational evidence of punishment following identification.

Faking warnings are widely employed in candidate assessment, and they broadly consist in informing candidates that faking can be detected and that there are serious consequences linked to faking, such as disqualification (Dilchert & Ones, 2012). The effectiveness of warnings in reducing faking varies quite substantially in the relevant literature, with estimates ranging between 30% and 50% in some

studies (e.g.: Fan, Gao, Carroll, Lopez, Tian, & Meng, 2012; Landers et al, 2012) to small, if at all significant, effect sizes in others (e.g.: Fisher et al, 2018). A validity study, for example, found that while warnings reduced mean scores for some personality dimensions, this did not have an effect on improving convergent validity between self and observer-rated personality on those dimensions, suggesting that the effect of warnings, despite being observed, might be rather inconsequential (Robson, Jones, & Abraham, 2007). Finally, Dwight and Donowan (2003) did a meta-analysis of warnings research, and estimated the average effect size in reducing faking to be about d=0.23 (which is small), however, they then ran an experimental study encompassing different types of warning and warning detection, and a range of personality scales, and concluded that the degree to which warnings reduce faking is the byproduct of an interplay between many different factors but, when optimised, it can often yield large effect sizes and significantly reduce the likelihood of fakers outscoring non-fakers.

This being said, considering that the magnitude of faking is larger in simulation studies than it is in field studies, and that participants in experimental studies, but not real candidates, might also be subject to demand characteristics effect (Hough, 1998), the effect sizes of deterrents on faking observed in experimental studies might be overly inflated compared to what would happen during real candidate selection. A study with a field sample of bus operator showed that warnings did have an effect on candidate scores, however, the effect was too small to have any impact on hiring decisions (Kuroyama, Wright, Manson & Sblynski, 2010); on the other hand, Lopez, Hou, and Fan (2019) found evidence from a field study that faking warnings did not just lower assessment scores but they also successfully prevented fakers from securing the top ranks of the score distribution.

Nevertheless, as Goffin & Boyd (2009) point out, candidates will be likely to fake when they feel it is an advantage but they won't do so when they believe that this will lead to negative consequences. The key to this otherwise obvious statement is belief. In order to *believe* that faking will result in negative consequences, candidates must believe that the faking deterrent is able to detect or prevent faking,

and this implies substantial overlap between the type of deterrent and the assessment format. For example, if the best way of faking an assessment is lying, then the threat of having to complete the assessment in proctored settings would be very low. However, this would change quiet drastically if the best way to fake was googling the right answers as proctoring would make that impossible. As it will be described below, this consideration is at the basis of the choice of deterrents used in this study, as each of them was selected for its relevance to a different type of assessment.

3.1.0.1 This Study

This study was designed to compare how four assessment formats (two BBAs and two non-BBAs) differed from one another in terms of how strongly participants intended to fake them. The study aims to replicate the first stage of the Integrated Model of Faking (McFarland & Ryan, 2006), *Intentions*, exploring whether - and how differently - personality traits and the constructs in the TPB predict intentions to fake across four different assessment formats, and whether and how three different faking warnings moderate those predictions.

The key purpose of this study is to test the assumption that BBAs will be less of a target of intentions to fake than non-BBAs due to the lower level of PBC this format offers for faking; this is expected to be the case whether faking deterrents are present or not.

3.1.0.2 Hypotheses

Three main hypotheses are advanced to explore how the TPB replicates across BBA and non-BBA assessment formats in predicting faking intentions.

Each of the hypotheses focus on a different antecedent of intentions - the first hypothesis is centred on individual-level factors that have previously been linked to intentions to fake, whilst the second hypothesis focuses on assessment-level factors such as perceived behavioural control, and the third hypothesis evaluates the effect of situation-level factors such as faking deterrents.

Where feasible, the study tests the assumptions at the overall level and then comparatively at the assessment format level.

The three hypotheses were formulated as follows:

- **Hypothesis 1: Do individual factors predict intentions to fake?**
  - *Hypothesis 1a*: Attitudes toward faking, subjective norms, moral obligations, and past behaviour will be positively correlated with overall intentions to fake
  - *Hypothesis 1b*: Attitudes toward faking, subjective norms, moral obligations, and past behaviour will be positively correlated with intention to fake self-reported and intelligence assessments but not with BBAs
  - *Hypothesis 1c*: Some HEXACO scores will be significantly correlated with general intentions to fake. More specifically, Honesty-Humility, Agreeableness and Conscientiousness will be negatively correlated with intentions to fake, whilst no significant correlations with other HEXACO scores are expected
  - *Hypothesis 1d*: The HEXACO scores that are significantly correlated with intention to fake will show this effect only in the context of self-reported and intelligence assessments but not with BBAs
- **Hypothesis 2: Does an individual's Perceived Behavioural Control over an assessment format predicts their intentions to fake it?**
  - *Hypothesis 2a*: Perceived Behavioural Control will be significantly correlated with intentions to fake all assessment formats in equal measures
  - *Hypothesis 2b*: Participants will perceive significantly less behavioural control over BBAs compared to self-reported and intelligence assessments
  - *Hypothesis 2c*: Participants will indicate that they are less intentioned to fake BBAs than self-reported and intelligence assessments
- **Hypothesis 3: Do faking warnings reduce intentions to fake?**
  - *Hypothesis 3a*: Faking warnings will significantly reduce perceived behavioural control and intentions to fake across all assessment formats

- *Hypothesis 3b*: Intentions to fake BBAs will be equally reduced by all deterrents
- *Hypothesis 3c*: Detection Algorithms will reduce intentions to fake self-reported assessments more than the other deterrents
- *Hypothesis 3d*: Proctored Testing will reduce intentions to face Intelligence tests more than the other deterrents

3.1.1 Method

A within-design study was carried out in two phases on Prolific Academic to assess whether a series of individual differences and perceptions about assessments predicted participants' intention to fake them if they encountered the same assessments in the context of a job application (both phases), and whether faking warnings might deter them from doing so (phase 2 only).

3.1.1.1 Sample

The samples (N=260 in phase 1; and N=200 in phase 2)were recruited from the research platform Prolific Academic with the only restrictions of residing in UK or US in order to ensure that the participants understood English very well. The samples in the two phases were comparable with regards to their demographic distributions, both featuring mostly white individuals (63% and 84.6%) and females (69% and 68.9%). The average age for the two samples was also similar - 28.56 and 29.59. The sample from phase 1 was blacklisted from phase 2 in order to maintain the data consistent; blacklisting is an option offered by Prolific Academic that hides a study opportunity from a specific group of participants.

3.1.1.2 Measures

The study used a series of self-reported measures and four videos.

The self reported measures were the HEXACO (Lee & Ashton, 2004; as used in Bourgade et al 2019 study), and the measured used in Cronan et al (2018) study on academic integrity violation, which were adapted for use in non-academic assessment faking by tweaking the language very slightly - see Appendix A.1 and list below.

Table 3.1: Descriptive statistics of all measures used in Study 1

		h	e	x	a	c	o	Attitudes	Sub norms	Mor Obl	Past Beh	Intent
Phase1	**M**	16.24	13.79	15.97	15.04	18.77	17.46	117.69	10.86	10.88	22.67	5.14
	SD	3.33	2.95	3.48	3.72	3.06	4.15	98.26	2.66	3.28	39.01	2.82
Phase2	**M**	16.17	17.59	15.39	14.86	18.68	17.65	11.74	14.53	13.45	2.12	6.36
	SD	3.14	3.64	3.41	3.61	3.14	3.86	5.52	3.78	4.64	1.96	4.18

- Intentions (Madden et al, 1992) - three items, such as *I intend to cheat on a test for a job application in the near future* rated from Definitely (1) to Definitely not(7)

- Attitudes (Bodur and Bringberg, 2000) - four items rating the statement *Overall, I see cheating on a test for a job application as...* on dimensions such as Favourable to Unfavourable

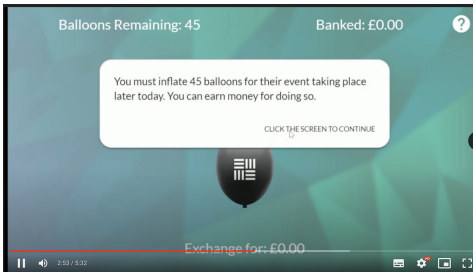
- Subjective norms (Ajzen, 1991) - four items rated 1-7 focusing on important people's perceptions of cheating (e.g.: *Most people who are important to me think I should not cheat on a test for a job application*)

- Perceived Behavioural Control (Ajzen, 2002) - five items, such as *I believe that I have the ability to cheat on <insert format>* rated from Strongly Disagree (1) to Strongly Agree (7)

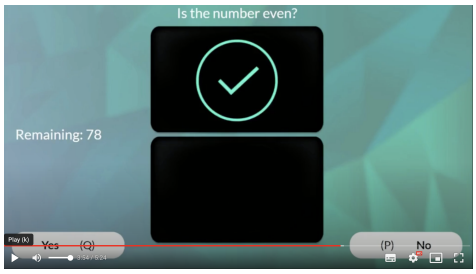
- Moral Obligations (Beck and Ajzen, 1991) - three items, such as *Cheating on an assessment goes against my principles* rated from Strongly Disagree (1) to Strongly Agree (7)

- Past Behaviour (Cronan et al, 2018) - Two items: *"How much have you cheated on assessment in the past few years?" Very little (1)...a lot (7)*, and *"In the past two years, how frequently did you cheat on assessments?"* Never (1)... every time (7)

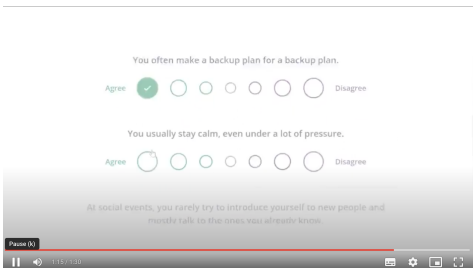
The four videos used in the study each described one of the four types of assessment under investigation: a self-reported personality assessment, an intelligence test, and two different BBAs - one featuring decision-making tasks and the other one featuring puzzle-like tasks.



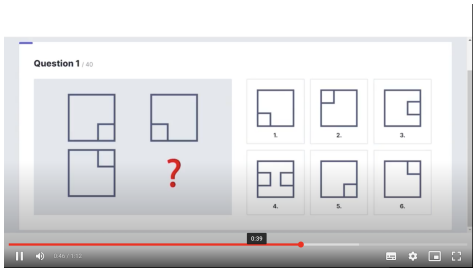
(a) Video - BBA-DM



(b) Video - BBA-Puzzle



(c) Video - Self-reported Test



(d) Video - Intelligence Test

Figure 3.1: Assessment format demo videos

Those four assessment formats were chosen as they offer different avenues for faking to potential applicants, and therefore are expected to vary with regards to how much behavioural control participants perceive over them. See Table 3.2 for a detailed description of each format and the scope they offer for faking.

The four videos were recorded on Screencastify by a UCL student who described the key methodological aspects of each assessment whilst demonstrating how to complete them without making any references to how to fake them (see Figure 3.1 for screenshots). The four videos can be accessed via the links in the Appendix A.2.

Table 3.2: Description of the four formats used in the study, with a note on the opportunity to fake they offer and a link to the faking deterrent which is the most aligned to the format's vulnerability to faking

Format	Description	Faking Opportunity	Format-Aligned Deterrent
BBA-DMT	Gamified assessment based on decision making tasks in which participants are faced with challenges to which they decide how to respond, usually untimed.	Participants should be able to modify their responses to appear in a different way but but this is limited by their understanding of the relationship between their task behaviour and the desired assessment scores, which is typically very hard to guess if not impossible.	All in equal (small) measures, however, online guides might contain useful information to improve task performance (which does not necessarily equate to test performance).
BBA-Puzzle	Gamified assessment based on puzzle-like tasks in which participants have to make timed responses to stimuli on the screen, usually requiring reactive rather than deliberate responses.	Participants have little avenues for faking this type of assessment as they tend to be predominantly based on measures of maximal performance, which, by definition, cannot be faked for the best.	All in equal (small) measures
Self-Reported	Question-based assessments in which participants have to indicate how much they agree with statements describing them, their behaviour, preferences, attitudes, etc	Participants can decide to endorse any response whether accurate or not. The relationship between the questions and successful performance is typically quite obvious.	Faking Detection Algorithm
Intelligence	Question-based assessments in which participants have to work out the single right answer to a problem tapping on a specific ability or area of knowledge	Participants can fake this type of assessment by knowing the answers to the questions either by accessing them or by getting help by somebody more able to answer than them.	Proctored Testing

3.1.1.3 Procedure

The two phases of the study were identical with the exception of a series of questions about deterrents added to phase 2, which were subsequently incorporated into the study to complement the existing evidence with a situational layer.

In both phases, participants responded to a study opportunity on Prolific Academic advertised as "Study on Assessment Faking". They were then directed to a survey hosted on the survey platform Gorilla where they were required to complete three blocks of questions: one featuring a demographic questionnaire, another featuring the HEXACO and the TPB scales, and the last one in which, for each assessment format, participants had to watch the video describing that format and then complete the PBC and the TPB scales.

In phase two, the last block was modified to include questions assessing the effectiveness of three different faking deterrents on reducing PBC and faking intentions for each assessment type.

The three deterrents were: proctored testing, fake-detecting algorithms, and a warning against the use of online guides and information found on online forums. Figure 3.2 reports the description of proctored testing, see the Appendix for the descriptions of the other two deterrents. Also, Table 3.2 on page 61 shows the alignment between the different deterrent with the faking opportunities afforded by the different formats.

> Within the instructions, it is stated that the employer is able to record your behaviour whilst you are completing the assessment. If your application is successful, this will be reviewed for markers of cheating; you might be invited to complete the assessment again in a supervised setting after providing photographic identification.
>
> The instructions also tell you that if cheating behaviour is detected and/or if your scores differ between the assessments, you will be rejected from the application process.

Figure 3.2: Example of deterrent description - Proctored Testing

For each assessment formats, participants watched the description video and completed the PBC and faking intention scales exactly as in phase one, but in ad-

dition to that they then read the descriptions of the three faking deterrent each followed by the PBC and intention scales.

3.1.2 Results

The data was cleaned and analysed with SPSS v.29 to address the three research questions, all collectively aiming at testing whether the intentions-related factors of the Integrated Model of Faking can be replicated across BBA and non-BBA assessment formats in the same way.

Overall, it is expected that there will be differences in the way in which the model replicates across formats, resulting in BBAs being less of a target for intentions to fake.

The analyses were performed separately for the two phases and compared to establish whether to combine the two sub-samples. A series of significant differences emerged between them, and the results are therefore reported side by side.

In more detail, a series of Fisher's r-to-z transformation tests revealed several significant differences in the correlation coefficients between the two phases. The correlation coefficients between intentions to fake and Past Behaviour were significantly different across all formats of assessment, with Phase 1 displaying stronger positive correlations than Phase 2 for BBA-DMT (r = .501 vs. r = .253, z = 3.086, p = .002), BBA-Puzzle (r = .667 vs. r = .214, z = 6.214, p ¡ .001), Self-reported measures (r = .478 vs. r = .286, z = 2.391, p = .017), and Intelligence (r = .594 vs. r = .344, z = 3.437, p = .001). Moral Obligations demonstrated significantly different correlations with Intentions to Fake in two formats, with Phase 2 showing stronger negative correlations for both BBA-DMT (r = -.354 vs. r = -.558, z = 2.747, p = .006) and Self-reported measures (r = -.348 vs. r = -.507, z = 2.066, p = .039). For Attitudes, only the BBA-Puzzle measure showed a significantly different correlation with Intentions to Fake, with Phase 1 displaying a stronger positive correlation (r = .542 vs. r = .373, z = 2.273, p = .023) than Phase 2. Factor X of the HEXACO showed a significant difference only in the context of intentions to fake Intelligence tests, with Phase 1 showing a slight positive correlation while Phase 2 showed a slight negative correlation (r = .090 vs. r = -.118, z = 2.207, p = .027). All other

comparisons between the two Phases did not reach statistical significance.

3.1.2.1 Hypothesis 1 - Do individual factors predict intentions to fake?

The first hypothesis investigated the relationship between a series of individual factors and intentions to fake. These relationships were tested at the overall level and then individually for each assessment format.

It is expected that, for BBAs, intentions to fake will not be associated with individual factors as participants might not possess strong intentions to fake this type of assessment and these consistently lower scores will dilute contribution of individual factors on intentions.

Hypothesis 1a - Attitudes toward faking, subjective norms, moral obligations, and past behaviour will be significantly correlated with overall intentions to fake A series of bivariate correlations showed that attitudes to faking, subjective norms, moral obligations and past behaviours all correlated significantly with intentions to fake, and this was true for both sub-samples. In phase one the correlation coefficients were r=.716, r=-.349, r=-.480, and r=.754 respectively (all significant at the .01 level), whereas in phase 2 they were r=.625, r=-.377, r=-.701, and r=.438 respectively (all significant at the .01 level).

Hypothesis 1a was fully supported showing that more positive attitudes towards faking and instances of past faking behaviours were positively correlated with intentions to fake, while higher scores on subjective norms and moral obligations were negatively correlated with intentions to fake.

Hypothesis 1b - Attitudes toward faking, subjective norms, moral obligations, and past behaviour will be significantly correlated with intention to fake self-reported and intelligence assessments but not with the two BBAs The same analyses were ran on the data, only this time the correlations were calculated with the intentions to fake each of the four assessment formats. The same pattern emerged from the analyses yet with less strong correlations compared to those observed with the overall intentions to fake.

The hypothesis was only partially supported as the individual factors were cor-

related with all the assessment formats without any pattern suggesting any systematic differences between BBA and non-BBA formats. Despite some overall magnitude difference between the two phases, the pattern of results was comparable between them.

Table 3.3 reports the bivariate correlations observed between the individual factors and intentions to fake - both at the overall and format level - across the two phases. It is worth noting that the correlations at the overall level are consistently stronger compared to the same correlations at the format level.

Table 3.3: Bivariate correlations between relevant individual factors (Attitudes, Subjective Norms, Moral Obligations, Past Behaviour, and HEXACO factors) and intention to fake - overall and by assessment format (Correlations significant at the .01**, .01* , and .05 levels).

	Overall		**BBA-DMT**		**BBA-Puzzle**		**Self-Report**		**Intelligence**	
	Ph1	*Ph2*	*Ph1*	*Ph2*	*Ph1*	*Ph2*	*Ph1*	*Ph2*	*Ph1*	*Ph2*
Attitudes	.716**	.625**	.391**	.394**	.542**	.373**	.487**	.488*	.495**	.383**
Sub Norms	-.349**	-.377**	-.203**	-.164*	-.235**	-.173*	-.214**	-.226**	-.258**	-.256**
Moral Obl	-.480**	-.701**	-.354**	-.558**	-.431**	-.435**	-.348**	-.507**	-.367**	-.488**
Past Beh	.754**	.438**	.501**	.253**	.667**	.214**	.478**	.286**	.594**	.344**
H	-.341**	-.339**	-.284**	-.298**	-.262**	-.289**	-.321**	-.254**	-.334**	-.359**
E	-.111 (ns)	-.143*	-.104 (ns)	-.105 (ns)	-.184**	-.082 (ns)	-.084 (ns)	.012 (ns)	-.105 (ns)	.051 (ns)
X	-.004 (ns)	-.111 (ns)	.054 (ns)	-.069 (ns)	.059 (ns)	-.050 (ns)	.073 (ns)	-.048 (ns)	.090 (ns)	-.118 (ns)
A	-.075 (ns)	-.097 (ns)	-.155 (ns)	-.108 (ns)	-.118 (ns)	-.065 (ns)	-.169**	-.058 (ns)	-.148*	-.060 (ns)
C	-.270**	-.291**	-.199**	-.251**	-.168**	-.137 (ns)	-.175**	-.317**	-.190**	-.272**
O	.117 (ns)	.013 (ns)	.056 (ns)	-.020 (ns)	.058 (ns)	.028 (ns)	.051 (ns)	.039 (ns)	.043 (ns)	.049 (ns)

Hypothesis 1c - Some HEXACO scores will be significantly correlated with general intentions to fake. More specifically, Honesty-Humility, Agreeableness and Conscientiousness will be negatively correlated with intentions to fake, whilst no significant correlations with other HEXACO scores are expected A series of bivariate correlations between intention to fake and all the HEXACO scores was run on the data to test the hypothesis and also to explore whether other relationships emerged with the other HEXACO factors. These showed that only some of the HEXACO factors correlated significantly with intentions to fake, and this was true for both sub-samples, albeit with one small inconsistency.

In both phases, as predicted, Honesty-Humility was negatively correlated with intentions to fake (r=-.341 and r=-.339 respectively, both at the .001 level), and so was Conscientiousness (r=-.270 and r=-.291 respectively, both at the .001 level), whilst Extraversion was not. Against expectations, Emotionality showed a sig-

nificant correlation with intentions to fake, albeit only in the phase 2 sub-sample (r=-.143, $p<.05$), and Agreeableness did not.

The hypothesis is therefore only partially supported, showing that two of the HEXACO's six factors, Honesty-Humility and Conscientiousness, are significantly and consistently correlated with intentions to fake, whilst the others are not.

Hypothesis 1d - The HEXACO scores that are significantly correlated with intention to fake will show this effect only in the context of self-reported and intelligence assessments but not with BBAs The same analyses were ran on the data, only this time the correlations were calculated with the intentions to fake each of the four assessment formats. The same pattern emerged from the analyses, yet with generally less strong correlations compared to those observed with the overall intentions to fake, and some small inconsistencies between formats and phases.

Similarly, the hypothesis was only partially supported as only Honesty-Humility consistently correlated with intention to fake across formats and phases, whilst Openness and Extraversion consistently did not.

The factors Emotionality and Agreeableness were almost consistently non significantly correlated with intentions to fake, however, some exceptions to this emerged whereby for BBA Puzzle in the phase 1 sub-sample a significant negative correlation was observed between Emotionality and intentions to fake(r=-.184, $p<.001$), and in the phase 1 sample, two negatively correlations were detected between Agreeableness and intentions to fake in Self-Report (r=-.169, $p<.001$) and Intelligence tests (r=-.148, $p<.001$).

Finally, Conscientiousness was almost universally negatively correlated with intentions to fake, however, one non significant correlation was detected between Conscientiousness and intention to fake BBA-Puzzles in the phase 2 sub-sample.

Table 3.3 reports the bivariate correlations observed between the HEXACO factors and intentions to fake - both at the overall and format level - across the two phases.

Hypothesis 1 - summary: Taken together these results show that the relationships between attitude towards faking, subjective norms, moral obligations, past

behaviour, the HEXACO, and faking intentions are generally consistent across assessment formats with the exception of some minor differences which are most likely due to sampling.

The hypothesis that intentions to fake BBAs were not associated with individual factors was not supported, suggesting that the same behaviours, attitudes, values, and personality traits underlie intentions to fake independent of assessment formats.

3.1.2.2 Hypothesis 2 - Does an individual's Perceived Behavioural Control over an assessment format predicts their intentions to fake it?

The second hypothesis investigated the relationship between the perceived behavioural control afforded by the four different formats and the participants' intentions to fake them. Perceived behavioural control is intrinsic to the specific format and therefore it was not possible to assess an overall baseline for this measure.

It is predicted that this relationship will be consistent for all the formats, but that both PBC and intentions to fakes will be significantly lower for BBAs.

Hypothesis 2a - Perceived Behavioural Control will be significantly correlated with intentions to fake all assessment formats in equal measures A series of bivariate correlations investigated whether each of the four assessment format's perceived behavioural control and the participants' intention to fake it were significantly correlated.

As predicted, the two variables were significantly correlated across all formats and this was consistent between the two phases.

Upon further investigation, a slight inconsistency was detected, whilst the PBC/intentions correlations for BBA-Puzzles and Intelligence Tests were consistent between phases (r=.424 vs r=.499, $p<.001$; and r=.435 vs r=.438, $p<.001$), they were less similar for BBA-DMT and Intelligence tests (r=.381 vs r=.548, $p<.001$; and r=.373 vs r=.591, $p<.001$).

Nevertheless, the correlations between perceived behavioural control and in-

tentions to fake were all significant for each of the four formats, and did not vary too drastically between them. The observed difference in the strength of coefficients is most likely due to sampling.

Hypothesis 2b - Participants will perceive significantly less behavioural control over BBAs compared to self-reported and intelligence assessments Whilst a consistent correlations between perceived behavioural control and intentions to fake across formats is providing support for the integrated model of faking as a whole, this might conceal significant differences between formats on the two variables.

To assess this, two one-way ANOVAs investigated the differences in perceived behavioural control and faking intentions between the two BBAs and the two non-BBA assessment formats.

The first one-way ANOVA was conducted to compare the degree of behavioural control participants perceived in BBAs and non-BBA assessment formats. There was a significant effect of format on perceived behavioural control for the four conditions, and this was true for phase 1 [$F(3, 1036) = 94.078$, $p<.001$] and phase 2 [$F(3, 796) = 79.791$, $p<.001$].

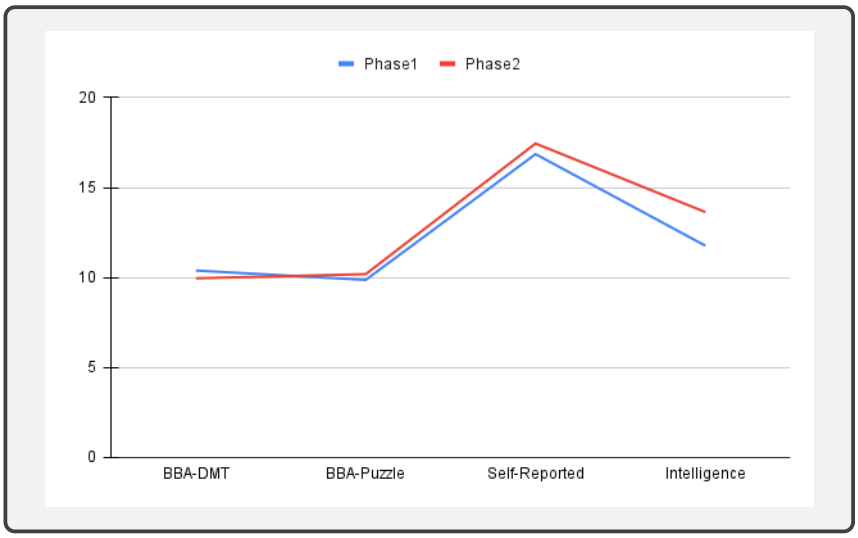


Figure 3.3: Level of PBC across formats over the two phases

Post hoc comparisons using the Bonferroni correction indicated that, in phase 1, the mean PBC score of the two BBAs [M (BBA-DMT) = 10.39, SD = 5.30; and M (BBA-Puzzle) = 9.87, SD = 3.99] were not significantly different from one another,

but they were each significantly lower than both Self-Reported (M = 16.86, SD = 6.04) and Intelligence Tests (M = 11.77, SD = 5.68). This supported the hypotheses.

Furthermore, it was observed that self-reported measures afforded participants the significantly highest degree of perceived behavioural control across all formats.

The same pattern was observed in phase 2 whereby BBA-DM (M = 9.96, SD = 5.00) and BBA-Puzzle (M = 10.19, SD = 5.29) did not differ from one another in terms of the PBC they afforded to participants but they differed from both non-BBA formats, showing significantly lower levels of PBC compared to both Self-Report (M = 17.45, SD = 5.87) and Intelligence Test (M = 13.64, SD = 6.08). Once again, Self-Reported assessment afforded significantly more PBC to participants than any other format.

Hypothesis 2c - Participants will indicate that they are less intentioned to fake BBAs than self-reported and intelligence assessments Similarly, the second one-way between subjects ANOVA was conducted to compare the degree of intentions participants declared to have to fake BBAs and non-BBA assessment formats. There was a significant effect of format on intentions to fake for the four conditions, and this was true for phase 1 [F(3, 1036) = 16.851, p<.001] and phase 2 [F(3, 796) = 14.421, p <.001].

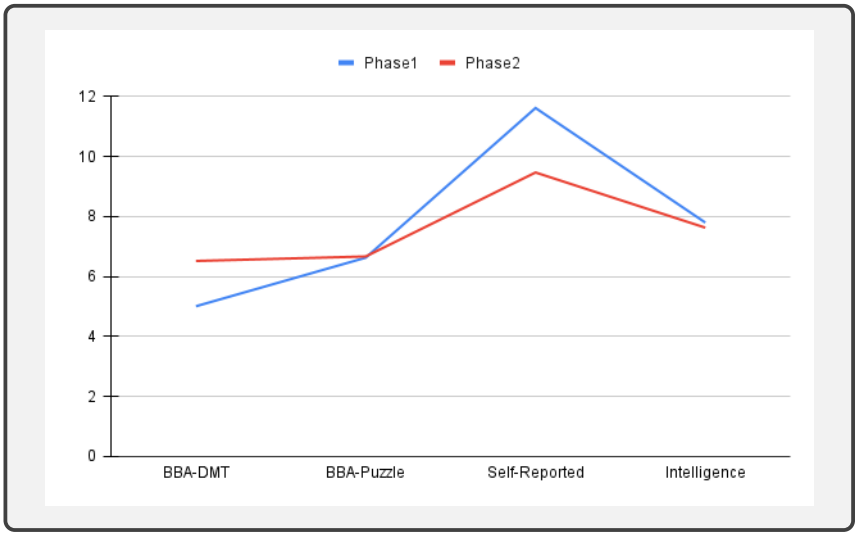


Figure 3.4: Level of intentions to fake across formats over the two phases

Post hoc comparisons using the Bonferroni correction indicated that, in phase

1, the mean intention to fake the two BBAs [M (BBA-DMT) = 5.01, SD = 4.75; and M (BBA-Puzzle) = 6.63, SD = 5.56] was not significantly different between one another, but in both cases it was significantly lower than Self-Reported Tests (M = 11.62, SD = 6.66). In the case of intention to fake Intelligence Tests (M = 7.80, SD = 6.23), those were significantly higher than BBA-DMT but not significantly higher than BBA-Puzzle. Consistently, Self-Reported test were the format for which participants had the highest faking intentions. This partially supported the hypotheses.

A similar pattern was observed in phase 2 whereby BBA-DM (M = 6.52, SD = 4.60) and BBA-Puzzle (M = 6.67, SD = 4.65) did not differ from one another in terms of how much participants intended to fake them but the mean intention to fake Self-Report assessment (M = 9.47, SD = 5.91) was significantly higher than both. Intentions to fake Intelligence Test (M = 7.63, SD = 4.98), on the other hand, were significantly higher than BBA-Puzzles but not significantly higher than BBA-DMT. Once again, Self-Report assessment was the format participants had the most intentions to fake.

Hypothesis 2 - summary: Taken together, these results partially supported hypothesis 2. There is a consistent relationship between perceived behavioural control and intentions to fake, however, the perceived behavioural control afforded by the assessment vary as a function of format and subsequently do intentions to fake. Whilst it is true that BBAs afford less perceived behavioural control than non-BBA formats, this only fully translates into significantly lower levels of intention to fake when compared to self-reported measures, whereas the comparison with Intelligence Test is less clear and consistent, despite a significant difference in perceived behavioural control between BBAs and the format.

3.1.2.3 Hypothesis 3 - Do faking warnings reduce intentions to fake?

One proposed situational moderator of the relationship between individual and assessment factors and intentions to fake is the presence of faking warnings. Phase 2 was designed especially to test Hypothesis 3, and to asses whether the addition of a

faking warning might reduce intentions to fake through the reduction of perceived behavioural control, and whereas this effect is consistent across formats.

Hypothesis 3a - Faking warnings will significantly reduce perceived behavioural control and intentions to fake across all assessment formats. To test this hypothesis, two one-way ANOVAs were performed on the data.

The first one-way ANOVA was conducted to compare the average degree of behavioural control participants perceived across four deterrent conditions: no deterrents, proctored testing, fake detecting algorithms, and warning about the use of online guides. There was a significant effect of deterrent on perceived behavioural control [F(3, 1036) = 26.822, $p<.001$].

Post hoc comparisons using the Bonferroni correction indicated that, the mean PBC score of the no deterrent condition (M = 12.22, SD = 3.72) was significantly higher than the mean PBC score in the proctored learning (M = 10.09, SD = 3.32) and detection algorithm (M = 9.95, SD = 3.63) conditions, but not significantly higher than the condition encompassing the online guides warning (M = 12.05, SD = 4.48), suggesting that this type of deterrent has no effect in reducing PBC compared to not having a faking warning when PBC data from all assessment formats was collated.

In terms of how the different deterrents differed from one another in terms of impacting PBC, proctored testing and detection algorithms did not differ from each other, whereas they both had a larger effect than online guides warning.

The second one-way ANOVA was conducted to compare the average degree of intentions to fake in the four deterrent conditions, and it showed a significant effect of deterrent on intentions to fake [F(3, 1036) = 28.426, $p<.001$].

Post hoc comparisons using the Bonferroni correction indicated that, the mean intention to fake in the no deterrent condition (M = 17.86, SD = 20.97) was significantly higher than the mean intention to fake in the proctored learning (M = 9.79, SD = 12.78), detection algorithm (M = 5.87, SD = 7.94), and the online guides warning (M = 13.90, SD = 17.70) conditions, suggesting that all types of deterrent are effective at reducing intentions to fake compared to not having a faking warning.

In terms of how the different deterrents compare to each other, detection algorithms significantly reduced intentions to fake more than any other warning, and online guides reduced it significantly the least compared to the others. Proctored testing reduced intentions to fake significantly more than warnings about online guides but significantly less than detection algorithms.

It is also hypothesised that the detection algorithms will have different degrees of impact in reducing intentions to fake depending on the format, the following sub-hypotheses are formulated to assess whether this is true.

Hypothesis 3b - Intentions to fake BBAs will be equally reduced by all the deterrents. Considering the already lower levels of perceived behavioural control over BBAs, it is expected that faking warnings will have a minimal effect on reducing intentions to fake, as participants might struggle to envision faking strategies for BBAs. For this, it can be expected that any deterred will be perceived equally as in/effective as the other.

A series of t-tests were used to test this hypothesis, and the results suggested that for BBA-DM, only proctored testing [M= 6.41 SD= 13.86; t(259)=5.084, $p<.001$] and detection algorithms [M=7.21 SD= 15.19; t(259)=4.341, $p<.001$] had a significant effect at reducing intentions to fake from the no deterrents baseline (M= 11.53 SD=20.63). Both significant effects were small, with Cohen's d values of .315 and .269 respectively.

For BBA-Puzzle, the same analyses were employed and the results showed that all the deterrents significantly reduced the intentions to fake the assessment. However, while proctored testing [M= 5.85 SD= 13.90; t(259)=7.660, $p<.001$] and detecting algorithms [M= 6.29 SD= 12.38; t(259)=7.017, $p<.001$] had a small effect sizes (Cohen's d =.475 and .385), the effect of online guides [M= 11.73 SD= 18.82; t(259)=2.854, $p<.005$] compared to baseline (M= 15.25 SD=24.19) was below the threshold for a small effect size (Cohen's d = .177).

This shows a similar pattern between the two BBA formats, with reduction of intentions to fake of small effects sizes for proctored testing and detection algorithms but not consistently for online guides. The hypothesis is therefore only

partially supported.

Hypothesis 3c - Detection algorithms will reduce intentions to fake self-reported assessments more than the other deterrents. As faking in self-reported assessments typically involves lying, the deterrent expected to have the largest effect on reducing intentions to fake is faking detection algorithms. This is because proctored testing is not meant to detect lying - as it is not evident to an observer - and warnings about online guides might be perceived as irrelevant in this context.

To test this assumptions, a series of t-test were used to compare baseline levels of intentions to fake self reported assessments in absence of deterrents (M= 26.73 SD=28.96) with intentions to fake the same type of assessment when proctored testing (M=19.81, SD= 24.75), detection algorithms (M=15.75, SD=22.77), and warning against the use of online guides (M=20.20, SD= 25.08) were employed.

All the t-tests showed a significant reduction of intentions to fake for each of the three deterrents, with detection algorithms having a larger effect [t(259)=9.189, $p<.001$; Cohen's d= .570] compared to proctored testing [t(259)=5.696, $p<.001$; Cohen's d= .353] and warnings against online guides [t(259)=5.493, $p<.001$; Cohen's d= .341], thereby supporting the hypothesis.

Hypothesis 3d - Proctored testing will reduce intentions to fake Intelligence Assessments more than the other deterrents. Considering that faking intelligence tests might rely more heavily on accessing information than other assessment formats, it is expected that proctored testing will reduce intentions to fake further than other faking warnings because this type of deterrent is centred on observing and preventing this type of behaviour specifically.

To test this assumptions, in line with previous analyses, a series of t-tests were used to compare baseline levels of intentions to fake Intelligence Tests in absence of deterrents (M= 17.93 SD=27.09) with intentions to fake the same type of assessment when proctored testing (M=7.10, SD= 14.87), detection algorithms (M=9.34, SD=17.57), and warning against the use of online guides (M=12.43, SD= 20.10) were employed.

All the t-tests showed a significant reduction of intentions to fake for each of

Table 3.4: Effect size of the reduction of intentions to fake each deterrent caused across the four assessment formats compared to no deterrents. **The medium effect size in of proctored testing on intentions to fake Intelligence test is borderline Medium*

	BBA-DM	**BBA-Puzzle**	**Self-Report**	**Intelligence**
Proctored	Small	Small	Small	Medium*
Algorithm	Small	Small	Medium	Small
Guides	-	-	Small	Small

the three deterrents, with proctored testing having a larger effect [t(259)=7.840, $p<.001$; Cohen's d= .486] compared to detection algorithms [t(259)=6.316, $p<.001$; Cohen's d= .392] and warnings against online guides [t(259)=4.217, $p<.001$; Cohen's d= .262], thereby supporting the hypothesis.

Hypothesis 3 - summary: Taken together, the data explored in hypothesis 3 shows that faking warnings do have an effect on reducing assessment faking intentions in general but also when investigating different assessment formats specifically. This effect is observed with different levels of intensity depending on the assessment format, its baseline levels of PBC and faking intentions, and the deterrent, suggesting that faking warnings and deterrents are the most effective when systematically aligned to the format of the assessment for which they attempt to deter faking. The effect sizes of all formats and warnings combinations are reported in Table 3.4

Nevertheless, faking intentions still vary widely between assessment formats even when format-specific deterrents are present. Table 3.5 shows how the different assessment formats and faking warning types ranked against each other as a function of the level of intentions to fake to which they are linked.

It is worth noting that self-reported measures and warnings against the use of online guides seems to be the least efficient format and deterrent in preventing faking, whereas BBAs, proctored testing and faking detection algorithms are linked to the lowest levels of faking intentions.

Table 3.5: Format/deterrent combinations ranked by corresponding intentions to fake (M) - Q1=7.155, Q2= 11.63, Q3=16.84.

Format	Faking Warning	Intentions to Fake	Quartile
BBA-Puzzle	Proctored Testing	5.85	1
BBA-Puzzle	Algorithm	6.29	1
BBA-DMT	Proctored Testing	6.41	1
Intelligence	Proctored Testing	7.10	1
BBA-DMT	Algorithm	7.21	2
Intelligence	Algorithm	9.30	2
BBA-DMT	Online Guides	11.22	2
BBA-DMT	No Warning	11.52	2
BBA-Puzzle	Online Guides	11.74	3
Intelligence	Online Guides	12.42	3
BBA-Puzzle	No Warning	15.25	3
Self-Report	Algorithm	15.75	3
Intelligence	No Warning	17.93	4
Self-Report	Proctored Testing	19.81	4
Self-Report	Online Guides	20.20	4
Self-Report	No Warning	26.73	4

3.1.3 Discussion

This study focused on the interface between the Theory of Planned Behaviour and Perceived Behavioural Control that underlies the Integrated Model of faking, with the aim of establishing whether BBAs might be less of a target of intentions to fake. More specifically, the study investigated how the TPB and PBC replicated across different assessment formats, and whether any observed differences in the relationship between the factors included in the models or the relative differences of the factors' strengths/levels across formats resulted in lower levels of intentions to fake BBAs compared to other assessment formats. Overall, the results of the study mostly support this hypothesis, showing that participants have generally lower intentions to fake BBAs compared to non-BBA formats, albeit with some ambiguities with intelligence tests. Interestingly, BBAs are linked to lower levels of intentions to fake compared to self-reported assessment when faking deterrents are taken into consideration, meaning that participants have stronger intentions to fake self-reported assessments when a faking detection warning is present compared to any BBA with no faking detection warning. Not all hypotheses in the study were

entirely supported. Despite the overall support of the core research question, several unexpected results emerged. The first hypothesis expected the relationship between individual factors and intentions to fake to not be significant for BBAs but only for non-BBAs formats. This was because the expected lower level of intentions to fake BBAs may have caused a floor effect, thereby reducing the variance too much for a significant correlation to emerge. In other words, the rationale behind hypothesis 1 was that, whilst it is expected - and demonstrated - that people with certain values, attitudes, traits, and having engaged in past faking behaviour will be more likely to fake assessments, BBAs possess some characteristics as a format able to transcend the effect of individual differences on faking intentions. The data revealed that this was not the case, and that individual-level factors constituting the TPB model and the HEXACO factors Humility and Conscientiousness were consistently correlated with intentions to fake across all formats, with only some inconsistencies between coefficient strengths observed between study phases rather than between formats. The results of hypothesis 1 did not preclude lower intentions to fake BBAs compared to non-BBA formats, which was tested in hypothesis 2. As expected, a consistent relationship between perceived behavioural control and intentions to fake was observed, with the perceived behavioural control afforded by the assessment varying as a function of format, and subsequently affecting the intentions to fake that format. The rationale behind this hypothesis was that PBC, being a byproduct of assessment characteristics related to faking scope, was expected to be lower for BBAs and to drive lower faking intentions. While BBAs afforded less perceived behavioural control than both non-BBA formats, this only converted into significantly lower levels of intentions to fake when compared to self-reported measures. The comparison with intelligence tests showed less consistency: despite a significant difference in perceived behavioural control between BBAs and intelligence tests in the hypothesised direction, only one BBA format in each phase showed significantly lower faking intentions compared to intelligence tests (BBA-DMT in phase one, BBA-Puzzles in phase two). Given that the two BBAs did not differ from each other, and the differences between formats were small, the safest inter-

pretation is that BBAs and intelligence tests do not significantly differ in terms of intentions to fake. The final hypothesis, examined only in phase 2, investigated the potential effect of faking deterrents on perceived behavioural control and intentions to fake. At this baseline, intentions to fake BBAs and intelligence tests were lower than for self-reported measures, with only BBA-DMT showing lower intentions to fake compared to intelligence tests. The results demonstrated that both perceived behavioural control and intentions to fake were lowered by faking deterrents, though not entirely consistently. While perceived behavioural control was only affected by faking detecting algorithms and proctored testing, all faking warnings successfully deterred intentions to fake. Because the relationship between perceived behavioural control and intentions was shown to be consistent across formats, it was hypothesised that deterrents targeting the most likely faking behaviour for each assessment type would show different effects aligned with each format's scope for faking. The effect varied in intensity depending on the assessment format and deterrent type, with medium-sized effects observed when the deterrent matched the format, and small or trivial effects in other cases. Notably, BBAs showed the smallest sensitivity to deterrents but maintained consistently lower intentions to fake across conditions. When the faking intentions of the four assessment formats with or without deterrents were ranked, BBAs dominated the top positions, taking up the first three places and sharing the first quartile only with "intelligence test-proctored test" and the second quartile with "intelligence test-algorithms." Self-reported measures consistently occupied the lowest quartile, except for intelligence tests with no deterrents. Intelligence tests were featured once per quartile, reflecting the varying effectiveness of different deterrents. The study revealed an interesting pattern regarding online guides - the only potential method for improving BBA scores. Warnings against their use had no effect on reducing faking intentions, but this may be inconsequential since using guides for BBAs is more likely to deteriorate scores than improve them. This is due to the complexity of personality scoring models, which involve intricate interactions of data from multiple assessment tasks, where task performance isn't uniformly desirable across traits. Additionally, the reliance

on guides tends to affect response speed and timing variance, which are crucial metrics for this type of assessment. The scoring models are based on proprietary models of individual differences that are anyway too complex to translate into deliberately reproducible task behaviour, even if such information were accessible. The primary concern about assessment faking typically centers on score improvement leading to unqualified individuals securing positions at the expense of qualified candidates. However, if the only deterrent relevant to BBAs' potential vulnerability proves ineffective at reducing faking intentions, but the attempted faking behavior actually reduces scores and opportunities, this ineffectiveness becomes less problematic. In conclusion, this study provides evidence that people demonstrate lower intentions to fake BBAs compared to other assessment formats, even in the presence of faking warnings and deterrents. This evidence supports the core thesis idea that implementing BBAs in assessment contexts may naturally reduce faking behavior.

3.1.3.1 Implications

The results of this study have some promising implications for practice.

First, as intentions to fake can be assumed to be quite substantial in high stakes situations such as candidate assessment, this study suggests that using BBAs might be able to mitigate that. Most importantly, by reducing people's intentions to fake, it should be expected that BBAs would also reduce faking attempts.

Second, the mechanism through which BBAs reduces intentions in PBC. Whilst this is nothing new, this study is the first comparing assessment formats in terms of how much PBC individual experience over them, and the results demonstrate how this can widely differ between assessment format with tangible effects on faking intentions. One important point about PBC is that, by definition, is a perception, and for this it might be easier to manipulate compared to redesigning a whole assessment. Whilst this thesis doesn't necessarily advocate for deception, there might be ways in which psychometric companies can frame their assessment to trigger low levels of PBC. On this note, the results of phase 2 show that with self-reported measures, intentions to fake were significantly reduced by a warning claiming to use a faking detection algorithm. This suggests that candidates may just

need to *perceive* less control, even if this does not reflect the control they actually have over an assessment.

Third, an interesting implication of this study comes from the results showing that warnings against online guides don't significantly reduce intentions to fake BBAs, and they are generally the worst performing deterrent across all assessment formats. Whilst small effect sizes in reducing intentions to fake BBAs might be partially due to a low baseline, warnings against online guides could be expected to have at least the same effect as other deterrents. This deterrent clearly states that online guides would lead to worse results, and it is true for BBAs. Online guides demonstrating how to complete a BBAs are entirely focused on maximising task performance, but this does not equate to better results as personality models are not so straightforward and sometimes "require" mistakes. Study 3 (section 4.3.4) features a thorough explanation of why this is the case from a methodological perspective, but the interesting fact in the context of this study is that participants most likely did not believe the deterrent. Future studies should investigate how candidates process this type of information to fully understand why they would discard a warning flagging a potential risk to their performance but they would not the information contained in an online guide. Nevertheless, this study shows that this type of warning might be not useful at all, and that at least some candidate may try to mimic the behaviours described in the online guide, with potentially detrimental effects on their psychometric scores.

3.1.3.2 Limitations

This study had some limitations. As per most studies in this field, the observed intentions to fake are entirely hypothetical. As far as the participants were instructed to think about how they would feel upon a job application, there is a very high chance that their motivational processes were not only quantitatively but also qualitatively different to what they would have been in high stakes. This means that their intentions could have been different, and that would also be true of the effect of the deterrent, and the relationship between them too. That is, candidates might be more motivated to fake but might also be more cautious about deterrents. Considering

that the field evidence on this is mixed, this study might also carry some of that qualitative discrepancy in its results.

The two study phases showed some significant differences despite being identical, suggesting that some confounding variables not considered in the study could have been at play. Considering that the scales showed good reliability in both studies and that all factors including sampling, recruitment, materials, and study design were equal, it would be difficult to infer a cause without erring on the side of guesstimation.

This study missed a good opportunity to investigate *why* people perceive differing levels of PBC across assessment types. As far as this study adds to the literature with a direct comparison of PBC and faking intentions between different assessment formats, the reasons underlying this effect remain unknown and can only be inferred theoretically. Future studies should investigate this directly to understand the mechanisms that lead to differences in PCB. Revealing these would add much value to practice in that the assessment features responsible for deterring intentions could be identified and leveraged elsewhere.

3.1.3.3 Conclusion

Overall, this study shows that the low levels of intentions to fake observed in BBAs are not the result of an interaction with individual-level antecedents of faking intentions which might suggest that a certain type of individuals could be more inclined than other to fake a specific assessment format, on the contrary, the results show that the lower intentions to fake are simply due to lower levels of PBC over the format. This highlights the possibility that people might simply feel that they *don't know how to fake BBAs*, and this might be the only reason why their intentions are lower.

The next chapter taps on the *Ability* block of the Integrated Model of Faking, and investigates if this assumption is true through two studies exploring the perceived and actual scope BBAs offer for faking. The first study explores faking strategies across formats, which is thematically linked to the amount of control individual feel over an assessment, and the second study explores whether in reality BBAs do offer much control over them by taking a close look at the type of data

and repose format used in BBAs.

Chapter 4

ABILITY

The previous chapter showed that the key reason why people have lower intention to fake BBAs compared to other types of assessment is a lower level of Perceived Behavioural Control (PBC). This chapter adds more depth to these results by investigating two dimensions of ability to fake: knowing how to fake, which is arguably linked to PBC, and possessing the objective ability to fake, which is less directly yet still linked to PBC. The results of the studies in this chapter should be useful to understand more about the way in which BBAs act on PBC and whether this perception is substantiated by actual constraints in the assessment.

PBC is a crucial antecedent of intentions but in fact it actually sits at the cusp between intentions and ability. It is difficult to uncouple subjective ability to fake from PBC, and *subjective* ability to fake from *objective* ability to fake, as one feeds into the other and vice versa. Furthermore, Ajzen and Madden (1986) pointed out that for PBC to predict behaviour, that is, to be accurate, a good degree of alignment between perceived and actual control over a behaviour must exist. In other words, PBC needs to be a valid perception of objective degree of control over an action in order to be associated with behavioural outcomes, and not just an ephemeral perception.

Whilst PBC is necessary to leverage actual control, if control is limited to perceptions, then as much as it might predict intentions, it won't predict behavioural execution. Sheeran, Trafimow, and Armitage (2003) proposed a proxy measure of actual control (PMAC) to address this accuracy concern by assessing participants'

actual control *post-behaviour*, whereby pre- and post-behaviour ratings of control might be used as a proxy to infer the accuracy of PBC in a given context. This measure implies testing volitional control twice, with one of the assessments occurring *post factum*, irreparably limiting its utility in predicting behaviour.

An alternative way of establishing whether PBC might be a realistic marker of actual control would be to ask candidates to describe how they would fake a given assessment, and evaluate whether the strategy they propose could be successful based on the opportunity to fake that the assessment offers. Whilst this might not be a perfect metric of accuracy for PBC, it will certainly narrow the gap between perceived and actual control by anchoring PBC to the control constraints and opportunities imposed and afforded by the assessment. This is because it would force the individual to process the situation at a deeper level to identify vulnerabilities in the assessment. Candidates would have to reflect on possible behaviours to include in their strategy and then evaluate the feasibility of the strategy in the context of what they believe they can do, and this should equate to a more ”educated” PBC.

Knowing how to fake an assessment can be considered the minimum requirement for ability to fake, but also an ability in itself: the ability to identify the right vulnerability in an assessment and devising a viable strategy to take advantage of it. This chapter explores both the subjective and the objective sides of ability to fake a BBA. The first being what people are able to suggest to fake it, and the second being what people are actually able to do to fake it.

4.1 Ability to Fake - General Introduction

The second block of the Integrated Model of Faking features the ability to fake an assessment, and is believed to mediate the relationship between intentions and behaviour. Whilst intentions is the trigger of behaviour, ability is the facilitator. Without ability, intentions, no matter how strong, can have no effect on behaviour. It follows that when an assessment does not offer candidates enough perceived and actual ability to fake it, it is highly unlikely that it will be faked, even if the candidates are highly motivated to do so. This means that ability could be the best target

for interventions aimed at preventing faking as it would be easier to address the intrinsic and extrinsic properties of the assessment than to control the antecedents of intentions and motivation.

The theoretical and experimental role of ability in faking is unpacked and deconstructed to a much deeper level in the VIE model of faking (Ellignson & McFarland, 2011) compared to the Integrated Model of Faking (McFarland & Ryan, 2006) where it is treated more as a monolithic factor rather than as a rich multifaceted source of individual-situational interactions.

First, in the VIE model of faking, a clear link between the concept of expectancy and ability is made which is not as obvious in the Integrated Model of Faking between PBC and ability, and second, expectancy is given a much more crucial role compared to PBC by proposing a multiplicative relationship between the antecedent of motivation to fake whereby if any of the three antecedents (valence, instrumentality, and expectancy) is too close to zero, intentions to fake are not expected to emerge.

The VIE model divides ability to fake into two strands: one focusing on the individual and the other focusing on the situation.

Whilst the authors mention several individual differences such as cognitive ability, emotional intelligence, and self-monitoring as potential determinants of ability to fake, they also point out that a key mechanism bridging those differences and ability to fake is being better at working out how to fake by making sense of situational requirements and how those translate into assessment behaviours. In other words, some people are more able than others to gauge what they need to do to fake assessments. However, when defining the situational characteristics associated to ability to fake, the authors make an important distinction between expectancy and ability - that is, expectancy is about conscious awareness of one's ability to fake (whether realistic or not), but ability is determined by situational factors of which the candidate is not aware. This highlights a subtle yet important link between the individual factors underlying ability to fake and expectancy which blurs the line between proximal motivational factors and ability. The individual factors underlying

ability are believed to warrant ability to fake through more reasonable perceptions of expectancy, that is, expectancy in people possessing higher levels of the ability-boosting individual differences identified in the model is more tightly aligned to the actual opportunities they have to fake compared to people who don't possess those abilities. Yet the fact that they are aware of this knowledge makes their ability a motivational factor - expectancy - more than an ability, of which by definition they should be unaware.

That said, the authors define the situational characteristics hampering faking much less ambiguously as being beyond the grasp of candidates; they mainly focus on scoring techniques, both in the sense of what constitutes a desirable response, and in the sense of identifying faking through diagnostic strategies. Either way, the authors don't consider an assessment's features, such as response options, to be relevant to ability. This is perhaps due to the fact that assessment features other than scoring would be linked to expectancy as they overtly signal the level of difficulty of the assessment.

However, it is unlikely that this is so clear cut for all types of assessment, as some formats might appear relatively easy to fake but offer no scope for faking, and this could be true in a way that is not related to scoring methods or faking detecting diagnostics. This, once again, challenges boundaries between expectancy and ability described in the model. This time by questioning whether there might be a discrepancy between how candidates perceive certain assessment features and what those features mean in reality for faking.

With this in mind, the simplicity of the Integrated Model of Faking pays dividends by not discriminating between antecedents of ability on the basis of the degree to which candidates are aware of them. Whilst no deliberate link exists between PBC and ability, the model nests the knowledge of constructs being measured, self-monitoring, item transparency, and opportunity to fake under ability. In this chapter, the first study uses a proxy measure of the *perceived* item transparency, knowledge of what is being measures, and the opportunities to fake that BBAs offer, whereas the second study investigate those directly on the assessment itself.

The first study (Study 2) focuses on the individual, and it investigates at a very basics prerequisite of ability to fake: knowing *how* to do that, that is, having a clear strategy in mind on how to fake a given assessment. Whilst this could arguably fall within the PBC/expectancy realm as much as within ability, the way in which the study is designed is markedly within the ability space. The study investigates what strategies participants *are able* to articulate to describe how they would fake four different assessment formats. This is conceived as a subjective yet relevant proxy of their ability to fake a specific assessment, which is in stark contrast with the decontextualised nature of PBC and expectancy at the basis of intentions and motivation. The expectation is that those strategies will differ across formats as a function of the varying levels of opportunity to fake and difficulty that they signal to candidates, and that participants will find it harder to advance viable strategies for the BBAs than for other formats. Being a within-participants design, this study mitigates the effect of the individual antecedents of ability identifies in the VIE models, and will be able to attribute the variance in ability to describe faking strategies to the assessment format.

The second study (Study 3) focuses on BBAs as a format, more specifically, it focuses on how likely it is that potential candidates might be able to improve their scores if they wanted and knew how to. The core focus of this study is the *type* of tasks, response options, and data used in BBA models of individual differences, which are believed to severely limit the scope for faking. By deconstructing a model of Extraversion and examining the nature of the variables it contains, the study investigates how much of of the model's variance is explained by ability-based response options, which by definition cannot be deliberately inflated, compared to response options that lend themselves to faking both in terms of scope and difficulty.

It is not possible to completely isolate these two studies from PBC and expectancy, and for this, from intentions. Most specifically, knowing how to fake an assessment is fundamental for perceiving behavioural control over it and develop perceptions of high expectancy, but it is also necessary for people to be perceive the ability to fake in order to act on their objective ability to fake, and this goes

well beyond intentions. The blurred distinction between perception of self-efficacy (i.e.: knowing how to fake an assessment, which is the core of the first study) and controllability (i.e: the degree to which faking is up to the person which is the core of the second study) echoes the sentiment of Aizen's (2002) paper on the conceptual and methodological ambiguities of PBC. Further adding to this complexity, Trafimov, Sheeran, Conner, and Finlay (2002) proposed a multidimensional view of PBC whereby perceived *control* and perceived *difficulty* are related yet distinct component of overall PBC, with perceived difficulty being the strongest predictor of behaviour between the two. This makes the difficulty of an assessment hard to pin neatly to intentions or ability from a theoretical perspective, although from a practical perspective, difficulty should be categorised within ability because of the direct link between the two.

It is beyond the scope of this thesis to successfully navigate this theoretical quagmire, and the focus of this chapter is to establish whether people might be able to fake BBAs based on their ability to define a strategy to fake them and the room for faking that BBAs offer. This is the core contribution to the faking literature that this thesis is aiming to make. In the interest of simplicity and parsimony, the focus of this chapter is ability to fake, and this is limited to four key assumptions: the individual needs to know how to fake, they need to feel able to fake (Study 2), and the assessment must provide scope for faking, and not be too difficult to fake (Study 3).

4.2 Study 2 - Perceived Ability To Fake.

"To him that will, ways are not wanting"

(George Herbert, 1640)

AKA

"Where there is a will there is a way".

With the potential life-changing opportunities a job can offer, people can be expected to actively wanting to find a way to enhance their chances to do well in a job application, and that often means finding a way to boost their assessment scores.

Ways of boosting assessment scores vary quite drastically between assessment types and situations, and successful cheating - or faking - is far from a one size fits all trick.

This is true from two perspectives: first, not everybody possesses the same ability to identify an assessment's vulnerabilities and leverage them successfully by modifying their behaviour according to the outcome they need to achieve. Christiansen, Wolcott-Burnam, Janovics, Burns, and Quirk (2005) for example developed the concept of "dispositional intelligence", a construct which encapsulates the contribution that multiple cognitive and non-cognitive characteristics make to a person's ability to link personality, behaviours, and situations. This type of ability is normally distributed and is linked to individual differences in the ability to fake assessment (Ellingson & McFarland, 2011). This means that some people are better than others at deciphering how they need to appear to increase their chances of being hired, what behaviours do they need to signal, and what they need to do to signal those behaviours successfully in the situation they have at hand. That is, some people are better at envisioning how to successfully fake an assessment.

The second reason why successful cheating is not a one size fits all is that assessment formats vary enormously between one another, and they all offer different combinations of faking opportunities and constraints. For example, Extraversion can be faked quite easily in a questionnaire but less so during a job interview, and coding skills can be easily faked with an assessment based on biodata, but not in a work sample. Similarly, lying can be successful in a self-reported personality measure but not in a knowledge test, and a calculator can help a numerical test but not a role play exercise. Furthermore, some assessment formats may signal their vulnerabilities more openly whilst others might be harder to decode, and some assessment formats are just objectively harder to fake than others.

This study seeks to eliminate the contribution of individual differences through a within-subject design in order to isolate the role of assessment format as the unique source of the observed differences in people's ability to advance viable faking strategies.

No study to date has investigated this if not largely indirectly. This study is a direct investigation of how assessment formats differ in how openly they signal their vulnerabilities to faking, and how well individuals are able to capture them and translate them into faking strategies. But, most importantly, this study aims to test if the above-mentioned proverb can withstand the challenge of new formats of psychometric assessments by asking:

Is there really always a way where there is a will?

4.2.1 Literature Review

The Integrated Model of faking (McFarland, & Ryan, 2006) lists several proximal factors of the ability to fake, such as item transparency, knowledge of what is being measured and opportunity to fake. Those are important in the context of this study in terms of their meta-cognitive value. That is, the ability to articulate a faking strategy is harboured, amongst other things not related to an assessment format, by the underlying belief that, first, it will be possible to develop the knowledge of what is being measured because the assessment format provides the information required to foster that knowledge, that, second, the items will be transparent enough to make faking relatively intuitive, and that, finally, the format will offer opportunities to produce responses as desired to achieve the intended psychometric outcomes.

In other words, faking strategies can be seen as a summary of the proximal factors of the ability to fake that has been contextualised to the idiosyncratic faking-relevant features of an assessment format.

4.2.1.1 Faking Strategies

Very few studies have focused on assessment faking strategies in the past, but the literature provides a good range of evidence from which to extract enough about them to inform this study. With very few exceptions, the faking strategy literature has focused on interviews, self-reported personality assessment, and ability assessment, but never on BBAs (or GBA).

There are many possible faking strategies that candidate can use, each more or less suitable for the context and type of assessment at hand. Josien and Broder-

ick (2013) for example counted 16 different techniques students use to cheat their way through a degree, and they summarised them in three categories: collaboration, plagiarism, and technology. That is, the active involvement of other people in completing the task instead of completing it personally, obtaining information to use in the task instead of producing novel content, and using technology to solve the task instead of solving it manually.

With regards to self-reported personality tests, the literature consistently focuses on dishonesty. Interestingly, when exploring faking strategies in self-reported measures, authors hardly ever feel the urge to clarify that those strategies imply lying and their studies focus on the specific patterns of distortion under scrutiny taking lying for granted and assumed. König, Mura, and Schmidt (2015) compared the incidence of extreme responses (i.e.: the endorsement of the minimum and maximum point of a rating scale) and midpoint responses (i.e.: responses gravitating around the middle of a rating scale) between participants led to believe that a hypothetical company was after candidates with either a ”strong” or a ”well-balanced” character. They found that participants were able to adapt their strategy to the demands of the job advertisement and engaged in the pattern of response that equated to the description they read.

A qualitative study investigating the cognitive processes leading to faking highlighted different perceptions of effectiveness between extreme and midpoint responding across participants, but generally all participants admitted to start by establishing what is expected of them (i.e.: what an organisation is looking for) in order to define a profile they intend to impersonate, identify which questions are measuring those qualities, and provide responses consistent to the intended profile (König, Merz, & Trauffer, 2012). This in-depth qualitative analysis of interview scripts show a very consistent pattern of well-thought and highly targeted lying in self-reported personality measures.

Dishonestly is not limited to personality, in fact, the Overclaiming Technique (OTC; Raubenheimer, 1925 in Ziegler, McCann, & Roberts, 2012) is a self-reported test meant to trick people into lying about their knowledge whereby individuals have

to indicate their familiarity with a series of items - some real and some made up. Whilst the OCT is not a faking strategy in itself, the test shows that people would just make things up to appear more knowledgeable. Dunlop, Bourdage, de Vries, McNeill, Jorritsma, Orchard, et al (2019) reported a study in which overclaiming was detected in a real applicant sample to firefighters positions. They found that overclaiming on job-relevant OTC items was correlated with other measures of faking but this did not happen when the items were job-irrelevant. This shows that overclaiming as a strategy can be very specific, deliberate, and targeted.

Ability tests, despite measuring maximal performance, are not immune to faking attempts, and several strategies have been identified in the literature. A study investigating the emergence and success of cheating in unproctored internet assessments of cognitive abilities found evidence of six different strategies (Bloemers, Oud, & Van Dam, 2016). They found that participants used a mix of calculators, dictionaries, internet, help from others, technical manipulation, and foreknowledge to improve their assessment scores, and that this equated to a large effect size (d=0.59) for overall test performance compared to a control group. Although, interestingly, they found that also three participants from the control group admitted to cheat so the effect size could be even higher. Most participants used several strategies, and the more strategies they used, the higher was their score. Overall, all the strategies employed by the participants to cheat on the ability assessments were aimed at working out the correct answer.

This shows differences between self-reported and ability measures on two grounds. First, in self-reported measures a *skill* is required to fake but nothing else, whereas in ability testing faking cannot rely on a skill - or it would not be faking - but instead relies on external aids. This means that faking in ability assessment is far more demanding and resources-/time-consuming than faking self-reported measures. Second, whilst faking can go completely unnoticed in self-report, albeit not entirely (as discussed in Study 4), using aids or collaborating with others would be a lot harder to conceal in ability assessments.

From a more behavioural perspective, a study investigating strategies to fake

the IAT identified several different strategies in which participants engage to manipulate their scores for the better or for the worse (Röhner, Schröder-Abé, & Schütz, 2013). The study observed instructed and naïve participants attempting to increase or decrease their IAT score and they summarised the strategies in two factors: speed, implying responding either faster or slower, and accuracy, implying increasing or reducing the number of errors. The results of the study showed that people can fake good and bad quite skilfully by adapting their speed and accuracy to the identity of the different trials (congruent and incongruent), and that being told how to fake improves faking outcomes. It is important to highlight the fact that the IAT is relatively simple to fake as is based on single right/wrong responses, and that the response features lending themselves to faking are only two: speed and accuracy. Nevertheless, this study is important for several reason: first, is one of the few studies directly comparing different avenues for faking within the same assessment (i.e.: speed and accuracy), second, it compares faking bad and faking good at the same time, and finally, it compares the faking performance of instructed and naïve fakers. This offers a good combination of knowing the right answers on one hand, and also of modifying behaviour in a way that would be difficult to pinpoint to either individual differences or faking and is nevertheless entirely up to the candidate to control.

All the strategies included above have a common thread - they imply an awareness of the opportunities offered by the assessment format, the choice of a strategy leveraging those opportunities, and a deliberate attempt to produce a specific response following that strategy. It is clear that faking doesn't *just happen* but it requires a conscious stream of evaluations and decisions in order to exist, and sometimes it requires external aids as well.

With this in mind, it can be argued that the ability to articulate a faking strategy is an obligatory step to envision a method to manipulate assessment scores that is aligned to the assessment and that is within the control and capacity of the individual. Whilst simply being able to articulate a faking strategy might not warrant the ability to fake *per se*, faking strategies are an imperative prerequisite for the ability

to fake.

The current study investigates how the ability to articulate a viable faking strategy varies across different assessment formats, expecting that people would find it significantly more challenging to figure out how to fake BBAs compared to self-reported measures and intelligence tests. This difficulty is expected to be the byproduct of the inability to identify a suitable faking vulnerability in the format that could be translated in deliberate actions and behaviours aimed at modifying the psychometric outcomes of the test.

4.2.1.2 This Study

Considering that, to be able to fake an assessment, candidates need to perceive an opportunity to improve their scores and have a clear strategy in mind to achieve that, and that each different assessment format presents a unique set of vulnerabilities to faking, this study investigates the strategies in which candidates would engage to improve their scores on a range of workplace assessment formats.

By simply asking a sample of participants how they would cheat on each of the four assessment formats described and demonstrated in a video, free-text qualitative data was used to test a series of assumptions suggesting that it is less likely that people would know a viable faking strategy for BBAs than for any of the other assessment formats in the study, suggesting that, when faced with a BBA they would not know what to do to fake.

4.2.1.3 Hypotheses

The hypotheses in this study are all advanced to provide evidence that one of the reasons why BBAs are harder to cheat might be that people don't actually know how to do that.

Due to the task-centred behavioural complexity of BBAs, and the intricacy and opacity of their scoring mechanisms, it is likely that this type format might not lend itself to manipulation as overtly as other formats, and for this their vulnerabilities might not be as obvious to candidates as they are in other formats.

This study endeavours to test the following four hypotheses:

- **Hypothesis 1: The most common cheating strategy suggested for each assessment format will be different between formats.**
  - *Hypothesis 1a:* When asked how they would cheat a BBA, people will be more likely to say that they don't know how to do that than to describe any other strategy
  - *Hypothesis 1b:* Most people will mention deception to describe how they would cheat Self-Reported measures
  - *Hypothesis 1c:* Most people will suggest finding the right answers as the best strategy for cheating Intelligence tests
- **Hypothesis 2: Each cheating strategy will be employed in different measures for different assessment formats.**
  - *Hypothesis 2a:* Most of the instances in which people are not able to suggest a strategy, they will be referring to a BBA
  - *Hypothesis 2b:* Most of the instances in which people suggest deception as a cheating strategy, they will be referring to a Self-Reported measure
  - *Hypothesis 2c:* Most of the instances in which people suggest finding answers as a cheating strategy, they will be referring to an Intelligence test
- **Hypothesis 3: The likelihood of suggesting any viable cheating strategy will differ across formats.**
  - *Hypothesis 3a:* In BBAs, participants will be more likely to say that they don't know a cheating strategy compared to suggesting any other viable cheating strategy
  - *Hypothesis 3b:* In formats other than BBAs, participants will be more likely to suggest a cheating strategy than to say they don't know how to cheat.

- **Hypothesis 4: The likelihood of suggesting a viable cheating strategy altogether will differ across formats.**
  - *Hypothesis 4a:* In most of the instances in which people are not able to suggest a viable cheating strategy of any type, they refer to a BBA (this is different from Hypothesis 2a as all the viable cheating strategies are collated)
  - *Hypothesis 4b:* In most of the instances in which people are able to suggest a viable cheating strategy, they don't refer to a BBA

4.2.2 Method

A qualitative study was conducted on Prolific Academic to explore the strategies in which candidates engage when attempting to inflate their scores in workplace assessment.

The data for this study was collected at the same time in which the data used for Study 1 was collected, therefore, in the interest of brevity, the sections below might refer the reader back to Chapter 3 on page 51 for detail that is not essential to the understanding of this study.

4.2.2.1 Sample

The sample in this study is the same used in Study 1, which consisted of a Prolific Academic sample of 460 participants. The demographic composition of the sample is not relevant to this study as there is no interest in individual factors, however, it is worth noting that the sample had a good representation of gender, age and ethnicity. The reader can refer to page 58 for detailed information about the sample.

4.2.2.2 Procedure

This study used the same participant recruitment procedure and some of the same materials described in Study 1 (see page 62 for further detail on participant recruitment).

The study consisted in showing four videos each describing and showing a different assessment format - a BBA-Puzzle, a BBA-DMT, a Self-Reported Assess-

ment, and an Intelligence Test (see Page 62 for detailed descriptions of the videos, and Page 246 for links to the videos). After each of which the participants were asked a question: *"Please indicate how you would cheat this type of assessment"* followed by a free-text box with no further instructions on how to approach the question.

They then proceeded with the rest of the study as described in Study 1, but the data collected after this point is not relevant to this study.

The free text answers were collated on a spreadsheet free from the assessment format to which they referred to avoid interpretation bias. A total of 1840 response fields were analysed and coded according to Braun and Clarke (2006) thematic analysis technique.

Thematic analysis is a qualitative research method used to explore and extract themes from text, the following steps should be followed where possible according to best practice:

- *Familiarisation with data*: The first step is to read and reread the text to become familiar with the content. This step is crucial to develop a sense of the context, meaning, and nuances of the text.

- *Coding*: Coding involves identifying and labelling sections of the text that are relevant to the research question. Coding can be done manually or using software. There are different types of coding, such as descriptive, inductive, or deductive coding.

- *Categorising codes*: Once the coding is complete, the researcher should organise the codes into categories that represent the themes in the data. Categories should be mutually exclusive, and each code should belong to only one category.

- *Developing a code-book*: A code-book is a document that defines each code and category and provides examples of text that illustrate them. A code-book ensures consistency and reliability across different coders and analysis stages.

- *Testing themes*: Testing the themes involves going back to the data to check whether they are supported by the evidence. The researcher should check whether the themes explain the data, are consistent across different sections of the text, and account for any outliers or exceptions.

- *Writing up the analysis*: Finally, the researcher should write up the analysis, using quotes and examples from the text to illustrate the themes. The analysis should be clear, concise, and accessible to readers.

These steps can be adapted and customised depending on the research question, methodology, and type of data. However, following these steps can ensure that the analysis is rigorous, transparent, and replicable, and that the themes extracted from the data are meaningful and valid.

In addition to the steps listed above, three themes were defined *a priori*: no strategy, deception, and finding answers. These themes are aligned to the strategies reported in the literature review above, and they were also expected to emerge as a result of the features of the formats featured in the study, as it was predicted that participants would predominantly indicate that they didn't know how to cheat (”no”), that they would engage in deceptive behaviour (”lie”), or that they would need to find the right answers before sitting the test (”find”), as these seem to be the most plausible options. Table 3.2 on page 61, describes how the four assessment formats are linked to these three strategies.

After reading all the participants' responses, three further themes were identified: ”social”, ”tech”, and ”mixed” (described below), as the frequency and distinctiveness of these themes warranted separating them from the *a priori* categories. Also, a number of responses were not deemed to provide a valid attempt at answering the questions so a further label ”null” was added to the list of possible themes.

The data set was coded by assigning one single label to each response, according to the following rules (listed in order of prevalence):

- **Theme 1: ”No”** - the participant indicates that they don't know how to cheat; this category also includes instances in which participants mention

non-cheating test taking techniques such as ”concentrate a lot” or ”do your best” as cheating strategy.

- **Theme 2: ”Find”** - the participant indicates that they would cheat by means of finding the right answer to the test. This includes instances in which participants suggest doing so by finding answers online, asking people who have taken the assessment before, or accessing the assessment for unauthorised practice before completing it in the context of the application. In order to be labelled as ”find” the strategy had to be centred around finding specific questions and answers, or on accessing the questions outside the context of the application to have more time and resources to work out the right answers. So the emphasis of this theme is on securing the correct answers to the test before sitting it.

- **Theme 3: ”Lie”** - the participant indicates that they would modify their behaviour or responses to mimic what they believe is expected of them. Many responses labelled in this category include an element of accessing information but in those cases ”lie” was chosen instead of ”find” when the ultimate goal was to gather insight, in the form of tips and tricks or company resources about desirable traits and values, in order to be able to *modify* behaviour or responses rather than to *access* the right answers to specific questions. The core of this theme is the intention to engage with the assessment in a deceitful way with the aim of portraying a different profile.

- **Theme 4: ”Mixed”** - the participant’s response can be coded with more than one other label, and it is not possible to assign it to a single category. It is worth mentioning that, unlike for the other formats in which ”mixed” referred to truly mixed strategies, a pattern emerged in responses referring to Intelligence tests whereby out of the 115 instances in which participants indicated more than one cheating strategy, 49 of them were a specific combination of ”social” and ”find”. This phenomenon was not taken into account for two reasons: first, taking this into account would mean applying the same rules to

all the other instances of ”mixed” strategy, thereby eliminating a whole category which is important in itself; and second, because the data did not lend itself to extrapolating the relative importance of the different strategies listed in the responses and this would result in an arbitrary fraction of a frequency assigned to each strategy depending on the total number of strategies listed in a single response, thereby artificially manipulating the frequency of the different strategies altogether.

- **Theme 5: ”Social”** - the participant indicates that they would ask somebody else to sit the test for them or to google the questions live whilst they are sitting the test themselves. The latter is different from ”find” because it requires another person to actively help whilst the assessment is in flight compared to providing the right answers before the assessment begins, making the strategy inherently dependent on another person, hence social. Requiring another person to physically do the test, either alone as test taker or in support capacity, at the time it takes place, is the key condition for a response to be labelled as ”social”.

- **Theme 6: ”Tech”** - the participant indicates that they would use technology to improve their scores. These responses typically include training models to answer questions correctly, hacking timers to get more time to respond, and manipulating scoring engines. It is unclear, however, whether any of the respondents would know how - or have the necessary resources and skills - to employ these tactics, or whether these tactics are even aligned to real vulnerabilities of the types of assessment they target. Nevertheless, whenever the cheating strategy in a response relies on tampering with the assessment or its scores, or producing responses by means of technology, it is labelled as ”tech”.

- **Theme 7: ”Null”** - the participant did not provide a plausible answer to the question. ”Null” responses might be blank text fields, random words, irrelevant responses (such as debating whether they should/would or

shouldn't/wouldn't cheat instead of indicating how, or stating *"I am great at this type of test, I don't need to cheat"*), or strategies that clearly wouldn't equate to improving assessment scores *per se* (such as *"bribe"* or *"answer randomly and very fast"*).

The answers were coded to reflect the ultimate goal of the strategy, independent from the specific tactic, meaning that ostensibly very similar answers might have been coded differently, and ostensibly different answers might fall in the same category.

For instance, many responses mentioned other people in the tactics (i.e.: the terms "friend/s", "family", "somebody" and "someone" were featured 100, 23, 300, and 21 times respectively) but those responses were only labelled as "social" when the tactic implied another person taking the test or making suggestions whist the test is happening. In some cases, other people were mentioned in the context of writing code to hack into an assessment's scoring engine, and those responses were labelled as "tech" because the ultimate goal was to use technology; in other cases, other people were mentioned in the context of sharing resources, and, depending on the specific type of resources and their intended use, those responses were labelled as "lie" or "find".

After completely the first coding, the data set was re-coded a week later blind to the first round of labelling. No discrepancies were found between the two rounds, so the responses were re-assigned to the question they answered, and the data set was prepared for quantitative analysis.

4.2.3 Results

The data from all participants was retained on the basis that all participants provided at least one valid (i.e.: not labelled as "null") response across the four questions, and this was true of all the participants. The rationale behind this retention approach is that a pattern might emerge showing that participants are more or less likely to provide "null" responses for a particular assessment format, and this is a noteworthy effect in its own.

The data was transformed into two variables: "format" and "strategy". Formats

were labelled as 1 (BBA-Puzzle), 2 (BBA-DMT), 3 (Self-Report), and 4 (Intelligence); and Strategies were labelled as 1 (no), 2 (find), 3 (lie), 4 (mixed), 5 (social), 6 (tech), and 7 (null). A third variable "NOvsANY" was computed in which only "no" (1) and "any" (2 - any viable cheating strategy, so all but "no" and "null") labels were coded.

Table 4.1: Frequency of proposed cheating strategies by assessment format.

	Cheating Strategy							
Format	no	find	lie	mixed	social	tech	null	any
BBA-Puzzle	203	101	18	51	68	17	2	255
BBA-DMT	197	128	29	33	52	16	5	258
Self-Report	38	41	341	31	4	0	5	417
Intelligence	78	206	3	115	39	4	15	367
Total	*516*	*476*	*391*	*230*	*163*	*37*	*27*	*1297*

Table 4.1 reports the frequencies of the different strategies across the four assessment formats; showing that the distribution of cheating strategies in the participants' responses vary quite widely depending on the assessment format to which they refer.

The frequency at which the strategies were featured in the responses also varied significantly regardless of the format to which they referred; Figure 4.1 depicts the overall proportions of the different strategies across the study as a whole.

The chart shows that three strategies account for roughly a quarter of the responses each, as did the remaining 4 combined. The most common response to the question "how would you cheat this assessment format?" was "I don't know" or similar responses in the "no" category, which accounted for 28% of the cases. This was closely followed by the two strategies "find", with 26.9% of the responses, and "lie", featured in 21.3% of the responses.

After removing the "null" cases, not knowing how to cheat accounted for 28.5% of all the valid responses, suggesting that the vast majority of the responses (75.5%) featured a suggestion or description of a cheating tactic (Figure **??**). Whether the tactic listed in the responses are useful, successful, or appropriate is beyond the scope of this study, these results however illustrate the fact that the majority of potential candidate would have ideas of how to cheat an assessment in

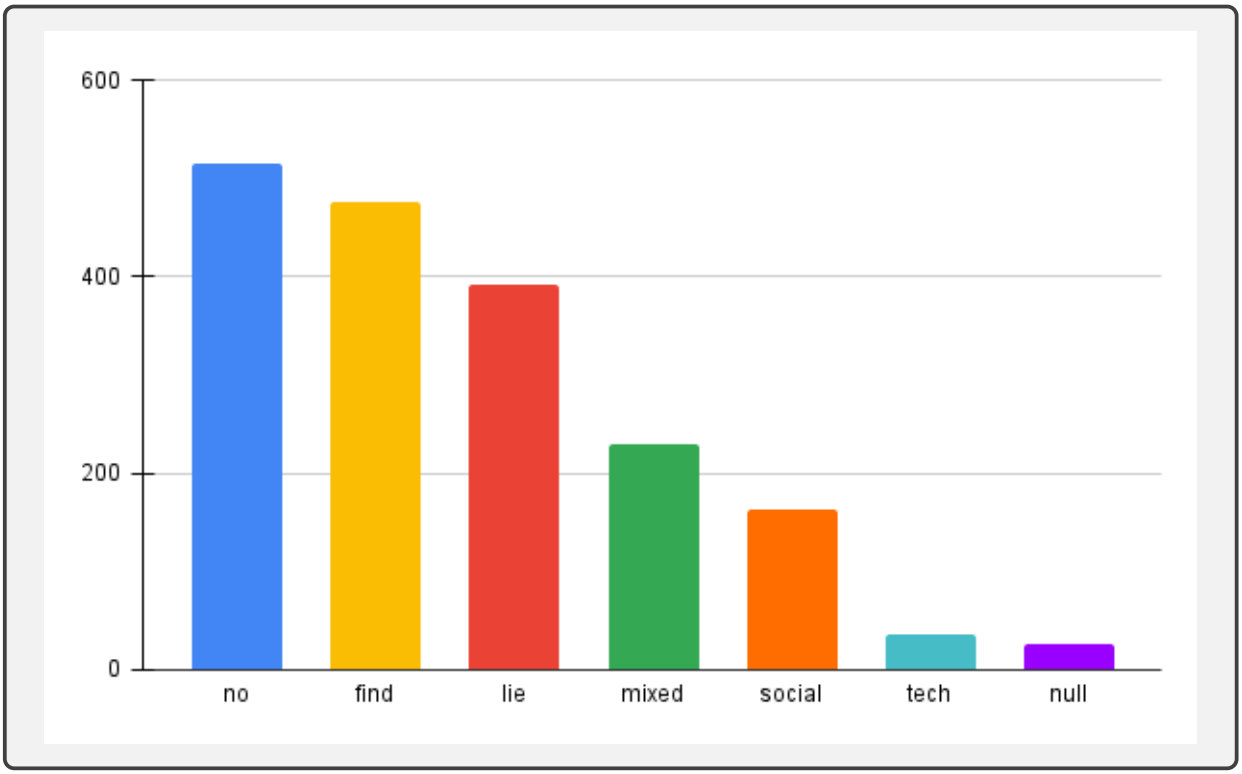


Figure 4.1: Comparative proportion of all the cheating strategies proposed throughout the study.

most cases.

4.2.3.1 Preliminary analysis

Before testing the four hypotheses, a preliminary analysis was performed on the data to establish whether the two types of BBA were interchangeable or should be considered a distinct formats, and whether their responses warranted combining and averaging.

BBA-Puzzles and BBA-DMT are quite similar on the surface but differ quite substantially from one another on the basis that the first type of BBA includes reactive time-sensitive tasks with right/wrong response options, and the second type includes ostensibly time-agnostic decision making tasks with a range of right/wrong and non-performance-based response options.

Because of this, the latter lends itself to deliberate behavioural manipulation much more than the former, and it is possible that participants might perceive different vulnerabilities between the two types of BBA.

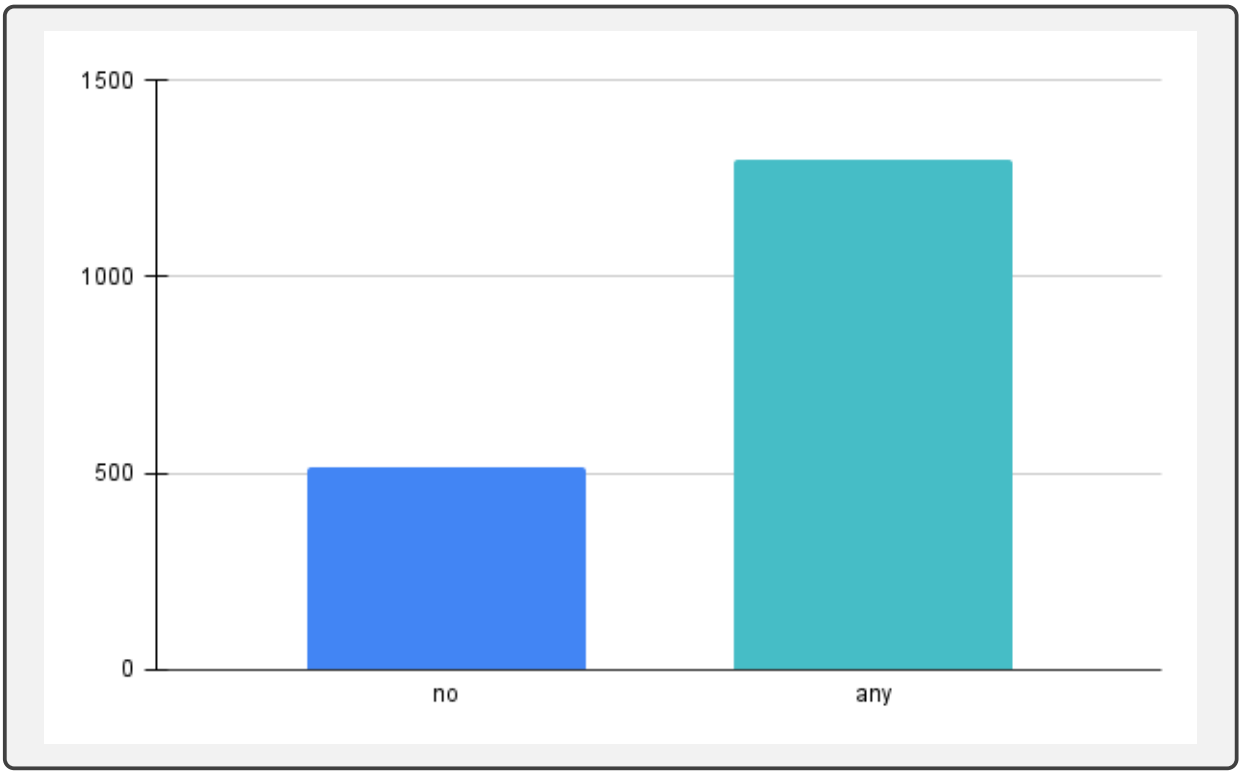


Figure 4.2: Comparison of responses featuring any viable cheating strategy and those in which participants declared to not knowing how to cheat.

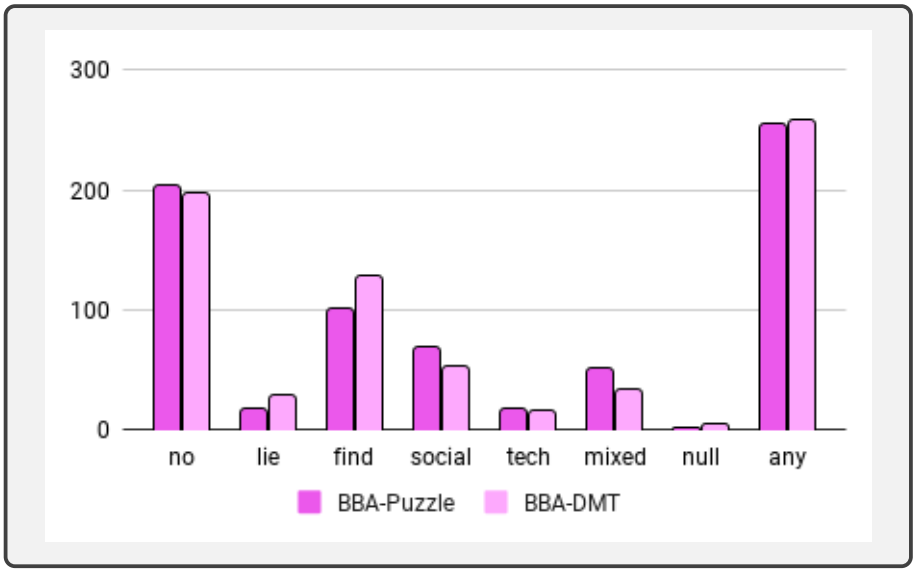


Figure 4.3: Comparison of cheating strategy frequencies between BBA-Puzzle and BBA-DMT.

To establish this, cheating strategy frequencies were compared between the two types of BBA through crosstabs analysis. The results of a chi-square test highlighted

some minor yet significant differences between the two sub-formats - $\chi^2(6)=13.154$, $p=.041$ - whereby more participants than expected (i.e.: with adjusted residuals larger than 1.96) indicated that they didn't know how to cheat a BBA-Puzzle, and more responses contained multiple strategies for BBA-DMT than it was expected.

Due to this, the two types of BBA were not combined for further analysis but retained as individual formats.

It is worth noting that the two significant differences recorded were very small (i.e.: with adjusted residuals of 2.1) and likely only due to sampling, therefore these results will not be featured in the discussion as any potential explanation would be purely speculative.

4.2.3.2 Hypothesis 1 - The most common cheating strategy suggested for each assessment format will be different.

The data was analysed through crosstabs, and a chi square test revealed that the participants suggested significantly different combinations of cheating strategies depending on the assessment they referred to - χ^2 (18) = 1264.786, p <.001 - overall supporting Hypothesis 1.

The distributions of the strategies described by participants across the different assessment formats are depicted in Figure 4.4.

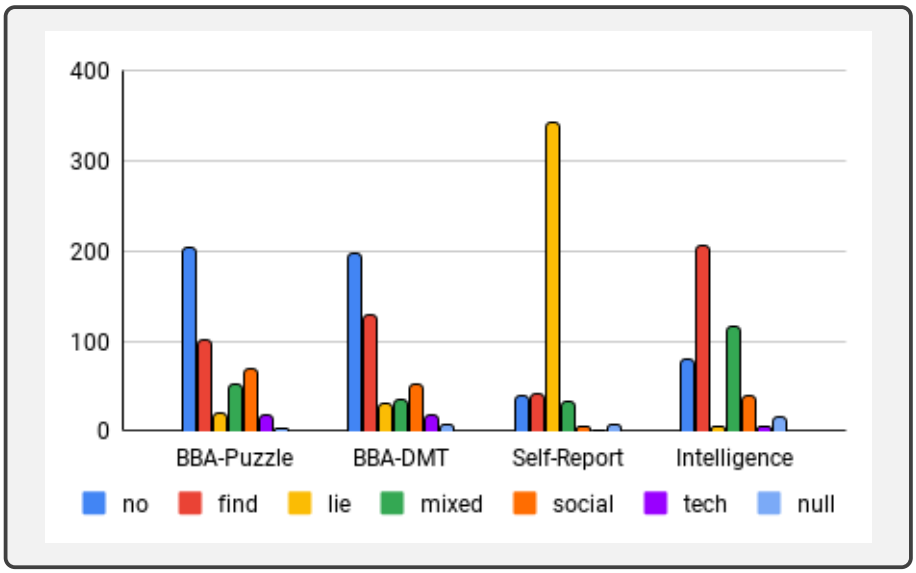


Figure 4.4: Distribution of cheating strategies across the different assessment format.

Hypothesis 1a: When asked how they would cheat a BBA, people will be more likely to say that they don't know how to do that than to describe any other strategy. When looking at the strategies suggested for BBAs, most people indicated that they didn't know how to cheat (BBA-Puzzle = 44.1% and BBA-DMT = 42.8%) compared to suggesting any of the other strategies. This was the strongest positive deviation from the model's expectation for both BBA types (adjusted residuals = 8.9 and 8.1 respectively), so the hypothesis is supported for both BBA types.

Another noteworthy result was the much lower than expected frequency of "lie", suggesting that deception is used as a strategy to cheat in BBA-Puzzle and BBA-DMT significantly much less frequently than any other strategy (3.9% and 6.3% of the times, with adjusted residuals of -10.5 and -9.0 respectively).

Hypothesis 1b: Most people will mention deception to describe how they would cheat Self-Reported measures. Contrary to what was expected for BBAs, it was hypothesised that the most popular cheating strategy in Self-Reported measures would be deception. The hypothesis was supported by the results which showed that 74.1% of the responses described cheating tactics aimed at inflating scores through deception for this format, and this was the strongest positive difference recorded for Self-Report (Adjusted residuals = 32).

In contrast to BBAs, the results showed that the participants who did not know how to cheat Self-Reported measures (8.3% with adjusted residuals of -10.9) were significantly fewer than expected.

Hypothesis 1c: Most people will suggest finding the right answers as the best strategy for cheating Intelligence tests. As expected, "find" was the most popular strategy included in the responses for Intelligence tests (44.8%) and also the largest significant difference compared to any other strategy (Adjusted residuals = 10.7), so hypothesis 1.3 was also supported.

Similarly to both BBAs formats, the largest negative difference found across strategies was with deception (adjusted residuals = -12.5), and lying was the least popular suggestion made for cheating Intelligence tests, in fact it was only made by 0.8% of the participants.

Summary of Hypothesis 1: Taken together, these results demonstrate that when participants encounter an assessment, they are able to adapt their cheating strategies to the key vulnerabilities of the assessment they encounter.

4.2.3.3 Hypothesis 2 - Each strategy will be employed in different measures for different assessment formats.

The same crosstabs analysis described in Hypothesis 1 was also used to address Hypothesis 2. It was expected that the formats for which people proposed a given type of strategy would significantly vary in proportion across strategies - as depicted in Figure 4.5.

This hypothesis was also supported as per Hypothesis 1's results (i.e.: χ^2 (18) = 1264.786, p <.001), and Figure 4.5 illustrates the different proportions of formats for which each strategy was suggested.

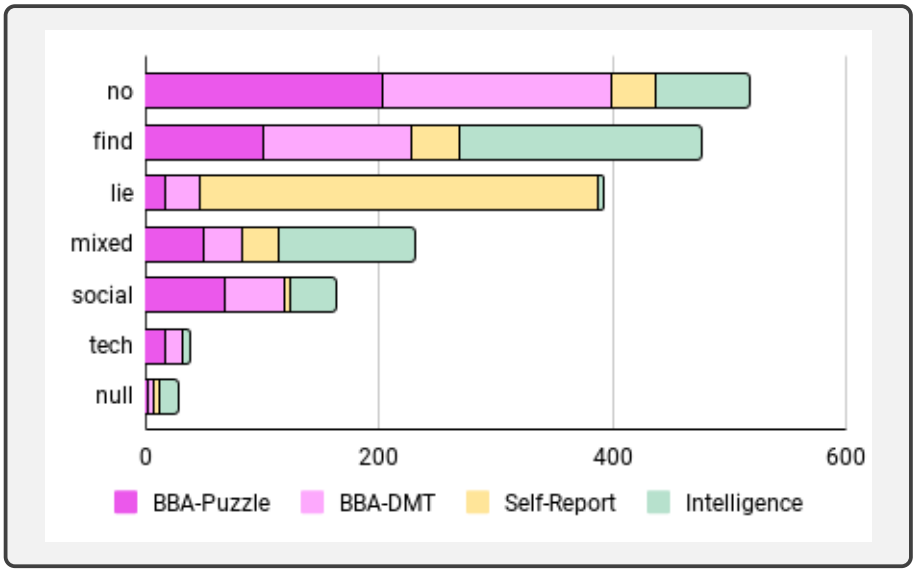


Figure 4.5: Prevalence of formats for which each strategy was suggested, by cheating strategy.

Hypothesis 2a: Most of the instances in which people are not able to suggest a strategy, they will be referring to a BBA. As predicted, the majority of cases in which people indicated that they were not able to think of a cheating strategy referred to a BBA - and this was true of both types. In fact, in 39.3% of the instances in which people reported to not know how to cheat, they referred to BBA-Puzzles,

and in 38.2% of the cases they referred to BBA-DMT, whereas they only referred to Self-Reported and Intelligence tests 8.3% and 17% of the times respectively.

In the ”no” category, both BBA types were recorded in significantly more instances (adjusted residuals = 8.9 for BBA-Puzzle and 8.1 for BBA-DMT) than expected whilst Self-Reported and Intelligence tests reported in significantly fewer cases than expected (adjusted residuals = -10.9 and -6.1 respectively).

Hypothesis 2b: Most of the instances in which people suggest deception as a strategy, they will be referring to a Self-Reported measure. Similarly, as predicted, in the majority of cases in which people described a cheating strategy based on deception, they referred to a Self-Reported measure. In fact, 87.2% of the times somebody mentioned deception as the sole strategy, they referred to Self-Reported measures, which was significantly more than expected (adjusted residuals = 32).

BBA-Puzzle, BBA-DMT, and Intelligence tests were the format to which participants referred to in deception-based strategies less than it was expected (4.6%, 7.4%, and 0.8% of the times, and with -10.5, -9, and -12.5 adjusted residuals respectively).

This effect is particularly pronounced, and it can be easily notices via visual inspection of Figure 4.5 which depicts and almost entire representation of Self-Report in the ”lie” bar chart.

Hypothesis 2c: Most of the instances in which people suggest finding answers as a strategy, they will be referring to an Intelligence test. Also as predicted, when people suggested finding answers to questions as a strategy to cheat an assessment, they were significantly more likely to refer to an Intelligence test than to any other format. In fact, 43.3% of the times somebody suggested finding correct answers to a test they referred to Intelligence tests, and this was significantly higher than expected (adjusted residuals = 10.7). BBA-DMT was featured in the strategy in line with expectations (26.9% of the times with adjusted residuals = 1.1, *n.s.*), whereas BBA-Puzzle and Self-Reported measures were featured significantly fewer times than expected (21.2% and 8.6% of the times and with adjusted residuals = -2.2 and -9.6 respectively) in the ”find” category.

Summary of Hypothesis 2: Taken together, these results suggest that while each cheating strategy is employed in every assessment format to a certain degree, participants tend to choose a given cheating strategy for the assessment format in which it is most effective.

4.2.3.4 Hypothesis 3 - The likelihood of suggesting a viable cheating strategy will differ across formats.

Hypothesis 3 postulates that the rate at which participants will be able to propose a viable cheating strategy of any kind, compared to not being able to suggest one at all, will be different across the four formats (see Figure 4.6).

To test this, a crosstabs analysis between strategies ”no” and ”any” (i.e.: all strategies but ”no” or ”null”) was ran on the responses to all formats, and this revealed a significant effect - χ^2 (3) = 222.247, p <.001 - thereby supporting the hypothesis.

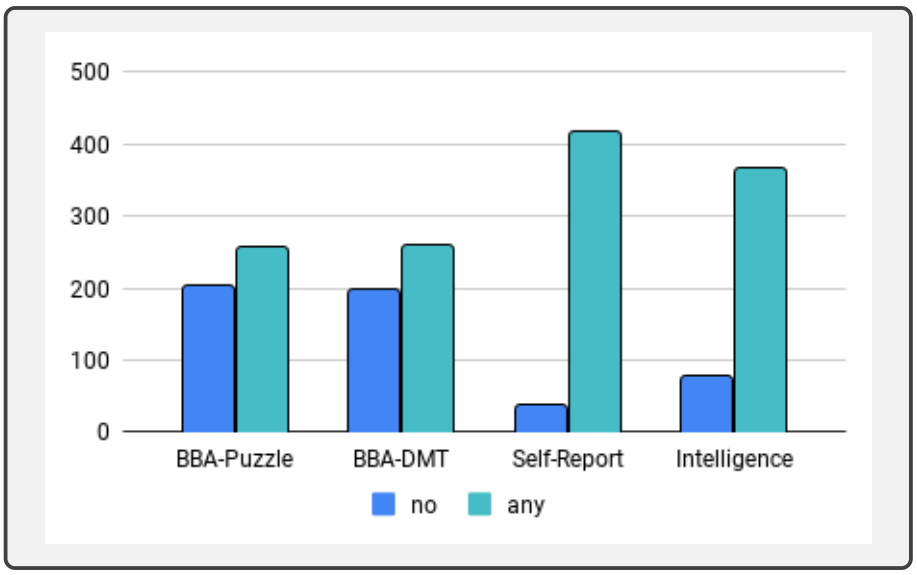


Figure 4.6: Comparison of number of time participants did or did not know how to cheat across assessment formats.

Hypothesis 3a: In BBAs, participants will be more likely to say that they don’t know a cheating strategy compared to suggesting any strategy. Whilst fewer people admitted to not know a way to cheat either type of BBAs (44.3% for BBA-Puzzle and 43.3% for BBA-DMT) compared to those who described a strategy,

compared to other formats, those were significantly more people than expected (adjusted residuals = 8.7 and 8.1 respectively), so the hypothesis is supported.

Whilst the analysis is very similar to the one performed for Hypothesis 1a, the results are not identical. In Hypothesis 1a the comparison was made with each individual cheating strategy, including ”null”, whereas Hypothesis 3a is comparing the incidence of not being able to describe a cheating strategy with the ability to describe any cheating strategy, whether appropriate to the format or not. The results from this analysis could have been even more different from those reported in Hypothesis 1a if only the magnitude of difference between ”no” and each of the other cheating strategies was not so large, so this hypothesis warranted separate testing.

Hypothesis 3b: In formats other than BBAs, participants will be more likely to suggest a cheating strategy than to say they don’t know how to cheat. Whilst most people were able to suggest a cheating strategy for BBA-Puzzle (55.7%) and BBA-DMT (56.7%), this was significantly lower than expected when compared to other formats (adjusted residuals = -8.7 and -8.1 respectively), thereby supporting the hypothesis.

Summary of Hypothesis 3: These results, taken together, show that the majority of participants are able to propose a viable cheating strategy for any assessment format. However, when the no vs any frequencies are compared across formats, it becomes evident that their ability to think of any cheating strategy for BBA is statistically lower than it is for the other formats.

4.2.3.5 Hypothesis 4 - There will be a different distribution of formats between responses containing ”no” and ”any” strategies.

Hypothesis 4 postulates that the format to which the most participants refer when making a strategy suggestion will be different from the format for which they are less more likely to not know how to cheat. This hypothesis was also supported by the same analysis used for Hypothesis 3 - i.e.: χ^2 (3) = 222.247, p <.001.

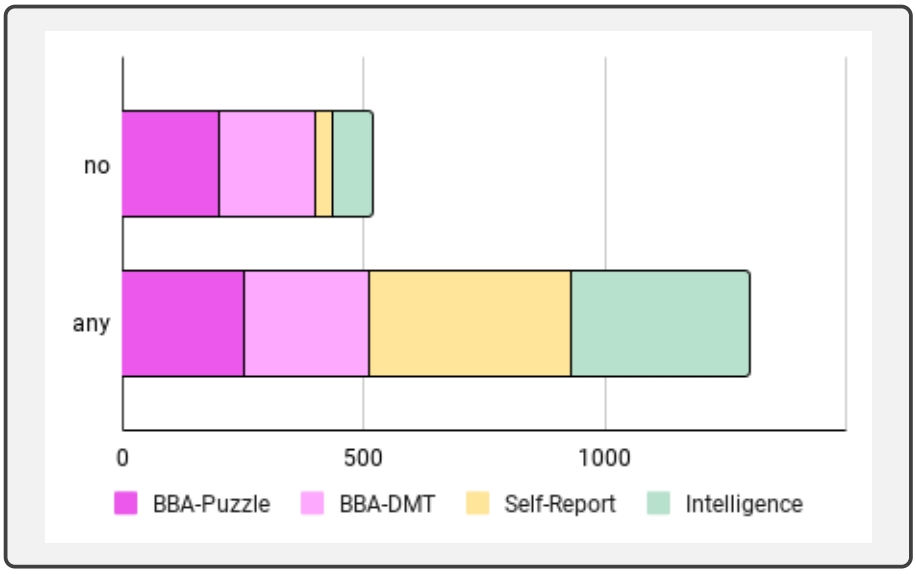


Figure 4.7: Distribution of formats across responses featuring a viable cheating strategy and not.

Hypothesis 4a: In most of the instances in which people are not able to suggest a viable cheating strategy of any type, they refer to a BBA. As predicted, most of the times in which participants admitted that they did not have a cheating strategy for an assessment format, they were referring to a BBA; in fact, 39.3% of the times they referred to a BBA-Puzzle and 38.2% of the times they referred to a BBA-DMT, which was more than expected (adjusted residuals = 8.7 and 8.1) compared to the 7.4% and 15.1% of cases referring to Self-Reported and Intelligence tests, which were less than expected (adjusted residuals = -11 and -5.9). This supported the hypothesis.

Hypothesis 4b: In most of the instances in which people are able to suggest a viable cheating strategy, they don't refer to a BBA. The last hypothesis was also supported as the results showed that most of the times in which participants described a cheating strategy they were referring to an assessment format other than a BBA, in fact they were most likely to refer to Self-Reported measures (32.2% of the times) or to Intelligence test (28.3% of the times), and in both cases this was higher than expected (adjusted residuals = 11 and 5.9 respectively).

Both BBAs were the least likely format participants referred to when describing a cheating strategy, only making up 19.7% (BBA-Puzzles) and 19.9%

(BBA=DMT) of the cases, and both being featured in the any category less than expected (adjusted residuals = -8.7 and -8.1 respectively).

Summary of Hypothesis 4: Taken together, these results suggest that while participants might refer to any type of assessment when describing a viable cheating strategy of any kind, they are less likely to refer to a BBA. The opposite is true for instances in which participants are not able to describe a strategy.

4.2.4 Discussion

This study employed a qualitative method to explore the faking strategies people might able to formulate and possibly follow when encountering different types of assessment formats. The study aimed at validating the hypothesis that people don't know how to fake BBAs, as this might explain the lack of behavioural control participants perceive over BBAs which was demonstrated in Study 1, but, most importantly, underlie their inability to produce adequate and successful faking behaviours even in context in which intentions to fake can expected to be very high.

Through the study, several categories of cheating strategies emerged, showing that people might approach cheating in very different ways. Deliberate deception, finding assessment answers, strategies including other people being active part of the assessment, and the use of technology were all suggested across the different formats along with not knowing how to cheat, null responses, and strategies containing a mix of more categories.

The distribution of the categories was remarkably different across the assessment formats, showing an overall alignment to the cheating opportunities each format affords.

The hypotheses only directly included the three categories of cheating strategies that were expected a priori given the formats included in the study - those were *not knowing how to cheat*, which assumed to be the most frequent answer for BBAs due to the behavioural complexity and item opacity of the format, *deception*, assumed to be the most frequent answer for self-reported assessment as lying is an easy and efficient way of manipulating scores in assessments asking questions that cannot be verified, and *finding the right answers*, assumed to be the most frequent

answer for intelligence tests as success in this type of assessment is based on right and wrong answer, and the only way to improve scores beyond ability is to have access to the right answers. The other categories were included in the analysis for exploratory purposes, and did not affect any of the hypotheses; that is, none of the categories extracted from the qualitative data were the most frequent category for any of the formats. Nevertheless they added interesting insight and variance to the results.

As predicted, the most common response to the question *"how would you cheat this assessment?"* was *"I don't know"* for both types of BBAs, meaning that people generally don't know how to cheat this type of assessment. This is likely to impact the levels of PBC over the format and hence intentions to fake, but also the actual ability to produce an adequate faking behaviour - should intentions be high despite low PBC - as the lack of a strategy implies the inability to engage in goal-driven behaviour.

Also, in line with predictions, the most common cheating strategy for self-reported assessment was deception, and whereas finding answers was the most popular strategy described to fake intelligence tests. Whilst these results don't add much insight to the perceptions of self-efficacy over cheating BBAs, they add in terms of the overall validity of the study, showing that participants did modify their response as a function of the assessment method for which they were asked to describe a cheating strategy. To further reinforce the difference, the results showed that deception was the least popular category for BBAs and intelligence tests, whereas the least amount of people said the did not know how to cheat a self-reported measure, and this was closely followed by finding the right answers. This shows a great degree of orthogonality between the different formats, supporting the assumption that participants can fine tune their strategies and successfully adapt to the vulnerabilities of each different format.

The choice of strategies was also observed through the lenses of how relatively often each of them was used across the different formats, and again, the hypotheses were all supported. Not knowing how to cheat was predominantly claimed for

BBAs, deception predominantly suggested as a strategy to cheat self-reported measures, and finding questions was predominantly used as a strategy to cheat intelligence tests. These results suggest that not only certain strategies are more popular in certain formats, but also that they are not popular for the formats in which they would not be as efficient.

Finally, all the viable cheating strategies were collated and comparisons were made to investigate whether once the granularity of the different categories of viable strategies was removed, not knowing how to cheat was still the dominant response for BBAs - and the least dominant for the other formats. Interestingly, this was only partially true when the distribution of viable and non-viable strategies (i.e.: not knowing how to cheat and null responses) was compared format by format, in fact, more viable strategies than non-viable strategies were suggested for all formats. However, when the data was analysed as a whole, this showed that, despite a numerically higher proportion of viable strategies for both BBAs, this was relatively much lower than what was observed in the other two format. In addition to this, the inability of suggesting a viable strategy was by far more used for BBAs than for any other format, and viceversa, the suggestion of any viable strategy was predominantly featured for assessment types other than BBAs. Both analyses yielded significant and very strong results, showing how important it was to include other types of assessment to make inferences about the degree of self-efficacy people might have over cheating a specific type of assessment.

This results are aligned to the literature in that they confirmed that the same faking strategies identified studies focusing on self-report and ability studies also emerged for those formats. Whist it is not surprising, this adds to the literature by means of replication and it further provides evidence of people's ability to modify faking behaviours on the basis of assessment format.

Furthermore, the results also corroborate the findings of Study 1 pertaining to PBC. In Study 1 lower levels of PBC were observed for BBAs compared to self-reported and ability assessments and this was reflected in the results of this study. When the ability to articulate a strategy was compared across formats, this showed

that significantly fewer people were able to do so for BBAs, and that the format for which people where most likely to have a faking strategy was self-reported assessment. This is aligned to the levels of PBC observed in Study 1, which suggests that the reason why participants experienced lower level of PBC for certain assessment formats is most certainly linked to their inability to envision how to fake them.

4.2.4.1 Implications

The results of this study have some positive implication for practice. First, people seem to be quite skilled at tweaking faking strategies on the basis of assessment formats. Overall, the results show that the majority of people had a different strategy for most assessment type which suggests that they are able to perceive both quantitative and qualitative differences in vulnerability from ostensible assessment features. This could be leveraged by assessment designers to conceal some obvious vulnerabilities from their measures to reduce people's ability to take advantage of them.

Most importantly, the outcome of this study implies that BBAs do impair people's ability to fake. As explained in the introduction and literature review, strategies are paramount to faking. Not knowing how to fake is highly unlikely to lead to faking. That is, successful faking. Whilst people may still decide to attempt some degree of faking not knowing how to do it, this might be a lot more infrequent compared to assessment for which people know faking strategies, and, most importantly, a lot less efficient as their faking attempt would lack purpose and alignment to the assessment's vulnerabilities.

As the subjective ability to fake decreases in BBAs so would be intentions due to perception of expectancy and control, but so would objective abilities. So the effect of not being able to describe a strategy should reduce faking from two different angles. However, this would be quite hard to disambiguate. Whilst a distinction in the literature is made about the comparative degrees of awareness between subjective and objective ability, it is possible that for BBAs the lines are even more blurred. Judging by the strategies described to fake BBAs in this study, people don't even know how unable they are to fake them and they assume they can

somehow ”crack the code”. That is, they believe that their avenues for faking BBAs are intrinsic to their ability to control the scoring mechanisms of the assessment, which is technically a situational factor pertaining to objective ability to fake of which candidates should not be aware (Ellignson & Mc Farland, 2011).

Most importantly, virtually all the strategies to fake BBAs described by the participants in this study would not equate to successful faking. Combining this to the fact that most people admit to not know how to fake, the implication of these results are that people would either not try to fake because they don't know how, or they would try and fail. This is a great piece of evidence which strongly suggests that people are unlikely to successfully modify their scores in BBAs, and that this type of assessment should be safe to use in high stakes settings.

4.2.4.2 Limitations

There are some limitations in this study. First, of course, the link between PBC and knowing how to fake an assessment is not established so this may or may not be the case. This means that intentions may still develop even in absence of a strategy. Future studies should address this directly and uncover the actual relationship between PBC and ability to articulate a viable faking strategy.

Second, not knowing how to fake might only be temporary, and once immersed in the assessment people could be able to identify vulnerabilities that they had not been able to consider before. Furthermore, they could be trying their best to fake and some of those method might result in success. As far as this is unlikely it is also untested, so future research should directly test the relationship between the ability to articulate a strategy, the theoretical viability of those strategy, and the objective ability to manipulate assessments.

Finally, this study was done in low stakes setting and it is possible that some of the ”I don't know” might just have emerged from lack of motivation. Being a new type of assessment, it would require comparatively more effort to think of a faking strategy than it would be for more familiar types of assessment. And this is further exacerbated by the relatively much higher degree of complexity as well. This means that if this study took place during a real job application and the candidate

were equally as familiar as they are with other types of assessment, the results might have been different. However, as it will be explained in the next study, this might not be the case as BBAs could objectively be much harder to fake.

4.2.4.3 Conclusion

In summary, it seems that the proverbial *"where there is a will there is a way"* is bound to struggle when faced with BBAs, and candidates will be significantly less likely to find a way to fake BBAs no matter the strength of their will.

This study asked participants to describe how they would fake four different types of assessments, and the results unambiguously demonstrated that BBAs are the type of assessment that they would struggle to fake the most. Whether the ability to describe a faking strategy is most closely linked to intentions or ability or both is beyond the scope of this study and thesis, however, this will certainly have an impact on candidate's objective ability to fake BBAs and the success rates of eventual attempts.

There are some very encouraging implications linked to this, as these results further suggest that BBAs might be a good alternative to other assessments for high stakes use as candidates may lack enough perceived ability to fake BBAs to attempt to do it, and if they opt for trying, they would do so without a clear idea in mind about how to do it.

4.3 Study 3 - Objective Ability To Fake.

The previous study explored the concept of Perceived Ability, or Expectancy, in the context of BBA, showing that one of the reasons why participants indicate lower levels of perceived behavioural control over BBAs could be that they generally don't know how to fake them. This study shifts the focus from the individual to the assessment itself, investigating whether BBAs even provide any scope for faking or if people's inability to imagine a faking strategy for BBAs is justified by features in this types of assessment which make it too hard to fake.

*****

Whilst the concept of personality as a summary of tendencies, preferences, and emotions is well aligned to lexical models of personality that describe individuals based on their observable differences (e.g.: the Five Factor Model - McCrae & Costa, 1999; the Hexaco model - Lee & Ashton, 2004), more recent differential psychology propositions based on experimental testing methods from cognitive science introduce the possibility that personality might also be conceptualized, at least partially, as an ability.

Whilst this is far from being a novel concept (see Paulhus and Martin, 1987, or Wallace, 1966), more recent Psychobiological models of personality (e.g: DeYoung, Hirsh, Shane, Papademetris, Rajeevan, & Gray, 2010) are centred around the notion that individual differences in brain structure and activation, which can be measured via ability-based lab tasks, are the main driver of the observable behavioural differences used to describe personality traits. Behaviour-Based Assessment is a great candidate to provide evidence for this idea, as its measurement principles reside in Personality Neuroscience, which is defined in the next section, and it uses experimental tasks to generate models of individual differences.

For instance, when using experimental tasks such as the Flanker Task (Eriksen & Eriksen, 1974) in BBAs, the data generated and used for psychological models of personality is predominantly made up of right/wrong responses and intra/inter-individual differences in response speed - that is, ability. This suggests that the emergence of observable behavioural patterns summarised as *personality traits* might inherently depend on the individual's *ability* to perform in a specific way in response to internal and/or external stimuli and events.

According to this hypothesis, observable expressions of extroverted behaviour, such as enjoying parties or having more friends, might depend on the ability to filter out information and to sustain engagement under increased level of stimulation that more introverted individuals would find exhausting, rather than an innate penchant for people.

For instance, Matz, Hofstedt and Wood (2008) found that Extraversion moderates the cognitive dissonance caused by disagreement, meaning that extroverts ex-

perience less discomfort when others don't agree with them, and they attributed this effect to a lower level of vulnerability to arousing experiences in line with Eysenck's early work on the biological substrates of individual differences (e.g.: Eysenck, 1967). In other words, this evidence suggests that the neural underpinnings of Extraversion are linked to individuals' *ability* to deal with the inevitable levels of disagreement that emerge within social groups, that is, extroverts are *able* to "put up with people" more than introverts. At the same time, Lieberman and Rosenthal (2001) suggested that the reason why introverts might struggle to work out who likes them is that they find it harder than extroverts to multitask, and this makes nonverbal encoding (i.e.: deriving social information through nonverbal cues) more challenging when performed as a secondary task, which is how this normally occurs in social settings. Again, this effect is explained on the basis of differences in arousal, whereby extroverts are more efficient in processing information under higher level of arousal compared to introverts, and for this they are better at working out who likes them because they are *able* to perceive and process secondary nonverbal cue more efficiently whilst performing primary task such as talking. This has likely got less to do with people than with information processing in general.

This study explores the possibility that trait Extraversion, as measured through a BBA, might be underlined by a substantial ability component. The role of this questions within this thesis is that if evidence is found that BBAs measure personality through ability, this would suggest that BBAs are objectively much harder to fake compared to other personality measures.

4.3.1 Literature Review

Extraversion was selected as a "case study" in this thesis to investigate the variable-level composition of a BBA model of personality because it's the construct that requires the most Skyrise City levels to be scored, and having access to data from several experimental tasks allows for richer exploration and better generalisability of the results.

This section summarises some of the experimental evidence linking individual differences in Extraversion to individual differences at the neurobiological level.

Whilst this is not meant to be a comprehensive review of the subject, the following pages should provide enough background to appreciate why this study expects that a BBA model of Extraversion might include enough ability-based data to suggest that BBAs generate scores of personality traits at least partially through individual differences in ability.

4.3.1.1 Extraversion

In the Five Factor Model of personality (McRae & Costa, 1999), which is the most widely used taxonomy in differential psychology, Extraversion is defined as the degree of willingness and motivation of an individual to interact externally. Whilst the lexical model of personality on which the Five Factor Model is based is useful to *describe* repeatable and observable behavioural patterns that co-occur in similar clusters among individuals, two earlier models of personality lend themselves much better to *understand* the mechanisms underlying Extraversion in the context of this study and thesis. This is because they focus on the biological underpinning of those observed behavioural patterns that are at the basis of the measurement approach used by BBAs described in the introduction.

The first model, Eysenck's (1967, 1970), postulates that personality can be understood in terms of different levels of arousal. According to this model, Extraversion is a result of low levels of baseline activity in the reticulo-cortical loop, which makes individual high on Extraversion more likely to seek stimulating environments and activities compared to introverts who possess higher baseline level of arousal and are therefore more easily overwhelmed by stimulation. Regarding the second model, Gray (1970) proposed the Reward Sensitivity Theory (RST) of individual differences, which encompasses two neural systems: the Behavioural Activation System (BAS), responsible for approach behaviours aimed at increasing reward, and the Behavioural Inhibition System (BIS) responsible for inhibitory behaviours aimed at decreasing punishment. Individual differences in Extraversion are mapped on the spectrum of the BAS, whilst the BIS is associated with individual differences in other traits such as Neuroticism.

Taken together, the two models paint a clear and complementary picture of

Extraversion whereby extroverts are characterised by low baseline levels of arousal, high tendency towards rewarding behaviours, and high tolerance of stimulation, whereas introverts are characterised by high baseline levels of arousal, low tendency towards rewarding behaviours, and low tolerance of stimulation. This suggests, in exceedingly simple terms, that the observed behavioural-level individual differences understood as trait-level individual differences in Extraversion are the result of neural-level individual differences in processing capacity and reward sensitivity.

4.3.1.2 Personality Neuroscience

Personality Neuroscience is an "*interdisciplinary approach to understanding brain mechanisms that produce relatively stable patterns of behaviour, motivation, emotion, and cognition that differ among individuals*" (DeYoung, 2013, p.1). While the exact neural basis of Extraversion is not entirely understood, research suggests that it may be related to differences in the structure and function of certain brain regions. Specifically, Extraversion has been associated with increased gray matter volume and/or glucose metabolism in several areas of the brain, including the medial orbitofrontal cortex, anterior cingulate cortex, and ventral striatum (e.g.: Kim et al, 2008; DeYoung, Hirsh, Shane, Papademetris, Rajeevan, & Gray, 2010; Umemoto & Holroyd, 2016). These regions are involved in a variety of processes related to goal directed behaviour, motivation, emotions, and reward, which may help to explain some of the individual differences associated with the trait, such as sociability and positive affect.

Furthermore, Extraversion is believed to depend on Dopamine - a neurotransmitter predominantly acting as neuromodulator, responsible for stimulating function in frontal/temporal cortices and basal ganglia from the midbrain (DeYoung, 2013) - both in terms of the reward mechanisms in which the neurotransmitter is involved, and in terms of its availability for and facilitation of Executive Functions (Campbell et al, 2011). Extraversion is the personality trait that has been linked to Dopamine most consistently, and it is believed encompass most of the behavioural expressions of sensitivity to reward (DeYoung, 2013).

Extraversion has been rather consistently linked to differences in performance and brain activity during Executive Function (EF) tasks requiring cognitive control, cognitive inhibition, or set-shifting. Cognitive control refers to maintaining goals and updating relevant information, whereas cognitive inhibition refers to the ability to suspend a reactive response, although there is a debate on whether this is an explicit (Miyake, Friedman, Emerson, Witzki, and Howerter, 2000) or an unconscious process (MacLeod, 2007), finally, set-shifting refers to the ability to efficiently shift between different task goals (Miyake et al. 2000). These cognitive processes are measured through different tasks, and, despite all depending on and resulting in dopaminergic activation in the prefrontal cortex, they do so through different pathways (Collette & Van der Linden, 2002).

Gray and Braver (2002) identified a correlation between Extraversion and Executive Functions. They found that extroverts performed better than introverts in a N-back working memory task, and they found that the correlation also intensified as a function of increases in task difficulty. Task difficulty was manipulated through the number of distractors within the task (i.e.: higher n-back value), suggesting an EF advantage with regards to both cognitive control and cognitive inhibition.

Dujmovic and Penezic (2017) compared introverts and extroverts on inhibition of return (IOR), a spatial task requiring the perceptual inhibition of a cued location in order to process the stimulus displayed shortly after in the goal location. They found a significant difference of refractory period for IOR whereby introverts are less efficient at reorienting their attention to the goal stimulus when the interval between the cue and the goal is short, whereas extroverts are able to reorient their attention much faster. However, performance on trials with longer cue-goal intervals was not significantly different between introverts and extroverts, and there was no main effect of Extraversion on IOR, suggesting that the differences in arousal are quantitative and not qualitative. The observed timeline differences suggest that introverts experience a stronger initial physiological response to the cue compared to extroverts which leads to requiring a longer period of time to redirect attention and resume optimal performance in a different location, whereas extroverts, having

a lower level of arousability, need less time to redirect focus. That is, Extraversion does not predict whether interference has an effect on performance, but the amount of time that it gets to adjust to it - which makes sense in the context of a trait vs type view of personality.

Campbell et al (2011) designed a comprehensive study investigating the relationship of EF and Extraversion. They selected a battery of cognitive tasks to measure the three main components of EF - update/control, inhibition, and set-shifting - and then compared the performance of introverts (i.e.: the lowest 16% of scores), extroverts (i.e.: the top 16%) and ambiverts (i.e.: those in between) on all the tasks. They observed some mixed results through an interaction of Extraversion and EF components, whereby extroverts performed better than ambiverts and introverts on update/control tasks but worse in set-shifting tasks, and no significant differences were observed in inhibition tasks. That said, when update/control and inhibition tasks where desconstructed on the basis of trial difficulty, it emerged that whilst performance in easy trials was equal for all Extraversion groups, extroverts were significantly better in difficult trials for both task types. The authors interpreted these results suggesting that there is a very complex interplay between Extraversion, specific EF task types, their difficulty, baseline dopamine activation, and the magnitude and speed of dopaminergic changes in response to task demand.

In terms of brain structure, frontal regions, such as the dorsolateral prefrontal cortex, have been consistently demonstrated to be involved in EF performance such as working memory and the suppression of irrelevant information in cognitive tasks featuring distractors or other forms of interference (e.g.: Bernal & Altman, 2009; Chen, Wei, & Zhou, 2006), and so are other areas of the brain such as the anterior cingulate, the posterior parietal cortex and the insula (Nee & Jonides, 2011).

For example, Gray and Burgess (2004) found that individuals high in Behavioural Activation Sensitivity (BAS) which is very similar to Extraversion, had an advantage over individuals lower in BAS on a working memory task with distractors that required high levels of both cognitive (working memory span) and inhibitory control (filtering the distractors). Furthermore, their fMRI data showed *lower* ac-

tivation in the prefrontal cortex of individuals higher in BAS, suggesting that not only they had higher level of cognitive control and inhibition control but they also performed the task more efficiently.

Furthermore, Herrmann and Wacker (2021) found that Agentic Extraversion was correlated with task performance on a distractor (3-back) working memory task and a cognitive flexibility task (set-switching), and they also found that the observed correlations were sensitive to sulpiride, a chemical that affects dopamine receptors in the stratium. The pharmacological manipulation generated a surge of dopamine in the stratium which disrupted performance for extroverts causing distractibility, but enhanced performance in introverts by making more dopamine available for the task, suggesting a U-shaped relationship between baseline dopamine in the mesolimbic system and task performance.

This study bridges the experimental evidence linking Extraversion to EF and the experimental evidence linking it to reward. In fact, an influential body of evidence has emerged suggesting that differences in the processing of rewards within the mesolimbic dopamine system might underlie individual differences in Extraversion (e.g.: DeYoung, 2013; Pickering & Gray, 2001), and this is confirmed by the association between genes involved in the coding of Dopamine and self-reported measures of Extraversion (Smillie, Cooper, Proitsi, Powell, & Pickering, 2010; Smillie, Cooper, & Pickering, 2011), and by how Extraversion facilitates learning under reward conditions in reinforced learning tasks (Pickering & Pesola, 2014).

This reward-processing-centric view of Extraversion explains observable behaviours through the lenses of the reward to which they are associated. For example, boldness, talkativity, and positive emotionality can be interpreted as the behavioural expression of an overarching pursuit of social reward, and their co-occurrence within the umbrella concept of Extraversion should be attributed to their association with the same neural mechanisms (Smille, Jack, Hughes, Wacker, Cooper, and Pickering, 2019).

Extroverts have also been found to show greater activation in the ventral striatum during reward processing tasks, suggesting that they may be more sensitive to

rewards and motivated by positive outcomes (Smillie, Cooper, & Pickering, 2011), and Smille et al (2019) found a positive association between several measures of Extraversion and Reward Positivity (i.e.: the magnitude of difference in dopaminergic cell firing between better and worse than expected rewards) suggesting that extroverts are much more sensitive to rewards than introverts.

However, some recent evidence (Zou, Su, Qi, Zheng, & Wang, 2018) has found a negative correlation between Extraversion and bilateral gray matter volume in the putamen, which is involved in reward processing, especially in reward anticipation. This contradicts previous finding as the putamen is part of the stratial region, and it was previously found that resting glucose metabolism in the stratium is more efficient in extroverts (Kim, 2008). Furthermore, Zou et al (2018) also found a negative correlation between Extraversion and the functional connectivity density in the precuneus, which is also in contradiction with existing evidence (e.g.: Pang et al 2017) showing positive associations between functional connectivity in the precuneus and Extraversion.

Overall, a recent systematic review and meta-analysis or voxel-based morphometry studies found good evidence for stable brain grey matter correlates of Extraversion in the frontal behavioural regulatory system, the limbic reward and emotional circuits, and the parietal mirror neuron sociable system, and the author settled the argument of heterogeneity of results across studies with evidence showing that age and gender influence the relationship between brain morphology and Extraversion (Lai, Wang, Zhao, Zhang, Yang, & Gong, 2019). For example, it emerged that gender dictates the type of neural pattern activation in response of social and monetary rewards in the cuadate nucleus, and that age moderate the correlation between Extraversion and grey matter volume in the superior temporal gyrus. Graham and Lachman (2014) also found that while Neuroticism was correlated with cognitive performance in young adults only, emotional aspects of personality, such as Extraversion, only begin showing associations with cognitive performance from older adulthood onwards. This means that some of the discrepancies observed in the literature might be the result of sampling differences but also the results of inconsis-

tencies in the stimuli and imaging techniques used in different studies, suggesting that there are many factors and interactions to consider when using neuroscientific and neurobiological evidence to infer individual differences.

The development of BBAs of Extraversion, such as Skyrise City, rely on this type of evidence to inform the choice of what experimental tasks to use to generate the behavioural differences necessary to generate valid trait scores. One of the reason why several experimental tasks are typically used to develop a single BBA trait model is to mitigate the heterogeneity of the results in the literature and minimise measurement error through repetition and convergence, but also to ensure that different aspects of the traits are fully captured. In fact, as it will be explained in more detail in the method section, several different tasks are necessary to generate a score of Extraversion from Skyrise City. The Extraversion model examined in this study uses data from tasks measuring reward sensitivity and each of the three key distinct processes of EF described by Miyake et al (2000): updating/control, set shifting, and inhibition.

One aspect that links all these tasks is that they can all be considered ability measures, that is, their main effects are typically computed as a function of relative differences in speed and accuracy between trials.

4.3.1.3 Typical and Maximal performance

The success of this study depends on the ability to demonstrate that at least some of the variance in the BBA model of Extraversion is explained by variables from task behaviours that, by nature, don't lend themselves to manipulation. In order to do this, it is necessary to use an established taxonomy to identify which of the BBA variables might fall into this category.

Cronbach (1949;1960) proposed a distinction between assessment measures to distinguish them on the basis of whether they test what people *can* do or what people *will* do - that is, *maximal* and *typical* performance.

Maximal performance describes type of assessments measuring how well people can perform at their best, or at their performance ceiling. This would include ability tests such as cognitive tests, but it can also pertain to other areas such as

sports for example. In this case, the speed at which a runner can compete would be a measure of maximal performance.

On the other hand, tests of typical performance are used to assess what people do, as in what they choose to do, what they tend to do - regardless of what they can do. A psychometric example of this would be situational judgement tests, that is, a type of assessment in which people are asked to indicate what they would do in response to a given scenario. Whilst typical performance might include some variance from maximal performance, one of the defining characteristics of typical performance is that it mainly depends on choice. In contrast, maximal performance relies on competence.

The reliance on competence, or ability, defining measures of maximal performance opens an interesting avenue for preventing assessment faking. If faking implies improving psychometric scores, then scores generated by means of measures of maximal performance would, by definition, be impossible to fake.

Experimental measures typically used in BBAs, such as those measuring the various aspects of EF, rely very heavily on maximal performance as they are scored through speed and accuracy, which are respectively limited by an individual's *ability* to respond fast and correctly. Whilst it would be possible to pretend to be less able in that type of tasks by deliberately responding slower and making more mistakes, it is not possible to pretend to be more able and operate at a level above one's performance ceiling - in the same way in which we can run slower than our personal best but not faster. Shoen, Williams, Reichin, and Meyer (2022) for example demonstrated that in a Conditional Reasoning Test (which in good part measures maximal performance) participants required extensive training to marginally improve their score but required no training at all to successfully "fake bad".

This means that any portion of variance accounted by variables measuring maximal performance in BBAs should be unaffected by faking, and, the larger that variance, the harder the faking.

4.3.1.4 This Study

This study investigates the relative contribution of maximal and typical performance data on a personality model generated through behavioural tasks, proposing that the use of measures of maximal performance provides protection from faking.

The rationale of this study is based on the assumption that measures of maximal performance are by definition a representation of the highest level of ability that the individual can achieve, and for this they do not lend themselves to deliberate manipulation.

4.3.1.5 Hypotheses

To serve as evidence for this, the current study will investigate the composition and predictive validity of a BBA model of Extraversion, which is made up of data coming from several different experimental tasks.

First, variable-level data from a BBA model of Extraversion will be divided according to the experimental task it replicates, and each task's main effect will be tested for correlations with a self reported measure of Extraversion. Subsequently, three machine learning (ML) models will be developed - one with all the variables, one only including typical performance variables, and one only including maximal performance variables. The models will be compared for how much variance in the Extraversion scores they explain, as the size of the unique variance explained by measures of maximal performance can be expected to be more resistant to cheating than the variance explained by measures of typical performance.

This study will test the following two hypotheses:

- **Hypothesis 1: All the tasks' main effects will correlate with the self reported measure of Extraversion.**
  - *Hypothesis 1a:* The Impossible Task main effect will correlate with Extraversion
  - *Hypothesis 1b:* The Flanker Task's main effects of speed and accuracy will correlate with Extraversion

  - *Hypothesis 1c:* The effect of distractors between the Corsi Block Task and Block Suppression Test will correlate with Extraversion

  - *Hypothesis 1d:* The speed and accuracy main effects of task switching will correlate with Extraversion

- **Hypothesis 2: The best performing ML model will be one featuring both maximal and typical performance data**

  - *Hypothesis 2a:* A ML model based on all data will predict Extraversion

  - *Hypothesis 2b:* A ML model based on typical performance data will predict Extraversion

  - *Hypothesis 2c:* A ML model based on maximal performance data will predict Extraversion

4.3.2 Method

An experimental study was conducted with participants recruited from a testing platform to investigate the relative contribution of typical and maximal performance data in a BBA model of Extraversion.

4.3.2.1 Sample

The sample in this study consisted of 292 participants recruited via Prolific Academic for the experiment. The sample was relatively balanced in its representation of gender, with 43.9% of participants identifying as females, 54.8% as male, and 2 who did not disclose a gender identity. Individuals identifying as white (65%) were over represented in the sample - the ethnic composition of the sample is summarised in Table 4.2 - and most people responded from the UK (48.3%) or the US (41.8%). The age of the participants ranged between 18 and 24 years with a mean of 20.8, and two participants did not report their age.

Table 4.2: Ethnicity distribution of the sample

Ethnicity	Count	Percentage
Caucasian	192	65.3%
South Asian	24	8.2%
East Asian	19	6.5%
Mixed	17	5.8%
Black African	14	4.8%
Missing	11	3.8%
Other	8	2.7%
Latino/Hispanic	6	2.0%
Black Caribbean	2	0.7%
Middle Eastern	1	0.3%

4.3.2.2 Materials

For materials, two assessment measures were used in this study: a BBA and a self-reported personality test. Both measures encompass assessments other than Extraversion, but in the interest of succinctness, only the Extraversion-related portions of the two assessments are described in this Chapter.

Additional descriptions of both measures can be found in Study 5 on Page 202 and in the Appendix where they are reported in full.

Self reported measure - Clear Perspectives (Extraversion Scale): The self-reported measure used in the current study is Clear Perspectives (CITE), a 90-item measure that generates 10 trait scores distributed over five factors, encompassing the whole spectrum of normal personality. This study used the 18 items measuring Extraversion (Table 4.3).

The measure is administered as 7-point Likert scale with a mix of positively and reversed keyed items, and it was chosen because it was constructed with a basis on neuropsychology, making it closer to BBA than other self-reported assessments. It used a simple and concise language and was constructed with modern analytics such as item response theory, making it more precise. Furthermore, it was validated on job performance, making it more relevant to use in workplace settings compared to other measures validated with other self-reported assessments.

Table 4.3: Clear Perspectives Extraversion Items (Rated from 1 to 7 as a Likert scale, R indicates that the item is reverse scored)

Scale Items
Dislike social gatherings. (R)
Socialise with my co-workers.
Keep to myself. (R)
Have a wide circle of friends.
Keep my distance. (R)
Am a warm colleague to work with.
Laugh a lot at work.
Get a buzz from interacting with others.
Am reserved in work situations. (R)
Like being in charge.
Enjoy leading others.
Prefer others be in charge. (R)
Find responsibility daunting. (R)
Am good at influencing colleagues.
Lack initiative. (R)
Speak up regardless of whom I'm around.
Have loads of energy.
Feel positive.

BBA - SkyRise City (Extraversion Modules) Skyrise City is a BBA developed by the HR Tech startup Arctic Shores Ltd (www.arcticshores.com), and it is made up of several modules which are used in different configurations to assess individual differences in candidates and employees. The BBA is set up like an imaginary building in which test takers are faced with a range of challenges, each representing a different module based on one or more experimental tasks.

Overall, the assessment is based on established laboratory tasks which are replicated by adding graphics and story lines whilst retaining the key cognitive and behavioural elements completely intact (Arctic Shores, 2016). For this, Skyrise City is a great measure for this study as it makes it possible to use a real workplace assessment tool whilst providing scope for isolating variables from established experimental tasks to test the hypotheses listed above.

In this study, only the modules used to generate Extraversion scores were used, those are listed in Table 4.4 along with the original task they replicate.

Navigation (Flanker Task)

Table 4.4: Skyrise City modules used to generate Extraversion scores, and the original experimental task each of them replicates.

Skyrise City Module	Original Task
Navigation	Flanker Task (Eriksen, 1974)
Promotional	Corsi Block (Corsi, 1972)
Promotional	Block-suppression task (Beblo et al, 2004)
Ticket Master	Task Switching (Alport, 1994)
Code Breaker	Impossible task (Green, 1981; in Arctic Shores, 2018)

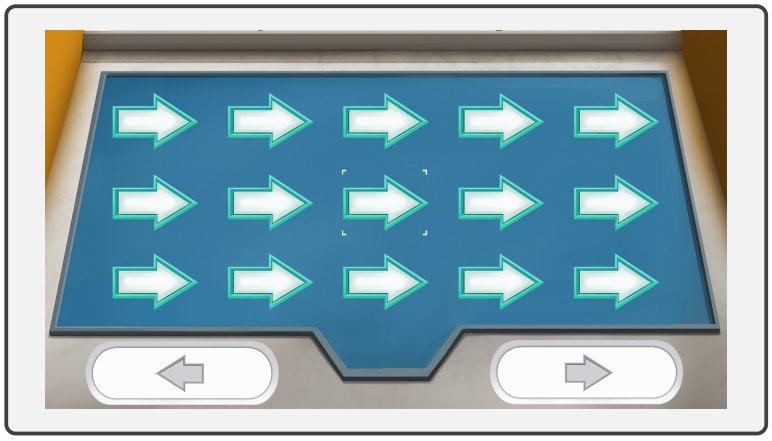


Figure 4.8: Navigation Module - Skyrise City

This module is an adaptation of the Flanker Task, which is a laboratory task in which participants are presented with an array of arrows and are required to indicate the direction of the arrow in the middle of the array by pressing left and right buttons (typically Q and P on a keyboard); the peripheral arrows around the central one might point in the same or opposite direction, thereby facilitating performance or generating interference. The task was designed to test inhibitory cognitive control, and performance is computed by comparing speed and accuracy between three types of trials: *congruent* (i.e.: trials in which the peripheral arrows pointing in the same direction to which the central arrow is pointing), *incongruent* (i.e.: trials in which the peripheral arrows pointing in the opposite direction from which the central arrow is pointing), and *neutral* (i.e.: trials with straight lines instead of arrows).

The task' main effect is calculated through means of maximal performance by comparing the average speed and accuracy between trial types.

Due to directional interference, effective processing of incongruent trials re-

quires a greater degree of inhibitory cognitive control than congruent trials, and, while performance on incongruent trials is generally affected by central/peripheral arrows incongruity as a whole, the degree of susceptibility to directional interference varies between individuals. As evidenced in the introduction to this study, inhibition control is a key neural correlate of Extraversion (e.g.: Campbell et al, 2011).

In Navigation, participants are instructed to act as receptionists and guide visitors around the Skyrise building using a computer terminal. They must tap one two buttons, each displaying either an arrow pointing to the left or to the right, to match the direction of the arrow presented in the middle of the screen. As in the original task, the central arrow is presented very briefly (around 500ms) in rows with other arrows - which may point to the same or the opposite direction - or with straight lines.

Promotional (Corsi Block and Block Suppression Task)

Promotional is based on the Corsi Block Task (Corsi, 1972), a spatial span test of visuo-spatial working memory, or processing capacity, which is a measure of maximal performance often used as a marker for fluid intelligence (e.g,: Engle, Kane, & Tuholski, 1999), and is a key EF correlate of Extraversion (e.g.: Campbell et al, 2011). The Corsi Block consists in observing a set of objects being briefly placed in different positions on a grid in a specific temporal sequence and having to recall and reproduce the sequence by placing the same objects in the same position in the same order. Digital adaptations (using touch-screen devices) and the traditional task (using wooden blocks) demonstrate equivalent performance in span length, with the additional benefit that the digital version is considered subjectively less stressful (Robinson & Brewer, 2016).

The Corsi Block Task is often supplemented with the Block Suppression Test (Beblo, Macek, Brinkers, Hartje, & Klaver, 2004). In this variation, some non-target items are added to the target sequence, however, this visual distraction must be actively suppressed when encoding and reproducing the target sequence, requiring additional input from executive areas of the brain involved in inhibitory control

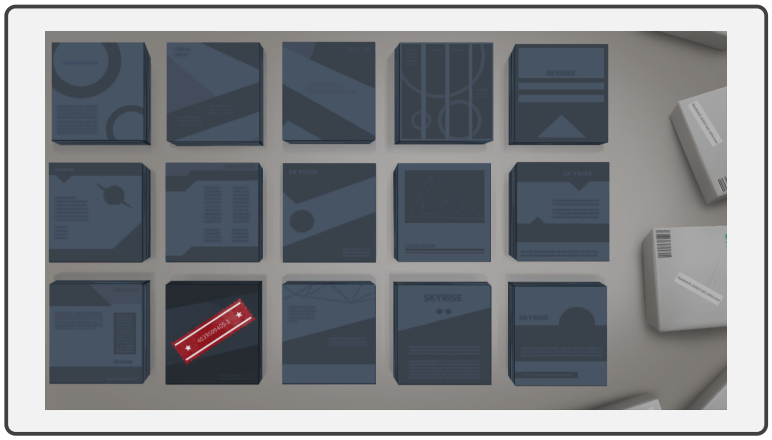


Figure 4.9: Promotional Module - Skyrise City

(Toepper et al, 2010), which, again, is a correlate of Extraversion (e.g.: Campbell et al, 2011).

In Promotional, leaflets for a Skyrise City event need to be distributed across the building, and they are organised in 16 piles on a 4x4 grid. A circular blue stamp is used to show the order in which the different piles should to be packed, and the participant must accurately reproduce the sequence by tapping the piles in the same order in which the stamps appear. Starting with three stamps, participants can progress to sequences of up to 12 stamps. If an error is made, participants will re-attempt the task starting with the number of stamps of their last correct sequence.

After a couple of trials, some long red stamps start appearing in addition to the original blue stamps. These are postage errors which must be ignored when trying to reproduce the sequence. Both types of stamps are shown for 500ms.

There are two task main effects coming from Promotional - the first, which is from the Corsi Block Task and is the length of the longest sequence correctly replicated, or memory span, and the second which measures the effect of adding the Suppression Block Task to the exercise, and is calculated as the maximum span difference between ”normal” and ”distractor” trials. The respectively represent the control/update and the inhibition components of Executive Functioning (see Miyake et al, 2000).

Ticket Master (Task Switching)

Ticket Master is based on the psychological paradigm of task-switching -

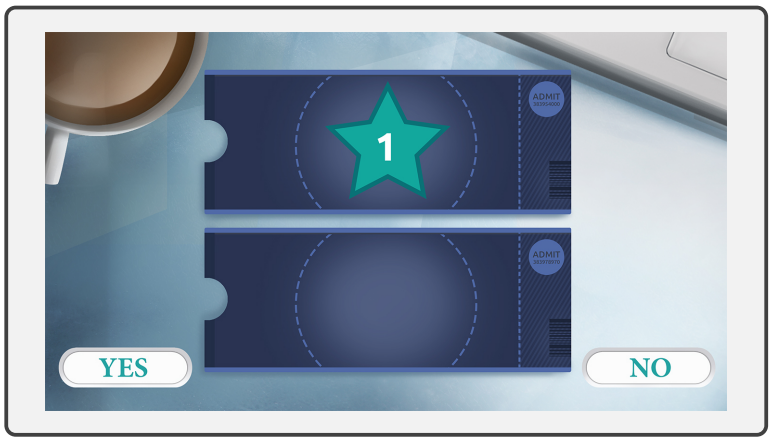


Figure 4.10: Ticket Master Module - Skyrise City

or set-shifting, which is the third key component of Executive Functioning (e.g.; Miyake et al, 2000). Experimental tasks measuring task/set-switching effects typically feature multiple stimulus-rule combinations within the same trial blocks, requiring participants to switch between appropriate responses as a function of stimulus identity. This requires rapid and reactive changes in attentional parameters that rely on Executive Functioning mechanisms situated in the prefrontal cortex (e.g.: Alport et al, 1994; Monsell, 2003)

Whilst performance is generally better on trials following the same stimulus type (repeat trials) compared to trials following a different stimulus type (switch trials), ”switch cost”, which is the task’s main effect and refers to the magnitude of difference in speed and accuracy between repeat and switch trials, which is a measure of maximal performance, varies as a function of individual differences in Executive Functioning. In addition to this, inter- and intra-individual variations in response time have been associated with sensitivity to punishment and reward, which are also significantly associated with Extraversion (e.g.; Umemoto & Holroyd, 2016).

In Ticket Master, participants act as receptionists of Skyrise City sorting tickets for attendees of an event according to the shapes and numbers displayed on them.

All the tickets feature a shape with a number in the middle, and they may appear in the top or bottom half of the screen. When a ticket appears in the top, the participant must indicate whether its number is even or odd, whereas, when it

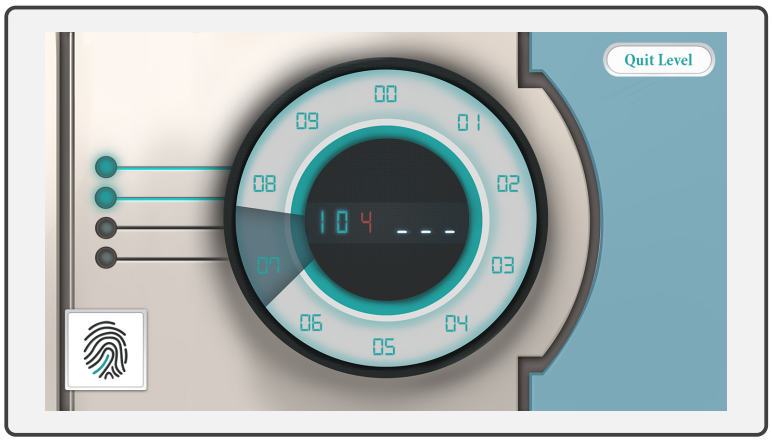


Figure 4.11: Code Breaker Module - Skyrise City

appears in the bottom, they must state whether the shape on it is rounded or edgy. This requires them to use the ticket's location as a prompt for guiding their attention toward the stimulus' feature that is relevant to their response.

Initially, the participant is presented with a block of top tickets and a block of bottom tickets, then the tickets start appearing at the bottom or at the top at random, requiring the participant to switch between the relevant stimulus parameters to respond accordingly.

Code Breaker (Impossible Task)

Code Breaker is based on the Impossible Task paradigm (Green 1981; in Arctic Shores, 2018), whereby participants are faced with a task on which they keep on failing but they are not aware that they are in fact attempting to solve a task that has been deliberately designed to be impossible to solve. Due to its link with reward and its extreme difficulty, performance on the impossible task is expected to correlate with Extraversion (e.g.: DeYoung, 2013; Smillie et al, 2019; Campbell et al, 2011), more specifically, it is expected that individuals with higher levels of Extraversion will persist on the task for longer than those lower on the trait.

In Code Breaker, the security system at the Skyrise building needs to be tested for vulnerability. There are four locks that protect the security system and each is deactivated with a six digit code entered by tapping on a finger print when a laser circles above the target digit on the lock's dial.

The task becomes increasingly more difficult as the dial moves at varying

speeds and directions making the pattern incrementally less predictable at each digit. The task is designed to be very difficult and to prevent any progress beyond the first few digits of the last lock. Any mistake results in starting the task from the very first lock, which moves at a rather slow pace.

At each failed attempt, the participant can opt to quit the level or to have another go - indefinitely.

Despite ostensibly being an ability-based task, the variables generated by Code Breaker are not considered maximal performance, instead, the task main effect is a measure of typical performance due to the fact that participants can decide to persist as long as they like even if they would have instinctively given up much sooner, and vice versa. The maximal performance element of the task - i.e.: the ability to solve all the locks - is mostly a red herring.

4.3.2.3 Procedure

A study was published on Prolific Academic asking participants to complete a BBA and a self reported questionnaire. The participants were instructed to approach both assessments with full attention and commitment; they were told that an algorithm would detect distracted and or unusual patterns of behaviour in the BBA, and that the self reported measure contained attention checks and was sensible to social desirability. To further improve the quality of the data, participants were told that those with the best quality of data (i.e.: with no flags from the algorithm or attention checks) would receive an extra payment of £50, £25, £10, or £5 depending on their data's quality ranking within the sample.

The participants completed a short demographic form, the items from the Clear Perspective questionnaire and the Skyrise City BBA. The whole study lasted about 30 minutes.

4.3.3 Results

4.3.3.1 Data Overview

Data from Clear Perspective was used to generate an overall score of Extraversion by adding all the items' scores (reversed where needed) together, ignoring the two

sub-facets. The reason for this decision is that the variable-level data making up the BBA model of Extraversion was offered by the test publisher without specifying which sub-facet of Extraversion the variables measured in order to prevent exposing too much information about proprietary models. Furthermore, Smille etc al (2019) found that overall Extraversion showed clearer relationships with neural markers of Reward Positivity compared to narrower sub-traits, further supporting the choice of using the overall Extraversion score.

Clear Perspective had a very high internal consistency in this study (α=.903), and the scores were normally distributed. The average trait score in the sample was 73.43 (SD= 16.50), with scores ranging between 26 and 122.

Data from Skyrise City was a collection of all the game variables used by the test publisher to generate a score of Extraversion, and those were divided between maximal performance and typical performance variables.

Maximal performance variables in this study (Table 4.5) can be one of two types: speed or accuracy, and they are further divided between task main effects and extra variables. The task main effects are the ”official” variables used by the original experimental task to score participants' performance, whereas the extra variables refer to data that is generated by the task which is not usually considered for analysis in the experimental literature.

One of the key pillars of BBA's measurement advantage is the granularity of data extracted from gameplay and it is very common for data which is traditionally overlooked to be incorporated in trait models. The Extraversion model generated by Skyrise City contains several extra variables from each of the modules reported in this study.

For example, these are variables from the Navigation module (Flanker Task) that are included in the BBA's model of Extraversion:

- **Task Main Effects**
  - *(Max1) Accuracy Main Effect:* Difference in number of correct responses between congruent and incongruent trials; a small difference signals a high level of inhibition control as the participant's performance

doesn't deteriorate too much as a function of interference.

- *(Max2) Speed Main Effect:* Difference of (average) reaction time for correct responses between congruent and incongruent trials; a small difference signals a high level of inhibition control as the participant's performance doesn't deteriorate too much as a function of interference.

- **Extra variables**

  - *(Max8) Variance in Reaction Time for correct trials:* This is the standard deviation of the reaction times for all the correct responses, the rationale behind including this variable is that there is an effect of individual differences in the amount of interference caused by incongruent trials and high variance equates to a higher susceptibility to it.
  - *(Max9) Mean Response Latency of correct trials:* This refers to the average time that participants take to make their response from when the stimulus appear; higher reaction times signal higher effect of interference in general.
  - *(Max10) Variance difference between congruent and incongruent trials' reaction times:* This variable provides a measure of the degree to which incongruence destabilise consistency in response time; the relative magnitude of variance in reaction times between congruent and incongruent trials is used as a proxy for levels of inhibition control as larger difference in standard deviation between trial types are a marker of the detrimental effect of inconguence on response consistency.

In the interest of Arctic Shores' intellectual property, the typical performance variables are anonymised and listed as TP1, 2, 3, 4, 5, and 6 (Table 4.6). Only the Impossible Task's main effect (TP6) is identified in order to test the relevant task main effect hypothesis, whereas the other variables refer to a range of different behaviours within Code Breaker and other Skyrise City modules that were deliberately not referenced and described in this study to further protect their proprietary trait models.

This degree of obfuscation is necessary because, while the relationship between task main effects and personality traits is relatively well known and documented in the literature, the development of predictive BBA models of personality is heavily dependent on defining the new relationships between tasks data and traits which constitute an assessment's intellectual property, and for this they cannot be disclosed.

As the current study is entirely focused on investigating and explaining the unique contribution of maximal performance variables on personality in a very specific applied context, not knowing the identity of the typical performance variables does not detract from the clarity of the results or their implications in the context of this research question. The point of the study is to establish whether BBA models contain enough variance explained by measures of maximal performance to warrant the claim that this might make them less susceptible to faking, and the identity of any of the variables is irrelevant to this pursuit.

Nevertheless, typical performance variables can be described as all the variables not representing reaction times and accuracy, and/or representing task behaviours that participants could decide to modify for the best - i.e.: actions and decisions not limited by a person's ability to perform them.

All the variables in this study are described in the tables featured in the follwing pages.

Table 4.5: List of maximal performance variables featured in the Skyrise City Extraversion model

Variable	Task	Effect	Description	Mean	SD
MAX1	Flanker	Inhibition Control	Difference in number of correct responses between congruent and incongruent trials.	12.95	2.63
MAX2	Flanker	Inhibition Control	Difference in reaction times of correct responses between congruent and incongruent trials.	-1.30	3.85
MAX3	Corsi/Suppr	Distraction Filtering	Longest span in distractor trials.	48.83	54.29
MAX4	Corsi/Suppr	Distraction Filtering	Span difference between regular and distractor trials.	5.90	1.378
MAX5	Task Switch	Cognitive Control	For trials where the participant responded, different in reaction time between switch and repeat trials.	-.31	1.534
MAX6	Task Switch	Cognitive Control	Difference in correct responses between switch and repeat trials.	147.60	83.47
MAX7	Task Switch	Cognitive Control	Standard deviation of reaction times for all correctly-answered trials.	-45.30	4.615
MAX8	Flanker	no	Standard deviation of reaction times for all correctly-answered trials.	194.01	38.00

Table 4.5: List of maximal performance variables featured in the Skyrise City Extraversion model

Variable	Task	Effect	Description	Mean	SD
MAX9	Flanker	no	Mean average reaction time for all correctly-answered trials.	152.89	56.43
MAX10	Flanker	no	Difference in standard deviation of correct responses between congruent and incongruent trials	565.60	46.69
MAX11	Corsi Block	no	Maximum span in regular trials	15.25	79.08
MAX12	Task Switch	no	Difference in standard deviation of reaction times between repeat and switch trials	6.21	1.24
MAX13	Task Switch	no	The mean average reaction time for correct responses trials.	23.95	68.65
MAX14	Task Switch	no	The proportion between 0 and 1 of switch trials that the participant answered correctly	731.83	71.31
MAX15	Task Switch	no	Number of correct responses in trials 31 to 80 of any trial type.	.68	.14
MAX16	Impossible	no	Maximum score achieved in the impossible task	40.03	4.90

Table 4.6: List of typical performance variables featured in the Skyrise City Extraversion model

Variable	Task	Effect	Description	Mean	SD
TP1	-	-	Anonymised measure of typical performance	1.28	1.15
TP2	-	-	Anonymised measure of typical performance	1.08	1.10
TP3	-	-	Anonymised measure of typical performance	4.34	2.72
TP4	-	-	Anonymised measure of typical performance	-2.15	6.93
TP5	-	-	Anonymised measure of typical performance	-6.09	20.49
TP6	Impossible	Persistence	Number of attempts at opening the locks	10.60	11.10

4.3.3.2 Hypothesis 1 - Results

A series of bivariate correlations were used to test the hypothesis that all the tasks' main effects will be correlated with the self-reported measure of Extraversion. More specifically, four sets of correlations investigated the relationship between Extraversion and the main effects of the Impossible Task, the Flanker Task, The Corsi/Suppression Block task, and Task Switching. Only one of the sub-hypotheses was supported.

Hypothesis 1a - The Impossible Task main effect will correlate with Extraversion: A bivariate correlation between the number of attempts at opening the locks in the security system of Skyrise City, which represents the main effect of persistence, and scores of Extraversion showed that the two measures are not correlated with each other (r= .077 *p*= .187), thereby rejecting the hypothesis.

Hypothesis 1b - The Flanker Task's main effects of speed and accuracy will correlate with Extraversion: Two bivariate correlation between the scores of Extraversion and the effect of congruency on accuracy and speed in Navigation, which represent the main effects of inhibition control, showed that the two measures are both correlated with scores of Extraversion (r= .120 *p*= .040 , and r= .147 *p*= .012 respectively) supporting the hypothesis on both speed and accuracy effects.

Hypothesis 1c - The effect of distractors between the Corsi Block Task and Block Suppression Test will correlate with Extraversion: A bivariate correlation between the scores of Extraversion and the magnitude of difference in working memory span between regular and distractor trials in Promotional, which represents the main effect of distraction filtering, showed that the two measures are not correlated with each other (r= .006 *p*= .919).

Hypothesis 1d - The speed and accuracy main effects of task switching will correlate with Extraversion: Two bivariate correlation between the scores of Extraversion and the switch cost to accuracy and speed in Ticket Master, which represent the main effects of cognitive control, showed that the two measures are not correlated with scores of Extraversion (r= -.075 *p*= .199 , and r= -.079 *p*= .176 respectively), rejecting the hypothesis.

Summary of Hypothesis 1: A series of hypotheses were advanced expecting that the main effects of the cognitive tasks replicated in Skyrise City would each individually correlate with scores of Extraversion. Whilst the main effect of inhibition control from the Flanker Task (Navigation) correlated with Extraversion both in terms of speed and accuracy, none of the other Skyrise City modules showed a link between Extraversion and the main effect of the task they replicated, meaning that the update/control and set-shifting aspects of Executive Functioning measured via the BBA task did not correlated with self-reported Extraversion, contradicting the experimental evidence reported in the introduction to this study.

4.3.3.3 Hypothesis 2 - Results

A series of ML models were developed to predict scores of Extraversion. Hypothesis 2 postulates that a predictive ML encompassing both typical and maximal performance variable would outperform ML models including only one type of performance variables. This hypothesis was advanced on the basis that personality traits are the byproduct of a complex multitude of factors, and a model limited to one type of data, while still expected to predict some elements of a construct, cannot be expected to account for the entire breadth of such a multifaceted personality factor.

To test this, Skyrise City and Clear Perspectives data were uploaded on SPSS Modeler 1.0 and three different versions of the dataset - one using all the Extraversion BBA variables as inputs, and the other two each featuring either maximal or typical performance variables as inputs - were screened for potential models. In all instances, the overall Extraversion score for Clear Perspective was set as the target for prediction, and the models were instructed to use 75% of the sample for training and 25% of the sample for testing at random.

SPSS Modeler 1.0 ”auto-numeric” function screens data for 15 different Machine Learning prediction, classification, and segmentation modelling techniques such as Random Forest, Neural Network, XG Boost and more, and it runs those appropriate to the type of data in the inputs (independent variables) and target/s (dependent variable/s), ranking them for performance.

Once developed, the best performing model for each version of the dataset was

retained, and the predicted scores were generated and saved for further analysis (see Table 4.7).

Table 4.7: Machine Learning models comparison - prediction of Extraversion (*** correlations significant at $p<.001$)

	All Variables		**Typical Perf**		**Maximal Perf**	
Model Type	Random Trees		XG Boost Linear		XG Boost Tree 1	
Partition	Train	Test	Train	Test	Train	Test
Occurrences	212	80	212	80	212	80
Error (Mean)	-0.001	2.969	2.524	4.396	-0.005	4.998
Predictions	.882	.227	.071	.185	.957	.207
Mean Score	72.62		72.86		72.63	
St Dev	8.85		1.93		9.60	
Correlations	.709***		.128***		.795***	

Hypothesis 2a - A ML model based on all data will predict scores of Extraversion: A Random Tree model was found to be the best performing ML model featuring all the variables. The model retained seven maximal performance variables and 3 typical performance variables, and predicted the Extraversion score with a coefficient of r=.882 in the training sample and r=.227 in the test sample.

Once generated and tested for correlation, a bivariate analysis showed a strong positive significant correlation between the model and scores of Extraversion (r=.709, $p<.001$)

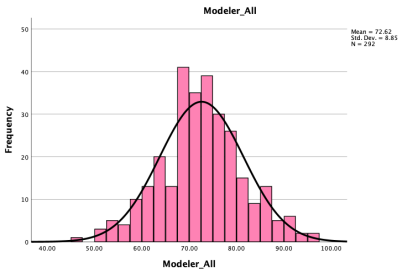
Hypothesis 2b - A ML model based on typical performance data will predict scores of Extraversion: A XG Boost Tree 1 model was found to be the best performing ML model featuring typical performance variables only. The model retained all the typical performance variables, and predicted the Extraversion score with a coefficient of r=.071 in the training sample and r=.185 in the test sample.

Once generated and tested for correlation, a bivariate analysis showed a weak positive significant correlation between the model and scores of Extraversion (r=.128, $p<.001$)

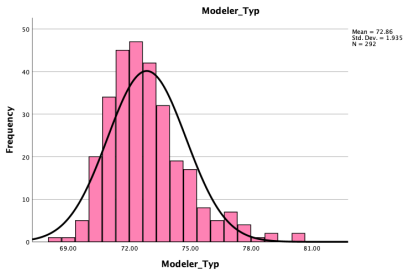
Hypothesis 2c - A ML model based on maximal performance data will predict scores of Extraversion: A XG Boost Linear model was found to be the best performing model featuring maximal performance variables only. The model retained

10 of the maximal performance variables, and predicted the Extraversion score with a coefficient of r=.957 in the training sample and r=.207 in the test sample.

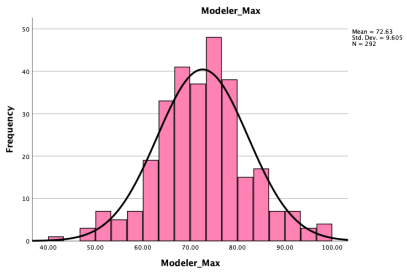
Once generated and tested for correlation, a bivariate analysis showed a strong positive significant correlation between the model and scores of Extraversion (r=.795, $p<.001$).



(a) All Variables Combined



(b) Typical Performance



(c) Maximal Performance

Figure 4.12: Distribution of predicted Extraversion scores generated by the three ML models.

Summary of Hypothesis 2: Hypothesis 2 predicted that the best performing ML model would be one containing all the variables of the Skyrise City Extraversion model.

Whilst the mixed model was indeed performing very well, the best performing model was the one only containing variables of maximal performance. Not only the model developed with maximal performance variables showed a stronger correlation with the Extraversion scores, the predicted scores were also better distributed.

By comparing the histograms in Figure 4.12, it is possible to see that the inclusion of just three typical performance variables is enough to affect the normality of score distribution compared to the model in which they are not featured.

The results of this hypothesis showcase the role that ability plays in individual

differences of personality, and supports the overall aim of the study demonstrating that a large portion of the variance in BBA-generated scores of personality, in this case Extraversion, is explained by measures of maximal performance which are by nature not possible to fake.

4.3.4 Discussion

This study was designed to assess whether BBAs, as a format, offered candidates enough scope to fake. The rationale of this study is that objective ability to fake an assessment is dictated by the type task requirement and by the type of performance data used to generate the scores, and that the more an assessment relies on maximal performance variance, the harder this will be to fake it. This is because maximal performance is by definition the limit to a person's ability to perform, and, with this, the limit to a person's objective ability to fake.

This piece of research used the trait Extraversion as a case study because the Extraversion models of personality in Skyrise City contains data from the widest range of levels/tasks, and this makes the results of this study more generalisable to BBAs as a whole than other trait models.

The study began by investigate the relationship between the main effects of the experimental tasks replicated in Skyrise City and a self-reported measure of Extraversion. Whilst, according to the existing body of evidence, the experimental tasks replicated in Skyrise City to generate Extraversion scores are great measures of the cognitive mechanisms linked to the trait, most of the hypotheses advanced to replicate those associations were not supported. The exception to this was the Flanker Task, measuring inhibition control, which showed positive correlations between the speed and accuracy main effects and the trait. However, there is no reason to speculate why this was observed in the Flanker Task and not in other tasks.

Interestingly, when other variables from these and other tasks were combined together, the data extracted by Skyrise City was able to predict scores of Extraversion quite well. Whilst it is not unusual for BBAs to include non-main effect data, trace data and paradata (Auer, Mercy, Marin, Blaik, and Landers, 2022), most of the results contradicts the existing experimental evidence pertaining the neuro-

cognitive basis of Extraversion, which may raise a concern.

This being said, the main interest of this paper was not so much to question the theoretical underpinnings of the Extraversion model as this has been extensively validated by its developers both in terms of construct and criterion validity (Arctic Shores, 2016), but it was to establish whether the model variance explained by measures of maximal performance was enough to suggest that BBAs would be too hard to fake.

To test this hypothesis, this study used all the variable-level data currently featured in the Skyrise City Extraversion model with no information about the model itself. That is, the variables were shared independednt from one another and with no model weights assigned. The variables were divided between maximal and typical performance, and three predictive models were developed to investigate whether maximal and typical performance both predicted a concurrent measure of Extraversion, and which type of data explained the most variance. One model included all the variables and the other two each included only one type of variables. The expectation was to observe a good portion of the variance being explained by maximal performance, possibly more than typical performance, but ultimately to find that a model containing both types of variables would outperform both single-variable-type models, as it should be expected that personality would encompass both. This did not emerge, as the model made up entirely by maximal performance variables outperformed the other models quite substantially. Furthermore, when the models were compared, it became clear that the inclusion of typical performance variables lowered the predictive validity of the model and also affected the distribution of the simulated Extraversion scores.

These results, whilst not perfectly aligning with expectations, fully provide support for the idea that due to the nature of the tasks used in BBA to generate personality scores, BBAa are objectively hard to fake. Of course this study does not directly investigate faking in action, but its focus on the type of response options and the nature of data used to generate the model, and especially the focus on measures of maximal performance, is enough to safely draw this conclusions from a purely

format perspective.

In addition to the fact that in BBAs the measurement of Extraversion, and personality in general, relies on measures of maximal performance heavily enough to prevent faking, and this will be explored more in depth in Study 4, there are more objective abilities that come into play when attempting to fake BBAs that go above and beyond the typical vs maximal performance comparison.

One is the ability to understand the relationship between task behaviour and psychometric scores. Looking back at the literature review, there is plenty of evidence linking task performance to individual differences in Extraversion, however, the literature itself is not exactly consistent. Furthermore, the results of this study shows that, despite overall being a valid measure of Extraversion, the way in which the experimental tasks are used in Skyrise City to generate trait scores is not perfectly aligned to the existing evidence, and it involves much more data than simply tasks' main effects. In order to be able to fake the BBA successfully, it would be necessary to have access to the proprietary models of personality defined by the test publisher, and these are likely to be highly complex behavioural models encompassing data, paradata and trace data in a way that is not necessarily intuitive to humans or easy to translate into actionable behavioural patterns.

Another way in which objective ability would be involved in faking this type of assessment is being able to reproduce the expected behaviour, and this assumes surpassing the hurdle described above, which is highly unlikely. Of course there is a non-negotiable limit imposed by BBAs' heavy reliance of maximal performance measures, but, some tasks used in Skyrise City are not entirely impossible to fake despite being maximal performance measures, especially if the nature of the task is known in advance and aids are available. Whilst it is not possible to fake higher working memory unaided, the Corsi Block is rather easy to fake with pen and paper; the candidate only needs to draw a grid and then write numbers in the corresponding squares to replicate the sequence in the right order, and this works with or without distractors. In the Flanker Task, the effect of incongruence can be completely eliminated by covering the peripheral arrows and focusing on the central one only - for

this, a piece of paper with a hole in the middle would be enough. This being said, other tasks are genuinely impossible to fake, Ticket Master for instance does require executive function effort no matter what a candidate may try, and Promotional is designed to be completely impossible to finish by the developers. Furthermore, it would be easy to identify fakers as their interaction with the assessment would be distinctively flagging those faking strategies via specific patterns of latencies and responses; for example, a complete lack of congruence effect in the Flanker Task is extreme unlikely, so if a candidate shows no speed or accuracy cost, this could be used very safely as a tampering flag. However, not all BBAs have in-built faking flags so some faking might go unnoticed.

That said, the real issue is different. First, the trait models are likely to be too complex to be deliberately *and* successfully reproduced, and, second, as demonstrated in the first batch of hypotheses, simple task performance is hardly ever linked to trait scores, and manipulating task behaviour may actually result in a detriment rather than an advantage, as paradata and meta data have a lot more influence in the scoring model that main effect data, and those are likely to be affected by artificial behaviours. Study 5 in this thesis shows evidence for this phenomenon at the outcome (score) level.

Finally, there is a Schrödingeresque semi-paradoxical twist in the role of objective ability in faking BBAs. As each experimental task is involved in many trait models and each trait model draws data from several experimental task, for any given behavioural pattern there might be a mutually exclusive degree of success. That is, whether faked or genuine, a given behaviour is equally desirable and undesirable. The degree of desirability is entirely dependent on the trait model that is using it, and trait scores are generated in tandem. So, in order to successfully tailor a psychometric profile, a candidate would have to be aware of several different model at once and somehow be able to mitigate this paradox behaviourally. Which is impossible.

4.3.4.1 Implications

The lack of correlations between tasks' main effects and Extraversion trait scores makes the results of this study even more remarkable, because the BBA would be a lot easier to fake if the easiest-to-manipulate behaviour involved in the assessment was predictive of trait scores. This means that even if some candidates might use strategic faking aids, they are highly unlikely to be able to improve their scores. The data suggests that it is in the candidates' interest to approach the BBA as naturally as they can to avoid affecting their scores for the worse. This means that BBAs are not only intrinsically resistant to faking, they may actually also penalise faking attempts. This echoes the practice of placing ink capsules inside security tags to prevent the theft of clothing, so that if the security tag is not enough to prevent and detect the theft, the ink in the capsule permanently damages the garment upon removal making theft pointless. Study 5 provides some interesting experimental evidence about this.

Furthermore, guides on how to fake BBAs are now widely available online and they all overly emphasise task performance. Whilst Study 1 showed that warnings against the use of online guides do not significantly reduce intentions to fake BBAs, this study suggests that the use of online guides not only is unlikely to be of much help to candidates, it also more likely to be a hindrance.

On another note, the lack of replication of the correlations between most tasks main effects and scores of Extraversion may add to our ability to navigate the ambiguity in the experimental evidence discussed in the literature review. It is plausible that simply focusing on task main effect may be too reductionist and more data should be consider in experimental studies that allows for more subtle nuances in performance to be captured. Considering the incredible complexity of the neurobiological structures and processes involved in individual differences in personality, experimental tasks constrained by highly arbitrary cut offs and over-relying on summary data (i.e.: total accuracy, or accuracy difference between congruent and incongruent tasks, etc) may conceal entire layers of trait-relevant variance that were previously overlooked and therefore not understood.

Nevertheless, this study was designed to provides sufficient evidence to suggest that when faking is a concern, BBAs are a viable alternative to self-reported personality assessment, and it was successful in this pursuit. Not only the results demonstrate that scope for faking is minimal, if any, and enable an explanation of why this is the case, the results also show that BBAs have high levels of convergent validity with existing measures of personality.

4.3.4.2 Limitations

Whilst the results of this study are strong in demonstrating that BBAs offer very little scope for objective ability to fake, it might be dangerous to overemphasise the use of ability-based measures for personality assessment, as cognition most certainly plays a role in this which might present the risk of discriminating unfairly against certain groups. Inter-individual differences in reaction times and response accuracy might be due to other factors which are not related to individual differences in personality, and it would be impossible to ascertain the specific sources of observable behaviour at the individual level. This means that the same behavioural pattern may be a genuine expression of a latent personality trait or of something completely unrelated to it. This is however not different from other assessment, as a response on a self-reported questionnaire or a situational judgement test might equally be the representation of the trait the assessment measures and/or the representation of many other factors. The difference is that, whilst it is easy to manipulate responses in other types of assessments to obtain a certain score, with BBA the concern is not around deliberate manipulation but around possible confounding factors beyond the responsibility and intention of the candidate. So it is more of an ethical than a methodological validity consideration. This however, as far as it does constitute a potential issue with validity, does not reduce BBAs' expected level of resistance to faking.

It is also important to reflect on the fact that the data used in this study is low stakes data, that is, it's not data that was collected during a job application, and it is not clear if and how the results would replicate in high stakes settings. It is likely that some of the differences would only be quantitative but some qualitative

differences in behavioural patterns might also emerge, meaning that the relative contribution of typical and maximal performance to the overall Extraversion model could be different. Whilst a qualitative shift radical enough to challenge the replicability of these findings is highly improbable, it is still possible that the role of maximal performance could be slightly inflated by the context in which this study was set, but it could also be the opposite.

Whilst this could be a concern, it is important to note that this remark only pertains to the relative influence of maximal performance over typical performance and does not imply that there could be a scenario in which maximal performance does not account for a good portion of the variance in the model. The model itself contains many more maximal performance than typical performance variables, and even if stakes may affect how much each contribute due to different levels of attention (see Study 4 which focuses on this specifically), there will still be a substantial contribution of ability measures on BBA personality models which will protect the assessment from faking.

4.3.4.3 Conclusion

This study shows that BBAs do not allow enough objective ability to fake because of the way in which they are developed, which is to include a very large proportion of measures of maximal performance. As maximal performance is by nature impossible to fake, this makes BBAs also impossible to fake, at least for as much of the variance explained by those measures.

This is possibly the most crucial piece of evidence of this thesis as it directly addresses its main research questions and purpose. No matter how strong the motivation and intention might be, or how confident somebody might be about their faking strategy and ability to fake, when faced with a BBAs, candidates will find themselves in a behavioural *cul de sac* imposed by the format forcing them to make a u-turn on their intentions and perceived likelihood to succeed in faking.

4.4 Ability General Discussion

This chapter features a unique combination of qualitative and quantitative data, and it explores the concept of ability from both the perspective of the individual and the perspective of the assessment, providing insight on subjective, or perceived, ability to fake, and also on objective ability, or scope, to fake.

Collectively, the results of this study make a strong case to suggest that candidates would not be able to fake BBAs. Considering people's inability to work out a viable strategy to fake this type of assessment and the fact that most of the variance in the BBA model of Extraversion is explained by ability-based measures, the only viable conclusion is that BBAs cannot be faked.

Whilst these results are rather unambiguous, they do not test faking behaviour in action. The focus of this chapter was to provide evidence pertaining the reasons why candidates may not be able to fake BBAs more than it was to provide experimental evidence of their inability to do so.

As the first chapter was entirely focused on observing differences in perceived behavioural control and intentions to fake, this chapter corroborates that evidence by adding two key pieces of the puzzle focusing on the *why*. First, as it was expected by observing differences in PBC across formats, participants found it more challenging to envision strategies to fake BBAs than any other formats, and the two are likely linked. While neither PBC nor the inability to generate strategy ideas are a definite marker of the objective ability to fake, the second major contribution of this study was to demonstrate that BBA as a format doesn't offer enough opportunities to fake. This adds to the validity of both study 1 and study 2 results in that it suggests that those perceptions were well founded.

However, to fully demonstrate that BBAs cannot be faked, one more step is necessary, which is to explore what happens when people complete the assessment in situation in which high intentions to fake may be expected. That is, the behavioural outcome of faking should be tested empirically to validate the theoretical assumptions of the existing models of faking whereby intentions/motivation and ability to fake are the necessary prerequisites of behaviour.

The following chapter focuses precisely on this. Each targeting different aspects of assessment performance, two studies investigate whether BBAs are susceptible to shifts in intentions and motivation. The two studies are both designed to assume that the moderating situational effect of valence towards doing well in the assessment will be high enough to override the antecedent effect of perceived behavioural control on intentions to fake for candidates but not for employees/control group, thereby isolating the contribution of objective ability to fake as the only plausible predictor of faking behaviour. That is, if BBA data does not differ across samples on the basis of motivation, that it can be confidently inferred that faking did not occur due to ability.

Chapter 5

BEHAVIOUR

5.1 Behaviour General Introduction

This chapter explores the final block of the Integrated Model of Faking, that is the observed faking *behaviour*. Faking behaviour is the outcome measure of the model, and it's believed to be the byproduct of the combination of individual and situational antecedents of intentions and ability to fake, whereby ability moderates the relationship between intentions to fake and the emergence of faking behaviour.

Most of the existing experimental literature on assessment faking focuses on behaviour and can be summarised as the observation of performance differences between candidates and employees, whether real ones from field studies or instructed ones in lab settings. This thesis, after addressing intentions/motivation and ability, will now feature a closing chapter entirely focused on behaviour to complete the experimental journey around the Integrated Model of Faking. In line with the existing body of evidence, two studies will compare the faking behaviours of groups differing on their (assumed) levels of intentions to fake.

In the first study (Study 4) the BBA trait scores of real candidates and employees are compared, and variable-level data is compared to establish whether any eventual trait-level differences can be explained by means other than faking. The second study (Study 5) assesses participants' ability to impersonate a specific psychological profile on two different types of assessments, one of which being a BBA, for a cash reward, and it attempts to reduce the moderating effect of subjective abil-

ity from the intention-to-behaviour link by providing clear information pertaining what traits were being assessed and what the desirable scores of those traits were.

In other words, this chapter investigates the faking behaviour in BBAs of individuals with supposedly high intentions to fake by interpreting their variable-level data to establish whether higher trait scores could be due to successful faking behaviour, and by enhancing their ability to fake by telling them what scores they are expected to produce for which traits.

5.2 Study 4 - Candidates vs Employees

According to research, BBAs may be effective in addressing the issue of faking in high stakes settings, and good theoretical points have been made in the literature that suggest how this type of assessment may be offering protecting from faking from several different angles across the most popular models of faking (see Landers & Sanchez, 2022; also see Section 2.3 on page 42).

However, the differences in performance between candidates and employees in these assessments are not well understood, nor they have been purposefully investigated.

5.2.1 Literature Review

5.2.1.1 Motivation and Assessment Performance

Previous studies have established a correlation between motivational differences and performance on assessments in general, as well as on some of the tasks replicated in BBAs (Manohar et al., 2015; Stoeber, Chesterman, & Tarn, 2010). These studies have also demonstrated that individuals perform better on assessments and tasks when provided with incentives (Bonner Hastie, Sprinkle, & Young, 2000; Dambacher, Hübner,& Schlösser, 2011; Hübner & Schlösser, 2010), and this is consistent with the expectancy-value theory, which suggests that an individual's perceived value of a task influences their effort and subsequent performance (Eccles, 1983).

Studies have also found that motivated individuals exhibit enhanced attentional control and cognitive control, leading to higher accuracy and lower reaction time in

task performance (Chiew & Braver, 2016; Engelmann & Pessoa, 2014; Padmala, Sirbu, & Pessoa, 2017). This is particularly important as BBAs are often developed using individuals in low stakes contexts, where they may be unmotivated to exert effort (Arctic Shores, 2016). However, these assessments are typically used in high stakes settings such as recruitment and selection, where individuals are highly motivated to perform well. Understanding the differences in assessment performance between individuals in different stakes settings can have important implications for the development and use of BBAs in various contexts.

A recent meta-analysis (Hu & Connelly, 2021) suggests that score discrepancies between high and low stakes completions of psychometric assessments might be due to faking. The proposed model implies that candidates tend to inflate their assessment scores but they don't do the same when they take the same assessment as employees. In the meta-analysis, low stakes assessment sessions are labelled as "honest", suggesting that psychometric assessments provide a more accurate view of the individual when they are not used in high stake scenarios. This makes perfect sense in the context of self-reported measure. As explained in Study 2, the most common strategy to improve scores in this type of assessment is deception, therefore it can be expected that when there is no motivation to lie, like in the case of employees, lower scores represent the un-inflated, honest representation of the test taker, whereas the opposite is true for candidates for whom higher scores are a combination of who they are and how they wish to appear and therefore not a valid representation of their psychometric qualities. A similar pattern can also be expected in ability tests, which, as referenced and demonstrated in Study 2, are best faked by means of producing correct answers artificially, hence resulting in higher and more dishonest scores in candidate samples compared to the lower and more honest scores generated by employees.

It is not clear how this effect would replicate in types of assessment, such as BBAs, for which deception and providing correct answers artificially are nor feasible or viable faking strategies, as demonstrated in Study 3.

This study is the first of its kind, and it investigates score-level and variable-

level differences between high and low stakes completions of a BBA, exploring whether the observed differences, if any, can be attributed to faking or if they warrant an alternative explanation. The rationale of this study is that, in BBAs, high stakes completions should set the bar for validity, whilst low stake completions are those affected by context and thereby less valid. This paradox is expected to emerge on the basis that BBAs, due to generating scores of personality at least partially through measures of maximal performance, do not lend themselves to deliberate performance enhancements, meaning that eventual stake-dependent differences should instead be attributed to employees having lower levels of concentration and motivation compared to candidates, and not to candidates faking.

5.2.1.2 Evidence of Score Differences Between Candidates and Employees

Many studies have investigated the differences in psychometric scores between candidates and employees in the past, all rather unanimously arriving at the same conclusion that candidates fake assessment whilst employees don't. Whilst this might sound obvious, it is important to establish this empirically in order to understand the potential impact of candidate faking and how to detect it, and comparing candidates and employees is an excellent way to isolate the effect of stakes on assessment variables and scores, as those groups, especially in within-participant studies, only differ by the context in which they completed the assessment.

Differences in scores between candidates and applicant have been detected across assessment types and constructs. Hu and Connoly (2021) found, in the meta-analysis referenced above, that studies comparing candidates and employees assessment scores observed large effect sized significant differences in all the Big Five constructs, and many other measures. For example, Lievens, Klehe, and Libbrecht (2011) found very large effect size differences (d=1.12) between the scores of candidates and employees on a measure of Emotional Intelligence, and they also found that the candidate sample had less variance than the employee sample (SD ratio = .86/1), and hence lower discriminant validity. Griffith, Chmielowski, and Yoshita (2007) found that, when they retested a sample of employees, between 30 and 50%

of them must have had inflated their scores on a scale measuring Customer Service Conscientiousness when they took the test as candidates. Furthermore, Birkeland, Mason, Kisamore, Brannick, and Smith (2006) ran a meta-analytic study finding that across all job types candidates score higher than employees, interestingly they found that the rank ordering of mean construct scores changed across job, suggesting that candidates are able to distort their profile according to the role to which they are applying - which is explored in Study 5.

Anglim, Molloy, Dunlop, Albrecht, Lievens, and Marty (2022) compared the Portrait Values Questionnaire (PVQ; measuring universal personal values) scores of a large sample of candidates to those of an employee sample matched on country, age and gender. They found some critical differences in scores suggesting a deliberate attempt in the candidate population to appear more responsible and considerate. Candidates reported valuing power, self direction, stimulation, hedonism, and achievement significantly less than employees, but conformity, universalism, security, tradition, and benevolence significantly more than them. This shows that candidate would not only try to appear more competent or suitable by endorsing role-relevant responses, but they would also pretend to be completely different people to secure a job.

A study also found that candidates would fake information on health questionnaires and life events. Bäccma, Sjöberg, and Almqvist (2015) tested a group of military recruits on a battery of measures first as candidates and then retested them months later when they were employed. They found that candidates scored higher on all the ”positive” measure, which included seven personality aspects and four aspects of the Sense of Coherence measure, with effect sizes up to .65 compared to when they repeated the assessment as employees. On the health questionnaires, they found that the candidates heavily downplayed ”negative” measures, such as depression, anxiety, and somatic problems, with effect sizes up to .74. A very interesting fact is that the study found that the odd ratio of reporting previous difficult life events and hardship was 2.07 times higher when this was reported as employees compared to when it was reported during the application - which was only 3 to 4

months prior. This degree of deception in such delicate and potentially critical personal aspects can be highly problematic especially in the context of military training and combat, and would most likely have serious implication for both employers and employees.

5.2.1.3 Implications

Faking would not be an issue if there were no implications with it but this is not the case. For instance, Jeong, Christiansen, Robie, Kung, and Kinney (2017) found that the large effect size of the difference in scores between candidates and employees (up to d=0.80) were responsible for substantial shifts in the validity of the selection measure. They found that sample membership was a strong moderator of the relationship between the selection measure and job performance, explaining 27% of the variance in the correlation.

Donovan and Dwight (2013) examined the prevalence of candidate faking and its impact of the psychometric properties of the measured that were used for selection, the quality of the resulting hiring decisions, and employee performance. They compared the scores on a self-reported measure that a group of candidates completed when they applied for a sales position with the scores on the same measure after they had been hired, and they also looked at their training performance and actual job performance 5 months after the training. They found that about half of them had faked their self-reported scores during their application, and this did have some serious consequences. By affecting the psychometric properties of the measure, faking affected the quality of hiring decisions, and this resulted in overall poorer training and job performance. The study showed that honest candidates performed better and as the selection measures would have expected, whereas those who faked exhibited lower levels of job performance both compared to the non faking group and to what the selection measure would have expected.

Harold, McFarland, and Weekley (2006) compared the predictive validity of verifiable and non-verifiable biodata. They found that verifiable biodata such as length of employment, qualification, and positions held predicted performance such as task performance, organisational citizenship behaviours, and customer service

much better than un-verifiable biodata such as the ability to work for long periods of time, one's leadership skills compared to others, etc. Interestingly, they did not find differences in overall biodata scores between candidates and employees, but the magnitude of difference of predictive validity between verifiable and non-verifiable biodata was larger in the candidate sample. This shows that the degree of verifiability of biodata items was sufficient to affect predictive validity of the assessment, suggesting that even small manipulations can have a serious implications on the overall validity of candidate selection decisions. Whilst there is an argument to be made about the differences in objectivity between verifiable and non-verifiable biodata which could suggest that faking is equally as a plausible as an explanation as the lack of self-insight or frame of references might be, the fact that the difference in predictive validity between verifiable and non-verifiable biodata was smaller in employees compared to candidates does suggest that faking was the reason behind this effect, and that biodata loses at least some predictive validity when used in high stakes.

Furthermore, the detrimental effect of faking on hiring decisions might also be partially responsible for adverse impact. Herde, Lievens, Jackson, Shalfrooshan, and Roth (2019) published a paper showing that subgroup level differences in SJT scores are exacerbated by faking. They found that the effect size of score differences between Black and White individuals almost doubled from an employee (d=0.38) to a candidate sample (d=0.66), which raises a very important flag on the seriousness of the implications of candidate faking.

In the Griffith et al (2007) cited in the previous sub-section, the author estimated how many employees in their sample - who were hired on the basis of their candidate score - should have not been hired on the basis of their employee scores. They found that with a 50% hiring ratio, 31% of people should not have been hired, and the figure raised to 33% with a 20% ratio, and 66% with a 10% ratio. This is similar to the results of a simulated faking study in which 65% of people hired with a 10% selection rates came from the "fraud" group but only 18 of them (compared to the 28 making up the 65% of the hired sample) should have qualified according

to their honest scores.

A practitioner-oriented paper (Bott, O'Connell, Ramakrishnan, & Doverspike, 2007) questions whether the common practice of setting pass marks on the basis of internal validation is a good idea. The paper compared the scores of candidates and employees in several cognitive and non-cognitive measures and they identified some significant group differences with effect sizes ranging from d=0.16 for cognitive measures to d=1.05 for non-cognitive measures, which influenced pass rates. In the study, three scoring models were compared, all of which based on employee data and each with a different pass mark and rules. The results showed that many more candidates than expected "passed", and in some cases this figure was more than 90%, which present a great challenge in terms of costs, logistics, and time required for further screening. In other words, the level of faking observed in this study made the assessment stage under scrutiny entirely redundant.

This type of evidence shows the importance of being able to prevent or at least identify faking to retain the validity of the selection measures and prevent fakers from securing jobs for which they are not qualified at the detriment of suitable honest candidates.

5.2.1.4 Faking Identification

Whilst preventing faking is not always possible, detecting faking might help reduce the negative impact of faking on hiring decisions. Several studies have described methods to identify faking by observing the ways in which candidates and employees approach assessments, and overall found more distinctive patterns of data in faked assessments compared to honest ones. For example, Salgado (2016) examined the data of 46 studies using the NEO-PI-R to assess the Big Five, and tested a model based on the assumption that faking generally results in the homogenisation of scores by increasing the mean and decreasing the standard deviation, and that this in turns affects the assessment's covariance, reliability and validity. Their results fully supported the model, showing that candidates' data is much more restricted in range and has in average more desirable mean scores.

In a large semi-experimental sample, Landers, Sackett, and Tuzinski (2010)

observed the evolution of a specific faking response pattern over time. The author collected personality data from a large sample of candidates and employees, and found that, following a coaching rumor, employees who failed the assessment as candidates started exhibiting blatant extreme responding (BER; the endorsement of extreme values in a rating scale) increasingly frequently upon retest. An interactive warning placed on the assessment a year after BER was identified reverted the effect of the coaching rumor, showing that using specific data pattern to identify aberrant responses can be useful to detect and deter faking.

However, endorsing extreme values may not always indicate faking, as honest respondents may also feel very strongly about some of their characteristics. Also, test fakers may be quite parsimonious with their extreme responding and limit it to the traits, or even items that they consider job-relevant so extreme responses might go unnoticed in skilled assessment fakers (Kiefer and Benit, 2016). Kuncel and Borneman (2007) also challenge the reliance on extreme values endorsement by proposing a new technique focusing on idisyncrating item responses. their technique implies focusing on items which generate dramatically different responses between honest and dishonest conditions which are not just as simple as a score increases. The authors suggest focusing on items showing bi-and tri-modal distributions in faking conditions but normal distributions in honest conditions, and then use the relative frequencies of a given response to infer the likelihood of faking and assign ”faking points” accordingly. Their approach is essentially targeting the level of oddity of the responses, and this is somehow also echoed in a method proposed by LaHuis and Copeland (2009) whereby multilevel logistic regression is used to establish how odd a given response is in comparison to what would be expected of an individual based on the rest of their data. Whilst this technique focuses on intra- rather than inter-personal abnormalities, both approaches bring attention to the importance of evaluating responses in context, demonstrating that the assumption that a specific response pattern might be a signal of faking in itself is reductionist at best.

Some researchers have focused on reaction times and latencies to investigate faking. For example, Robie, Brown and Beaty (2007) reported a verbal protocol

analysis study in which they found that honest test takers took in average less time to complete an assessment and went back to correct their responses fewer time, suggesting that "lying takes time" (p.504). However, the relationship between faking and response time is not entirely clear cut. In fact, Callegaro, Yang, Bhola, Dillman, and Chin (2009) investigated differences in response latencies between candidates and employees, and found some interesting differences. They were interested in understanding whether the cognitive processes underlying different approach to assessment behaviour were responsible for the differences and why. In the paper, they describe a four step cognitive process that respondents perform to produce an answer of a close-ended question: comprehension, information retrieval, judgement of the information retrieved, and selection of the appropriate answer option. Individuals might be *optimizers* and carefully perform all the steps, or *satisficers*, either performing all the steps but less thoroughly (weak) or skipping some steps (strong), and this distinction leads to different assessment results. The author believe that employees should be more likely to be satisficer than candidates as they don't have much at stake, whereas candidates should be more likely to be optimiser as they have more at stake. They also expected fakers to be more likely to be optimisers as the cognitive processes of faking require all the four stages to produce the desired response. They found that in general, responses became incrementally faster as the assessment unfolded, and that candidates were slower than employees. However, they also found that the difference between the groups reduced in the last half of the assessment, until disappearing completely in the last quarter. Whilst the results of the study show that candidates may be ore likely to be optimisers, this does not provide unambiguous evidence about the emergence of faking as the candidates could have simply have taken more time to increase accuracy of their responses.

Finally, in a recent review of modelling techniques aimed at identifying faking fingerprints, Kiefer and Beint (2016) stress the importance of collecting strong evidence before flagging an assessment session for faking, as this could have very serious consequences, pointing out that most faking detection strategies are based on response patterns that are not necessarily exclusive to faking. Whilst it is rel-

atively safe to infer faking at a group level, individual-level faking is hard to pin, especially with limited amount of data points as it is the case of self-reported assessment, and, considering the potential repercussions of false positives, caution must be always used.

The current study brings value to this argument by suggesting that, since faking should not be a problem in BBAs, a model able to identify high and low stake completions could be used to flag candidates who have not performed well enough due to low effort or distraction and invite them to repeat the assessment.

5.2.1.5 This Study

This non-experimental study uses secondary candidate and employee data to investigate whether the contextual performance differences observed in self reported measures in the literature are also observed in BBAs. Most importantly, this study investigates whether this effect might warrant, on the basis of the assessment format, an alternative explanation to the traditionally held belief that employees provide a honest picture of themselves due to the inconsequential nature of assessment in low stakes contexts, whereas candidates systematically inflate their scores to increase their probabilities of being hired.

The proposed alternative explanation stems from the idea that stakes-dependent differences in the perceived valence of an assessment might trigger different sets of behaviours depending on the opportunities afforded by the assessment. That is, depending on the scope for scores inflation an assessment affords, the observed score differences between candidates and employees might be attributed to different underlying mechanisms. It is possible that when the scope for boosting performance is severely limited by the assessment format itself, more motivated individuals outperform less motivated individuals through the exertion of maximal effort and not through self-presentation, and that less motivated individual relatively under perform due to a comparatively lower investment of resources that results in assessment outcomes below their potential, rather than a more accurate representation of themselves based on honesty.

This assumption is justified by the type of tasks used in BBAs. According

to several studies, performance on tasks involving attentional control, such as the Flanker Task, declines over time as indicated by decreasing accuracy and slower reaction time (RT) (Demars, 2007; Kato, Hendo, & Kizuka, 2009; Thomson, Seli, Besner, & Smilek, 2014). This decrease in performance over time is thought to be due to mental fatigue as attention deteriorates with repeated engagement in the task (Langner, Steinborn, Chatterjee, Sturm, & Willmes, 2010). However, motivation levels may moderate this performance deterioration. In a Flanker Task study by Bonnefond, Doignon-Camus, Hoeft, and Dufour, (2011), neural activity indicating decreased alertness was observed for both highly and less motivated individuals over time, but only the less motivated group showed actual performance deterioration. This suggests that highly motivated individuals are more likely to apply effort to maintain performance over time despite decreased alertness. Another study by Brosowsky, DeGutis, Esterman, Smilek, and Seli (2020) found that mind wandering was negatively related to task performance over time for individuals with low motivation, but this relationship was not significant for highly motivated individuals, and less motivated individuals are more likely to engage in random tapping due to lower levels of preoccupation with the outcomes of the task. This suggests that while performance declines over time for both groups, highly motivated individuals in high stakes settings may be less susceptible to mind wandering and random responses, leading to generally better performance but also to an less steep performance deterioration compared to less motivated individuals in low stakes settings.

This study draws from this body of evidence, and it also links all the elements of the Integrated Model of Faking together. The sample distribution of the study includes two groups differing on a situational factor potentially underlying faking intentions, that is, the context in which the assessment was taken and resulting valence attached to it. The task used for this study is almost entirely made of measures of maximal performance, thereby limiting the ability of the participants to deliberately improve their scores, and this taps on the ability block of the model. Finally, the assessment data of candidates and employees is compared, providing evidence of the interplay between intentions and ability on faking behaviour.

5.2.1.6 Hypotheses

To test whether the same score differences observed in traditional psychometric assessments between candidates and employees are replicated in BBA, and whether it is possible to explain those score differences as the byproduct of effort rather than self-presentation, this study scrutinises the variable-level data of a BBA replication of the Flanker Task, and five hypotheses are advanced:

- **Hypothesis 1: Personality trait scores will vary across samples** Scores of Determination, Self-Discipline, Self-Belief, and Stability generated from a BBA will show a stake-dependent performance effect whereby candidates' scores will be more desirable than employees' scores.
- **Hypothesis 2:** Task performance on the Flanker task will be better in high stakes than in low stakes conditions, more specifically:
  - *Hypothesis 2a:* Candidates will make fewer mistakes than employees
  - *Hypothesis 2b:* Candidates will respond faster than employees
- **Hypothesis 3: The Flanker Task's main effect (i.e.: the performance difference between congurent and incongruent trials, also known as cost) will be more pronounced in low stakes completions than in high stakes completions**
  - *Hypothesis 3a:* Independent of stakes, proportionally fewer mistakes will be observed in congruent trials compared to incongruent trials
  - *Hypothesis 3b:* Independent of stakes, faster responses will be observed in congruent trials compared to incongruent trials
  - *Hypothesis 3c:* Candidates will show a lower accuracy cost than employees
  - *Hypothesis 3d:* Candidates will show a lower speed cost than employees
- **Hypothesis 4: Candidates and employees will approach the assessment with a different attitude, generally showing higher levels of motivation and concentration, and more precisely:**

  - *Hypothesis 4a:* Candidates will spend more time reading instructions and will be more likely to repeat the instructions than employees
  - *Hypothesis 4b:* Candidates will show fewer incidence of early responses (suggesting random tapping) and missed responses (suggesting lack of attention) than employees
  - *Hypothesis 4c:* The performance of candidates will begin to deteriorate faster than the performance of employees
  - *Hypothesis 4d:* Employees will show less performance deterioration following a mistake compared to candidates

- **Hypothesis 5: A predictive model based on all the BBA task data available will be able to identify low stakes completions more accurately than high stakes completions, as those will be markedly more similar between them.**

5.2.2 Method

This study used secondary assessment data supplied by Arctic Shores to investigate difference in the behavioural patterns observed in a BBA task completed in real-life high and low stakes conditions.

5.2.2.1 Sample

The sample used in this study is a collection of different samples of live candidates and employees shared by Arctic Shores anonymously. The test publisher provided data from a good selection of high and low stakes client samples, overall matching the sessions from the two conditions by age, gender, demographic distribution, and seniority of the role which the person either held or applied for. The high stakes group included assessment data from job candidates, whereas the low stakes group included assessment data from current employees who took the assessment either during an internal validation of Skyrise City aimed at identifying the qualities of low, medium, and high performers within a company, or in the context of collecting psychometric data to support development or coaching conversations within the employees' company.

The high stakes group is made of 973 participants, whereas the low stakes group features 1497 participants.

5.2.2.2 Materials

The materials used in this study were the variable-level data generated by one module of Skyrise City (Arctic Shores, 2018) - Navigation, which is a replication of the Flanker Task (Eriksen & Eriksen, 1974) and is described in Study 1 and also summarised below - and four psychometric scores generated from the same level. Those were scores of Determination, Self-Discipline, Self-Belief and Stability, and they were chosen on the basis that data from the Navigation module constituted at least 30% of the variables in the trait model.

It is worth noting that the Navigation level is only one of several levels contributing to the scores of those psychological constructs, and that other trait models also use some data from Navigation, albeit in much smaller proportions.

Navigation, and the Flanker Task (Eriksen & Eriksen, 1974) on which it is based, broadly consists in indicating the direction of a target arrow within an array of arrows by pressing a button at either side of the screen (or keyboard) whilst ignoring the direction of the non-target arrows, which may or may not point in the same direction, or in any direction at all, depending on the trial type - i.e.: congruent, incongruent, and neutral.

The dataset was shared clean from outliers and missing data, and it included 207 variables.

For each of the 68 trial in the task (28 congruent, 32 incongruent, and 8 neutral) the dataset contained two variables: the trial outcome and the response's reaction time. The trial outcome could be either correct, wrong, early (i.e.: occurring before the stimulus was presented), or missed (i.e.: not occurring or occurring after the response window, ranging between 500 and 2000ms, ended). The variables were labelled with the type of trial they referred to and the ordinal number of presentation for that trial type. Each BBA session delivers a slightly different pattern of trials which is chosen at random during the delivery of the assessment, so it is not possible to establish the exact order of presentation of all the trials. That is, the numbering

of the three trial types each starts with 1 so it is only possible to analyse outcome and speed trends over time for one trial type at the time. However, the number of trial types is identical in every session, and so is the relative proportion of trial types across the four quarters of trials.

In addition to the single-trials' outcomes and RTs, composite variables representing sums, averages, and standard deviations were also shared, and so were variables pertaining other aspects of the tasks such as the amount of time spent on reading instructions or whether the candidate/employee asked to read the instructions again.

In addition to the variables shared by Arctic Shores, other variables were computed: the level of accuracy of each trial type (as the total number of trials differed between types and therefore the total number of correct trials for each trial type was not a reliable measure), the average number of outcome types and reaction times per quarter of trials over time for each trial type, and two scores representing the speed and accuracy cost of incongruence (i.e.: the difference in performance between congruent and incongruent trials).

5.2.3 Results

A series of analyses were ran on the data to test the five hypotheses advanced to compare the assessment data of candidates and employees, and identify stake-relevant differences in the way in which they approach the assessment.

5.2.3.1 Hypothesis 1 - Personality trait scores from high stakes completions will be more desirable than scores in low stakes completions

This hypothesis is observing differences between the two groups at the trait level and it is expected that candidates' scores will be more desirable than employees' scores.

Four Independent Samples T-Tests were used to compare the mean trait scores of of Determination, Self- Discipline, Self-Belief, and Stability between low and high stake completions.

The results showed that the hypothesis was supported for three of the four constructs. Determination scores were significantly higher in the candidate sample (M=-.1599, SD=.98875) than they were in the employee sample (M=-.2432, SD=.96523), t(2450)= -9.971, $p<.001$, and the same pattern was observed for Self-Discipline [(M=-.0655, SD=1.00281 low stakes),(M=.1005, SD=.97463 high stakes),t(2450)= -4.005, $p<.001$] and Self-Belief [(M=-.0587, SD=1.00054 low stakes),(M=-.0896, SD=.98571 high stakes),t(2450)= -3.611, $p<.001$]. Whereas the scores of Stability did not significantly differ between the two samples [(M=.0086, SD=.98584 low stakes),(M=-.0086, SD=1.00295 high stakes),t(2450)= -9.971, $p=.338$].

However, out of the three significant differences, only Determination scores differed with a small effect size (Cohen's d=.412) between the two groups, whereas the Self Discipline (Cohen's d=.167) and Self-Belief (Cohen's d=.149) did not surpass the threshold for small effect sizes (i.e.: Cohen's d=.2), therefore, the hypothesis is only fully supported for one out of four constructs.

5.2.3.2 Hypothesis 2 - Task performance on the Flanker task will be better in high stakes than in low stakes

This hypothesis focus on general task performance, advancing two sub-hypotheses - the first investigating differences in wrong and correct responses, and the second one investigating the speed of correct responses. Speed was only observed in correct responses in order to eliminate possible confounds (such as a variable incidence of random tapping or missed responses between stakes conditions).

Hypothesis 2a - Participants in high stakes conditions will averagely make fewer mistakes than participants in low stakes conditions This hypothesis was tested by comparing the mean number of correct trials and wrong trials between the two groups with two Independent Samples t-test.

Both analyses showed significant effects as predicted, whereby employees (M= 59.04, SD= 6.68) responded correctly to significantly fewer trials than candidates (M=60.32, SD= 6.26) t(2540)=-4.4766, $p<.001$ - Cohen's d=-.197, and they also made significantly more mistakes [(M=2.23, SD=2.29 low stakes),(M=1.68,

Table 5.1: High and Low Stake sample comparisons for performance in construct scores and Flanker Task

		High Stakes		**Low Stakes**		**Sample Differences**	
		M	sd	M	sd	t-test	Cohen's d
Constructs	*Determination*	-0.1599	0.98875	-0.2432	0.96523	-9.971***	d=.412 (S)
	Self-Discipline	-0.0655	1.00281	0.1005	0.97463	-4.005***	d=.167 (na)
	Self-Belief	-0.0587	1.00054	-0.8960	0.98571	-3.611***	d=.149 (na)
	Stability	0.0086	0.98584	-0.0086	1.00295	ns	—
Flanker Task	*Mistakes*	1.68	2.04	2.23	2.29	6.188***	d=.249 (S)
	Correct	60.32	6.26	59.04	6.68	-4.476***	d=198 (S)
	Speed	556.949	750.079	553.628	40.316	ns	—

SD=2.04 high stakes),t(2243.510)= 6.188, $p<.001$ - variance not assumed, Cohen's d=.249].

This supported the hypothesis fully, albeit with the number of mistakes showing a slightly more marked group effect than the number of correct trials.

Hypothesis 2b - Participants in high stakes conditions will averagely respond faster than participants in low stakes conditions The same analysis was used to compare the average reaction time for correct trials between the two group, and it showed a non significant difference between them [(M=553.628, SD=40.316 low stakes),(M=556.949, SD=750.079 high stakes),t(2450)= -1.193, p=.116] which was also in the opposite direction compared to what was expected.

Hypothesis 2 - summary of results: Hypothesis two investigated the differences in general task performance between the two groups, revealing that the speed at which candidates and employees responded did not differ between them but the number of mistakes and correct trials, and especially mistakes, did, suggesting that while the two group might approach the task with the same energy, their attention level might differ slightly, making employees generally more prone to mistakes than candidates.

5.2.3.3 Hypothesis 3 - The Flanker Task's main effect will be more pronounced in low stakes completions than in high stakes completions

The third hypothesis focuses on task effect. The Flanker Task was designed to demonstrate that systematic stimulus feature inconguence generating perceptual interference between stimulus and response affects task performance. The task effect

of Flanker Task is conceptualised as the relative difference between congruent and incongruent trials in terms of speed and accuracy - i.e.: speed cost and accuracy cost.

As mentioned in the Method section, accuracy was computed as the proportion of correct responses over total responses for each trial type as those differed in number and therefore the total number of correct responses would have been misleading.

Task effect was first investigated independent of stake condition to ensure that the BBA version of the Flanker Task did indeed replicate the task's main effect, and then the magnitude of the task effect was compared between the two groups, with four sub-hypotheses being advanced in total.

To establish whether the Skyrise City replication of the Flanker Task generated the same task effect, two Paired Samples t-tests were used to compared accuracy and speed across trial types.

Hypothesis 3a -Independent of stakes, fewer mistakes will be observed in congruent trials compared to incongruent trials The first t-test revealed a large significant task effect of accuracy whereby the mean accuracy of responses was significantly higher in congruent trials (M=.9361, SD=.08102) than it was in incongruent trials (M=.8202, SD=.13345), t(2451)= 54.768, $p<.001$ - Cohen's d=1.106, supporting the hypothesis that the interference caused by incongruence has an impact on response accuracy.

Hypothesis 3b - Independent of stakes, faster responses will be observed in congruent trials compared to incongruent trials Similarly, the second t-test revealed a large significant task effect of speed whereby the mean speed of responses was significantly higher (i.e.: smaller reaction times in ms) in congruent trials (M=498.7799, SD=44.03348) than it was in incongruent trials (M=537.2216, SD=44.83453), t(2451)= -60.313, $p<.001$ - Cohen's d=-1.218, supporting the hypothesis that the interference caused by incongruence has an impact on response speed.

Once it was established that the Flanker Task's main effect was replicated in

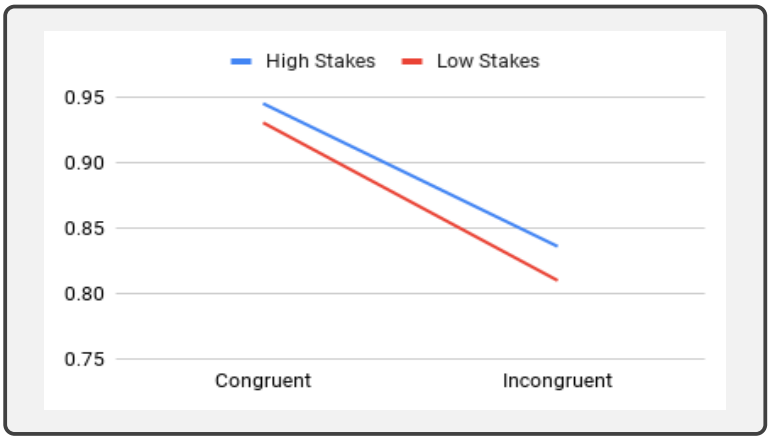


Figure 5.1: Accuracy across samples and trial types

Navigation, two further analyses were ran on the data to compare the magnitude of the task effect between the two samples.

Accuracy and speed costs refer to the differences of accuracy and speed between the two groups, by subtracting the incongruent accuracy scores from the congruent accuracy scores, and the congruent speed score from the incongruent speed score, so that the larger the difference, and hence the cost, the stronger the task effect.

Hypothesis 3c - Candidates will show a lower accuracy cost than employees Whilst hypothesis 2 revealed that candidates make fewer mistakes than employees, and hypothesis 3a revealed that accuracy is higher in congruent trials compared to incongruent trials, a two-way ANOVA replicated the two main effects [$F(1, 4900)=39.918$, $p<.001$; and $F(1, 4900)=1276.822$, $p<.001$], but failed to reveal a significant interaction of trial type and stakes on accuracy $F(1,4900)=3.118$, $p=.078$, $\eta_p^2=.001$, suggesting that the main effect of stakes was consistent on accuracy no matter the trial type, and that the main effect of trial type was consistent on accuracy no matter the stakes in which the assessment was taken. This is clearly depicted in figure 5.1 which shows two not exactly parallel, yet well-separated lines.

The data was further analysed to investigated whether the size of the task effect of accuracy, described as accuracy cost, differed as a function of stakes. To test this, an Independent Samples T-Test was performed on the data comparing the accuracy

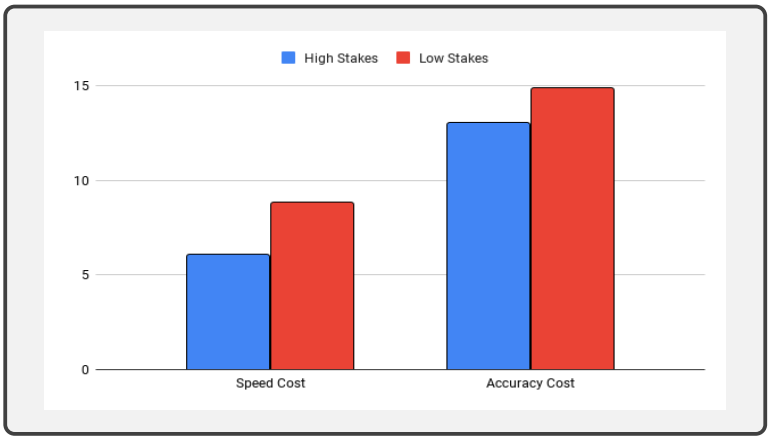


Figure 5.2: Flanker Task Main Effects across samples as percentage of performance deterioration in incongruent trials

cost observed in employees (M=.1203; SD=.10839) to the accuracy cost observed in candidates (M=.1090, SD=.09855), and this revealed a significant difference between the two groups t(2212.940)=2.676, p=.004 (variance not assumed, Cohen's d=.108), showing that incongruent tasks affected employees' performance significantly more than candidates' performance by causing relatively more mistakes. See Figure 5.2.

Hypothesis 3d - Candidates will show a higher speed cost than employees Contrary to accuracy, hypothesis 2 did not show a significant difference in speed between candidates and employees, whereas hypothesis 3b showed a large effect of trial type on speed. A two-way ANOVA was performed on the data to investigate this further and this revealed two significant main effects showing that speed did actually significantly differ between candidates and employees when corrected for trial type F(1, 4900)=9.296, p=.002, and it also significantly differed as a function of trial type when corrected for stakes F(1, 4900)=821.964, p<.001. The analysis of variance also showed a significant interaction of trial type and stakes on speed F(1,4900)=27.22, p<.001, η_p^2=.005, suggesting that the effect of trial type did not affect speed equally across groups.

Figure 5.1 in fact shows that while candidates (M= 500.456) responded more slowly than employees (M=497.678) in congruent trials, the opposite was observed in incongruent trials where candidates (M=530.792) responded faster than employ-

Figure 5.3: Speed across samples and trial types

ees (M=541.451), showing that interference did not affect speed in the same way across different stakes conditions.

The data was further analysed to investigated whether this meant that the size of the task effect of speed, described as speed cost, was larger for employees than it was for candidates, and when an Independent Samples T-Test compared the speed cost observed in employees (M=43.7739; SD=32.87709) to the speed cost observed in candidates (M=30.3366, SD=27.55349), it revealed a significant difference between the two groups t(2311.787)=10.931, $p<.001$ (variance not assumed, Cohen's d=.435), showing that incongruent tasks affected employees' performance significantly more than candidates' performance by slowing down responses more aggressively. See Figure 5.2 on page 176.

Hypothesis 3 - summary of results: taken together, the results reported above show that whilst Navigation generally replicated the Flanker Task effect independent of stakes, there are significant differences in how strongly the interference caused by the directional incongruence between the target and the non-target arrows affect speed and accuracy performance as a function of the condition in which the assessment is taken.

5.2.3.4 Hypothesis 4 - Candidates and employees will approach the assessment with a different attitude

Hypothesis 4 was advanced to test whether candidates and employees generally approach the task differently from each other in a way that might suggest a difference in the valence they attach to the outcome of the assessment, echoing the situational factors included in the model as precursors of faking intentions.

Five groups of variables were identified as potential places where to observe differences in effort, attention, or motivation; and five sub-hypotheses were advanced to investigate whether those differences emerged.

Hypothesis 4a - Candidates will spend more time reading instructions and will be more likely to repeat the instructions than employees One hypothesis is that candidates will be more interested in understanding what the task entails and they will therefore spend more time than employees reading the instruction, and they will also be more likely than employees to re-read the instructions a second time.

To test this, the average time spent on reading the instructions was compared between candidates (M=103.6712, SD=1079.93763) and employees (M=58.8769, SD=2.62) with an Independent Sample t-test, but the analysis did not show a significant difference between the two samples [t(983.148)=-1.585, p=.099, equal variance not assumed], despite the ostensibly very large difference in mean time spent reading the instructions.

Further to this, a cross-tabs analysis comparing the incidence of instruction repeats between two groups was performed on the data, and the results showed a significant difference [χ^2 (1) = 39.547, p <.001] whereby significantly more candidates than employees repeated the instructions (i.e.: 52 times compared to 16 times, with adjusted residuals= 6.3).

Taken together, the hypothesis is only partially supported as candidates do not seem to spend more time than employees reading instructions but they are more likely to go back to them and read them again.

Hypothesis 4b - Candidates will show fewer incidence of early responses (suggesting random tapping) and missed responses (suggesting lack of attention)

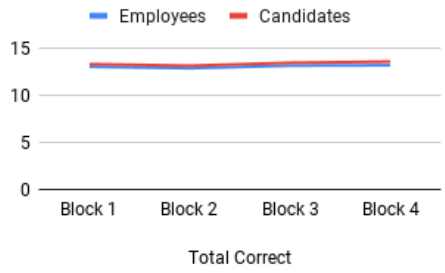
than employees Another hypothesis of a potential difference in how candidates and employees might have approached the task differently is that employees might be more likely than candidates to randomly tap the screen (or keyboard) to go through the assessment quickly rather than taking their time to produce the correct response, but at the same time they might be more likely to get distracted and miss the response window.

Hypothesis 4b compares the incidence of early and missed responses between candidates and employees to investigate if these two proxies for carelessness and distraction occur more frequently in employees than in candidates through two separate Independent Samples t-tests. In both cases, the analyses revealed a significant difference in the expected direction, in fact, employees (M=6.6626, SD 5.83215) were significantly more likely than candidates (M=5.9661, SD 5.49653) to engage in random responses [t(2450)=2.960, p=.002, Cohen's d=.122], and they (M=0.0703, SD=0.61920) were also more likely than candidates (M=0.0288, SD= 0.25898) to miss responses altogether [t(2138.148)=2.293, p=.022, Cohen's d=.082].

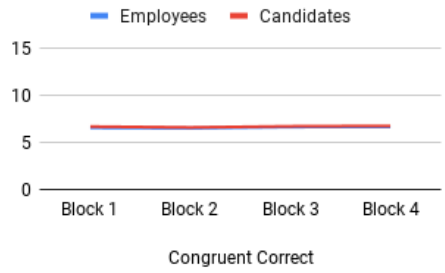
Taken together, these results suggest that employees might be less interested than candidates in the outcome of the task and that they don't pay as much attention to their responses. It should be noted that these differences were rather small.

Hypothesis 4c - The performance of candidates will begin to deteriorate faster than the performance of employees It can be assumed that candidates will strive to sustain attention and exert effort consistently throughout the assessment, whilst employees might "allow" their attention and effort to drop sooner. To investigate this, data from congruent and incongruent trials were each divided into quarters on the basis on order of appearance, and the performance of candidates and employees was compared over time.

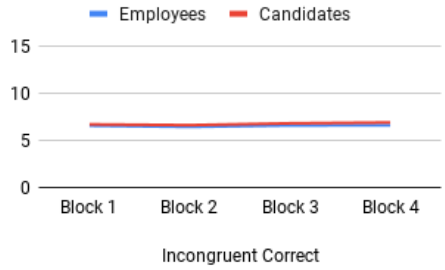
A series of two-way ANOVAs were performed on the data, comparing the evolution of the four response outcomes across time and trial types. The analyses generally showed consistent patterns showing a parallel evolution of performance between the two groups, and confirming the group differences revealed in previous



(a) Evolution of Correct Responses across time



(b) Evolution of Correct Congruent Responses across time



(c) Evolution of Correct Incongruent Responses across time

Figure 5.4: Correct Responses over time divided in four blocks.

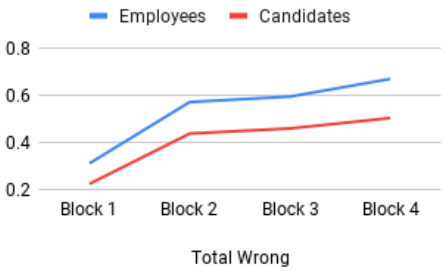
hypotheses, but no interaction between time and stake conditions.

Figure 5.4 depicts the trends observed in correct trials over time, which show no effect of time or group.

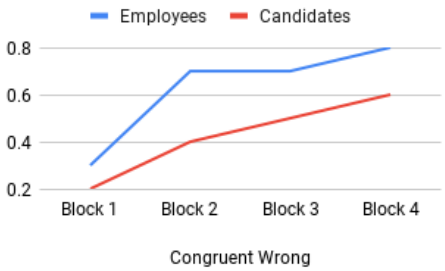
In two cases, however, the pattern in which performance evolved, and in this case deteriorated, significantly differed between the two samples.

Unlike for the overall number of wrong responses and the number of wrong responses in congruent trials, which increased in the same way over time for both samples, the incidence of wrong responses in congruent trials followed a markedly different timeline between candidates and employees whereby the performance of candidates deteriorated linearly over time, whereas the performance of employees deteriorated steeply between the first and the second block only to then plateau afterwards (Figure 5.5b).

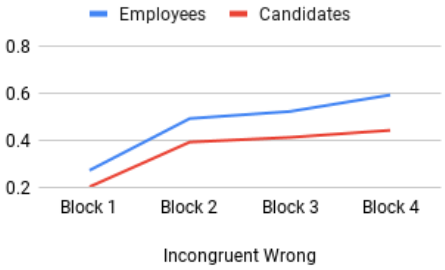
Similarly, yet oppositely, the incidence of early responses was distributed differently between candidates and employees over time - in this case, instances of early responses increased linearly over time for employees, suggesting decreasing



(a) Evolution of Wrong Responses across time

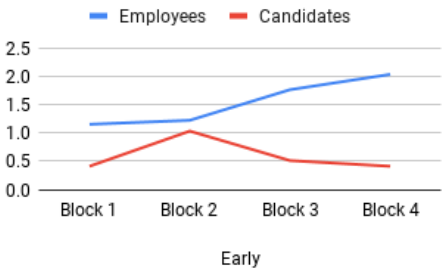


(b) Evolution of Wrong Congruent Responses across time

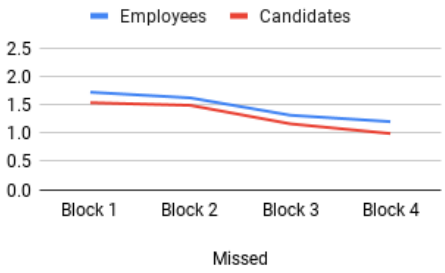


(c) Evolution of Wrong Incongruent Responses across time

Figure 5.5: Wrong Responses over time divided in four blocks.



(a) Evolution of Early Responses over time



(b) Evolution of Missed Responses over time

Figure 5.6: Missed and Early Responses over time divided in four blocks.

interest in the assessment results, whereas candidates showed a peak in the second block of trials which then decreased back to almost zero and plateaued (Figure 5.6).

Contrary to what was expected, employee did not seem to incrementally miss more trials over time compared to candidates, instead the two groups missed trials at proportionally similar rates over time.

Hypothesis 4d - Employees will show less performance deterioration following a mistake compared to candidates By looking at the impact that mistakes have on subsequent trials, it may be possible to infer that the more the individual cares about the outcome of the assessment, the more their performance might be affected by previous mistakes.

Hypothesis 4d compares the size of accuracy and speed performance deterioration caused by mistakes in the previous trial between the two groups to investigate whether mistakes affect candidates to a larger extent than employees. The two markers of performance deterioration were respectively measured as the proportion of correct responses in trials following a mistake, and the difference between average reaction times and post-mistake reaction times. The last variable should have ideally been the reaction time difference between post-correct and post-mistake trials, but this information was not supplied by Arctic Shores, meaning that the effect of mistakes on subsequent trials' reaction times is most likely larger than what is reported below.

Two Independent Samples t-tests were performed on the data to compare the accuracy and speed performance deterioration caused by previous trials' mistakes between candidates and employees. Whilst accuracy did not seem to deteriorate more for candidates (M=.9613, SD=.64655) than it did for employees (M=.9393, SD=.48159) as an effect of mistakes in the previous trial [t(2444)=-.966, p=.167], mistakes slowed down the response time in subsequent trials significantly more for candidates (M=50.3319, SD=750.10224) than for employees (M=17.1716, SD=35.31084), suggesting that, after mistakes, candidates became relatively more cautious and deliberate in their responses compared to employees [t(2444)=-1.695, p=.045].

As mentioned, due to the limitation of the data available for analysis, this last comparison might appear less remarkable than it is, as the post-mistakes reaction times are included in the total reaction times, thereby reducing the observed difference between the two variables in this study.

5.2.3.5 Hypothesis 5

Hypothesis 5 postulates that a predictive model would be more accurate at identifying data coming from employees than data coming from candidates. This is expected because candidates can be assumed to exert optimal effort to achieve good results and their ability to produce said good results will differ from one another, generating a healthy distribution of scores. Conversely, it can be expected that employees, being less adamant to obtain a good result from the assessment, might consistently under-perform and show similar patterns of data due to lack of attention and motivation.

For this it is expected that the higher degree of similarity in the employee sample will make their multivariate data distribution more distinctive and easier to classify compared to the more heterogeneous data generated by candidates.

A binary logistic regression was ran on the data which indicated that out of the 68 variables entered in the analysis, eight were significant predictors of stakes conditions [χ^2 (48) = 415.197, p<.001], as reported on Table 5.2.

Another 40 variables were not individually considered significant predictors of stakes, however, when combined with the other eight variables by the analysis, the 48 retained variables collectively explained 21.1% (Nagelkerke R^2) of the variance in participant's stakes group membership.

Table 5.2: Model impact statistics for significant variables in the Equation

Variable	**B**	**S.E.**	**Wald**	**Sig.**
Repeat Instructions	1.504	.307	24.057	<.001
Speed Variance in Correct Trials	.010	.002	19.361	<.001
Mean Speed in Inconguent Correct Trials	-.013	.005	6.590	.010
Speed Variance in Correct Incongruent Trials	-.006	.002	6.988	.008
Mean Speed in Neutral Correct Trials	-.008	.002	10.997	<.001
Post Mistake Correct Responses (not accuracy)	-.105	.022	23.669	<.001
Post Mistake Speed (not difference in speed)	-.008	.002	25.699	<.001
Accuracy in first block of Neutral Trials	-.352	.169	4.372	.037

Overall, the model was accurate in 67.8% of its predictions, however, as expected, the accuracy of the model varied quite substantially in predicting whether a BBA session belong to a candidate or an employee; in fact, the model can correctly

identify employees 82.1% of the times whereas it is only able to correctly identify candidates 46.2% of the times - which is below chance for a binary outcome.

This supports the hypothesis as it was predicted that, while the data from candidates is expected be relative normally distributed to reflect genuine individual differences between people, the data from employees would be more tightly bunched together in a way that reflects a specific and consistent pattern of low effort and motivation, making it easier to identify employee data than candidate data.

That said, surprisingly, the best predictor of the model was ”repeat Instructions” which is not part of any of the psychometric models and it’s also a binary variable, which by definition does not have a distribution.

5.2.4 Discussion

This non-experimental study was designed to explore the differences between real candidates’ and employees’ assessment behaviour in BBA to propose an alternative explanation to the observed trait-level score differences between the two groups whereby the higher scores of candidates are explained as the result of higher degrees of effort rather than self-presentation.

Score differences between candidates and employees have traditionally been explained in terms of self-presentation, whereby candidates have been demonstrated to systematically boost their scores compared to employees (reference?) due to differences in the valence towards the outcomes of the assessment. Due to this effect, it is commonly held (ref) that assessments taken in low stakes conditions paint a more accurate picture of the individual whereas the scores generated from assessments taken in low stakes need to be interpreted with the assumption that the candidate has made an attempt at inflating their scores.

Considering the results of the previous studies which shows that people have generally very low intentions to fake BBAs, have significant difficulties describing how they would do that, and, due to the large proportion of measures of maximal performance in BBAs, they are highly unlikely to be able to fake, this study looked at the variable-level BBA data of candidates and employees to test the hypothesis that, while BBA trait-level scores might differ as a function of stakes, these differ-

ences should not be attributable to deliberate score inflation but to a higher degree of effort.

The study focused on a single BBA module based on an established experimental task - the Flanker Task. Data from this module is used to generate scores of several traits, and the study started by investigating whether those varied as a function of stakes.

The results showed that out of four constructs using data from the Flanker Task, only the scores of one trait, Determination, were significantly better in the candidate sample. This means that the hypothesis is rejected but this outcome should be interpreted in favour of BBAs as it seems that the data available for scoring and the scoring models themselves are not as sensitive to testing conditions as other formats.

Whether the sensitivity equates to susceptibility to deliberate distortion or to sensitivity to variance in effort, these results suggest that BBAs are *mostly* reliable enough to withstand the challenge of context-dependent differences in assessment behaviour.

Interestingly, the only construct in which a small significant effect was recorded was Determination. As far as it is not possible to validate this speculation, the BBA model of Determination might use data and data patterns closely linked with sustained effort and attention throughout the assessment, which, in this study, is one of the hypothesised differences between the two groups.

If this was true, then these results would open a case for questioning the validity of the scoring model as the demarcation between *trait* Determination and *state* Determination would be too blurred to trust the results as the representation of a stable individual difference.

Nevertheless, the core aim of this study was to examine and compare the variable-level behaviours underlying the trait scores to demonstrate that wherever a difference was observed, this would be better explained as a byproduct of motivational differences and not faking.

The rest of the analyses focused on this and the results showed behavioural and

performance differences that were generally in line with expectations.

When looking at simple cognitive performance, the data showed that candidates and employees approached the task at the same speed as each other yet their levels of accuracy were significantly different from each other. This means that while both candidates and employees might have approached the task with the same energy, their attention levels were different and employees made significantly more mistakes than candidates. Being accuracy a measure of maximal performance, it is fair to attribute this difference to a low level of attention in employees and not to a potential attempt at improving scores in the candidate sample.

The main task effect was also investigated and the analyses showed that on both speed and accuracy accounts, the cost of interference was significantly larger for employees than for candidates, meaning that candidates engaged more inhibitory cognitive control than employees when approaching the task. Again, this difference is clearly the byproduct of higher degree of effort as inhibitory control cannot be improved above the individual's capacity.

In addition to this, the reaction times data showed an interesting interaction effect whereby not only the speed cost was higher for employees but the speed at which employees and candidates responded in congruent and incongruent trials were significantly different but in opposite direction depending on trial type. That is, candidates were faster than employees at responding in incongruent trials, whereas employees were faster at congruent trials. This shows a different behavioural approach whereby, in congruent trials, candidates employ a slightly slower and more deliberate approach than employees whereas employees performance is so heavily impacted by interference in incongurent trials that the effect is tipped, making them significantly slower than candidates. Results like this signal a marked difference in how the task is approached on the basis of stakes; candidates approach the task with more focus and deliberation, whereas employees comparatively lack motivation and perform less well than candidates.

The study explored differences in BBA data beyond the immediate scope of the Flanker Task. Hypothesis 4 proposed that there would be attitude differences in the

way in which candidates and employees approach the assessment, whereby candidates are expected to care more about instructions, engage less in careless response patterns such as random tapping and missed trials, maintain the same performance levels for longer, and be more affected by mistakes than employees.

Some of the sub-hypotheses were only partially supported, yet, several interesting results emerged.

First, whilst there was no difference in the time spent reading the instructions between candidates and employees, candidates were significantly more likely to repeat the instructions than employees. Upon further investigation, it appears that the way in which the instructions are displayed in Navigation doesn't lend itself to the expected reading time effect because the instructions are delivered as a series of very short sentences displayed in popups on the screen in a way that is a lot more similar to a tutorial than to "traditional" static instruction. What seems to be more relevant is the fact that candidates were significantly more likely to go back to the instructions a second time (or more), and this unequivocally signals a higher degree of intentions to understand the requirements of the task compared to spending more time reading the instructions. This is aligned to the hypothesis that candidates obtain better scores because the way in which they approach the assessment implies higher motivation; by ensuring that they understand the instructions and requirements of the tasks, candidates are able to set themselves up for more successful assessment sessions as they are more familiar than employees with what is required of them to succeed in the task, meaning that they are more likely to be faster and more accurate.

In terms of careless responding, two patterns of responses were investigated. The first, labelled as early responses, refers to instances in which candidates produce a response between the end of one trial and the beginning of the next one, that is, before the stimulus is displayed. This type of response is considered careless because it implies that the person might be tapping the screen (or keyboard) randomly in order to get through the assessment as quick as possible, disregarding the possibility of making a mistake. The second, labelled as missed responses, refers to instance in which candidates fail to produce a response in a trial, and this is con-

sidered careless because it implies that the person is not focused enough to produce a response at all. In both cases, candidates showed significantly fewer instances of careless responding than employees, supporting the hypothesis that high stake assessment behaviour is not more successful because it is inflated but because it is less affected by carelessness.

Another hypothesis was that candidate would be able to sustain the same levels of effort throughout the assessment - or at least for longer than employees - whereas employees' effort would decline faster. Timeline trend data was plotted for all the possible trial outcomes each averaged in four equivalent blocks. Most timelines looked very similar if not identical in terms of how candidate and employee performance deteriorated over time, and this was confirmed by ANOVA showing main effect of time and group but no interactions, but some exceptions to this emerged by inspecting the graphs.

Whilst the total number of errors and errors in incongruent trials increased in the same way for candidates and employers, errors for congruent trials showed a different pattern for the two stakes groups: while the number of errors increased linearly for the candidates, they peaked a lot sooner for employees and then plateaued. This pattern revealed that they begun making more *preventable* mistakes sooner than employees and they then maintained the same level of performance throughout. By the second block of trials, they more than doubled their rate of mistakes and surpassed the rate that candidate would only reach at the end of the task by 20%. It is peculiar that this difference emerged in congruent trials and not in incongruent ones. One possible explanation for this is that the relatively lower level of cognitive demand of congruent trials triggered even less attentional control, resulting in more mistakes emerging sooner. However, a more plausible explanation is that candidates' performance deteriorated at a slower and more consistent pace for less demanding trial, making fewer incremental mistakes. This is supported by the higher degree of dissimilarity between the evolution of candidates' mistakes in congruent trials and the evolution of mistakes in other types of trials for both groups, whilst employees made mistakes with the same timeline pattern in all the trial trial

types.

Another areas in which performance evolved in different ways for the two groups was the incidence of early responses. Here, employees' performance declined over time linearly, whereas for candidates it remained very close to zero with the exception of a hiccup in block two. It must be noted that the incidence of early responses is so low for both groups that it would be unwise to attach any strong claims to this data. Nevertheless, candifdates seemed to be able to control early response over time whereas employees engaged in this type of behaviour incrementally more as the task unfolded, signalling a potential crescendo of eagerness to finish the task which was not observed in the candidates.

One way to infer how much an assessment taker cares about the outcome of the assessment is by looking at the level of performance deterioration caused by mistakes. The results of the analysis performed to investigate this showed that whilst accuracy did not seem to deteriorate more for candidates than for employees, candidates approached the trials following a mistake significantly slower than employees, and also with a higher reduction in speed compared to them. This clearly signal higher sensitivity to the mistake, and a performance change aimed at increasing accuracy through increased caution in the candidate group that was absent in the employee group. Furthermore, due to the way in which the data was shared, it was not possible to remove the post mistakes reaction times from the total reaction times meaning that this effect might be even more robust than what is reported. Once again this shows a markedly different behavioural pattern between the two groups which in this case was limited to speed and did not equate to higher levels of accuracy, however, the absence of a relative disruption to accuracy might be due to the significant decrease of speed. Nevertheless, this signals a higher level of eagerness to perform well compared to employees, and not an attempt at manipulating scores.

Finally, a last hypothesis was advanced postulating that employee data would be easier to identify compared to candidate data. This is because candidate data is expected to reflect a more realistic variance of data scores and patterns representing

real individual differences between people compared the data of employees which is expected to suffer from restriction of range due to lower level of effort having a homogenising effect on task behaviour.

This is the opposite pattern observed in self-reported measures whereby high stakes data show skewness and restriction of range as candidates inflate their scores in the most desirable range of the scale, severely limiting the occurrence of low scores, whereas low stakes data is much more normally distributed as employees don't feel the need to improve their scores and low and high scores emerge with similar frequencies.

In a way, this final hypothesis addresses the core aim of the study - that is to demonstrate that in BBAs high stakes assessment data represents a more realistic overview of the individual compared to low stake data, and not a "curated" psychometric profile aimed at impressing employers. This is because candidates are not able to fake assessment but higher degrees of valence towards doing well in the assessment equate to optimal performance, which in turn equate to more normally distributed data, whereas lower degrees of valence equates to employees underperforming compared to their full potential.

A binary logistic regression supported the hypothesis showing that whilst a predictive model of stakes was able to identify low stakes conditions most of the times, it was worse than chance at identifying high stake conditions. Whilst it is not entirely clear from the model *why* this discrepancy emerged, and the high proportion of categorical data makes it difficult to compare the normality of the distributions, it is very plausible that the hypothesis was supported because of the rationale explained above.

The results of the study collectively support the hypothesis that in BBA candidate data represent the true scores of the individuals free of contextual influence and not the other way around. This is important because it demonstrates that this assessment format not only is not affected by cheating but it's also more likely to provide a valid overview of the individual to inform high stakes decisions than other types of assessment might be.

This is also important because, by being able to identify low stakes completions, it is possible to identify candidates who may not have sufficiently engaged with the assessment and invite them to repeat it, whereas this would be difficult with other types of assessment.

5.2.4.1 Implications

This field study has some important and highly reassuring implications for the use of BBAs in real candidate selection. The results makes two crucial contributions to the assessment literature and practice alike.

First, this study demonstrates clearly and unambiguously that candidates cannot fake BBAs. The data shows optimal performance and no signs of attempts to fake in the candidate sample compared to the employee sample which shows signs of carelessness and low effort instead. This field study used data from unproctored assessment sessions, meaning that this level of resistance to faking was achieved despite all motivational and situational odds.

Second, contrary to expectations, candidates and employees did not differ very much in their scores, and the small differences observed were remarkably smaller compared to those reported in the relevant literature. This means that BBAs may be able to retain the same validity coefficients across stakes situations, making them a great tool to use not only for selection but also for low stakes purposes.

This means that employers don't need to worry about choosing re-test or high stakes norm for selection as much as they would need to worry for other types of assessment. And it also suggests that BBA data is highly reliable and invariant across samples, which is quite rare.

One other positive implication of the results of this study is that low stake data can be easily be identified and with this candidates who may have not been able to give the assessment full attention. This shifts the usefulness of diagnostics into a much more constructive space whereby instead of being used to disqualify cheating candidates, with the risk of legal repercussions, they can be used to give candidates a second opportunity to fully showcase their quality. Furthermore, whilst this is not necessarily advised as it has not been tested directly, this information could also be

used as a marker of a candidate's motivation to obtain the job for which they have applied.

One less than positive possible limitation is linked to the fact that Determination was the only significant and fairly sized difference in scores. Whist the difference small (d=0.412) and anyway not as large as some of the figures reported in the body of evidence referenced in the literature review, the nature of the constructs raises a concern on the validity of the measure. Whilst the scoring model for this trait is not known, and the way in which it is described in Skyrise City technical manual provides no clues on this, it is possible that some of the data that differentiates high and low stakes performance are included in the model, such as slower reaction times, higher accuracy etc. This means that the model might have been "intuitive" rather than based on experimental evidence, and due to this it may tap on at least some situational rather than solely on dispositional variance. As BBA is a very new assessment method, not enough evidence is available to date to rule out this possibility. Considering the otherwise robust nature of this type of assessment, future research should investigate whether BBAs might be more suitable for some constructs rather than other, and whether lack of experimental and theoretical evidence supporting the link between certain tasks and individual differences in personality should really be subsidised with data-driven models.

5.2.4.2 Limitations

One limitation of this study is that it was between participants, however, with such large samples and with how closely they were matched on all the most crucial demographic factors, this should not be a cause of great concern. Of course, future research should address this by means of re-testing candidates in low stakes, however, as far as this sounds ideal, there could be a risk that this would reveal poor test-re-test reliability. The fact that reliability statistics of BBA/GBAs are typically absent in otherwise quite comprehensive technical documentations, it might be legitimate to assume that this type of assessment might have some issues around stability. Which calls even louder for this type of investigation.

Potentially, whilst high motivation can be assumed for candidates, motivation

to fake cannot. Especially considering the results of the first two studies in this thesis, it is possible that candidates, not knowing how to fake, didn't even try to do that. Little this matters though as the main point was to prevent faking and this happened. Also, considering the results on Study 3 it is to be expected that eventual faking attempts would have not been successful. That said, the nature of the differences observed between the two groups pointed very strongly towards the employee group been not entirely motivated, and there is no justified reason to suspect that faking attempts were made. The most plausible explanation is that proximal motivational factors and subjective ability played a role strong enough to preventing faking and, if any faking occurred, it was so unsuccessful that it did not make itself noticeable.

5.2.4.3 Conclusion

This study provides remarkable evidence showing that even in high stakes situation, BBAs cannot be faked. Data from the two samples show clear signs of low motivations in employees and optimal performance in candidate, suggesting a completely reverse pattern compared to the existing evidence from similar studies but with different types of assessment.

The point of this study was to demonstrate that BBAs can go as far as fostering optimal performance but not beyond the threshold of a person's dispositional boundaries. This echos the results of study 3 which showed that much of the variance in BBAs is explained by measures of maximal performance, and provide direct evidence for this from a high stakes behavioural perspective.

There is a small chance that the candidates whose data was used in this study did not want to fake, and this could be for a range of reasons including that they may not have known how to do that and they did not want to gamble their chances to get a job. To rule out the possibility that the results might be confounded by this type of interference, the next study is set up as a simulated job application. It will actively instruct participants to fake a psychometric profile and will also provide clear information about the desirable qualities of the ideal candidate through job descriptions. Furthermore, to increase motivation, the participants will be offered a

cash prize ten times higher than their baseline compensation if they ”get the job”. This should control for all the possible confounding variables that was not possible to control in the current study.

5.3 Study 5 - Job Application Simulation

Are people able to fake BBAs? At this point of the thesis, several factors have been established that strongly suggest that the answer is no: first, Study 1 demonstrated that intentions to fake BBAs are underlied by the same antecedents underlying other assessment types but they are generally lower due to lower levels of perceived behavioural control, so in general, people are less motivated to fake BBAs. Second, both subjective and objective ability to fake BBAs are impacted by its format so that, on one hand, individuals find it much harder to think of a strategy to fake BBAs compared to any other format type (Study 2), and on the other hand, the format itself offer very little scope for faking as it is almost entirely made of maximal performance measures (Study 3). This means that those with enough motivation may not know how to fake, and despite that, they have very little opportunities to do that. Finally, Study 4 in this chapter compared candidate and employee data and found no evidence of faking in candidates, suggesting that motivation to fake is not enough without a viable strategy or when the assessment does not invite faking.

Whilst it is highly unlikely that the candidates in Study 4 had no motivation to fake, this was not tested or manipulated experimentally, and it is also possible that they might have been motivated to fake but not knowing how to made it too risky. This study closes the loop by turning the last stone to evidence that BBAs are resistant to fake. Designed as a simulated job application, Study 5 instructed a group of participants to maximise their BBAs and self-reported assessment scores to obtain a job in a role described in much detail in a job description and win a cash prize, thereby insuring a good level of motivation to fake but also at least some subjective ability.

As Ajzen & Madden (1986) noted, one condition that determines the strength of the relationship between PBC and actual behaviour is the alignment between

perceived and actual control. That is, a person's perception of how much control they have over a behaviour needs to be substantiated with the real degree of control they have in order to be a good predictor of behaviour. So, if the same information about the successful outcome of an assessment is provided for two different types of assessment, then a difference in observed faking behaviour between assessment types should be attributed to the actual degree of control that the assessment offers.

5.3.1 Literature Review

Many studies before this have reported experiments in which participants were asked to modify their assessment behaviour to obtain scores aligned to different job roles. However, this is the first study comparing self-reported measures and BBAs in this context.

Ziegler (2011) defined a complex model to describe the cognitive processes underlying faking, showing that faking is a multi-step and multi-faceted cognitive process requiring conscious effort an motivation, so it should not expected to emerge simply because of instructions. Most studies asking participants to fake have used incentives to ensure that participants would be motivated to fake, and this has typically resulted in the desired outcomes. Successfully generating the intended assessment scores also require a good degree of metacognition in order to decide whether to lie or to respond accurately. Furnham (1997) tested whether participants could accurately rate their own Big Five score, and he found that they could predict their scores on Extraversion, Conscientiousness, and Emotional Stability reasonably accurately, but Agreeableness and Openness to Experience less so.

The most cited study in this field is the one from Viswesvaran and Ones (1999), which meta-analysed a large number of studies in which participants were asked to fake their Big Five assessment scores. They found that between-participant faking good studies observed effect sizes ranging between .48 and .64, whereas within-participants faking good studies observed much larger effect sizes ranging between .47 and .93. They also reported faking bad studies, and the effect sizes were even larger - between -.1.00 and -1.95 for between-participants studies and between .91 and -3.34 for within-participants studies.

Several other studies have demonstrated that faking can also be highly targeted to the specific requirement of a situation. For example, Furnham (1990) was one of the first to test people ability to impersonate different profiles through personality assessments, and he found that self-reported measures were highly susceptible to faking and enabled the accurate self-presentation of several professional profiles.

Roulin and Krings (2019) provided different instructions to their participants making them think that the organisational culture in a simulated application was either competitive or collaborative, and they found that Honesty-Humility and Agreeableness scores on the HEXACO reflected the group to which they were assigned; similarly, portraying a company as innovative resulted in higher Openness scores. They also tested their participants' opinion of what the ideal profile for those companies would be and they found that their scores on this measure mediated their scores on the relevant traits, suggesting that the perception of what an ideal employee would look like guided their faking strategy.

Targeted faking is not limited to experimental setting but is also observed in candidate samples. In their meta-analysis, Birkeland et al (2006) found good evidence of faking in candidate samples, and they also found that the rank ordering of faked mean scores varied as a function of the perceived relevance of a trait in a specific scenario, suggesting that individuals are able to distort their assessment behaviour with a specific aim in mind.

Overall, it seems that individuals' ability to identify the criteria (ATIC) used to evaluate their performance has been demonstrated to increase their chances of faking successfully (König, Melchers, Kleinmann, Richter, & Klehe, 2006; Melcher, Kleinmann, Richeter, König, & Kelhe, 2004), and it is also believed that it facilitates job performance by providing an better understanding of what others consider important (König, Melchers, Kleinmann, Richter, & Klehe, 2007), and also because it is highly correlated with cognitive ability (Melchers et al, 2009). König et al (2007) found that ATIC scores were consistent across assessment types, and that partially accounted for the performance in the selection process, and that ATIC predicted performance in other selection processes above cognitive ability.

However, faking attempts don't always equate to success. o'Brien and LaHuis (2011) analysed data from two large samples, one made of candidates and the other made of employees, and found that a sizeable proportion of items showed differential item functioning between the two groups but in the opposite direction compared to what would have been expected. Furthermore, Boss, König, and Melchers (2015) found evidence of faking good and faking bad in the same compulsory military recruitment sample, and they attributed the incidence of faking good or bad to whether the recruit wanted to do the military service or not.

Furthermore, faking may occur with different degree of success across assessment formats. Allinger and Dwight (2000) reviewed a large number of studies on faking Integrity measures, and they found that whilst Integrity measures can generally be faked, certain types of assessment were far more susceptible than other to the effect of fake-good instructions and coaching. Overt measures, which taps directly on attitudes related to integrity and make no attempt to disguise the purpose of the assessment, showed very large effect sizes up to d=1.32, whereas personality-based measures, tapping on individual differences related to integrity, showed small effect sizes around d=0.37.

A study comparing the validity of theory of planned behaviour (TBP) constructs in predicting faking behaviour across two assessment types differing on the basis of fidelity. Scores on an interview (low fidelity) and a role play group exercise (high fidelity) in a mock selection procedure showed that TBP predicted faking in the interview but not in the role play exercise, suggesting that when TBP, and hence indention, is equal, the format is key in differentiating behaviour (Dürr & Klehe, 2018).

Similarly Martin, Bowen, and Hunt (2002) found that their participants were able to successfully fake the Occupational Personality Questionnaire to match a specific job by generating scores in line with what HR professionals described as ideal for the role, however, this was only possible in a normative version of the assessment and not in the ipsative version. This shows that the ability to identify criteria alone is not sufficient to facilitate faking, and that format plays a crucial role

in enabling behaviour.

Taken together, the two studies reported above both show that assessment formats are the most decisive factor in preventing faking, and this study aims to add to this strand of evidence. No instructional faking studies has yet been published to date that included BBAs, and this is the first time that participants have been instructed to impersonate a specific psychometric profile through this assessment format.

5.3.1.1 This Study

Like the provious study, this study focuses on the outcome block of the Integrated Model of Faking, and it investigates how well people can modify their assessment *behaviour* to impersonate a specific profile in a context in which their intentions and subjective ability to fake have been deliberately increased to isolate the role of the opportunities to fake offered by the assessment.

By framing the experiment as a job application simulation, participants are encouraged to impersonate the psychological qualities listed in a job advertisement for the chance to receive a cash prize. Their ability to generate the intended psychometric profile is tested and compared on two assessment methods measuring the same personality traits - a BBA and a self-reported questionnaire.

By maintaining motivation and knowledge of constructs being tested equal throughout the study, the expected differences in faking success should be safely attricuted to the effect of assessment format.

It is expected that the BBA will lend itself to the task significantly less than the self-reported questionnaire, resulting in smaller effects sized differences between experimental and control groups, and less accuracy in generating the correct psychometric profiles as instructed.

5.3.1.2 Hypotheses

The main purpose of this study is to demonstrate that BBAs are less susceptible to deliberate distortion than self-reported questionnaires. It is expected that participants' ability to manipulate their assessment behaviours well-enough to impact their trait scores as they intend will vary between the two methods, yielding larger

effect sizes in the self-reported measure.

In this study, the concept of *susceptibility to distortion* is operationalised as the ability of participants to generate the scores they intend to generate through an assessment. More specifically, the participants of this study are divided between a control group and an experimental group with five conditions each representing the job application simulation for a different job, and these five jobs require completely different sets of personal qualities from one another. This means that if the participants can modify their assessment behaviour to produce the right scores for their job condition, they will receive a higher "fit" score for that job than for the other four, whereas if they are not able to do that, they won't necessarily receive a better or a worse fit score for any of the five roles in particular.

Considering the cash prize incentive, participants can be expected to try their best to modify their scores according to the job description, and their ability to do so will depend on the assessment's susceptibility to distortion.

To provide evidence that BBAs are less susceptible to distortion than self-reported questionnaires, the current study is testing two hypotheses:

- **Hypothesis 1: The effect of manipulation will be visible at the trait level but this will not be consistent across assessment formats.**
  - *Hypothesis 1a*: There will be significant differences in self reported scores between the control and the experimental groups
  - Hypothesis 1b: There will be no significant differences in BBA scores between the control and experimental groups
- **Hypothesis 2: The effect of manipulation will be visible at the fit level but this will not be consistent across assessment formats.**
  - Hypothesis 2a: Participants will be able to improve their scores to match the profile in their experimental condition with the self-reported measure
  - Hypothesis 2b: Participants will not be able to improve their scores to match the profile in their experimental condition with the BBA

5.3.2 Method

A sample of participants was invited to complete a BBA and a self-reported questionnaire in a job application simulation experiment. They were divided in two groups: a control group, which was instructed to do their best to provide "good data" and was told that the mock employers were interested in honest candidates, and the experimental group, divided into five conditions each given a different job description and told to maximise their chances to get the job it described. All the groups, including the control group were incentivised to follow the instructions by a cash prize.

5.3.2.1 Sample

This study used a paid participant sample from Prolific Academic which consisted of two groups, the control group and the experimental group, which were comparable in terms of most of their demographic distributions, which are reported in Table 5.3.

Table 5.3: Demographic distribution of control group (left) and experimental group (right)

	Control Group				**Experimental Group**			
Age	*Mean*	*SD*	*Range*		*Mean*	*SD*	*Range*	
	21	1.72	18-24		28.63	9.27	16-65	
Gender	*Male*	*Female*	*Other*		*Male*	*Female*	*Other*	
	125	149	0		197	149	4	
Ethnicity	*White*	*Black*	*Asian*	*Other*	*White*	*Black*	*Asian*	*Other*
	178	15	41	40	205	23	78	44

5.3.2.2 Procedure

Both samples were recruited via Prolific Academic, yet through two separate studies advertised as *Job Application 1* and *Job Application 2* respectively. *Job Application 1* recruited the control group, and was launched first; once the first part of the study was completed, participants were blacklisted from taking part in *Job Application 2* and the second part of the study was launched to recruit the experimental sample, which was randomly assigned to one of five job conditions.

For both samples, the study took about 40 minutes to complete, and all the participants were paid £5 for their time.

Control Sample

The control sample was instructed to complete two assessments, a BBA and a self reported questionnaire. The study was described as a job application simulation, and the participants were encouraged to complete both assessments with as much effort as possible and by responding as honestly as possible. To increase the chances of honest responding, the participants were told that the study was testing a ”honesty algorithm” for use in candidate selection, and that the researchers were looking for honest candidates who provided ”good data”. To ensure as much as possible that the participants approached the assessments honestly to provide a valid control group, they were not given a job description and they were told that the person with the best data, i.e.: achieving the highest ”honesty score”, would receive an extra cash prize of £50 in addition to their payment.

Experimental Sample

The experimental sample was also instructed to complete the same two assessments, and their study was also described as a job application simulation. Differently from the control group, the participants were required to carefully read a job description before completing the assessments which listed the key responsibilities and personal qualities of the ideal candidate for the role to which they were applying. They were instructed to complete the assessments with as much effort as possible, and to try to maximise their chances of being hired by the mock employers. To further motivate the participants to distort their responses, a £50 cash prize in addition to their payment was offered if they were selected for the job in the simulation.

Participants in the experimental sample were randomly assigned to one of five role conditions via Qualtrics survey logic: School Librarian (N=80), Creative Director (N=75), Primary Teacher (N=64), Executive Salesperson (N=65), and Software Engineer (N=65).

For each of the five conditions, a winner was selected based on a pre-established ”job profile” which was defined on the basis of the content of the personal qualities section of the job adverts. See *Job Profiles* in Section 5.3.2.3 for

more detail.

5.3.2.3 Materials

This study used two assessments (a BBA and a self-reported measure), five job descriptions, and five job profiles.

BBA - the Behaviour Based Assessment used in this study was Skyrise City (CITE), which has been described in previous studies in this thesis and it's also described in full in the Appendix.

Whilst Skyrise City features 35 scales in total, to simplify the comparison between the two measures, this study only features the traits overlapping with the self-reported measure and used in the fit profiles. The traits are listed in Table 5.4 along with their reliability coefficients, the corresponding traits in the self-reported measure, and the correlation coefficients between the two measures for each trait. Due to the experimental manipulation, the data reported in Table 5.4 refers to the control group.

Table 5.4: Self-Report and BBA traits - corresponding side by side - with their own reliability coefficients and correlation coefficients all calculated in the control group (Correlations' significance levels: $p<.05$*, $p<.01$**, and $p<.001$*** - N=271.)

	Reliability		**Reliability**	**Correlation**
Self-Report	*Cronbach's* α	***BBA***	*Stratified* α	*r, sig*
Power	0.84	Social Dominance	0.66	.289***
Warmth	0.86	Sociability	0.72	.182**
Altruism	0.85	Altruism	0.75	.300***
Drive	0.82	Determination	0.87	.179**
Order	0.77	Self-Discipline	0.74	.242***
Vulnerability	0.81	Self-Belief	0.62	.274***
Emotionality	0.84	Stability	0.74	.223***

Self-Reported Measure - the self-reported measure used in the current study is Clear Perspectives (CITE), a 90-item measure that generates 10 trait scores mapped on the five factor model (CITE).

As mentioned in Study 3, Clear perspective was chosen because it was constructed with a basis on neuropsychology, making it closer to BBA than other self-reported assessments. It used a simple and concise language and was constructed

with modern analytics such as item response theory, making it more precise. Furthermore, it was validated on job performance, making it more relevant to use in workplace settings compared to other measures validated with other self-reported assessments.

The measure is administered as 7-point Likert scale with a mix of positive- and reverse-keyed items. Below is an example of some items in the assessment, the full measure can be found in the Appendix:

"I get easily distracted"

"I respect tradition"

"I need positive feedback"

"I speak up regardless of whom I am around"

Measures Equivalence - The two measures used in the study have a good overlap of traits, and seven of them were featured in the five fit profiles and were therefore included in the current study. Table 5.4 reports the trait-level correlation coefficients between the two measure. The observed correlations between the two measures were all significant and small, which is within the expected expected range of coefficient size between self-reported and task-based assessment formats (Toplak, West, and Stanovich, 2013). Nevertheless, the study is not aiming at establishing a strong equivalence between the two measure or to collate the scores generated by them to make generic claims, the degree of shared variance between the two measures in this study is enough to justify using them in parallel to generate comparable trait and fit scores across samples yet reporting and discussing them separately, and, most importantly, comparing them for their degree of susceptibility to distortion.

Job Descriptions Five job roles were selected to represent a good mix of ideal individual differences featured in the personal qualities sections, those were School Librarian, Creative Manager, Primary Teacher, Executive Salesperson, and Software Engineer. This was done to observe people ability to impersonate specific yet different profiles so to prevent limiting the results to one or few roles and to justify generalisation.

The job descriptions were developed to provide three pieces of information to participants:

- *Role Specifications* - illustrating the key responsibilities and activities of the role to provide a credible background.

- *Person Specifications* - listing the desired experience, and the personal qualities necessary to succeed in the role. The latter is where trait-relevant information was embedded to guide response distortion. Each job description contained three bullet points, each describing the desirable expression of one of the traits measured by the two assessments.

- *Success Criteria* - listing objective markers of expected performance to further corroborate role and person specifications.

Job Profiles The job profiles for the five roles were each developed by:

- Selecting the three traits most relevant to the role

- Defining the desirable range for each trait between low (stens 1-3) and high (stens 8-10). The Medium ranges (stens 4-7) were deliberately never chosen to prevent ambiguity

- Translating those into three work-related bullet points describing the personal qualities in the Person Specification portion of the job description

To ensure that participants were encouraged to respond differently across the five conditions, the five roles profiles were designed so that no two role profiles shared more than one bullet point with each other.

An example of a job description/profile is depicted in Figure **??**, and Table 5.5 reports the three desirable traits with their expected ranges for each of the five experimental conditions.

Table 5.5: Traits' desirable ranges across the five job descriptions and fit profiles.

	Librarian	**Creative**	**Teacher**	**Salesperson**	**Engineer**
Power/Social Dom	Low	High	-	High	-
Warmth/Sociability	Low	High	-	-	Low
Altruism	-	-	High	Low	-
Drive/Determination	-	-	High	-	High
Oder/Discipline	High	Low	-	-	High
Vulnerability(R)/Self-Belief	-	-	-	High	-
Emotionality(R)/Stability	-	-	High	-	-

5.3.3 Results

A series of analyses were performed on the data to establish whether the participants were able to impersonate the intended psychological profiles as instructed through the job advertisement, and how this ability was expressed across the two test formats.

5.3.3.1 Hypothesis 1: The effect of manipulation will be visible at the trait level but this will not be consistent across assessment formats.

The first hypothesis focuses on trait-level differences between the control group and the experimental groups, investigating whether the participants are able to skew their trait scores towards the intended range of the scales.

Seven one-way ANOVAS for each assessment were used to compare the mean score of each trait between the control group and each of the experimental conditions.

Hypothesis 1a: Significant differences in self reported scores between the control and the experimental groups will emerge in line with the job descriptions
The first set of ANOVAS revealed significant main effects of group on all the self-reported trait scores - Altruism [F(5, 618)=7.031, p¡.001], Order [F(5, 618)=13.330, p¡.001], Drive [F(5, 618)=19.213, p¡.001], Power [F(5, 618)=25.322, p¡.001], Sociability [F(5, 618)=13.895, p¡.001], Withdrawal [F(5, 618)=3.052, p=.010], and Volatility [F(5, 618)=11.978, p¡.001]. However, the seven ANOVAS all revealed the same pattern upon post-hoc comparisons whereby, for each trait, the only sig-

nificant group differences were found between the control group and each of the experimental groups, but never between the experimental groups. The only exception to this was Vulnerability where the only significant difference was detected between the control group and the Creative Director condition but not between the control group and the other experimental conditions, or between any of the experimental conditions - creative director included.

The data shows that, while the experimental sample participants consistently scored differently in every traits compared to the control group, no matter the role condition, they all tried to obtain higher scores on all the traits measured by the assessment, even when this was not necessary or actually the opposite of what the job description indicated.

So, while some of the observed differences between the mean trait scores of the control group and the experimental groups emerged in line with the job descriptions, the backdrop against which they emerged suggests that these significant effects should not considered the byproduct of a deliberate manipulation of assessment behaviour aimed at them specifically but just part of a generic shift in scores towards the stereotypical desirable range.

For this reason, these significant results, despite being in line with expectations, will not be used to support the hypothesis.

Hypothesis 1b: There will be no significant differences in BBA scores between the control and experimental groups in line with the job descriptions The second set of ANOVAS revealed four significant main effects of group on BBA scores of Sociability [F(5, 618)=2.629, p=.023], Altruism [F(5, 618)=2.411, p=.035], Determination [F(5, 618)=6.340, p<.001], and Stability [F(5, 618)=24.726, p=<.001], whereas no significant group effects were detected for BBA scores of Social Dominance [F(5, 618)=1.012, p=.410], Self-Discipline [F(5, 618)=1.416, p=.217], and Self-Belief [F(5, 618)=2.136, p=.060].

The post-hoc comparisons in the BBA traits showing significant group main effect paint an inconsistent picture which makes it hard to draw any logical conclusions from the data. While a main effect of group was identified for Sociability,

none of the group comparisons - even with the control group - revealed significant differences, and the same was true for Altruism. For Determination, the control group differed significantly from all experimental group with the exception of the Teacher, while the experimental groups did not differ from each other. Finally, mean scores of Stability showed a significant difference between the control group and each of the experimental groups, but the experimental groups did not have significantly different Stability scores from one another.

All the significant differences identified in the post-hoc comparisons show that scores in the experimental groups were lower than the scores in the control group, whereas they were all expected to be higher, suggesting that none of the observed differences was in line with the job description.

The hypothesis that participants will not be able to modify their BBA scores to fit the description of a job advertisement is therefore fully supported.

To further explore the data, a series of trait-level t-tests were used to compare mean trait scores between the control group and and the whole experimental sample.

The results showed that scores of all the traits in the self reported measure were significantly different between the two groups whilst this was only partially observed in the BBA measure. In Table 5.6, significant differences of medium and large effect sizes are marked in bold.

The self-reported data showed that all the trait scores shifted toward the same direction, with scores in the experimental conditions being generally more desirable than the scores in the control condition, suggesting that participants opted for a blanket approach of inflating all scores.

This was not true in the BBA data where the trait scores that significantly differed between conditions showed the opposite pattern. Trait scores were somehow *less* desirable in the experimental group than they were in the control group.

Hypothesis 1 - summary of results: Taken together the results paint an interesting picture whereby the participants were able to significantly improve their self-report trait scores yet not in line with the job descriptions. Instead of modifying their assessment behaviour to match the description in the job advert, they opted for a

	Mean (SD)		*Mean Comparison*	*Effect Size*	
	Control	Experimental	t (df)	Cohen's d	size
Self-Report					
Power	36.99 (8.48)	45.14 (8.69)	**-11.177 (622)*** **	-0.902	*L*
Warmth	35.96 (9.40)	42.30 (9.70)	**-8.223 (622)*** **	-0.663	*M*
Altruism	44.27 (8.75)	48.42 (8.84)	-5.837 (622)***	-0.471	*S*
*Drive^*	43.43 (8.48)	50.27 (9.27)	**-9.756 (607.084)*** **	-0.778	*M*
*Order^*	36.04 (9.06)	42.94 (11.77)	**-8.273 (621.844)*** **	-0.647	*M*
Vulnerability (R)	29.95 (8.54)	32.71 (8.54)	-3.704 (622)***	-0.299	*S*
Emotionality (R)	41.21 (9.03)	46.77 (9.03)	**-7.633 (622)*** **	-0.616	*M*

	Mean (SD)		*Mean Comparison*	*Effect Size*	
	Control	Experimental	t (df)	Cohen's d	size
BBA					
Social Dominance	5.45 (1.98)	5.14 (2.28)	1.796 (620)*	0.145	-
*Sociability^*	5.53 (2.01)	5.08 (2.89)	2.596 (614.487)*	0.206	*S*
Altruism	5.49 (1.96)	5.01 (1.83)	3.179 (622)***	0.256	*S*
Determination	5.51 (1.94)	4.76 (1.91)	4.884 (622)***	0.394	*S*
*Self-Discipline^*	5.45 (1.91)	5.30 (2.16)	0.960 (612.837) ns	*n.a.*	*n.a.*
Self-Belief	5.48 (1.95)	4.99 (2.22)	2.911 (622)**	0.235	*S*
Stability	5.56 (1.93)	3.77 (2.03)	10.992 (622)***	0.887	*L*

Table 5.6: Trait-level t-tests between control and whole experimental group across the two measures. Significance levels: $p<.05$*, $p<.01$**, and $p<.001$*** - (TraitName)^ denotes Equal Variance not Assumed for that trait.

blanket approach whereby they just pretended to be overall better, as demonstrated through a comparison with the norm group.

On the other hand, as expected, participants were not able to modify their BBA scores in the same way. Assuming that they intended to modify their BBA scores with the same approach they used for the self-reported measure, that is, just for the better, they were not able to achieve the same results. Even if dropping this assumption, they were still not able to manipulate their results in a way that suggested a deliberate - and successful - attempt to mimic the qualities listed in the job avert. When the experimental data was compared in aggregate to the control group, this revealed that the participants' (assumed) attempt to modify their behaviour resulted in a deterioration of their scores, rather than in an improvement.

5.3.3.2 Hypothesis 2: The effect of assessment manipulation will be visible at the fit level but this will not be consistent across assessment formats.

Further analysis in the experimental group were performed to investigate whether participants obtained better fit-profile scores in the role to which they were assigned compared to the roles to which they were not assigned.

For each participants a fit score was calculated for each of the five roles by applying a scoring equation on the data whereby for the traits expected to be high their score would get multiplied by zero for stens 1-3, by one for stens 4-7, and by two for stens 8-10, whereas for the traits expected to be low their score would be multiplied by two for stens 1-3, by one for stens 4-7, and by zero for stens 8-10. Then the sum of the scores for each of the three traits in a fit profile were added to generate the fit score for that role.

In order to do this, the scores of the self-reported assessment were transformed into stens, whereas the BBA scores did not need transformation because they are automatically generated as sten scores as default.

The fit scores for all the roles were calculated for each participants, and a variable representing the average fit score each participant obtained on roles *other* than that of the condition to which they were assigned was computed. So, each participants was associated with four scores: the self-reported and BBA fit score of the job for which they applied, and the self-reported and BBA score representing their average fit profile for all the other jobs combined.

Hypothesis 2a: Participants will be able to improve their scores to match the profile in their experimental condition with the self-reported measure To establish weather participants were able to improve their chances of being selected for the job by purposefully modifying their performance on the self-reported assessment to match the intended profile, a t-test compared the experimental sample's fit scores for the condition to which they were assigned (M=6.34, SD= 3.34) to the average of the fit scores of the conditions to which they were not assigned (M=6.29, SD=2.42). The analysis showed that there was no significant difference between

the two scores [t(349)=.321, p=.374], suggesting that participants were not able to modify their scores well enough for this to result in a higher fit with their targeted profile.

When the same analysis was performed separately for each of the conditions, one of the t-test showed a significant difference between own (M=7.69, SD=4.18) and others' (M=6.34, SD=1.49) fit scores, t(63)=-1.925, $p<.001$ - however, this was the teacher condition, which was the only fit score in which the desirable range for all the traits was the highest.

The results of this analysis are not corroborated by the trait-level differences necessary to attribute this effect to a successful deliberate attempt of the participants in the teacher condition to modify their behaviour according to the job description, as they equally increased the scores of all the other traits in the assessment, and it was just by chance that this resulted in a positive outcome for their condition.

Therefore the hypothesis is fully rejected.

Hypothesis 2b: Participants will not be able to improve their scores to match the profile in their experimental condition with the BBA The same analysis was used on BBA data for the same reason, and the t-test revealed the same results. The BBA fit scores participants obtained for their assigned condition (M=5.01, SD= 4.13) did not significantly differ from the average BBA fit scores they obtained from the other conditions (M=4.98, SD=2.01), suggesting that they were not able to mimic the personal qualities listed in the job descriptions through modified BBA behaviour [t(349)=.142, p=.444].

Similarly, when the same analysis was performed separately for each of the conditions, one of the t-test showed a significant difference between own (M=4.57, SD=4.26) and others' (M=5.41, SD=2.07) fit scores, t(63)=3.135, p=.029 - again, this was the teacher condition, which was the only fit score in which the desirable range for all the traits was the highest. In this case, the significant difference was in the opposite direction from what was expected, so it was even more obvious for the BBA than for the self-report that this results were the byproduct of the trait score ranges assigned to the teacher condition and not the results of a targeted behavioural

manipulation in one of the groups.

Therefore the hypothesis is fully supported.

Hypothesis 2 - summary of results: overall, the results of hypothesis 2 are not surprising, considering the results on hypothesis 1. This further reiterates that the participants were not able to *consistently* modify their assessment behaviour well enough to increase their chances of being hired in neither assessment formats, and the single significant result observed in this hypothesis is therefore disregarded as a fluke.

So, hypothesis 2 is partially supported and partially rejected - supported in that it is true that it is not possible to manipulate BBA scores to increase chances of being hired, and rejected in that it seems that this is also impossible to do with a self-reported assessment.

5.3.4 Discussion

This study was designed to manipulate both intentions and ability to fake to drive the participants towards a successful faking behaviour. The rationale for this was that if intentions are manipulated by means of a cash reward, and perceived ability is also manipulated by providing very clear information of what is measured and what is desirable, then the observed differences in the degree of successful faking behaviour can safely be attributed to the opportunities to fake that the assessment offers.

The core aim of the study was to demonstrate that people are not able to successfully modify their assessment behaviour to increase their chances of being hired when assessment is done by means of a BBA, even when they are motivated to do so and when they have insight on what is expected of them. This was fully confirmed, so the study was successful in providing evidence for the main research question underlying this thesis.

However, the results of the study paint an interesting picture with plenty of room for interpretation, which warrants a more in-depth commentary encompassing a broader perspective on what the results represent beyond the core aim of the study.

The first part of the study focused on trait-level behaviour. The self-reported

and BBA trait scores of the five experimental groups were compared to each other and with the control group both individually and in aggregate. This revealed some consistencies between the two assessment formats but mostly some crucial discrepancies.

Consistently across the two formats, participants in the experimental sample were able to modify their assessment scores, and, also consistently in both cases, the scores were all modified towards the same direction regardless the instructions. So in both assessments, the participants adopted a blanket approach with all the constructs, whether that was desirable, irrelevant, or undesirable for their experimental condition.

The main discrepancy between the two groups was the direction towards which the scores were modified. Whilst the default direction for the self-report was to improve the scores, the opposite was true for the BBA. This was revealed, albeit with some mild discrepancies, from the trait-level analyses, but was made very clear with the fit score level analysis.

Hypothesis 2 showed that the only case in which a significant difference between the participants' own fit score and the average fit scores of the other groups was observed, this was in the only experimental group in which all the desirable desirable ranges (i.e.: high or low) of the traits listed in the job description were in the same direction. For the teacher group, high scores of Altruism, Drive/Determination, and Emotionality(R)/Stability were listed as the desirable personal qualities of the ideal candidate, whilst the traits in all the other job descriptions had a mix of high and low desirable ranges, which, considering the participants' blanket approach, were bound to generate a different fit score compared to the teacher group.

What is interesting is that, while the significant difference indicated a successful faking behaviour in the self-report, it suggested the opposite for the BBA. And, although both significant results are disregarded because they are clearly the byproduct of chance and not of deliberate behaviour, this does introduce an important layer of differentiation to the interpretation of the results of the two assessment formats. That is, in both assessment formats participants were not able to produce a

consistently successful, and deliberately accurate faking behaviour, but their inability must be interpreted as a marker of two completely different sets of underlying causes.

When looking at the self-reported data, it is clear that participants are able to consistently improve their scores. Each of the trait scores is better in the experimental sample than it is in the control group, and this is true of all the condition, but the trait scores are consistent across conditions. This means that it can be implied that there was an attempt to fake, just not a great one.

People are definitely able to be honest in a self-report, and they are definitely able to fake good - as demonstrated in this study and many more before it. It is legitimate to assume that they can also fake bad when needed (Viswesvaran amd Ones, 1999; Boss et al, 2015), so the fact that this was not observed should not be interpreted as a lack of ability.

The same *item transparency* that enabled the participants to fake for the better should have enabled them to fake for the worse, and if this was not enough, the way in which the personal qualities were described in the job descriptions had the same degree of overlap with the self-reported scale items whether the desirable range was high or low. Nevertheless, the participants were able to fake good in self-reported questions that were not mentioned in the job description, and they were able to even go against what was described in the ads, so it can be assumed that the scale's item transparency and the opportunity to fake that the format offered were enough for them to work out what the best response was and produce it without needing a detailed job description, and this should have be the equally enough for them to work out how to respond in the opposite direction. They just didn't.

The most likely reason for this effect is that while the reward on offer for matching the condition's profile was incentivising enough to motivate *generic* faking behaviour, it might not have been motivating enough for the participants to go the extra mile and fine tune their faking behaviour to generate the trait scores described in their condition. Or perhaps, the reward was not enough for them to even read the job description for enough time and with enough attention to understand

what they had to do beyond increasing their chances of being hired.

Whilst the results do reject the hypothesis that people should be able to impersonate a specific profile with a self-reported measure, it is safe to attribute this outcome to a flaw in the study whereby the incentive offered for achieving this was not attractive enough to manipulate the perceived valence of doing well in the assessment and through that increase intentions to fake, and to not attribute this outcome to the inability to *potentially* do so.

When looking at the BBA data, the outlook is different, and, whilst it can be assumed that the same motivational mechanisms were at play - after all, with regards to format, this was a within-participant study - there is more than just motivation to the participants' inability to impersonate the psychological profile of the ideal candidate in their condition.

While the self-reported data show a rather consistent skew towards the desirable score range, the opposite is true in the BBA. Participants (almost) consistently scored worse in the experimental conditions compared to the norm group, raising a crucial question on whether they were actually even trying to do that in the first place.

As per the assumption about motivation being consistent across the two formats, it can also be assumed that the intended outcome would be consistent. Yet the results were exactly the opposite. Since it is highly unlikely that in both cases participants intended to appear worse that they were and they failed at this in the self-reported measure, it is safely implied that the failed faking attempt was observed in the BBA scores. This means that the participants were not able to produce the same results in the BBA, that is, they were *not able to improve their scores* in the BBA, and their attempt to do so actually backfired.

The participants' hiring potential for the teacher's role was a welcome fluke in the data which showed that where people might be successful at increasing their hiring potential with self-reported measures, attempting the same with a BBA not only doesn't provide an advantage, but results in outcomes decreasing the participants' hiring potential.

This means that not only the hypothesis that people cannot fake BBAs is supported, the results show that faking BBAs might actually be penalising, which echoes the same conclusions made in previous studies.

As Ellingson and Mc Farland (2011) point out, objective ability is hampered by factors beyond awareness, and this is likely what happened in this case. High instrumentality and expectancy perception might have increase motivation, triggering the faking attempt. However, proximal situational factors of ability, such as the limits imposed by the response options and the way in which models of personality are scores, while unable to prevent faking, were able to prevented the faking attempt from being successful.

5.3.4.1 Implications

There were three key elements in this experimental study: the first was to use a cash incentive to increase valence to boost intentions to fake, and, as explained, this partially failed; the second was to increase faking ability by describing the content of what was measured and what scores were desirable, and this most likely failed completely. Whilst the first two elements were expected to be consistent across formats, the last element, which was the core of the study, was the opportunities to fake each format offers and this was expected to *differentiate* the formats from one another.

What emerged from the study was that with enough motivation to trigger intentions to fake to the degree in which participants are able to exert enough effort to significantly improve their trait scores but not enough for them to go the extra mile and fine tune them to the intended profile, the two assessment formats showed widely different degrees of opportunities to fake so remarkably in contrast with each other that one of them enabled successful faking across the whole assessment, whereas the other made the participants' scores even less desirable than the control group's.

These results imply that, whilst self-reported measures are relatively easy to fake and depending on the right motivation and access to a job description people should be likely to increase their hiring potential, BBAs do not afford the same

degree of manipulation. In fact, attempts at faking BBAs result in a consistent deterioration of scores, meaning that not only those who try to improve their scores are not able to do so, but that doing so might actually decrease their hiring potential as the desirable range of trait scores is most likely to be high rather than low.

This has important implications for employers and their choices of assessment when a reliable ranking if candidates is important to them. By opting for a BBA, not only there is no risk of cheaters artificially outranking other candidates, cheating is most likely to lower the rank of cheating applicant, thereby reducing the chances of them being hired.

These results also confirm the suspicion that the candidates in study 4 might have actively decided to not engage in faking because not knowing how to do it would have made it too risky in high stakes. This shows a peculiar interaction suggesting that the relationship between PBC, job desirability, and intentions might not be exactly linear. Whilst valence is consistently depicted as a positive antecedent of intentions, it is likely that in the case of BBAs really high job desirability might become a hindering factor for intention if PBC is low, whereas if it's relatively high but not the highest (as in study 5) it can override low PBC and still trigger intentions.

The results of this study also provide support for study 2, confirming that participants did not have any viable strategy to fake. Whether they did have a strategy or not is unclear as this was not testes, but the outcome of their faking attempt unambiguously demonstrate that they did not have a *viable* strategy in mind. In addition to this, the results of study 5 could be used as evidence for the speculations made in study 3 whereby attempting to fake would most likely result in deteriorating the data. This study shows very clearly that this is what happen when participants try to tamper with the assessment, and it suggests that if it's true that candidates in study 4 did not attempt faking because they were aware of not being able to, then their perception was very well founded.

5.3.4.2 Limitations

This study had a series of limitations. Similarly to study 4, this was an in-between participant study, so claims should be limited to this constraint. That said, the data is

rather clearly showing a consistent effect and also corroborates and expands many of the results in previous studies, so this might not be a huge limitation after all, whilst the risk of poor test re-test reliability remains a concern.

The results suggest that motivation to fake, despite the cash incentive, was not enough to completely induce faking as participants did not engage in purposeful faking aligned to the job description but only in a rather mild and generic faking good behaviour. It is possible that they did not feel that the cash incentive warranted the effort required to fake a particular profile, but they felt that it was enough to fake at least a bit and make an attempt at enhancing their chances to be hired. The job descriptions very clearly defined the profiles of the ideal candidate for each of the roles, so it is highly unlikely that lack of clarity was the underlying cause of this outcome. However, it is very plausible that the participants were not motivated enough to read the instructions with full attention, and the reason why they opted for a generic fake good approach could have been that they did not know what was expected of them other than that they had to pretend that the assessment was part of a job application. Either way, low motivation seems to be the root cause of the observed results.

Also, similarly to the existing literature, the simulated job application yielded a larger effect compared to similar field studies. In this case, the artificiality and inconsequential nature of the study setting generated an effect that did not occur at all in field setting, and this makes the ecological validity of this study somewhat dubious.

5.3.4.3 Conclusion

This study was the last piece of the puzzle needed to assess BBAs susceptibility to faking in accordance with existing models of faking. Participants were instructed to approach a self-reported assessment and a BBA as if they were applying for a specific job which was described to them with in depth information about the psychological profile of the ideal candidate. The results showed that while participants were generally able to modify their scores on both assessments, they were only successful in doing so with the self-reported measure as they BBA scores actually

deteriorated as a result of the attempted faking.

The results of this study have some important implications, especially when considered in the context of previous studies. One one hand, the results showed that attempting to fake BBA is not just not beneficial but is actually detrimental to a person's psychometric scores, and, on the other hand, they show that the reason why real candidates in a previous study did not show signs of faking was most likely due to low expectancy, and not to lack of objective ability to fake.

Both pieces of evidence are remarkable in demonstrating that BBAs cannot be successfully faked but also that candidates are fully aware of this, meaning that the are unlikely to even try.

5.4 Behaviour General Discussion

This chapter focused on behaviour with two studies observing (potentially) faking behaviour in action, but in reality it did so through the lenses of intentions and ability.

Most of the faking literature consists of studies comparing candidates and employees, or simulated candidates and control groups to understand how well people can fake assessment. This chapter includes evidence of both types of studies, exploring differences in performance at the variable and score level, and also differences across assessment types.

Two studies, one field and one lab-based, compared BBA completions in high and low stake. In both studies, stake-dependent differences were recorded which collectively provide striking evidence to suggest that BBAs cannot be faked.

More specifically, the first study was designed to provide evidence that, unlike other types of assessment in which low stakes completions are believed to represent more of a person true score compared to high stakes completion, in BBAs high stake completions represent the true score much more accurately than low stake completions.

The study compared the score and variable-level data of candidates and employees, finding that scores did not generally differ significantly across samples but

variables did. At the variable level, most of the expected differences in performance were observed. Those were not differences showcasing faking attempts in the candidate sample, on the contrary, they were differences marking clear patterns of low motivation in the employee sample. The exploration of the data showed that candidates, who had high motivation, performed on the task optimally, making fewer errors and maintaining the same level of performance for longer. They also did not engage in rapid careless responding and did not show markers of distraction. One fascinating difference between the two samples was that candidates did not spend any more time than employees reading the instructions but they were significantly more likely to repeat them. The reason why candidate data is not believed to show markers of faking is that the data from all variables was normally distributed and the task main effect was successfully replicated. On the other hand, employees data showed the tell signs of distraction and low motivation, with many more incidences of random tapping, missed trials, and overall more mistakes. The task main effect was also replicated in this sample, showing that they did engage with the task enough to produce viable data, however, their variables were not as normally distributed.

Another reason why the results of this study support the idea that high stakes data is a much more realistic representation of the individual than low stake data is that low stakes data is significantly easier to identify with predictive modelling. This is because high stake data is a faithful representation of the vast individual differences between people and because of this no one assessment session is like the other at the variable level. With low stakes, the behavioural patterns in which people engage due to low motivation, care, and attention make the session data a lot more similar between individuals, making it easier to detect.

This is a paradoxical effect because previous evidence show exactly the opposite pattern, whereby high stakes data is much more distinctive than low stakes data, as high stake data show markers of inflation such as skew and kurtosis and for that has less variance.

It is however not clear from this study *why* the candidate data showed no signs

of tampering. Whilst the candidates were most certainly motivated to get their jobs, it cannot be assumed that they were motivated to fake, and it is possible that the lack of faking markers in the data is not due to a lack of ability but to a lack of intentions.

The second study settled this doubt. A group of research panel participants were instructed to approach a BBA and a self-reported measure in order to maximise their chances to be ”hired”, and win a cash prize, in a job application simulation. The results of the study showed that, whilst the participants were able to change their scores in both types of assessment, the score shifts went in opposite directions. The participants were able to shift their scores towards the desirable range compared to a control group, albeit not following the instructions well enough to suggest they did so with enough purpose to mimic the ideal candidate profile described in the job description. Unlike that, the scores in the BBA were consistently deteriorated by the faking attempts, and they all show small effect sizes differences compared to the control group in the opposite direction compared to the self-reported data.

Whilst it can be assumed that the participants had the same faking intentions across assessment formats, it could be argued that it is not possible to tell which of the two directions they intended to fake toward and which one was the failed attempt. However, it is extremely unlikely that they meant to fake bad in the self-reported measure and failed, and it is equally as extremely likely that the did not know how to fake the BBA so they just attempted what they could and upset the data. So the results of this study are interpreted with this assumption.

The two studies provide good evidence independently from one another, but they provide great evidence combined.

The difference between real and simulated candidates is that valence and instrumentality (see Ellingson & McFarland, 2011; in section 2.1). Valence and instrumentality work alongside expectancy to fuel motivation to fake. Valence refers to the perceived importance of the outcome, whereas instrumentality refer to the degree of causal relatedness that the behaviour has with the outcome, and expectancy refers to the individual’s perceived ability to perform the behaviour in the way in which is associated with the intended outcome.

Although real candidates *could* be expected to invest more resources to inform their expectancy levels compared to simulated candidates, there are no reasons to expect that expectancy would be terribly different between real and simulated candidates. Instrumentality can be considered a (semi-)fixed property of the assessment, however, the bar of what constitute an acceptable level of instrumentality might change on the function of what is at stake. That is, an average level of instrumentality might be sufficient if there isn't much at stake, like in the case of simulated candidates, however, the bar might raise substantially if there is a lot at stake and individual need to make a much safer bet on their behaviour. Valence is the key factor here, and this is what differed the most between the two studies. For real candidate, there was a job at stake, for the simulated ones there was a £50 cash price at stake. This means that the instrumentality bar must have been significantly higher for the real candidates, and this in turn meant that low expectancy had a much stronger effect for real candidates than it had for simulated ones.

In other words, expectancy and valence seemed to have interacted as a moderator for the threshold of instrumentality, whereby low expectancy and high valence placed the bar very high, whereas low expectancy and medium-low valence placed it at a lower level. At some point between the two level at which the bar of instrumentality was place, there must be a level at which a go/no-go pivot occurs whereby people either do or do not attempt faking.

So, taken together, the studies in this final chapter provide not only evidence that BBAs are a good solution against assessment faking in candidate selection, but also provide an explanation of why this might be the case.

Faking in BBAs does not occur in real candidates because they don;t trust their ability to do so successfully.

Chapter 6

General Discussion

Psychometric assessments are widely used in candidate selection, with approximately 70% of companies globally employing them. However, assessment faking is highly prevalent and can lead to biased and inaccurate scores, resulting in the hiring of less suitable candidates and an increased likelihood of inviting deviant behaviour into the workforce. Self-reported personality measures are particularly vulnerable to faking, but they are still popular despite their limitations. Developing high-stakes assessments that are resistant to faking is crucial, and researchers must focus on creating methods that can reliably prevent or detect faking.

This thesis was designed to test whether a new type of assessment format might be a solution to this problem. Behaviour-Based Assessment, also known as Game-Based Assessment, is a relatively recent psychometric approach which leverages experimental paradigms from cognitive science to generate more objective and unbiased scores of individual differences.

6.1 Scientific Rationale

This thesis touch on two key strands of scientific literature: Assessment Gamification and Faking.

Game- and Behaviour-Based Assessment belong of course to the assessment literature, but also overlap with the field of Gamification, which is defined most commonly as the use of game mechanics and game elements in context other than gaming. One of the most consistently cited benefits of this new type of assessment

measures is their low susceptibility to faking.

6.1.1 Gamification

Gamification in candidate selection involves adding game elements such as badges, points, and storylines to the testing process to increase its attractiveness, ease of use, engagement, and motivation. Gamification can be versatile and can range from enhancing the appeal of recruitment campaigns to making onboarding activities more engaging. The most common game elements examined in the gamification literature are points, badges, and leaderboards.

There are three theoretical frameworks from psychology that are relevant to gamification: Flow theory, Need satisfaction theories, and Goal-setting theory. Flow theory suggests that an individual enters a condition called 'flow' when they work on a task that has a level of complexity and difficulty that is appropriate for their skillset. When this flow state is achieved, the task elicits a state of interest, concentration, and enjoyment for the individual. Need satisfaction theories propose that individuals have three basic psychological needs: autonomy, competence, and relatedness. Goal-setting theory suggests that individuals perform better when they have specific and challenging goals.

The use of gamification in selection and assessment processes can be a promising way to attract and retain high-quality talent, but psychometric assessment need to be carefully designed and implemented to also their validity.

6.1.1.1 Psychometrics and Gamification

Gamification can be used in assessment and selection to develop or redesign assessment methods that evaluate candidates using game design elements. Landers and Sanchez (2022) define three key concepts in gamification: gameful design, assessment gamification, and game-based assessment (GBA). Gameful design is a strategy that combines psychometrics and game mechanics to develop new assessments, while assessment gamification involves adding game elements to existing assessments without changing their psychometric properties. This leads to Gamified assessments, which refer to traditional measures that have been altered to look

like a game while maintaining their psychometric properties intact. Studies have shown that gamifying assessments can improve candidate performance and satisfaction, particularly when game mechanics are implemented around assessments. GBA, on the other hand, requires candidates to play a game, with their psychometric profile computed from data generated through the core gameplay loop.

GBAs use gamification as the core of their assessment model, which not only improves applicant reactions but also capitalises on the psychometric properties of games. GBAs measure individual differences through the adaptation of experimental tasks measuring cognitive abilities, general intelligence, sustained attention, memory, and problem-solving skills which are not only useful constructs in their own right, but they have also been demonstrated to underlie individual differences at the personality level. In GBAs, measurement occurs by presenting players with a stream of choices during gameplay, and most of the response options imply elements of ability and maximal performance. Recording both player choices and the game's paradata allows GBAs to analyse information concerning users' decision-making processes and gain a good overview of how their brain respond to carefully designed stimuli, which is difficult, if not impossible, to measure through traditional psychological assessments.

The way in which personality is derived from GBAs data provides, in theory, a valuable tool for measuring individual differences more safely from a faking perspective, especially for traits that are the most prone to deliberate distortion. For example, it is possible to infer levels of neuroticism by leveraging some differences in visual attention, and also to use accuracy in visual search as a marker for conscientiousness. GBAs have also been used to examine personality and decision-making in safety-critical environments, such as air traffic control, yielding encouraging results.

While GBAs seem to offer several advantages over traditional assessment methods, careful evidence-based design and thorough validations are still necessary, and the lack of reliability information available on commercial GBAs raises possible concerns with regards to their stability over time and repeated administra-

tions.

6.1.2 Theoretical Models of Faking

In terms of GBA's resistance to faking, there is very little published evidence available to date, but what has been found so far shows great promises, unanimously suggesting that this type of measures do not lend themselves to faking.

In addition to this, Landers and Sanchez (2022) propose a theoretical framework for how GBAs can reduce faking based on established models of faking, whereby they expect different features and properties of this type of assessment to interact with the factors known to be responsible for faking.

Theoretical models of faking broadly suggest that intentions/motivation, ability, and behaviour play crucial roles in faking. Two recent models, the Integrated Model of Faking and a later iteration drawing from Vroom's Expectancy Theory of Motivation, provide a framework for this thesis. The Integrated Model of Faking incorporates the Theory of Planned Behavior and argues that intentions and ability interact to trigger successful faking behaviour. The later model substitutes the theory of planned behaviour with Vroom's Expectancy Theory and suggests that motivation, rather than intentions, drives faking behaviour.

Vroom's theory explains why people behave in certain ways to achieve their goals and how their expectations about the outcomes of their actions influence their motivation. It is based on three concepts: Expectancy, which refers to the belief that effort will lead to desired performance, instrumentality, which refers to the belief that desired performance leads to rewards, and valence, which refers to the value placed on rewards associated with achieving desired performance. Motivation is based on individuals' expectations of the outcomes of their actions, and they are more likely to engage in faking if they believe it will result in a higher score or more favourable outcome. However, Vroom's theory also suggests that individuals will weigh the costs and benefits of faking before engaging in it, and this aspect of the theory is crucial to understanding the motivation behind assessment faking.

Landers and Sanchez (2002) identify the three main antecedents to faking behaviour as motivation, ability, and opportunity, and suggest that gameful design and

gamification can target these to inhibit faking. To address motivation, they propose using game mechanics that trigger a flow state and increase immersion, shifting motivation away from faking and towards engaging with the game. For ability, they suggest leveraging cognitive load and low transparency design. Regarding opportunity, they note that GBAs' highly digitalized nature and complexity of scoring models make them difficult to fake, as the tasks used are measures of maximal performance that don't lend themselves to deliberate improvement. Simons et al. (2023) also found evidence supporting this idea.

Experimental and theoretical evidence presented above suggests that game-based assessments (GBAs) may be less susceptible to faking than other methods. However, it is necessary to first establish whether this is the case by taking a holistic experimental approach to investigation providing insight from different angles, and, most importantly, it is necessary to understand *how* GBAs prevent faking in order to safely assume that this effect would replicate in high stakes real life settings, and also to add to existing body of evidence.

6.1.3 This Thesis

While Landers and Sanchez's (2002) framework emphasises game elements as key moderators of motivation, this thesis is much more focused on how and why the psychometric properties and response options of Behaviour-Based Assessments are relevant to faking. On one hand, this thesis is rooted in the idea that this type of assessment target faking though ability and opportunity to fake much more substantially than through motivation. On the other hand, this focuses entirely on the core psychometric approach underlying the behavioural aspects of the assessment and not on game elements and mechanics *per se*.

This thesis was designed to test the hypothesis that workplace assessments using experimental laboratory tasks to generate psychometric scores are not susceptible to faking, and to this purpose, gamification is a lot less relevant.

The term Behaviour-Based Assessment (BBA) was used throughout this thesis to describe assessments, whether gamified or not, which use data collected by replicating experimental laboratory tasks in digital form to build machine learning

models of individual differences.

The underlying rationale of this thesis was that due to the way in which psychometric data is collected and used in BBAs, this type of assessment does not enable faking. The core idea is that in BBAs, the objective ability to fake is too heavily impacted by both the response options and the scoring models.

This thesis' ultimate goal was to demonstrate that BBAs are a solution against response distortion in workplace assessment, and it was successful in doing so.

6.2 Discussion

The thesis consisted of five studies loosely mapped on the Integrated Model of Faking (McFarland & Ryan, 2006) and the subsequent model of faking featuring the VIE factors of Expectancy Theory (Ellingson & McFarland, 2011), both of which are broadly divided into three main blocks: intentions/motivation to fake, ability to fake, and faking behaviour.

6.2.1 Summary of Results

The first study (Study 1) explored the individual-level factors affecting intentions to fake, and investigated whether the same factors correlated with intentions to fake traditional assessment formats also predicted intentions to fake BBAs. In particular, this first study investigated whether the perceived behavioural control people have (or think they have) on a specific assessment format impacts their intentions to fake it, and whether the presence of a selection of faking warnings affects that perception and with it the intentions to fake.

The aim of this study was to demonstrate that people have less strong intentions to fake BBAs compared to other types of assessment because they perceive less behavioural control over them, and this is true even when faking deterrents are factored in. This study supported this hypothesis, showing that people have less intentions to fake BBAs because they perceive less control over them, and that the same mechanisms that underlie intentions also apply to BBAs.

Two studies were designed address two different aspects of ability, untangling the dissociation between subjective and objective ability to fake proposed in the

VIE model of faking.

The first of these two studies exploring ability (Study 2) investigated the subjective perception of ability to fake, by asking participants one single question about four different assessment formats (two of which were BBAs): ”how would you cheat on this assessment?” The faking strategies they described across formats were categorised to understand whether people were able to suggest viable cheating strategies, and whether they adapted their strategies to the assessment format. The aim of this study was to demonstrate that people don't know how to fake BBAs, meaning that they lack the subjective ability to fake BBAs. This was also supported and the results of this qualitative study showed that while people knew exactly how to fake other types of assessments, when it came to BBAs, they did not have a strategy in mind.

In Study 3, the focus was placed on the assessment itself and the type of response options and data used in a BBA model of personality was examined. For this study, a BBA model of Extraversion was unpacked, and its variables divided between measures of maximal performance and measure of typical performance. The rationale behind this study was that, because measures of maximal performance by definition don't lend themselves to faking, their presence in a model of personality would substantially reduce the objective ability to fake. The aim of this study is to demonstrate that there is a lot of variance explained by maximal performance in BBAs, and for this they don't offer people the objective ability to be faked. The investigation confirmed that most of the variance explained by the model came from maximal performance, hence supporting the hypothesis that BBAs don't offer scope for objective ability to fake.

The final chapter included two studies, both focusing on the emergence faking behaviour in situation where high intentions/motivation to fake should be expected to be high.

The first study (Study 4) compared candidates and employees to investigate two factors. First, whether the trait scores of the two groups differed in the same way in which they would fro other assessment types, and second, whether those

score-level differences were caused by variable-level differences that suggested faking or if an alternative explanation could be more feasible. The aim of the study was to demonstrate that while there might be differences in scores between candidates and employees, this is not due to faking but rather to the effect of low motivation (in employees) on cognitive tasks which equates to lower scores. This is the opposite to what the existing literature has shown to be the case for traditional assessment, whereby candidates' scores are expected to include at least some element of inflation whereas employees' scores are considered much more representative of their true qualities. This was supported and the data demonstrated that first there were very few rather trivial differences in assessment scores, but, most importantly, that the variable-level differences responsible for those effects aligned perfectly with the hypothesis in BBAs high motivation equates to optimal performance, and not inflated scores, whereas low(er) motivation equates to relatively less optimal performance, and not true scores.

The last study (Study 5) was framed as a simulated job application in which people had to pretend to apply to one of five different jobs for which they have been told the selection criteria and complete a self-reported assessment and a BBA. The aim of this study was to generate enough intentions and motivation to fake through a cash prize to trigger faking behaviour, and to increase subjective ability to fake by providing a very detailed description of the ideal candidate profile. The aim of this study was to demonstrate that people are not able to systematically manipulate their BBA scores well enough to generate an intended profile. The results showed that faking behaviour occurred on both types of assessment, however, whist the participants were able to fake their scores for the best through the self-reported assessment, their BBA scores were negatively affected by the faking attempt, showing that they were not able to modify their assessment behaviour in a way that equated with the scores that they intended to achieve.

Taken collectively, the results of the study unanimously supported the main hypothesis of this thesis, providing experimental evidence demonstrating that BBAs are a solution against response distortion in workplace assessment.

This thesis was designed to provide experimental evidence to test this assumption on the basis of the existing theoretical models of faking, which are broadly composed of three key factors, and the results summarised above provide good evidence for each of them.

6.2.2 Intentions

This study focused on the interface between the Theory and Planned Behaviour and Perceived behavioural Control that underlies the Integrated Model of faking, with the aim of establishing whether BBAs might be less of a target of intentions to fake. More specifically, the study investigated how the TPB and PBC replicated across different assessment formats, and whether any observed differences in the relationship between the factors included in the models or the relative differences of the factors' strengths/levels across formats resulted in lower levels of intentions to fake BBAs compared to other assessment formats.

Overall, the results of the study mostly support this hypothesis, showing that participants have generally lower intentions to fake BBAs compared to non-BBA formats, albeit with some ambiguities with intelligence tests. Interestingly, BBAs are linked to lower levels of intentions to fake compared to self-reported assessment when faking deterrents are taken into consideration, meaning that participants have stronger intentions to fake self-reported assessments when a faking detection warning is present compared to any BBA with no faking detection warning.

Not all hypotheses in the study were entirely supported. Despite the overall support of the core research question, several unexpected results emerged. The hypothesis was that there would be no significant relationship between individual factors and intentions to fake in BBAs but only in non-BBA formats due to very lower levels of intentions to fake expected for BBAs. However, the data showed that individual-level factors in TPB and two HEXACO factors (Honesty-Humility and Conscientiousness) were consistently correlated with intentions to fake across all formats, including BBAs, suggesting that there is nothing special in BBAs that could challenge the relationship between intentions to fake and individual-level antecedents.

The study tested a second hypothesis that the perceived behavioural control (PBC) afforded by an assessment format would vary, and subsequently the intentions to fake that format would also vary. The hypothesis was partially supported, as a consistent relationship between PBC and intentions to fake was observed across all formats. BBAs were found to afford less PBC than non-BBA formats in general, but the differences in intentions to fake between BBAs and intelligence tests were inconsistent and small, with only one of the two BBAs significantly differing from intelligence tests in each study phase. Therefore, the second hypothesis can only be supported for one of the non-BBA formats but not for the other.

The study investigated the effect of faking deterrents on perceived behavioural control and intentions to fake across four different assessment formats: self-reported measures, puzzles, intelligence tests, and business behavioural assessments (BBAs). The results showed that all the faking deterrents successfully deterred intentions to fake, with detection algorithms and proctored testing having the greatest impact on perceived behavioural control. BBAs consistently had the lowest intentions to fake even without any deterrents, and none of the BBAs-deterrent combinations were featured in the lowest quartile of faking intentions. The study suggests that faking warnings and deterrents are most effective when systematically aligned with the format of the assessment for which they attempt to deter faking.

Interestingly, warning against the use of online guides had no effect on reducing intentions to fake, and using online guides to modify behaviour in BBAs is more likely to cause undesirable effects than to improve scores. Scoring models of personality are complex, and task performance may be desirable for one trait but undesirable for another. Reliance on guides can slow down response speed and increase variance in reaction times. The fear surrounding assessment faking is that people may secure jobs they are not suitable for, but the lack of effectiveness of warning against faking on BBAs may not be a bad thing if it results in a deterioration of scores and fewer opportunities for unqualified candidates.

Taken together, this study provides evidence to suggest that when faced with BBAs, people will be less inclined to fake compared to when they encounter any

other assessment format, which is true even in the presence of faking warnings and deterrents. This evidence can be used to infer that people will be less likely to fake on assessment if BBAs are used. The study also shows that reducing PBC can be an effective way to reduce faking intentions, and warnings claiming to use a faking detection algorithm can help reduce intentions to fake. However, warnings against online guides are not effective in reducing intentions to fake BBAs, possibly because candidates may not believe the warning or may try to mimic the behaviors described in the online guide, which can have a negative impact on their scores.

6.2.3 Ability

This chapter focused on two distinct yet related aspects of ability to fake: subjective and objective. The first study used a qualitative approach to investigate the faking strategies that people might use when encountering different types of assessment formats, in order to validate the hypothesis that people do not know how to fake BBAs, that is, they lack the subjective ability to fake. The study found several categories of cheating strategies that emerged across different formats, including deliberate deception, finding assessment answers, and the use of technology. The distribution of these categories differed across the assessment formats and showed a good degree of alignment with the cheating opportunities each format affords. While the study's hypotheses only included three categories of cheating strategies expected a priori, the other categories that emerged from the qualitative data analysis were included for exploratory purposes and did not affect the results as none of these categories were the most frequent for any of the formats. They still provided interesting insights on what people think when trying to devise a cheating strategy. The study found that the most common response to the question "how would you cheat this assessment?" was "I don't know" for BBAs, indicating that people generally don't know how to cheat this type of assessment. As expected, the most common cheating strategy for self-reported assessments was deception, and finding answers was the most popular strategy for faking intelligence tests. The results showed a great degree of difference between the formats, supporting the assumption that participants can fine-tune their strategies and adapt to the vulnerabilities of each

format. When comparing viable and non-viable cheating strategies across formats, more viable strategies than non-viable strategies were suggested for all formats, but the proportion of viable strategies was relatively much lower for BBAs. The study also found that virtually all strategies people suggested to fake BBAs would not result in successful faking, suggesting that people are highly unlikely to modify their scores in BBAs if they try. This suggests that BBAs can be used safely in high stakes settings. The results of Study 2 further corroborated these findings. Participants were less able to articulate a cheating strategy for BBAs compared to self-reported and ability assessments, consistent with the lower levels of PBC observed for BBAs in Study 1. The format for which participants were most likely to have a cheating strategy was self-reported assessment, also consistent with the higher levels of PBC observed for this format in Study 1. These results suggest that the lower levels of PBC observed for BBAs may be due to participants' difficulty in envisioning how to cheat them, implying that BBAs impair people's ability to fake. The goal of the second study was to determine if BBAs provided enough opportunity for candidates to fake their performance. The rationale was to establish whether the type of response options and data used in BBAs would lend themselves to faking; specifically, the assumption was that measures of maximal performance should substantially limit faking emergence, so if personality models are largely influenced by ability measures, they would be harder to fake. The study used Extraversion as a case study because it contained data from a wide range of tasks. Models of individual differences in BBAs are developed by combining several experimental tasks adapted from the scientific literature to function as assessments. Most hypotheses testing the established relationships between experimental task performance and Extraversion were not supported. However, all the combined main effect and trace data from multiple tasks allowed for accurate prediction of Extraversion scores, suggesting that BBAs' computation of individual differences scores might not be linked to simple task performance. The fact that the results partially contradicted existing evidence regarding the neuro-cognitive basis of Extraversion did not affect the study's scope, as the main objective was to determine if BBAs would be too difficult

to fake due to the proportion of maximal over typical performance measures. The study used all variable-level data currently featured in the BBA Extraversion model of Skyrise City and developed three predictive models. The results showed that a model entirely made of maximal performance variables outperformed both a model containing only typical performance variables and a mixed model. Furthermore, including typical performance variables in the mixed model lowered predictive validity compared to the maximal performance model and affected the distribution of simulated Extraversion scores. These results fully support the idea that BBAs are objectively hard to fake due to the nature of their tasks. While this study doesn't directly investigate faking in action, its focus on response options, data types, and especially maximal performance measures allows safe conclusions from a format perspective. Additional aspects of objective ability come into play when attempting to fake BBAs. One is understanding the relationship between task behaviour and psychometric scores. Since Skyrise City's data convergence with Extraversion isn't perfectly aligned with experimental evidence, successful faking would require access to proprietary personality models, which are likely highly complex behavioral models encompassing data, paradata, and trace data in ways not easily translated into actionable patterns. Another consideration is the ability to produce required behaviours. While some Skyrise City tasks could be faked with well-designed aids, others don't offer this option. The role of aids in cognitive tasks would be to maximize task performance, which doesn't equate to desirable trait scores. This suggests that trait models are too complex to be deliberately reproduced, and manipulating task behaviour may be detrimental, as paradata and metadata have more influence than main effect data and are likely negatively affected by artificial behaviours. Finally, BBAs' scoring means that the same behavioral pattern might simultaneously be positive and negative, depending on the precise model being scored. This Schrödingeresque paradox makes successful faking impossible across the board. One notable finding is that the lack of correlation replication between task main effects and Extraversion scores suggests that neuroscience research focusing solely on task main effects may be too simplistic. Experimental studies should consider

more nuanced performance measures to capture the complexity of neurobiological structures in personality differences. This study provides sufficient evidence that BBAs are a viable alternative to self-reported personality assessment when faking is a concern. The results demonstrate minimal faking scope, explain why this is the case, and show high convergent validity with existing personality measures. The chapter's unique combination of qualitative and quantitative data, exploring both subjective and objective ability to fake, makes a strong case that candidates would not be able to fake BBAs, considering both their inability to devise viable faking strategies and the ability-based nature of BBA personality models.

6.2.4 Behaviour

The last chapter of this thesis focused on behaviour with two studies observing (potentially) faking behaviour in action, but in reality it did so through the lenses of intentions and ability, ultimately testing all the evidence accumulated throughout the previous studies. Most of the faking literature consists of studies comparing candidates and employees, or simulated candidates and control groups to understand how well people can fake assessment. This chapter includes evidence of both types of studies, exploring differences in performance at the variable and score level, and also differences across assessment types. This thesis included one field and one lab-based study, and compared BBA completions in high and low stake. In both studies, stake-dependent differences were recorded which collectively provide striking evidence to suggest that BBAs cannot be faked. The first study (Study 4) was designed to provide evidence that, unlike other types of assessment in which low stakes completions are believed to represent more of a person's true score compared to high stakes completion, in BBAs high stake completions represent the true score much more accurately than low stake completions. Study 4 compared the score and variable-level data of candidates and employees for one of the Skyrise City modules (Navigation), which is an adaptation of the Flanker Task, finding that scores did not generally differ significantly across samples but variables did. At the variable level, there were several expected differences between the two groups, most of which were observed. The data showed that can-

didates performed optimally and did not engage in rapid careless responding or show markers of distraction, while employees had more incidences of distraction and low motivation, resulting in more mistakes. Speed and accuracy data comparison showed both groups approached the task at the same speed, yet their accuracy levels differed significantly. This means that while both candidates and employees approached the task with similar energy, their attention levels differed and employees made significantly more mistakes. Being accuracy a measure of maximal performance, this difference can be attributed to low attention in employees rather than score improvement attempts in candidates. The main task effect analyses showed that the cost of interference was significantly larger for employees than candidates in both speed and accuracy, meaning candidates engaged more inhibitory cognitive control than employees. This difference clearly stems from higher effort, as inhibitory control cannot exceed individual capacity. Interestingly, comparing congruent and incongruent trials revealed candidates were faster in incongruent trials, whereas employees were faster in congruent trials. This shows different approaches: candidates used a more deliberate approach in congruent trials, while employees' performance was heavily impacted by interference in incongruent trials. Data beyond the main task effect revealed that employees were significantly more likely to engage in rapid response, showing carelessness and low interest in assessment outcomes, and missed more trials, demonstrating less commitment. The fatigue timeline analysis showed similar patterns of performance decline, but with some notable differences. While total errors and incongruent trial errors increased similarly for both groups, congruent trial errors showed earlier and steeper increases for employees, suggesting earlier preventable mistakes. A clear valence marker was the effect of mistakes on subsequent performance. Candidates took longer to respond after mistakes, suggesting increased caution, whereas employees showed no such pattern. Interestingly, candidates didn't spend more time reading instructions than employees but were significantly more likely to repeat them. This might be due to instruction display format, as Navigation's tutorial includes minimal text. All observed differences indicated low motivation in the employee sample rather than fak-

ing attempts by candidates. Candidates performed optimally with high motivation, while employees exerted less effort. Candidate data showed normal distribution and successful task main effect replication, whereas employee data showed signs of distraction and low motivation, with more random tapping, missed trials, and mistakes. Though employees engaged enough to produce viable data, their main effect was more pronounced, suggesting lower executive functioning engagement. This made low stakes data significantly easier to identify through predictive modeling than high stakes data, as high stakes data better represented individual differences, making each session unique at the variable level. Low stakes sessions showed similar behavioral patterns due to reduced motivation and attention. This presents a paradoxical effect compared to previous literature, where high stakes data typically shows more distinctiveness through inflation markers like skew and kurtosis. While Study 4 showed no faking in candidates, it wasn't clear whether this resulted from lack of ability or intentions. Study 5 provided insight into this question. Research panel participants were instructed to approach a BBA and a self-reported measure to maximize their chances of winning a cash prize in a job application simulation. The study investigated whether their "match score" for their assigned job condition was significantly higher than their average scores for other conditions. Role profiles were designed to support this analysis, with substantially different ideal candidate profiles. The results showed that experimental group participants obtained different scores in both assessment types compared to control groups, but in opposite directions. While participants could shift their self-reported scores toward the desirable range (though not precisely matching the ideal profile), BBA scores consistently deteriorated with faking attempts, showing small effect size differences in the opposite direction from self-reported data. Though participants presumably had similar faking intentions across formats, the results suggest they failed at faking BBAs rather than intentionally performing poorly. The study also showed that despite motivation to attempt faking, participants didn't invest enough resources to successfully modify their scores to match the ideal candidate description, with no evidence of strategic score manipulation between experimental groups. The two studies provide

compelling evidence both independently and combined. The key motivational difference between real and simulated candidates lies in valence (outcome importance) and instrumentality (behavior-outcome causality). While real candidates might invest more in understanding their ability to fake (expectancy), there's no reason to assume vast expectancy differences between real and simulated candidates. Instrumentality can be considered a semi-fixed assessment property, though acceptability thresholds may vary with stakes. The chapter showed expectancy and valence interacting to moderate instrumentality thresholds. In the field study, low expectancy and high valence set a high bar, deterring faking attempts. In the experimental study, low expectancy with moderate valence set a lower bar. This suggests a threshold point where faking becomes too risky to attempt. Together, these studies demonstrate that BBAs effectively resist assessment faking in candidate selection, explaining why this occurs. The fact that faking can occur unsuccessfully but doesn't emerge in high-stakes settings suggests BBAs signal low expectancy enough to offset valence's motivational influence. This corroborates earlier findings about lower faking intentions due to reduced PBC (Study 1), lack of faking knowledge (Study 2), and faking attempts' counterproductive effects (Study 3), completing the thesis's mapping of BBAs' faking resistance. Real candidates don't fake BBAs because they doubt their ability to succeed, and those who try fail.

6.3 Contributions

This thesis provides a comprehensive experimental validation of Behavioral-Based Assessments (BBAs) as an effective solution to response distortion in workplace assessment. Through five rigorous studies, the research has demonstrated that BBAs successfully prevent faking through multiple complementary and interconnected mechanisms.

The format itself reduces faking attempts by creating uncertainty about effective strategies, while the perceived risks outweigh potential benefits in high-stakes settings. The reliance on maximal performance measures inherently limits faking opportunities, and the dissociation between task performance and score calculation

invalidates traditional faking strategies. Furthermore, the complexity of the models exceeds human information processing capacity, leading faking attempts to consistently result in performance deterioration. The simultaneous use of multiple scoring models means that attempts to optimize one score inevitably compromise others.

These findings lead to two crucial conclusions: First, candidates are unlikely to attempt faking in high-stakes situations due to the perceived risks, resulting in more accurate assessment data. Second, when candidates do attempt to fake, they typically perform worse than honest respondents, minimizing the risk of poor hiring decisions.

6.3.1 Theoretical

From a theoretical perspective, this thesis support the main tenets of the key theoretical models of assessment faking. It provides evidence that ability to fake is the most important factor in preventing faking and this works from two different perspectives. The first through perceptions, affecting expectancy, perceived behavioural control and subjective ability. Assessments that don't show vulnerabilities to faking do not allow candidates to envision faking strategies, and this lowers their PBC in turn leading to lower level of intentions. Expectancy and subjective ability to fake are also affected by the same factors and this reduces motivations to fake in a similar fashion.

Secondly, this thesis also provides evidence for different ways in which objective ability may be reduced by an assessment format. Response options requiring maximal performance or behavioural patterns that are too complex to enact proved useful to prevent *objective* ability to fake, whilst of course sending signals of low vulnerability that inform *subjective* ability. This thesis also suggests that objective ability may be hampered by the faking attempt itself whereby faking attempts assuming the wrong strategy may further reduce a candidate's ability to fake (successfully).

The most remarkable set of results was the comparison of Study 4 and study 5, showing that the interaction between valence, instrumentality and expectancy is highly intricate. Elligson & Mcfarland (2011) argue that the three proximal factors

in motivation should be conceptualised as having a multiplicative relationship so that if one of them is too close to zero, then motivation to fake does not emerge. The last chapter of this thesis challenges this concept, as it proves to be too simplistic to explain the results of the two studies combined. The key difference between real and simulated candidates is that valence and instrumentality interact in a different way. Valence and instrumentality work alongside expectancy to fuel motivation to fake. Valence refers to the perceived importance of the outcome, whereas instrumentality refer to the degree of causal relatedness that the behaviour has with the outcome, and expectancy refers to the individual's perceived ability to perform the behaviour in the way in which is associated with the intended outcome. Although real candidates *could* be expected to invest more resources to inform their expectancy levels compared to simulated candidates, there are no reasons to assume that expectancy would be markedly different between real and simulated candidates. Instrumentality can be considered a (semi-)fixed property of the assessment, however, the bar of what constitute an acceptable level of instrumentality might change on the function of what is at stake. That is, an average level of instrumentality might be sufficient if there isn't much at stake, like in the case of simulated candidates, however, the bar might raise substantially if there is a lot at stake and individual need to make a much safer bet on their behaviour. Valence is the key factor here, and this is what differed the most between the two studies. For real candidate, there was a job at stake, for the simulated ones there was a £50 cash price at stake. This means that the instrumentality bar must have been significantly higher for the real candidates, and this in turn meant that low expectancy had a much stronger effect for real candidates than it had for simulated ones. In other words, expectancy and valence seemed to have interacted in tandem to set the threshold of instrumentality, whereby low expectancy and high valence placed the bar very high but low expectancy and medium-low valence placed it at a lower level.

Somewhere between the two levels at which the bar of instrumentality was placed in the two studies, there must be a point at which intentions to fake are switched off.

Finally, this thesis support the theoretical framework described by Landers and Sanchez (2022) whereby BBAs would reduce ability through items opacity and reduced opportunities to fake. As Sanchez and Langer (2020) pointed out, BBAs might be too novel and complex for candidates to have enough knowledge of what faking behaviour would be appropriate to use, and they explain that the format, presentation, and response modes in the assessment can be purposely designed avoiding clear desirable response patterns. This thesis is the first to test those theoretical assumptions experimentally, and it provides strong evidence to support them.

6.3.2 Methodological

From a methodological perspective this thesis is adding to the existing assessment literature by providing evidence for the use of concepts and evidence in personality neuroscience for psychometric.

This thesis explored how models of personality in BBAs are built adapting experimental tasks and evidence, and how this not only provide valid measures of personality but also more reliable ones.

Personality neuroscience can provide extensive experimental evidence to support the development of a wide range of BBAs, and the results are very promising. By harnessing the existing body of literature, it would be possible to leverage behavioural tasks able to generate psychometric scores in a much more objective, pleasant, and valid way.

Furthermore, this thesis makes a methodological contribution in that it provides an explanation of the reasons why BBAs are resistant to faking and this should be an useful evidence for the development of future assessments. The features that make BBAs resistant to faking might not be entirely exclusive to BBAs or GBAs, and can be leveraged to reduce susceptibility to faking in other assessment types.

Finally, this thesis suggests that the methods used in personality neuroscience might be excluding some important variance from the tasks they use. The models of Extraversion used in this study shows that there is more to individual differences in personality than just task main effects. The lack of replication of the correlations between most tasks main effects and scores of Extraversion may add to our

ability to navigate the ambiguity in the experimental evidence discussed in the literature review. It is plausible that simply focusing on task main effect may be too reductionist and more data should be considered in experimental studies that allows for more subtle nuances in performance to be captured. Considering the incredible complexity of the neurobiological structures and processes involved in individual differences in personality, experimental tasks constrained by highly arbitrary cut offs and over-relying on summary data (i.e.: total accuracy, or accuracy difference between congruent and incongruent tasks, etc) may conceal entire layers of trait-relevant variance that were previously overlooked and therefore not understood.

6.3.3 Practical

There are several practical implications. First, this thesis demonstrates that with regards to faking, BBAs are safe to use in high stake settings, and this is the first real claim of its kind. The depth, breadth and the outcomes of experimental testing that this assumption has been subjected to in order to substantiate this claim warrants its justification.

Second, the results of Study 4 suggests that BBAs are equally as valid in low stake settings, and it is likely that the same norms can be used across sample. Furthermore, low stake data can be easily identified and this may enable candidates to be invited to repeat the assessment should they have not be able to concentrate enough.

Third, due to their nature BBAs rely very little on language and prior knowledge which makes them excellent candidates to further reduce adverse impact in assessment. This, combined to fake resistance, makes their level of fairness unmatched.

However, despite being demonstrably resistant to faking, this theses highlighted some potential shortcomings for the use of BBA.

First, whilst the reliance on measures of maximal performance is great to prevent faking, this casts a serious doubt to the overall validity of the measures across the spectrum of ability. Whilst BBAs are more reliant on intra individual than inter individual differences, there might be a risk that fairly high and low levels of

cognitive ability may reduce the validity of these measures.

Second, for the same reason, it is possible that these measure might be carry more adverse impact than they claim as the large percentage of variance explained by ability most certainly links with some known demographic effects

Third, the fact that tasks main effects did not correlated with trait scores as expected from the existing evidence may actually signal poor construct validity. Whilst this does not seem to be the case from the data used in this study, and the existing literature is not consistent enough to be blindly relied upon, it is rather unexpected that most of the task main effect did not show correlations with trait scores.

Fourth, the results of this thesis unveiled a potential shortcoming in the way in which some trait models are developed. Determination was the only significant difference between candidates and employees in the field study, and this raised a flag. The *situational* level of determination between the two groups differed quite clearly due to the context in which they completed the assessment, but this is not expected to be *dispositional* and consistent difference between the two groups. It is likely that the data used to compute the score of determination relied on data patterns that were intuitively and not empirically linked to individual differences in determination, and this means that the measure might be too sensible to determination as a state in addition to measuring determination as a trait.

Finally, it must be noted that, due to the above point and more, BBAs may not be a viable assessment option for every trait. First, the field of personality neuroscience is dominated by the big five, and this mean that experimental evidence on other constructs might be limited if not entirely absent. Second, BBAs use individual differences at the cognitive level to infer individual differences at the personality level, and this means that there is no scope for incorporate learned capabilities or acquired knowledge in this type of assessment.

6.4 Limitations and Future Research

Whilst successful at demonstrating the intended outcomes, this thesis was not free of limitations.

First, access to data for the most of this thesis was severely limited by what was offered by the test publisher, meaning that most BBA data was only available at the score level, and the little available variable level data was mostly obfuscated and severely limited in breadth. Future research should have access to full data sets both at the variable and score level to enable a thorough and unbiased investigation of the assessment.

Second, the data was collected at different times and with several different samples, which made it slightly disjointed, and possibly introduced some confounding effects. Future research should attempt to replicate the same experimental protocol with as few different samples as possible, leveraging testing opportunities more deeply and broadly.

Third, most of the experiments in this thesis were too heavily focused on the emergence of the desired effects and missed many opportunities to investigate the underlying reasons for those effects. Further research should incorporate more emphasis on *why*, in addition to *what*, effects occur.

6.5 Conclusions

This thesis set up to take BBAs on a journey through the different factors surrounding assessment faking, and it unfolded over five different experiments encompassing field and lab studies, quantitative and qualitative designs, different types of psychometric assessment, variable and score-level data, and a range of statistical and machine learning technique.

The main purpose of this thesis was to demonstrate that a different type of assessment collecting and generating psychometric data in a completely different way could overcome the threat of response distortion by limiting people's objective ability to fake and by signalling features reducing subjective ability to fake, and with that intentions and motivation to do so.

This results in two key implications: the first is that candidates are much less likely if not unlikely to attempt faking as they don't trust their ability to do so successfully, and this means that BBAs data provides an unprecedented level of accuracy compared to other high stakes assessment. Second, in the unlikely scenario in which candidates may decide to attempt faking, evidence shows that this would not give them an advantage over non-faking candidates and it is more likely that they will be excluded from the candidate pool than that they will be at the top of the ranking, and this means that faking should not be a worry for candidate selection when BBAs are used.

With this in mind, this thesis can conclude stating that BBAs have been thoroughly investigated in their ability to tackle candidate faking and the results demonstrate that they are a very good solution against response distortion in workplace assessment.

Appendix A

Links to Everything

All the hyperlinks in this documents will take you to the relevant folders and files in my Google Drive. Access is currently restricted to a very few people, but if you want access, or need files in an alternative format, if the links are broken, if you want to chat about my thesis, but also if you need me to discourage you from applying for a PhD, please email me HERE.

A.1 Scales

Click HERE to access the scales used in Study 1.

A.2 Videos

Click HERE to access the videos mentioned in Study 1 and 2.

A.3 Technical Manuals

Click HERE to access Skyrise City and Clear Perspectives Technical Manuals.

A.4 Job Descriptions

Click HERE to access the Job Descriptions used in Study 5.

A.5 Datasets and Outputs

Click HERE to access the datasets and outputs used in this thesis. They are divided by study, you need SPSS to open them.

Bibliography

Ajzen, I. (1991). The theory of planned behavior. Organizational behavior and human decision processes, 50(2), 179-211.

Ajzen, I. (2020). The theory of planned behavior: Frequently asked questions. Human Behavior and Emerging Technologies, 2(4), 314–324. https://doi.org/10.1002/hbe2.195

Ajzen, I., & Fishbein, M. (1975). A Bayesian analysis of attribution processes. Psychological bulletin, 82(2), 261.

Ajzen, I., & Madden, T. J. (1986). Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control. Journal of experimental social psychology, 22(5), 453-474.

Alport, Alan, Elizabeth A. Styles, and Shulan Hsieh. "17 Shifting intentional set: Exploring the dynamic control of tasks." (1994).

Andrade, F.R.H., Mizoguchi, R. and Isotani, S. (2016), "The Bright and Dark Sides of Gamification", in Micarelli, A., Stamper, J. and Panourgia, K. (Eds.), Intelligent Tutoring Systems: 13th International Conference, ITS 2016, Zagreb, Croatia, June 7-10, 2016. Proceedings, Springer International Publishing, Cham, pp. 176–186.

Ángeles Quiroga, M., Escorial, S., Román, F.J., Morillo, D., Jarabo, A., Privado, J., Hernández, M., et al. (2015), "Can we reliably measure the general factor of intelligence (g) through commercial video games? Yes, we can!", Intelligence, Elsevier Inc., Vol. 53, pp. 1–7.

Anglim, J., Molloy, K., Dunlop, P. D., Albrecht, S. L., Lievens, F., & Marty, A. (2022). Values assessment for personnel selection: Comparing job applicants to non-applicants. European Journal of Work and Organizational Psychology, 31(4), 524–536. https://doi.org/10.1080/1359432X.2021.2008911

Arctic Shores. (2018), Arctic Shores Technical Manual, Unpublished manuscript.

Armitage, C. J., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. British journal of social psychology, 40(4), 471-499.

Attali, Y. and Arieli-Attali, M. (2015), "Gamification in assessment: Do points affect test performance?", Computers and Education, Elsevier Ltd, Vol. 83, pp. 57–63.

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

Bäccman, C., Sjöberg, L., & Almqvist, K. (2015). A Comparison of Applicants' and Incumbents' Mean Scores on Health Constructs and Personality Constructs. A Follow-up Study of Military Recruits in a Selection Setting: Comparison of Applicants' and Incumbents' Mean Score. A Follow-Up Study. International Journal of Selection and Assessment, 23(2), 120–130. https://doi.org/10.1111/ijsa.12101

Bakan, D. (1966). The duality of human existence: An essay on psychology and religion. Rand Mcnally.

Barends, A. J., & de Vries, R. E. (2023). Construct validity of a personality assessment game in a simulated selection situation and the moderating roles of the ability to identify criteria and dispositional insight. International Journal of Selection and Assessment, 31(1), 120–134. https://doi.org/10.1111/ijsa.12404

Barends, A. J., de Vries, R. E., & van Vugt, M. (2022). Construct and Predictive Validity of an Assessment Game to Measure Honesty–Humility. Assessment, 29(4), 630–650. https://doi.org/10.1177/1073191120985612

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. Journal of applied psychology, 81(3), 261-272.

Beblo, T., Macek, C., Brinkers, I., Hartje, W., & Klaver, P. (2004). A new approach in clinical neuropsychology to the assessment of spatial working memory: the block suppression test. Journal of clinical and experimental neuropsychology, 26(1), 105-114.

Bedwell, W.L., Pavlas, D., Heyne, K., Lazzara, E.H. and Salas, E. (2012), "Toward a Taxonomy Linking Game Attributes to Learning: An Empirical Study", Simulation
Gaming, Vol. 43 No. 6, pp. 729–760.

Bernal, B., & Altman, N. (2009). Neural networks of motor and cognitive inhibition are dissociated between brain hemispheres: an fMRI study. International Journal of Neuroscience, 119(10), 1848-1880.

Berry, C. M., & Sackett, P. R. (2009). Individual differences in course choice result in underestimation of the validity of college admissions systems. Psychological science, 20(7), 822-830.

Biggs, A.T., Clark, K. and Mitroff, S.R. (2017), "Who should be searching? Differences in personality can affect visual search accuracy", Personality and Individual Differences, Elsevier Ltd, Vol. 116, pp. 353–358.

Birkeland, S. A., Manson, T. M., & Kisamore, J. L. (2006). A Meta-Analytic Investigation of Job Applicant Faking on Personality Measures. International Journal of Selection and Assessment, 14(4).

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A Meta-Analytic Investigation of Job Applicant Faking on Personality Measures. International Journal of Selection and Assessment, 14(4), 317-335.

Bledlow, R. and Frese, M. (2009), "a Situational Judgment Test of Personal Initiative and Its Relationship To Performance", Personnel Psychology, Vol. 62 No.2, pp. 229–258.

Bloemers, W., Oud, A., & Dam, K. V. (2016). Cheating on unproctored internet intelligence tests: Strategies and effects. Personnel Assessment and Decisions, 2(1), 3.

Bonnefond, A., Doignon-Camus, N., Hoeft, A., & Dufour, A. (2011). Impact of motivation on cognitive control in the context of vigilance lowering: An ERP study. Brain and cognition, 77(3), 464-471. https://doi.org/10.1016/j.bandc.2011.08.010

Bonner, S. E., Hastie, R., Sprinkle, G. B., & Young, S. M. (2000). A Review of the Effects of Financial Incentives on Performance in Laboratory Tasks: Implications for Management Accounting. Journal of management accounting research, 12(1), 19-64. https://doi.org/10.2308/jmar.2000.12.1.19

Boss, P., König, C. J., & Melchers, K. G. (2015). Faking Good and Faking Bad Among Military Conscripts. Human Performance, 28(1), 26–39. https://doi.org/10.1080/08959285.2014.974758

Bott, J. P., O'Connell, M. S., Ramakrishnan, M., & Doverspike, D. (2007). Practical Limitations in Making Decisions Regarding the Distribution of Applicant Personality Test Scores Based on Incumbent Data. Journal of Business and Psychology, 22(2), 123–134. https://doi.org/10.1007/s10869-007-9053-x

Boudon, R. (2003). Beyond Rational Choice Theory. Annual Review of Sociology, 29(1), 1–21. https://doi.org/10.1146/annurev.soc.29.010202.100213

Bourdage, J. S., Schmidt, J., Wiltshire, J., Nguyen, B., & Lee, C. (2020). Personality, interview faking, and the mediating role of attitudes, norms, and perceived behavioral control. International Journal of Selection and Assessment, 28(2), 163–175. https://doi.org/10.1111/ijsa.12278

Bowen, C., Martin, B. A., & Hunt, S. T. (2002). A COMPARISON OF IPSATIVE AND NORMATIVE APPROACHES FOR ABILITY TO CONTROL FAKING IN PERSONALITY QUESTIONNAIRES. The International Journal of Organizational Analysis, 10(3), 240–259. https://doi.org/10.1108/eb028952

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101.

Bredemeier, K., Berenbaum, H., Most, S.B. and Simons, D.J. (2011), "Links between neuroticism,emotional distress, and disengaging attention: Evidence from a single-target RSVP task",Cognition and Emotion, Vol. 25 No. 8, pp. 1510–1519.

Brosowsky, N. P., DeGutis, J., Esterman, M., Smilek, D., & Seli, P. (2020). Mind wandering, motivation, and task performance over time: Evidence that motivation insulates people from the negative effects of mind wandering. Psychology of Consciousness: Theory, Research, and Practice.

Brühlmann, F. (2016), The Effects of Framing in Gamification: A Study of Failure, Springer Nature, Wiesbaden, Germany, available at:https://doi.org/10.1007/978-3-658-16926-8.

Buford, C.C. and O'Leary, B.J. (2015), "Assessment of Fluid Intelligence Utilizing a Computer Simulated Game", International Journal of Gaming and Computer-Mediated Simulations, Vol. 7 No. 4, pp. 1–17.

Burke, K. (1969). A rhetoric of motives. Univ of California Press.

Burns, G. N., & Christiansen, N. D. (2011). Methods of Measuring Faking Behavior. Human Performance, 24(4), 358–372. https://doi.org/10.1080/08959285.2011.597473

Callegaro, M., Yang, Y., Bhola, D. S., Dillman, D. A., & Chin, T.-Y. (2009). Response Latency as an Indicator of Optimizing in Online Questionnaires. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 103(1), 5–25. https://doi.org/10.1177/075910630910300103

Campbell, A. M., Davalos, D. B., McCabe, D. P., & Troup, L. J. (2011). Executive functions and extraversion. Personality and Individual Differences, 51(6), 720-725.

Chamorro-Premuzic, T., & Furnham, A. (2010). The psychology of personnel selection. Cambridge University Press.

Chiew, K. S., & Braver, T. S. (2016). Reward Favors the Prepared: Incentive and Task-Informative Cues Interact to Enhance Attentional Control. Journal of experimental psychology. Human perception and performance, 42(1), 52-66. https://doi.org/10.1037/xhp0000129

Chou, Y. (2015), "A comprehensive list of 90+ gamification cases with ROI stats", Youkaichou.Com.

Christiansen, N. D., Wolcott-Burnam, S., Janovics, J. E., Burns, G. N., & Quirk, S. W. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. Human Performance, 18(2), 123-149.

Christy, K.R. and 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, Elsevier Ltd, Vol. 78, pp. 66–77.

Church, A. H., & Rotolo, C. T. (2013). How are top companies assessing their high-potentials and senior executives? A talent management benchmark study. Consulting Psychology Journal: Practice and Research, 65(3), 199.

Collette, F., & Van der Linden, M. (2002). Brain imaging of the central executive component of working memory. Neuroscience & Biobehavioral Reviews, 26(2), 105-125.

Collmus, A. B., & Landers, R. N. (2019). Game-Framing to Improve Applicant Perceptions of Cognitive Assessments. Journal of Personnel Psychology, 18(3), 157–162. https://doi.org/10.1027/1866-5888/a000227

Corr, P.J., DeYoung, C.G. and McNaughton, N. (2013), "Motivation and Personality: A Neuropsychological Perspective", Social and Personality Psychology Compass, Vol. 7 No. 3, pp. 158–175.

Corsi, Philip Michael. "Human memory and the medial temporal region of the brain." (1972).

Cronan, T. P., Mullins, J. K., & Douglas, D. E. (2018a). Further Understanding Factors that Explain Freshman Business Students' Academic Integrity Intention and Behavior: Plagiarism and Sharing Homework. Journal of Business Ethics, 147(1), 197–220. https://doi.org/10.1007/s10551-015-2988-3

Cronbach, L. J. (1946). Response Sets and Test Validity. Educational and Psychological Measurement, 6, 475-494.

Cronbach, L. J. (1960). Essentials of psychological testing (2nd ed.). Harper.

Csikszentmihalyi, M. (1990), Flow: The Psychology of Optimal Experience, 1st ed., Harper
Row, New York, NY.

Dambacher, M., Hübner, R., & Schlösser, J. (2011). Monetary incentives in speeded perceptual decision: Effects of penalizing errors versus slow responses. Frontiers in psychology, 2, 248.

Deci, E.L. and Ryan, R.M. (2000), "The 'What' and 'Why' of Goal Pursuits: Human Needs and the Self-Determination of Behavior", Psychological Inquiry, Vol. 11 No. 4, pp. 227–268.

Demars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. Educational Assessment, 12(1), 23-45.

Deterding, S., Dixon, D., Khaled, R. and Nacke, L. (2011), "From Game Design Elements to Gamefulness: Defining 'Gamification'", Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, ACM, New York, pp. 9–15.

Deterding, S., Khaled, R., Nacke, L. and Dixon, D. (2011), "Gamification: toward a definition", Chi 2011, pp. 12–15.

Deterding, S., Sicart, M., Nacke, L.E., O'Hara, K. and Dixon, D. (2011), "Gamification: Using game-design elements in non-gaming contexts", Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '11, p. 2425.

DeYoung, C. G. (2013). The neuromodulator of exploration: A unifying theory of the role of dopamine in personality. Frontiers in human neuroscience, 762.

DeYoung, C. G., Hirsh, J. B., Shane, M. S., Papademetris, X., Rajeevan, N., & Gray, J. R. (2010). Testing predictions from personality neuroscience: Brain structure and the big five. Psychological science, 21(6), 820-828.

DeYoung, C.G. (2013), "The neuromodulator of exploration: A unifying theory of the role of dopamine in personality", Frontiers in Human Neuroscience, Vol. 7 No. November, pp. 1– 26.

DeYoung, C.G. and Gray, J.R. (2009), "Personality neuroscience: Explaining individual differences in affect, behavior, and cognition", in Corr, P. and Matthews, G. (Eds.), Corr, PJ.; Matthews, G., Editors. The Cambridge Handbook of Personality Psychology, Cambridge University Press, New York, NY, pp. 323–346.

DeYoung, C.G., Hirsh, J.B., Shane, M.S. and Papademetris, X. (2010), "Testing Predictions From Personality Neuroscience: Brain Structure and the Big Five", Psychological Science, Vol. 21 No. 6, pp. 820–828.

Dhinakaran, J., De Vos, M., Thorne, J.D. and Kranczioch, C. (2014), "Neuroticism focuses attention Evidence from SSVEPs", Experimental Brain Research, Vol. 232 No. 6, pp. 1895– 1903.

Dicerbo, K.E. (2014), "Game-Based Assessment of Persistence", Educational Technology
Society, Vol. 17 No. 1, pp. 17–28.

Dilchert, S., & Ones, D. S. (2011). Application of Preventive Strategies. In New Perspectives on Faking in Personality Assessment. Oxford University Press.

Docebo. (2016), Elearning Market Trends and Forecast 2017-2021, Docebo.

Donovan, J. J., Dwight, S. A., & Schneider, D. (2014a). The Impact of Applicant Faking on Selection Measures, Hiring Decisions, and Employee Performance. Journal of Business and Psychology, 29(3), 479–493. https://doi.org/10.1007/s10869-013-9318-5

Doty, D.H. and Glick, W.H. (1998), "Common methods bias: Does common methods variance really bias results?", Organizational Research Methods, Vol. 1 No. 4, pp. 374–406.

Dujmović, M., & Penezić, Z. (2017). To Do or not to Do: Inhibiting Attention and Action Depending on the Level of Extraversion. Psihologijske teme, 26(1), 47-60.

Dunlop, P., Bourdage, J. S., de Vries, R. E., McNeill, I. M., Jorritsma, K., Orchard, M., ... & Choe, W. K. (2019). What does overclaiming represent? Well, it depends!. In International Society for the Study of Individual Differences.

Dürr, D., & Klehe, U.-C. (2018a). Using the Theory of Planned Behavior to Predict Faking in Selection Exercises Varying in Fidelity. Journal of Personnel Psychology, 17(3), 155–160. https://doi.org/10.1027/1866-5888/a000211

Dwight, S. A., & Donovan, J. J. (2003). Do Warnings Not to Fake Reduce Faking? Human Performance, 16(1), 1–23.

Edwards, A. L. (1957). The social desirability variable in personality assessment and research.

Ellingson, J. E., & McFarland, L. A. (2011a). Understanding Faking Behavior Through the Lens of Motivation: An Application of VIE Theory. Human Performance, 24(4), 322–337. https://doi.org/10.1080/08959285.2011.597477

Engelmann, J. B., & Pessoa, L. (2014). Motivation sharpens exogenous spatial attention.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. Models of working memory: Mechanisms of active maintenance and executive control, 4, 102-134.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. Perception & psychophysics, 16(1), 143-149.

Eysenck, H. J. (1967). Personality and extra-sensory perception. Journal of the Society for Psychical Research.

Eysenck, S. B., & Eysenck, H. J. (1970). Crime and personality: an empirical study of the three-factor theory. The British Journal of Criminology, 10(3), 225-239.

Fan, J., Gao, D., Carroll, S. A., Lopez, F. J., Tian, T. S., & Meng, H. (2012). Testing the efficacy of a new procedure for reducing faking on personality tests within selection contexts. Journal of Applied Psychology, 97(4), 866–880.

Fell, C. B., & König, C. J. (2016). Cross-cultural differences in applicant faking on personality tests: A 43-nation study. Applied Psychology, 65(4), 671-717.

Ferrell, J.Z., Carpenter, J.E., Vaughn, E.D., Dudley, N.M. and Goodman, S.A. (2016), "Gamification of Human Resource Processes", in Gangadharbatla,

H. and Davis, D. (Eds.), Emerging Research and Trends in Gamification, IGI Global, Hershey, PA, pp. 108–139.

Fetzer, M., Mcnamara, J. and Geimer, J.L. (2017), "Gamification , Serious Games and Personnel Selection", The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection and Employee Retention, John Wiley Sons, pp. 293–309.

Fleeson, W. and Jayawickreme, E. (2015), "Whole Trait Theory", Journal of Research in Personality, Elsevier Inc., Vol. 56, pp. 82–92.

Foroughi, C.K., Serraino, C., Parasuraman, R. and Boehm-Davis, D.A. (2016), "Can we create a measure of fluid intelligence using Puzzle Creator within Portal 2?", Intelligence, Elsevier Inc., Vol. 56, pp. 58–64.

Fu, J., Zapata, D. and Mavronikolas, E. (2014), Statistical Methods for Assessments in Simulations and Serious Games, ETS Research Report Series, Princeton, NJ, available at:https://doi.org/10.1002/ets2.12011.

Furnham, A. (1990a). Faking personality questionnaires: Fabricating different profiles for different purposes. Current Psychology, 9(1), 46–55. https://doi.org/10.1007/BF02686767

Furnham, A. (2016), "Eysenck at work: The application of his theories to work psychology", Personality and Individual Differences, Vol. 103, pp. 148–152.

Furnham, A. F. (1997a). Knowing and Faking One's Five-Factor Personality Score. Journal of Personality Assessment, 69(1), 229–243. https://doi.org/10.1207/s15327752jpa6901$_1$4

Gilmore, D. C., Stevens, C. K., Harrell-Cook, G., & Ferris, G. R. (1999). Impression management tactics. The employment interview handbook, 321-336.

Godin, G., & Kok, G. (1996). The theory of planned behavior: a review of its applications to health-related behaviors. American journal of health promotion, 11(2), 87-98.

Goffin, R. D., & Boyd, A. C. (2009). Faking and personality assessment in personnel selection: Advancing models of faking. Canadian Psychology/Psychologie canadienne, 50(3), 151.

Goffman, E. (1959). The presentation of self in everyday life: Selections.

Goldberg, L.R. (1990), "An Alternative 'Description of Personality': The Big-Five Factor Structure", Journal of Personality and Social Psychology, Vol. 59 No. 6, pp. 1216–1229.

Graham, E. K., & Lachman, M. E. (2014a). Personality traits, facets and cognitive performance: Age differences in their relations. Personality and Individual Differences, 59, 89–95. https://doi.org/10.1016/j.paid.2013.11.011

Gray, J. A. (1970). The psychophysiological basis of introversion-extraversion. Behaviour research and therapy, 8(3), 249-266.

Gray, J. R., & Braver, T. S. (2002). Integration of emotion and cognitive control: A neurocomputational hypothesis of dynamic goal regulation.

Gray, J. R., & Burgess, G. C. (2004). Personality differences in cognitive control? BAS, processing efficiency, and the prefrontal cortex. Journal of Research in Personality, 38(1), 35-36.

Gray, J.A. (1981), "A critique of Eyesenck's theory of personality", A Model for Personality, pp.246–276.

Griffin, B., Hesketh, B., & Grayson, D. (2004b). Applicants faking good: Evidence of item bias in the NEO PI-R. Personality and Individual Differences, 36(7), 1545–1558. https://doi.org/10.1016/j.paid.2003.06.004

Griffith, R., & Converse, P. (2012) The rules of evidence and the prevalence of applicant faking. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), New perspectives on faking in personality assessment (pp. 34–52). Oxford University Press.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007b). Do applicants fake? An examination of the frequency of applicant faking behavior. Personnel Review, 36(3), 341–355. https://doi.org/10.1108/00483480710731310

Griffith, R. L., Yoshita, Y., Peterson, M. H., & Malm, T. (2006). Addressing elusive questions: Investigating the faking-performance relationship. In Griffith, RG, & Yoshita, Y.(Chairs). Deceptively simple: Applicant faking behavior and prediction of job performance. Symposium conducted at the 21 st Annual conference for the Society for Industrial and Organizational Psychology, Dallas, TX.

Hamari, J., Koivisto, J. and Sarsa, H. (2014), "Does gamification work? - A literature review of empirical studies on gamification", Proceedings of the Annual Hawaii International Conference on System Sciences, pp. 3025–3034.

Hao, J., Smith, L., Mislevy, R., von Davier, A. and Bauer, M. (2016), "Taming Log Files From Game/Simulation-Based Assessments: Data Models and Data Analysis Tools", ETS Research Report Series, No. March, pp. 1–17.

Harman, J. L. (2022). Game-Like Personality Measures Reduce Faking and Careless Responding. [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/5n92w

Harman, J. L., & Brown, K. D. (2022). Illustrating a narrative: A test of game elements in game-like personality assessment. International Journal of Selection and Assessment, 30(1), 157–166. https://doi.org/10.1111/ijsa.12374

Harman, J. L., & Purl, J. (2022). Advances in Game-Like Personality Assessment. Trends in Psychology. https://doi.org/10.1007/s43076-022-00162-x

Harold, C. M., McFarland, L. A., & Weekley, J. A. (2006a). The Validity of Verifiable and Non-verifiable Biodata Items: An Examination Across Applicants and Incumbents: BIODATA VALIDITY. International Journal of Selection and Assessment, 14(4), 336–346. https://doi.org/10.1111/j.1468-2389.2006.00355.x

Harrison, D. A., McLaughlin, M. E., & Coalter, T. M. (1996). Context, cognition, and common method variance: Psychometric and verbal protocol evidence. Organizational

Behavior and Human Decision Processes, 68(3), 246-261.

Hawkes, B., Cek, I. and Handler, C. (2017), "The Gamification of Employee Selection Tools", Next Generation Technology-Enhanced Assessment, pp. 288–314.

Hedge, C., Powell, G. and Sumner, P. (2018), "The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences", Behavior Research Methods, Behavior Research Methods, Vol. 50 No. 3, pp. 1166–1186.

Helfinstein, S.M., Schonberg, T., Congdon, E., Karlsgodt, K.H., Mumford, J.A., Sabb, F.W., Cannon, T.D., et al. (2014), "Predicting risky choices from brain activity patterns", Proceedings of the National Academy of Sciences, Vol. 111 No. 7, pp. 2470–2475.

Herde, C. N., Lievens, F., Jackson, D. J. R., Shalfrooshan, A., & Roth, P. L. (2020a). Subgroup differences in situational judgment test scores: Evidence from large applicant samples. International Journal of Selection and Assessment, 28(1), 45–54. https://doi.org/10.1111/ijsa.12269

Herrmann, W., & Wacker, J. (2021). The selective dopamine D2 blocker sulpiride modulates the relationship between agentic extraversion and executive functions. Cognitive, Affective, & Behavioral Neuroscience, 21, 852-867.

Hoffmann, F., Puetz, V.B., Viding, E., Sethi, A., Palmer, A. and McCrory, E.J. (2018), "Risk-taking, peer-influence and child maltreatment: A neurocognitive investigation", Social Cognitive and Affective Neuroscience, Vol. 13 No. 1, pp. 124–134.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. Journal of Applied Psychology, 92(5), 1270-1285.

Hogan, R. & Chamorro-Premuzic, T. (2012) Personality and career successs. in Cooper, L. & Larsen R. (Eds.), APA Handbook of personality and social psychology, Vol. III. Washington, DC: American Psychological Association

Holtrop, D., Oostrom, J. K., Dunlop, P. D., & Runneboom, C. (2021a). Predictors of faking behavior on personality inventories in selection: Do indicators of the ability and motivation to fake predict faking? International Journal of Selection and Assessment, 29(2), 185–202. https://doi.org/10.1111/ijsa.12322

Hope, D.A. (2012), The Influence of Attention, Learning, and Motivation on Visual Search, available at:https://doi.org/10.1007/978-1-4614-4794-8.

Hu, J., & Connelly, B. S. (2021a). Faking by actual applicants on personality tests: A meta-analysis of within-subjects studies. International Journal of Selection and Assessment, 29(3–4), 412–426. https://doi.org/10.1111/ijsa.12338

Hübner, R., & Schlösser, J. (2010). Monetary reward increases attentional effort in the flanker task. Psychonomic bulletin & review, 17(6), 821-826. https://doi.org/10.3758/PBR.17.6.821

Hughes, M. and Lacy, C.J. (2016), ""The Sugar'd Game before Thee": Gamification Revisited", Libraries and the Academy, Vol. 16 No. 2, pp. 311–326.

Ihsan, Z. and Furnham, A. (2018), "The new technologies in personality assessment A Review", Consulting Psychology Journal: Practice and Research, Vol. 70 No. 2, pp. 147–166.

Jeong, Y. R., Christiansen, N. D., Robie, C., Kung, M.-C., & Kinney, T. B. (2017a). Comparing applicants and incumbents: Effects of response distortion on mean scores and validity of personality measures: RA JEONG et al. International Journal of Selection and Assessment, 25(3), 311–315. https://doi.org/10.1111/ijsa.12182

Jia, Y., Xu, B., Karanam, Y. and Voida, S. (2016), "Personality-targeted Gamification", Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16, ACM Press, San Jose, California, pp. 2001–2013.

Johnson, J. A., & Hogan, R. (2006). A socioanalytic view of faking. In R.Griffith & M. H. Peterson (Eds.), A closer examination of applicant faking (pp. 209 –231). Greenwich, CT: Information Age.

Josien, L., & Broderick, B. (2013). Cheating in higher education: The case of multi-methods cheaters. Academy of Educational Leadership Journal, 17(3), 93.

Kanai, R. and Rees, G. (2011), "The structural basis of inter-individual differences in human behaviour and cognition", Nat. Rev. Neuroscience, Nature Publishing Group, Vol. 12 No. April, pp. 231–242.

Karren, R. J., & Zacharias, L. (2007). Integrity tests: critical issues. Human resource management review, 17(2), 221-234.

Kato, Y., Endo, H., & Kizuka, T. (2009). Mental fatigue and impaired response processes: Event-related brain potentials in a Go/NoGo task. International Journal of Psychophysiology, 72(2), 204-211.

KELLY, E. L., MILES, C. C., & TERMAN, L. M. (1936). Ability to influence one's score on a typical pen and paper test of personality. Journal of Personality, 4(3), 206-215.

Kennis, M., Rademaker, A.R. and Geuze, E. (2013), "Neural correlates of personality: An integrative review", Neuroscience and Biobehavioral Reviews, Elsevier Ltd, Vol. 37 No. 1, pp. 73–95.

Kiefer, C., & Benit, N. (2016) What is Applicant Faking Behavior? A Review on the Current State of Theory and Modeling Techniques.. Journal of European Psychology Students, 7(1), 9–19. https://doi.org/10.5334/jeps.345

Kiili, K. (2005), "Digital game-based learning: Towards an experiential gaming model", Internet and Higher Education, Vol. 8 No. 1, pp. 13–24.

Kiili, K. (2006), "Evaluations of an Experiential Gaming Model", Human Technology: An Interdisciplinary Journal on Humans in ICT Environments , Vol. 2 No. 2, pp. 187–201.

Kim, B. (2015), "Understanding Gamification", Library Technology Reports, Vol. 51 No. 2, pp. 1– 34.

Kim, S. H., Hwang, J. H., Park, H. S., & Kim, S. E. (2008). Resting brain metabolic correlates of neuroticism and extraversion in young men. Neuroreport, 19(8), 883-886.

Kim, Y.J. and Shute, V.J. (2015), "The interplay of game elements with psychometric qualities , learning , and enjoyment in game-based assessment", Computers & Education, Elsevier Ltd, Vol. 87, pp. 340–356.

Klein, H.J., Wesson, M.J., Hollenbeck, J.R. and Alge, B.J. (1999), "Goal commitment and the goal-setting process: conceptual clarification and empirical synthesis.", The Journal of Applied Psychology, Vol. 84 No. 6, pp. 885–896.

König, C. J., Melchers, K. G., Kleinmann, M., Richter, G. M., & Klehe, U. C. (2006). The relationship between the ability to identify evaluation criteria and integrity test scores. Psychological Test and Assessment Modeling, 48(3), 369.

König, C. J., Merz, A. S., & Trauffer, N. (2012). What is in applicants' minds when they fill out a personality test? Insights from a qualitative study. International Journal of Selection and Assessment, 20(4), 442-452.

König, C. J., Mura, M., & Schmidt, J. (2015). Applicants' strategic use of extreme or midpoint responses when faking personality tests. Psychological Reports, 117(2), 429-436.

Kuncel, N. R., & Borneman, M. J. (2007a). Toward a New Method of Detecting Deliberately Faked Personality Tests: The use of idiosyncratic item responses. International Journal of Selection and Assessment, 15(2), 220–231. https://doi.org/10.1111/j.1468-2389.2007.00383.x

Kuroyama, J., Consulting, L., Wright, C. W., Manson, T. M., & Sablynski, C. J. (n.d.-a). The Effect of Warning Against Faking on Noncognitive Test Outcomes: A field Study of Bus Operator Applicants.

Kurz, R. (2016), "Test User, Adaptor and Developer Perspectives on the British Psychological Society (BPS), European Federation of Psychologists (EFPA) and the Interna-

tional Test Commission Psychometric Assessment Qualifications and Guidelines", 31st International Congress of Psychology, Yokohama, pp. 1–24.

LaHuis, D. M., & Copeland, D. (2009a). Investigating Faking Using a Multilevel Logistic Regression Approach to Measuring Person Fit. Organizational Research Methods, 12(2), 296–319. https://doi.org/10.1177/1094428107302903

Lai, H., Wang, S., Zhao, Y., Zhang, L., Yang, C., & Gong, Q. (2019a). Brain gray matter correlates of extraversion: A systematic review and meta-analysis of voxel-based morphometry studies. Human Brain Mapping, 40(14), 4038–4057. https://doi.org/10.1002/hbm.24684

Landers, R. N., & Collmus, A. B. (2022a). Gamifying a personality measure by converting it into a story: Convergence, incremental prediction, faking, and reactions. International Journal of Selection and Assessment, 30(1), 145–156. https://doi.org/10.1111/ijsa.12373

Landers, R. N., & Sanchez, D. R. (2022a). Game-based, gamified, and gamefully designed assessments for employee selection: Definitions, distinctions, design, and validation. International Journal of Selection and Assessment, 30(1), 1–13. https://doi.org/10.1111/ijsa.12376

Landers, R. N., Auer, E. M., Mersy, G., Marin, S., & Blaik, J. (2022a). You are what you click: Using machine learning to model trace data for psychometric measurement. International Journal of Testing, 22(3–4), 243–263. https://doi.org/10.1080/15305058.2022.2134394

Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011a). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. Journal of Applied Psychology, 96(1), 202–210. https://doi.org/10.1037/a0020375

Landers, R.N. (2015), "An introduction to game-based assessment: Frameworks for the measurement of knowledge, skills, abilities and other human characteristics us-

ing behaviors observed within videogames.", International Journal of Gaming and Computer-Mediation Simulations, Vol. 7 No. 4, pp. iv–viii.

Landers, R.N. (2018), "Gamification Misunderstood: How Badly Executed and Rhetorical Gamification Obscures Its Transformative Potential", Journal of Management Inquiry, p. 105649261879091.

Landers, R.N., Armstrong, M.B. and Collmus, A.B. (2017), "Empirical validation of a general cognitive ability assessment game", SIOP Conference, Orlando.

Landers, R.N., Bauer, K.N. and Callan, R.C. (2017), "Gamification of task performance with leaderboards: A goal setting experiment", Computers in Human Behavior, Vol. 71, pp. 508– 515.

Langner, R., Steinborn, M. B., Chatterjee, A., Sturm, W., & Willmes, K. (2010). Mental fatigue and temporal preparation in simple reaction-time performance. Acta psychologica, 133(1), 64-72.

Lanyon, R. I., & Goodstein, L. D. (1997). Personality assessment. John Wiley & Sons. Chamorro-Premuzic, T., & Furnham, A. (2010). The psychology of personnel selection. Cambridge University Press.

Latham, G.P. and Locke, E.A. (1991), "Self-regulation through Goal Setting", ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES, Vol. 50, pp. 212–247.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. Multivariate behavioral research, 39(2), 329-358.

Lejuez, C.W., Richards, J.B., Read, J.P., Kahler, C.W., Ramsey, S.E., Stuart, G.L., Strong, D.R., et al. (2002), "Evaluation of a behavioral measure of risk taking: The balloon analogue risk task (BART)", Journal of Experimental Psychology: Applied, Vol. 8 No. 2, pp. 75–84.

Leutner, F., & Chamorro-Premuzic, T. (2018). Stronger together: Personality, intelligence and the assessment of career potential. Journal of Intelligence, 6(4), 49.

Lieberman, M. D., & Rosenthal, R. (2001). Why introverts can't always tell who likes them: multitasking and nonverbal decoding. Journal of personality and social psychology, 80(2), 294.

Lieberman, M. D., & Rosenthal, R. (2001a). Why introverts can't always tell who likes them: Multitasking and nonverbal decoding. Journal of Personality and Social Psychology, 80(2), 294–310. https://doi.org/10.1037/0022-3514.80.2.294

Lieberoth, A. (2015), "Shallow Gamification: Testing Psychological Effects of Framing an Activity as a Game", Games and Culture, Vol. 10 No. 3, pp. 229–248.

Lievens, F. (2017), "Assessing Personality–Situation Interplay in Personnel Selection: Toward More Integration into Personality Research", European Journal of Personality, Vol. 31 No. 5, pp. 424–440.

Lievens, F., Klehe, U.-C., & Libbrecht, N. (2011a). Applicant Versus Employee Scores on Self-Report Emotional Intelligence Measures. Journal of Personnel Psychology, 10(2), 89–95. https://doi.org/10.1027/1866-5888/a000036

Lievens, F., Lang, J.W.B., De Fruyt, F., Corstjens, J., Van de Vijver, M. and Bledow, R. (2018), "The predictive power of people's intraindividual variability across situations: Implementing whole trait theory in assessment", Journal of Applied Psychology, Vol. 103 No. 7, pp. 753– 771.

Lim, J. and Furnham, A. (2018), "Can Commercial Games Function as Intelligence Tests? A Pilot Study", The Computer Games Journal, Springer New York, Vol. 7 No. 1, pp. 27–37.

Locke, E.A. (1968), "Toward a theory of task motivation and incentives", Organizational Behavior and Human Performance, Vol. 3 No. 2, pp. 157–189.

Locke, E.A. and Latham, G. (2006), "New directions in goal-setting theory", Current Directions in Psychological Science, Vol. 15 No. 5, pp. 265–269.

Locke, E.A., Shaw, K.N., Saari, L.M. and Latham, G.P. (1981), "Goal setting and task performance: 1969-1980.", Psychological Bulletin, Vol. 90 No. 1, pp. 125–152.

Lopez, F. J., Hou, N., & Fan, J. (2019a). Reducing faking on personality tests: Testing a new faking-mitigation procedure in a U.S. job applicant sample. International Journal of Selection and Assessment, 27(4), 371–380. https://doi.org/10.1111/ijsa.12265

MacDonald, S.W.S., Li, S.C. and Bäckman, L. (2009), "Neural Underpinnings of Within-Person Variability in Cognitive Functioning", Psychology and Aging, Vol. 24 No. 4, pp. 792–808.

MacKenzie, W. I., Ployhart, R. E., Weekley, J. A., & Ehlers, C. (2009a). Contextual Effects on SJT Responses: An Examination of Construct Validity and Mean Differences Across Applicant and Incumbent Contexts. Human Performance, 23(1), 1–21. https://doi.org/10.1080/08959280903400143

MacLeod, C. M. (2007). The concept of inhibition in cognition. In Inhibition in cognition. (pp. 3-23). American Psychological Association.

Madden, T. J., Ellen, P. S., & Ajzen, I. (1992). A comparison of the theory of planned behavior and the theory of reasoned action. Personality and social psychology Bulletin, 18(1), 3-9.

Maeda, E., Yoshikawa, T., Nakashima, R., Kobayashi, K., Yokosawa, K., Hayashi, N., Masutani, Y., et al. (2013), "Experimental system for measurement of radiologists' performance by visual search task", SpringerPlus, Vol. 2 No. 1, pp. 1–6.

Malone, T.W. (1980), "What Makes Things Fun to Learn? Heuristics for Designing Instructional Computer Games", Proceedings of the 3rd ACM SIGSMALL Symposium and the First SIGPC Symposium on Small Systems, ACM, New York, NY, USA, pp. 162–169.

Manohar, Sanjay G., Chong, Trevor T. J., Apps, Matthew A. J., Batla, A., Stamelou, M., Jarman, Paul R., Bhatia, Kailash P., & Husain, M. (2015). Reward Pays the Cost of Noise Reduction in Motor and Cognitive Control. Current Biology, 25(13), 1707-1716.

Marcus, B. (2006). Relationships between faking, validity, and decision criteria in personnel selection. Psychological Test and Assessment Modeling, 48(3), 226.

Marcus, B., Goldenberg, J., Fine, S., Hummert, H., & Traum, A. (2020a). Self-Presentation in Selection Settings: The Case of Personality Tests. Journal of Business and Psychology, 35(5), 557–571. https://doi.org/10.1007/s10869-019-09642-x

Martin, B. A. (2002a). How e€ective are people at faking on personality questionnaires?

Matz, D. C., Hofstedt, P. M., & Wood, W. (2008). Extraversion as a moderator of the cognitive dissonance associated with disagreement. Personality and Individual Differences, 45(5), 401-405.

Mavridis, A. and Tsiatsos, T. (2016), "Game-based assessment: investigating the impact on test anxiety and exam performance", Journal of Computer Assisted Learning, Vol. 33 No. 2, pp. 137–150.

McCrae, R. R., & Costa Jr, P. T. (1999). A Five-Factor theory of personality.

Mccrae, R.R. and Costa, P.T. (1987), "Validation of the Five-Factor Model of Personality Across Instruments and Observers", Journal of Personality and Social Psychology, Vol. 52 No. 1, pp. 81–90.

McDaniel, R. and Fanfarelli, J. (2016), "Building Better Digital Badges", Simulation Gaming, Vol. 47 No. 1, pp. 73–102.

McFarland, L. A., & Ryan, A. M. (2006a). Toward an Integrated Model of Applicant Faking Behavior1: APPLICANT FAKING. Journal of Applied Social Psychology, 36(4), 979–1016. https://doi.org/10.1111/j.0021-9029.2006.00052.x

Mekler, E.D., Brühlmann, F., Opwis, K. and Tuch, A.N. (2013), "Do points, levels and leaderboards harm intrinsic motivation?", Proceedings of the First International Conference on Gameful Design, Research, and Applications - Gamification '13, ACM Press, New York, New York, USA, pp. 66–73.

Mekler, E.D., Brühlmann, F., Tuch, A.N. and Opwis, K. (2015), "Towards understanding the effects of individual gamification elements on intrinsic motivation and performance", Computers in Human Behavior, Vol. 71, pp. 525–534.

Melchers, K.G., Kleinmann, M., Richter, G.M., Konig, C.J.and Klehe, U.-C. (2004) Do Selection Interviews Measure What They Intend to Measure? The impact of applicants' cognitions about evaluated behavior. Zeitschrift fur Personalpsychologie, 3, 159–169.

Mitroff, S.R., Ericson, J.M. and Sharpe, B. (2018), "Predicting Airport Screening Officers' Visual Search Competency With a Rapid Assessment", Human Factors, Vol. 60 No. 2, pp. 201– 211.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. Cognitive psychology, 41(1), 49-100.

Monsell, S. (2003). Task switching. Trends in cognitive sciences, 7(3), 134-140

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. Personnel Psychology, 60(3), 683-729.

Mueller-Hanson, R., Heggestad, E. D., & Thornton III, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. Journal of Applied Psychology, 88(2), 348-355.

Murphy, K. R., & Dzieweczynski, J. L. (2005). Why don't measures of broad dimensions of personality perform better as predictors of job performance? Human Performance, 18, 343–358.

Nee, D. E., & Jonides, J. (2011). Dissociable contributions of prefrontal cortex and the hippocampus to short-term memory: evidence for a 3-state model of memory. Neuroimage, 54(2), 1540-1548.

Nov, O. and Arazy, O. (2013), "Personality-targeted design: theory, experimental procedure, and preliminary results", Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work and Social Computing, ACM, San Antonio, TX, pp. 977–984.

O'Brien, E., & LaHuis, D. M. (2011a). Do Applicants and Incumbents Respond to Personality Items Similarly? A Comparison of Dominance and Ideal Point Response Models: Ideal Point Response Models. International Journal of Selection and Assessment, 19(2), 109–118. https://doi.org/10.1111/j.1468-2389.2011.00539.x

O'Neill, T. A., Lee, N. M., Radan, J., Law, S. J., Lewis, R. J., & Carswell, J. J. (2013). The impact of "non-targeted traits" on personality test faking, hiring, and workplace deviance. Personality and individual differences, 55(2), 162-168.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. Journal of Applied Psychology, 81, 660-679.

Padmala, S., Sirbu, M., & Pessoa, L. (2017). Potential reward reduces the adverse impact of negative distractor stimuli. Social Cognitive and Affective Neuroscience, 12(9), 1402-1413. https://doi.org/10.1093/scan/nsx067

Pang, Y., Cui, Q., Duan, X., Chen, H., Zeng, L., Zhang, Z., Lu, G., & Chen, H. (2017a). Extraversion modulates functional connectivity hubs of resting-state brain networks. Journal of Neuropsychology, 11(3), 347–361. https://doi.org/10.1111/jnp.12090

Parker, D., Charlton, J., Ribeiro, A. and Pathak, R.D. (2013), "The interactive effect of attention control and the perceptions of others' entitlement behavior on job and health outcomes", Journal of Managerial Psychology, Vol. 22 No. 5, available at:https://doi.org/10.1108/IJRDM-06-2016-0103.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. Journal of Personality & Social Psychology, 46, 598-609.

Paulhus, D. L. (1988). Measurement and control of response bias. In J. P. Robinson, P. Shaver, & L. Wrightsman (Eds.), Measures of personality and social psychological attitudes (Vol. 1, pp. 17-60). New York: Academic Press.

Paulhus, D. L. (1989). Socially desirable responding: Some new solutions to old problems. In D. M. Buss & N. Cantor (Eds.), Personality psychology: Recent trends and emerging directions (pp. 201-209). New York: Springer-Verlag.

Paulhus, D. L., & Martin, C. L. (1987). The structure of personality capabilities. Journal of Personality and Social Psychology, 52(2), 354.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. Journal of Personality and Social Psychology, 60, 307-371.

Paulhus, D.L. (1986), "Self-Deception and Impression Management in Test Responses", in Angleitner, A. and Wiggins, J.S. (Eds.), Personality Assessment via Questionnaires: Current Issues in Theory and Measurement, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 143–165.

Paulus, M.P., Rogalsky, C., Simmons, A., Feinstein, J.S. and Stein, M.B. (2003), "Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism", NeuroImage, Vol. 19 No. 4, pp. 1439–1448.

Pelt, D. H., van der Linden, D., & Born, M. P. (2018). How emotional intelligence might get you the job: The relationship between trait emotional intelligence and faking on personality tests. Human Performance, 31(1), 33-54.

Peterson, M. H., Griffith, R. L., & Converse, P. D. (2009). Examining the role of applicant faking in hiring decisions: Percentage of fakers hired and hiring discrepancies in single-and multiple-predictor selection. Journal of Business and Psychology, 24, 373-386.

Pickering, A. D., & Pesola, F. (2014). Modeling dopaminergic and other processes involved in learning from reward prediction error: contributions from an individual differences perspective. Frontiers in Human Neuroscience, 8, 740.

Pickering, A., & Gray, J. A. (2001). Dopamine, appetitive reinforcement, and the neuropsychology of human learning: An individual differences approach. Advances in individual differences research, 113-149.

Ramsay, P.S. (2017), Can Selection Tests Administered via Video Games Reduce Faking?, ProQuest Dissertations and Theses, University of South Florida.

Reinecke, J., Schmidt, P., & Ajzen, I. (1996). Application of the theory of planned behavior to adolescents' condom use: A panel study 1. Journal of applied social psychology, 26(9), 749-772.

Robie, C., Brown, D. J., & Beaty, J. C. (2007a). Do people fake on personality inventories? A verbal protocol analysis. Journal of Business and Psychology, 21(4), 489–509. https://doi.org/10.1007/s10869-007-9038-9

Robie, C., Zickar, M. J., & Schmit, M. J. (2001b). Measurement Equivalence Between Applicant and Incumbent Groups: An IRT Analysis of Personality Scales. Human Performance, 14(2), 187–207.

Robins, R.W., Fraley, R.C. and Krueger, R.F. (2009), Handbook of Research Methods in Personality Psychology, Guilford Publications,

Robins, R.W., Noftle, E.E., Trzesniewski, K.H. and Roberts, B.W. (2005), "Do people know how their personality has changed? Correlates of perceived and actual personality change in young adulthood", Journal of Personality, Vol. 73 No. 2, pp. 489–521.

Robinson, D. and Bellotti, V. (2013), "A Preliminary Taxonomy of Gamification Elements for Varying Anticipated Commitment", ACM CHI 2013 Workshop on Designing Gamification: Creating Gameful and Playful Experiences., ACM, Paris, France, pp. 1–6.

Robinson, S. J., & Brewer, G. (2016). Performance on the traditional and the touch screen, tablet versions of the Corsi Block and the Tower of Hanoi tasks. Computers in Human Behavior, 60, 29-34.

Robson, S. M., Jones, A., & Abraham, J. (2007a). Personality, Faking, and Convergent Validity: A Warning Concerning Warning Statements. Human Performance, 21(1), 89–106. https://doi.org/10.1080/08959280701522155

Röhner, J., Schröder-Abé, M., & Schütz, A. (2013). What do fakers actually do to fake the IAT? An investigation of faking strategies under different faking conditions. Journal of Research in Personality, 47(4), 330-338.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. Journal of Applied Psychology, 83(4), 634.

Rothmann, S. and Coetzer, E. (2003), "The Big Five Personality Dimensions and Job Performance", SA Journal of Industrial Psychology, Vol. 29 No. 68–74

Roulin, N., & Krings, F. (2020a). Faking to fit in: Applicants' response strategies to match organizational culture. Journal of Applied Psychology, 105(2), 130–145. https://doi.org/10.1037/apl0000431

Rupp, A. a., Gushta, M., Mislevy, R.J. and Shaffer, D.W. (2010), "Evidence-centered design of epistemic games: Measurement principles for complex learning environments", The Journal of Technology Learning and Assessment, Vol. 8 No. 4, pp. 3–41.

Ryan, R.M., Rigby, C.S. and Przybylski, A. (2006), "The motivational pull of video games: A self-determination theory approach", Motivation and Emotion, Vol. 30 No. 4, pp. 347–363.

Sailer, M., Hense, J., Mandl, H. and Klevers, M. (2013), "Psychological Perspectives on Motivation through Gamification", Interaction Design and Architecture(s) Journal, Vol. 19, pp. 28–37.

Sailer, M., Hense, J.U., Mayr, S.K. and Mandl, H. (2017), "How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction", Computers in Human Behavior, Vol. 69, pp. 371–380.

Salgado, J. F. (2016a). A Theoretical Model of Psychometric Effects of Faking on Assessment Procedures: Empirical findings and implications for personality at work: A Theoretical Model of Faking Psychometric Effects. International Journal of Selection and Assessment, 24(3), 209–228. https://doi.org/10.1111/ijsa.12142

Salgado, J.F. (1997), "The five factor model of personality and job performance in the European Community.", Journal of Applied Psychology, Vol. 82 No. 1, pp. 30–43.

Sanchez, D. R., & Langer, M. (2020). Video game pursuit (VGPu) scale development: Designing and validating a scale with implications for game-based learning and assessment. Simulation & Gaming, 51(1), 55-86.

Sanchez, D. R., Langer, M., & Kaur, R. (2020). Gamification in the classroom: Examining the impact of gamified quizzes on student learning. Computers & Education, 144, 103666. https://doi.org/10.1016/j.compedu.2019.103666

Sanchez, D. R., Weiner, E., & Van Zelderen, A. (2022). Virtual reality assessments (VRAs): Exploring the reliability and validity of evaluations in VR. International Journal of Selection and Assessment, 30(1), 103–125. https://doi.org/10.1111/ijsa.12369

Schoen, J. L., Williams, J. L., Reichin, S. L., & Meyer, R. D. (2022). IT'S A TRAP! Faking and faking detection on conditional reasoning tests. Personality and Individual Differences, 198, 111803.

Seaborn, K. and Fels, D.I. (2015), "Gamification in theory and action: A survey", International Journal of Human Computer Studies, Vol. 74, pp. 14–31.

Seok, S. and Dacosta, B. (2015), "Predicting Video Game Behavior: An Investigation of the Relationship Between Personality and Mobile Game Play", Games and Culture, Vol. 10 No. 5, pp. 481–501.

Sheeran, P., Trafimow, D., & Armitage, C. J. (2003). Predicting behaviour from perceived behavioural control: Tests of the accuracy assumption of the theory of planned behaviour. British journal of social psychology, 42(3), 393-410.

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

Shute, V.J., Wang, L., Greiff, S., Zhao, W. and Moore, G. (2016), "Measuring problem solving skills via stealth assessment in an engaging video game", Computers in Human Behavior, Elsevier Ltd, Vol. 63, pp. 106–117.

Simons, A., Wohlgenannt, I., Zelt, S., Weinmann, M., Schneider, J., & vom Brocke, J. (2023a). Intelligence at play: Game-based assessment using a virtual-reality application. Virtual Reality. https://doi.org/10.1007/s10055-023-00752-9

Smillie, L. D., Cooper, A. J., & Pickering, A. D. (2011). Individual differences in reward–prediction–error: extraversion and feedback-related negativity. Social Cognitive and Affective Neuroscience, 6(5), 646-652.

Smillie, L. D., Cooper, A. J., Proitsi, P., Powell, J. F., & Pickering, A. D. (2010). Variation in DRD2 dopamine gene predicts extraverted personality. Neuroscience Letters, 468(3), 234-237.

Smillie, L. D., Cooper, A. J., Wilt, J., & Revelle, W. (2012). Do extraverts get more bang for the buck? Refining the affective-reactivity hypothesis of extraversion. Journal of Personality and Social Psychology, 103(2), 306–326. https://doi.org/10.1037/a0028372

Smillie, L. D., Jach, H. K., Hughes, D. M., Wacker, J., Cooper, A. J., & Pickering, A. D. (2019). Extraversion and reward-processing: Consolidating evidence from an electroencephalographic index of reward-prediction-error. Biological psychology, 146, 107735.

Smillie, L.D. (2013), "Extraversion and Reward Processing", Current Directions in Psychological Science, Vol. 22 No. 3, pp. 167–172.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. Human Resource Management Review, 9(2), 219-242.

Snow, E.L., Likens, A.D., Allen, L.K. and McNamara, D.S. (2016), "Taking Control: Stealth Assessment of Deterministic Behaviors Within a Game-Based System", International Journal of Artificial Intelligence in Education, International Journal of Artificial Intelligence in Education, Vol. 26 No. 4, pp. 1011–1032.

Stieger, S. and 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, Elsevier Ltd, Vol. 26, pp. 1488–1495

Stoeber, J., Chesterman, D., Tarn, T.-A. (2010). Perfectionism and task performance: Time on task mediates the perfectionistic strivings–performance relationship. Personality and individual differences, 48(4), 458-462.

Tales, A., Leonards, U., Bompas, A., Snowden, R.J., Philips, M., Porter, G., Haworth, J., et al. (2012), "Intra-Individual reaction time variability in amnestic mild cognitive impairment: A precursor to dementia?", Journal of Alzheimer's Disease, Vol. 32 No. 2, pp. 457–466.

Thomson, D. R., Seli, P., Besner, D., & Smilek, D. (2014). On the link between mind wandering and task performance over time. Consciousness and cognition, 27, 14-26. https://doi.org/10.1016/j.concog.2014.04.001

Toepper, M., Gebhardt, H., Beblo, T., Thomas, C., Driessen, M., Bischoff, M., ... & Sammer, G. (2010). Functional correlates of distractor suppression during spatial working memory encoding. Neuroscience, 165(4), 1244-1253.

Trafimow, D., Sheeran, P., Conner, M., & Finlay, K. A. (2002a). Evidence that perceived behavioural control is a multidimensional construct: Perceived control and perceived difficulty. British Journal of Social Psychology, 41(1), 101–121. https://doi.org/10.1348/014466602165081

Umemoto, A., & Holroyd, C. B. (2016). Exploring individual differences in task switching: Persistence and other personality traits related to anterior cingulate cortex function. Progress in brain research, 229, 189-212.

Valois, P., Desharnais, R., Godin, G., Perron, J., & Lecomte, C. (1993). Psychometric Properties of a Perceived Behavioral Control Multiplicative Scale Developed According to Ajzen's Theory of Planned Behavior. Psychological Reports, 72(suppl), 1079–1083.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. Educational and psychological measurement, 59(2), 197-210.

Vroom, V. H. (1964). Work and motivation.

Wacker, J., & Smillie, L. D. (2015a). Trait Extraversion and Dopamine Function: Extraversion and Dopamine. Social and Personality Psychology Compass, 9(6), 225–238. https://doi.org/10.1111/spc3.12175

Wallace, J. (1966). An abilities conception of personality: Some implications for personality measurement. American Psychologist, 21(2), 132.

Wang, X., Zhen, Z., Xu, S., Li, J., Song, Y., & Liu, J. (2022a). Behavioral and neural correlates of social network size: The unique and common contributions of face recognition and extraversion. Journal of Personality, 90(2), 294–305. https://doi.org/10.1111/jopy.12666

Waters, L. J., Sanchez, D. R., Garcia, K. M. & Rueda, A. (2022). Exploring faked scores in a Game-Based Assessment for personality [Unpublished Manuscript]. San Francisco State University.

Weiss, B., & Feldman, R. S. (2006). Looking good and lying to do it: Deception as an impression management strategy in job interviews. Journal of Applied Social Psychology, 36(4), 1070-1086.

Wilson, K.E., Lowe, M.X., Ruppel, J., Pratt, J. and Ferber, S. (2016), "The scope of no return: Openness predicts the spatial distribution of Inhibition of Return", Attention, Perception, and Psychophysics, Vol. 78 No. 1, pp. 209–217.

Zibarras, L.D. and Woods, S.A. (2010), "A survey of UK selection practices across different organization sizes and industry sectors", Journal of Occupational and Organizational Psychology, Vol. 83 No. 2, pp. 499–511.

Ziegler, M. (2011). Applicant faking: A look into the black box. The Industrial and Organizational Psychologist, 49(1), 29-36.

Ziegler, M., & Buehner, M. (2009). Modeling socially desirable responding and its effects. Educational and Psychological Measurement, 69(4), 548-565.

Ziegler, M., MacCann, C., & Roberts, R. D. (2012). Faking: Knowns, Unknowns, and Points of Contention. In M. Ziegler, C. MacCann & R. D. Roberts (Eds.), New Perspectives on Faking in Personality Assessment (pp. 3 - 16). New York, NY: Oxford.

Zou, L., Su, L., Qi, R., Zheng, S., & Wang, L. (2018a). Relationship between extraversion personality and gray matter volume and functional connectivity density in healthy young adults: An fMRI study. Psychiatry Research: Neuroimaging, 281, 19–23. https://doi.org/10.1016/j.pscychresns.2018.08.018